ISSI 2015

15th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

29 June - 4 July, 2015

BOĞAZİÇİ UNIVERSITY • ISTANBUL-TURKEY

www.issi2015.org

PROCEEDINGS OF ISSI 2015

BOĞAZİÇİ UNIVERSITY 1863

HACETTEPE UNIVERSITY

TÜBİTAK ULAKBİM

PROCEEDINGS OF ISSI 2015 ISTANBUL

15th International Society of Scientometrics and Informetrics Conference

Istanbul, Turkey

29th June to 4th July 2015

Editors

Albert Ali Salah, Yaşar Tonta, Alkım Almıla Akdağ Salah,
Cassidy Sugimoto, Umut Al

Partners

Boğaziçi University, Turkey
Hacettepe University, Turkey
TÜBİTAK ULAKBİM, Turkey

Sponsors

Thomson Reuters
Springer
EBSCO Information Services, USA
Emerald
Elsevier B.V.

Boğaziçi University Cataloging –in-Publication Data
Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015 / editors Albert Ali Salah, Yaşar Tonta, Alkım Almıla Akdağ Salah, Cassidy Sugimoto, Umut Al.

1275 p.; 29 cm.
ISBN 978-975-518-381-7
ISSN 2175-1935
1. Bibliometrics - Congresses. 2. Information science - Congresses. 3. Communication in science - Congresses. 4. Scientific literature - Congresses.
Z669.8|b.I58

Printed at the Boğaziçi University Printhouse
First Printing: June 2015

Boğaziçi Universitesi, ETA B Blok, Zemin Kat, Kuzey Kampüs, İstanbul / TÜRKİYE
Tel ve Fax: (90) 212 359 44 06

ORGANISATION AND COMMITTEES

Conference Chairs

Albert Ali Salah
Yaşar Tonta
M. Mirat Satoğlu

Programme Chairs

Alkım Almıla Akdağ Salah
Cassidy Sugimoto
Umut Al

Doctoral Consortium Chairs

Andrea Scharnhorst
Judit Bar-Ilan

Workshops & Tutorials Chair

Caroline Wagner

Local Organization Chair

Heysem Kaya

Scientific Program Committee

Giovanni Abramo
Isidro Aguillo
Isola Ajiferuke
Alkım Almıla Akdağ Salah
Umut Al
Jens-Peter Andersen
Eric Archambault
Clément Arsenault
Joaquin Azagra-Caro
Judit Bar-Ilan
Aparna Basu
Sada Bihari-Sahu
Maria Bordons
Lutz Bornmann
Hamid Bouabid
Kevin W. Boyack
Guillaume Cabanac
Juan Miguel Campanario
Chaomei Chen
Andrea D'Angelo
Hans-Dieter Daniel
Cinzia Daraio
Hamid Darvish
Koenraad Debackere
Gernot Deinzer
Brad Demarest
Swapan Deoghuria
Fereshteh Didegah
Ying Ding
Güleda Doğan
Tim Engels
Claire François
Jonathan Furner
Antonio Garcia
Aldo Geuna
Wolfgang Glänzel
Alicia Gomez
Isabel Gomez
Juan Gorraiz
Philippe Gorry
Abdullah Gök
Raf Guns
Nabi Hasan
Stefanie Haustein
Sybille Hinze
Michael Hofer
Marianne Hörlesberger
Zhigang Hu
Peter Ingwersen
Siladitya Jana
Evaristo Jiménez-Contreras
Milos Jovanovic
Yuya Kajikawa
Hildrun Kretschmer

ORGANISATION AND COMMITTEES

J P S Kumaravel
Benedetto Lepori
Jacqueline Leta
Jonathan Levitt
Loet Leydesdorff
Liming Liang
Judith Licea
Junwan Liu
Yuxian Liu
Szu-Chia Lo
Carmen Lopez Illesca
Bob Losee
Terttu Luukkonen
Marc Luwel
Domenico Maisano
Wolfgang Mayer
Kate McCain
Eustache Megnigbeto
Lokman Meho
Raul Mendez-Vasquez
Alexis Michel Mugabushaka
Ulle Must
Anton Nederhof
Ed Noyons
Michael Ochsner
Carlos Olmeda-Gómez
José Luis Ortega
Maria Antonia Ovalle-Perandones
Adèle Paul-Hus
Antonio Perianes-Rodríguez
Bluma C. Peritz
Fernanda Peset
Anastassios Pouris
Ismael Rafols
Emanuela Reale
John Rigby
Nicolas Robinson-Garcia
Ivana Roche
Jürgen Roth
Ronald Rousseau
Santanu Roy
Jane Russell
Bibhuti Sahoo
Albert Ali Salah
Ulf Sandström
Elias Sanz
Andrea Scharnhorst
Edgar Schiebel
Christian Schloegl
Jesper Schneider
Torben Schubert
Robert Shelton
Gunnar Sivertsen
Stig Slipersæter
Henry Small
Andreas Strotmann
Cassidy Sugimoto
Yuan Sun
Zehra Taşkın
Mike Thelwall
Bart Thijs
Yaşar Tonta
Andrew Tsou
Saeed Ul Hassan
Peter van den Besselaar
Nees Jan Van Eck
Thed Van Leeuwen
Bart Van Looy
Anthony Van Raan
Benjamin Vargas-Quesada
Liwen Vaughan
Peter Vinkler
Cathelijn Waaijer
Lili Wang
Xianwen Wang
Jos Winnink
Matthias Winterhager
Dietmar Wolfram
Paul Wouters
Qiang Wu
Yishan Wu
Erjia Yan
Dangzhi Zhao
Michel Zitt
Alesia Zuccala

CHAIRS' WELCOME

The 15th International Society of Scientometrics and Informetrics Conference took place at Boğaziçi University in Istanbul, from June 29 to July 4, 2015. The Conference was jointly organised by Boğaziçi University, Hacettepe University, and the TÜBİTAK ULAKBIM (Turkish Academic Network and Information Center – The Scientific and Technological Research Council of Turkey) under the auspices of ISSI – the International Society for Scientometrics and Informetrics.

The ISSI biennial conference is the premier international forum for scientists, research managers, authorities and information professionals to discuss the current status and progress in informetric and scientometric theories, concepts, tools, platforms, and indicators. In addition to theoretical and quantitative focus of the conference, the participants had the opportunity to discuss practical, cross-cultural, and multi-disciplinary aspects of information and library science, R&D-management, and science ethics, among other related topics.

The focus theme of ISSI2015 was "**the future of scientometrics**". Scientometrics and informetrics together represent a broad field with a rich history. Scientometrics has been responsible for creating tools for research assessment and evaluation, as well as for use in charting the flow of scientific ideas and people. Today, with the advancements of computing power, technology, and database management systems, the impact of scientometrics has become ubiquitous for scientists and science policy makers. However, the high diffusion of scientometric and informetric research has also brought a new wave of criticism and concern, as people grapple with issues of goal displacement and inappropriate use of indicators. The question facing the field is how best to move forward given the computational opportunities and the sociological concerns. Therefore, the goal of ISSI2015 was to highlight the best research in this field and to bring together scholars and practitioners in the area to discuss new research directions, methods, and theories, and to reflect upon the history of scientometrics and its implications.

The keynote given by Loet Leydesdorff demonstrated the potential of thinking of science as a complex institution. By building on the Triple Helix Model of University-Industry-Government relations, Dr. Leydesdorff showed that innovation systems can provide institutional mediation between knowledge production, wealth generation, and governance.

The second keynote, by Kevin Boyack, directly answered the challenge of the focus theme of ISSI2015, and proposed several opportunities to expand the field of scientometrics. Dr. Boyack called for increasing attention to funding, workforce, data and instrumentation, research objects, and innovation.

The conference included four special sessions on a range of topics, including performance indicators, algorithms for topic detection, empirical evaluation of education, research and innovation, and how scientometrics can be used to improve and inform university rankings. These special sessions included poster presentations, panel discussions, invited speakers, and public debates.

The increasing number of open-source software for scientometrics presents great opportunities for researchers. Four tutorials, organized on the first day of the conference, aimed to introduce a number of tools in depth: open source data analysis and visualization tools, citation exploration software, measurement of scholarly impact, and on social network analysis with the popular R software.

The Doctoral Forum, organized by Andrea Scharnhorst and Judit Bar-Ilan, is a meeting of senior researchers and selected doctoral students for presenting and discussing research projects and an

excellent way for students of getting valuable feedback, along with strong networking opportunities. This is the sixth ISSI Doctoral Forum and we are extremely happy about the interest it continues to receive from the community. Additionally, the prestigious Eugene Garfield Doctoral Dissertation Scholarship is given by the Eugene Garfield Foundation.

During the Conference, the Derek de Solla Price Award of the International Journal Scientometrics was given to Mike Thelwall, Professor of Information Science at the University of Wolverhampton (UK), in a special session organized for this purpose. This award recognizes excellence through outstanding, sustained career achievements in the field of quantitative studies of science and their applications.

The satellite workshops of the conference reflected the diversity of the field. In "**Mining Scientific Papers: Computational Linguistics and Bibliometrics",** researchers in bibliometrics and computational linguistics were brought together to study the ways bibliometrics can benefit from large-scale text analytics and sense mining of scientific papers, thus exploring the interdisciplinarity of Bibliometrics and Natural Language Processing. The workshop on "**Grand challenges in data integration for research and innovation policy**" dealt with problems of big, open and linked data. The "**Forecasting science: Models of science and technology dynamics for innovation policy"** workshop discussed methodology for predicting the circumstances leading to scientific or technological innovation. "**Workshop on Bibliometrics Education**" brought together educational institutions, employers, professional societies, and Bibliometrics researchers and professionals to tackle this problem. Finally, "**Google Scholar and related products"** was a highly interactive workshop on the benefits and limitations of some of the most important citation tools.

All contributions for the conference were evaluated by at least two reviewers of the Scientific Program Committee. The papers that required additional reviews were discussed by the Program Chairs before a decision was reached. From 228 full and research in progress paper submissions, 123 papers were accepted for publication (54 percent acceptance rate). 82 of these papers were full papers, and 41 were research in progress. There was a large number of paper submissions on social media, technology transfer, science policy and research assessment. From 123 poster and ignite talk submissions, 68 posters and 13 ignite talks were accepted (66 percent). The ignite talks were to increase discussion of underrepresented topics and novel ideas. Because of the large number of papers, and to allow proper discussion for each paper, four parallel sessions were implemented. Several poster sessions were organized, each containing a relatively manageable number of posters. The conference brought together researchers from 42 countries and the works of 458 researchers were presented.

We thank all our contributors for their submissions, the members of the Organizing Committee for their work, the Scientific Program Committee for their reviewing effort, the ISSI board for their trust and guidance, the Rectorate and the Faculty of Engineering of Boğaziçi University for their constant assistance and support, as well as the sponsors for their generous financial contributions. We particularly thank Metin Tunç (Thomson Reuters), Elif Gürses (formerly of TÜBİTAK ULAKBİM), Juan Gorraiz (Universitat Wien), Figen Atalan (Boğaziçi University), Orçun Madran (Hacettepe University) and Büşra Şahin (DEKON Congress & Tourism) for their help in organizing ISSI2015.

Albert Ali Salah, Yaşar Tonta, Mirat Satoğlu, Alkım Almıla Akdağ Salah, Cassidy Sugimoto, Umut Al

TABLE OF CONTENTS

2015

KEYNOTES

Increasing the Relevance of the ISSI Community in Today's Changing Scientific Landscape

Kevin W. Boyack

kboyack@mapofscience.com
SciTech Strategies, Inc., 8421 Manuel Cia Pl NE, Albuquerque, NM 87122 (USA)

The call for papers for the ISSI 2015 conference set forth a bold agenda by specifically asking for papers related to the "Future of Scientometrics". Many fields in science are what one might call primary fields in the same sense that there are primary colors. These are fields with a self-contained base and upon which other fields build. Scientometrics is not one of these primary fields, but rather operates on theories and data about the processes and outputs of science. We are, in essence, a service industry, and as a service industry we have the potential to exercise great influence on science as a whole. We also have the potential to flounder and die a slow death, or to be overtaken and replaced by another industry. In my opinion, our best opportunity to flourish as a field and community is to truly understand the structure, dynamics, and interactions of science as a whole and in parts, at multiple levels of detail, and to not only measure things but to develop predictive capacities. Opportunities exist for us to expand our view beyond traditional roles, if we can but see what they are.

In this talk I will propose that our opportunities to expand and flourish as a community can be enhanced in several ways. First, it is time for most of us to become far more acquainted with the work done by the pioneers in our field, and in related fields, than we currently are. Scientometrics is a melting pot in many ways, populated to a large degree by those trained in other fields – physics, chemistry, engineering, etc. Many of us are lacking in historical knowledge. We hear the names of Kuhn, Price, Merton, Crane, Latour, and many others, but how many of us are really familiar with not only their popular contributions, but also their smaller experiments that are less well known? There is much to be learned from the work started (and often abandoned due to lack of resources) by these giants that is perhaps even more relevant today than before.

Second, as a community we are highly focused on measuring the "arguments" (documentation) of science, whether using citation data or altmetrics. The science system, however, is comprised of far more than "arguments". Some of us do, to a lesser degree, address other parts of the science system – funding, workforce, data and instrumentation, research objects, and innovation. However, it is rare to see analyses that integrate multiple parts of the science system and explore their interactions. Our influence as a community can certainly be increased if we focus more on these interactions.

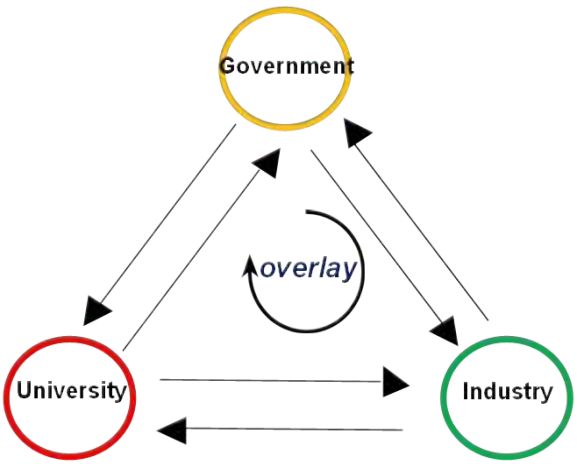
Third, and perhaps most controversially, I suggest that we seek to understand the effect of motivations on science. Perhaps the best way to do this is to start with ourselves, and reflect on "Why do we do what we do?" Are our motives aligned with the purest motives of society? Are we seeking, as individuals and as a community, to serve science and society, or are we seeking for self-aggrandizement? Each of us is many things in life, among which being a researcher or policy maker or scientometrician is only one facet. Often our choice of a career, and of the particular topics we research and for which we advocate are directly tied to these motives. Each of us has a story. Once we understand how our stories drive us to do what we do, then perhaps we can extend that knowledge to better understand science as a whole and how it is driven by the interacting motivations of researchers and institutions. Dick Klavans and I recently created a map of altruism, and were amazed at how much the motives in that map reflect why we do what we do. The parts of the science system mentioned above are all motivated differently. Do we consider this in our models and analyses? How would our analyses change if we were to consider motivation?

Although this talk will use some examples from my current research, it will be largely philosophical, and will raise far more questions than it will give answers. I fully expect many to disagree with much of what will be presented. Nevertheless, I submit that raising these questions at this time has the potential to cause us all to think critically, and that such critical thinking is the first step toward increasing our relevance as a community in the scientific world of the future.

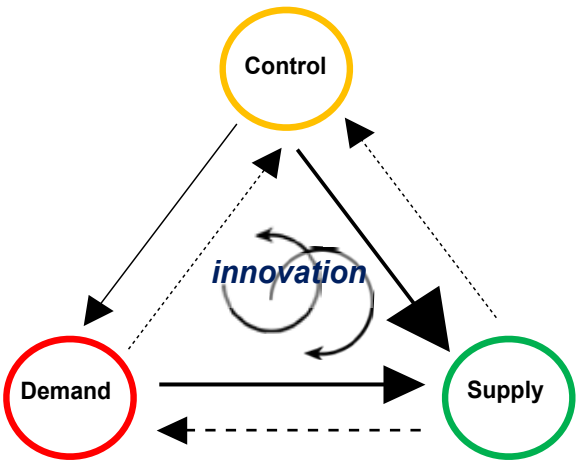
The Triple Helix of Knowledge Production, Wealth Generation, and Normative Control: A Neo-evolutionary Model of Innovation Ecosystems

Loet Leydesdorff

loet@leydesdorff.net
University of Amsterdam, Amsterdam School of Communication Research, Amsterdam, The Netherlands



The neo-institutional model of the Triple Helix of University-Industry-Government Relations (Etzkowitz & Leydesdorff, 1995).



The neo-evolutionary model of innovation in three dimensions (Leydesdorff & Zawdie, 2010; Lawton-Smith & Leydesdorff, 2014).

When three sub-dynamics can operate as selection environments on the variations among one another, a communication field can be generated that proliferates auto-catalytically using each third actor as a feedback or feed forward operating on mutual relations in clockwise or counter-clockwise rotations. This model improves on the neo-Schumpeterian models of innovation systems in evolutionary economics and technology studies, while these models assume a dialectics or co-evolution, for example, between trajectories and selection environments. By extending the Lotka-Volterra equations from two to three dimensions, Ivanova & Leydesdorff (2014) proved the possible emergence of a communication field ("overlay") as an emerging (fourth) subdynamic. In the communication field new options can be generated by sharing meaning provided to the events (Leydesdorff & Ivanova, 2014). This extension of innovative options can be measured as redundancy in terms of bits of information. Petersen, Rotolo & Leydesdorff (in preparation) analyzed Medicals Subject Headings (MEDLINE/PubMed) of approximately 100,000 articles in three research areas including technological breakthroughs in medical innovation (honored with Nobel Prizes in Physiology and Medicine) in terms of "Diseases" (demand), "Drugs and Chemicals" (supply), and "Techniques" (control). Periods of synergy (operationalized as redundancy) can be distinguished from periods in which outward exploration prevails. Innovation systems (e.g., at national or regional levels, but also sectorial ones such as in medicine) provide institutional mediation between wealth generation, knowledge

production, and governance as different perspectives. In the case of China, Leydesdorff & Zhou (2014) found, for example, that the four municipalities play a mediating role above expectation between knowledge production and wealth generation. Note that the three dimensions can differently be operationalized depending on the research design (e.g., as "university," "industry," and "government"); but the dimensions have to be specified as analytically independent so that the three co-variations can be measured (Leydesdorff, Park, & Lengyel, 2014).

References:

Etzkowitz, H., & Leydesdorff, L. (1995). The Triple Helix---University-Industry-Government Relations: A Laboratory for Knowledge-Based Economic Development. *EASST Review 14*(1), 14-19.

Ivanova, I. A., & Leydesdorff, L. (2014). Redundancy Generation in University-Industry-Government Relations: The Triple Helix Modeled, Measured, and Simulated. *Scientometrics, 99*(3), 927-948. doi: 10.1007/s11192-014-1241-7

Lawton Smith, H., & Leydesdorff, L. (2014). The Triple Helix in the context of global change: dynamics and challenges. *Prometheus* (ahead-of-print), 1-16.

Leydesdorff, L., & Ivanova, I. A. (2014). Mutual Redundancies in Inter-human Communication Systems: Steps Towards a Calculus of Processing Meaning. *Journal of the Association for Information Science and Technology, 65*(2), 386-399.

Leydesdorff, L., Park, H. W., & Lengyel, B. (2014). A Routine for Measuring Synergy in University-Industry-Government Relations: Mutual Information as a Triple-Helix and Quadruple-Helix Indicator. *Scientometrics, 99*(1), 27-35. doi: 10.1007/s11192-013-1079-4

Leydesdorff, L., & Zawdie, G. (2010). The Triple Helix Perspective of Innovation Systems. *Technology Analysis & Strategic Management, 22*(7), 789-804.

Leydesdorff, L. & Zhou, P. (2014). Measuring the Knowledge-Based Economy of China in terms of Synergy among Technological, Organizational, and Geographic Attributes of Firms, *Scientometrics 98*(3), 1703-1713.

Petersen, A., Rotolo, D. & Leydesdorff, L. (in preparation), The Interaction of 'Supply', 'Demand', and 'Technology' in terms of Medical Subject Headings: A Triple Helix Model of Medical Innovations.

ALTMETRICS

WEBOMETRICS

Who Publishes, Reads, and Cites Papers? An Analysis of Country Information

Robin Haunschild[1], Moritz Stefaner[2], and Lutz Bornmann[3]

[1] *R.Haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart (Germany)

[2] *moritz@stefaner.eu*
Eickedorfer Damm 35, 28865 Lilienthal (Germany)

[3] *bornmann@gv.mpg.de*
Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich (Germany)

Abstract
The research field of altmetrics has gathered increased attention within scientometrics. Here, we pay particular attention to the connection between countries of readers of papers (at Mendeley) and countries of authors as well as citers of papers (from Web of Science). This study uses the Mendeley application programming interface to gather Mendeley reader statistics for the comprehensive F1000Prime publication set (n_r=149,227 records, n_p = 114,582 papers). F1000Prime is a post-publication peer-review system for papers of the biomedical research. The F1000 papers are rated by experts as good, very good, or exceptional. We find no significant differences between authorship, readership, and authorship of citing papers broken down into countries across quality levels. Most authors, citers, and readers are located in the USA followed by UK and Germany. Except for a few cases, we find that percentages of readers, citers, and authors are rather well balanced. Although Russia and China host many large research groups with a large publication output, both countries are below the top 10 countries ordered according to readership percentages.

Conference Topic
Altmetrics

Introduction
Online reference managers can be seen as the scientific variant of social bookmarking platforms, in which users can save and tag web resources (e.g. blogs or web sites). The best known online reference managers with a social networking component are Mendeley (www.mendeley.com) and CiteULike (www.citeulike.org), which were launched in 2004 (CiteULike) and 2008 (Mendeley), and can be used free of charge (Li et al., 2012). Mendeley – in 2013 acquired by Elsevier (Rodgers and Barbrow 2013) – has developed since then into the most popular product among the reference managers (Haustein 2014), and most empirical studies involving reference managers have used data from Mendeley. Mendeley has obtained a rather unique position as an online reference manager with desktop and mobile app versions. Furthermore, Mendeley offers social networking services, which go beyond the capability of most reference managers.

The platforms allow users to save or organize literature, to share literature with other users, as well as to save keywords and comments on a publication (or to assign tags to them) (Bar-Ilan, et al., 2014, Haustein et al., 2014). Even if it is literature that is mainly saved by the users, they can also add to a library other products of scientific work (such as data sets, software and presentations). The providers of online reference managers make available a range of data for the use of publication by the users: The most important numbers are the user counts, which provide the number of readers of publications via the saves of publications (Li et al., 2012). The readers

can be differentiated into different status and country groups as well as scientific sub-disciplines. The readers' data from Mendeley is also evaluated to make suggestions to the users for new papers and potential collaborators (Priem & Hemminger, 2010, Galloway et al., 2013). Although it is not quite known what Mendeley reader counts mean exactly, they can be viewed as citations to be. Many Mendeley users bookmark a paper in Mendeley with the intend to cite this paper in a forthcoming manuscript. As this is not the only reason to bookmark a publication in Mendeley, it is clear that Mendeley reader counts measure also something different than citations. This additional part of a publication's impact is another means to measure its usage.

In this study, the country information of Mendeley readers is used to compare the readers of papers with their authors as well as those authors who have cited the papers. We are interested in differences and similarities between the countries worldwide: Which are the countries in which the scientists read (or cite) more than publish and vice versa? In which countries are the numbers of authors, readers, and citers similar? As publication set, we used papers from the post-publication peer review system of F1000. It is an advantage of this dataset that each paper is classified according to its quality (based on expert scores). Thus, we are able to investigate the distribution of authors, readers, and citers for papers with different quality.

Literature review

Mendeley is used chiefly by science, technology, engineering and mathematics researchers (Neylon et al., 2014). According to a questionnaire in the bibliometric community (Haustein et al., 2014), 77% of those questioned know Mendeley. But Mendeley is actually used by only 26% of those questioned. However, with respect to the number of saved papers there are large differences between disciplines: Thus, for example, only about a third of the humanities articles indexed in the Web of Science (WoS) can also be found in Mendeley; however, in the social sciences, it is more than half (Mohammadi & Thelwall, 2013). Among the reference managers, Mendeley seems to have the best coverage of globally published literature (Haustein et al., 2014, Zahedi et al., 2014). The large user population and coverage result in Mendeley being seen as the most promising new source for evaluation purposes (among the online reference managers) (Haustein, 2014). Priem (2014) sees Mendeley already as a rival to commercial databases (such as Scopus and WoS).

With a view to the use of the data from online reference managers in research evaluation, bookmarks to publications (i.e. the saving of bibliographic data about publications in libraries) express the interest of a user in a publication (Weller & Peters 2012). But this interest is very variable; the spectrum extends from simple saving of the bibliographic data of a publication up to painstaking reading, annotation and use of a publication (Shema et al., 2014, Thelwall & Maflahi, in press). According to Taylor (2013), the following motives could play a role in the saving of a publication: "Other people might be interested in this paper … I want other people to think I have read this paper … It is my paper, and I maintain my own library … It is my paper, and I want people to read it … It is my paper, and I want people to see that I wrote it" (p. 20). The problem of the unclear meaning of the saving (or naming) of a publication is common to bookmarks in reference managers and also many other traditional and alternative metrics: Thus, for example, traditional citations can mean either simple naming citations in the introduction to a paper, as well as extensive discussions in the results or discussion sections (Bornmann & Daniel, 2008). Traditional citations can also be self-citations.

The data from online reference managers is seen as one of the most attractive sources for the use of altmetrics in research evaluation (Sud & Thelwall, in press). The following reasons are chiefly given for this:

- The collection of literature in reference managers is – similar to the way this is the case with citations and downloads of publications – a by-product of existing workflows (Haustein 2014). This is why saves are appropriate as an alternative metric chiefly for the measurement of impact in areas of work where literature is collected and evaluated (such as with researchers in academic and industrial research, students and journalists).
- Whereas the impact of classical papers can be measured very well via citations in databases (such as the WoS), this is hardly possible with other types of publication such as books or reports.
- According to Mohammadi and Thelwall (2014), usage data of literature may be partially available (i.e. from publishers); but there is a shortage of global and publisher-independent usage data.
- Data sets of online reference managing platforms are highly accessible. The data may be available via API or database dumps (Priem & Hemminger, 2010).

However, the use of data from online reference managers is not only seen as advantageous, but also as problematic:

- Since not everybody who reads and uses scientific literature works with an online reference manager (and Mendeley, particularly), there is the problem that the evaluation of saved data only takes into account a part of the actual readership. Among researchers this part is probably younger, more sociable and more technologically-oriented than average for researchers (Sud & Thelwall, in press).
- The data which are entered by users into the online reference managers are erroneous or incomplete. This can lead to saves not being able to be associated unambiguously with a publication (Haustein, 2014).

Similar to Twitter citations, readership counts can also be manipulated relatively simply (for example with artificially generated spam) (Bar-Ilan et al., 2014).

Many of the empirical-statistical studies into social bookmarking – according to Priem and Hemminger (2010) – deal with tags and tagging. Seen overall, the studies come to the conclusion that exact overlaps of tags and professionally created metadata are rare; most matches are found when comparing tags and title terms (Haustein & Peters, 2012). A large part of the studies into online reference managers has evaluated the correlation between traditional citations (from Scopus, Google Scholar and the WoS) and bookmarks in Mendeley and/or CiteULike. The meta-analysis of (Bornmann, 2015) shows that the correlation is medium to large (CiteULike pooled r=0.23; Mendeley pooled r=0.51).

Two studies have already investigated country information from Mendeley: (1) Haustein and Larivière (2014) analyzed the journal *Aslib Proceedings* (AP) with a set of indicators from several perspectives. The results show that the largest share of AP papers in the last eight years were written by authors affiliated to UK (58 %), Iran (6 %), South Africa and USA (both 5 %). In contrast, Mendeley readers of AP articles were mainly from the USA (14 %), UK (12 %), Spain (6 %), India (4 %), Canada (3 %), South Africa (3 %) and Malaysia (2 %). (2) For some WoS categories, Thelwall and Maflahi (in press) downloaded all article (article meta data) that were written in English from 2011. The country affiliation of the authors was extracted from the WoS affiliation field; each article was searched for in Mendeley to receive the number of readers from each country. The results of the study show that there is a tendency for articles to be more read in

countries with a higher share of their authorship. Possible reasons for the tendency are that authors are often readers of their own articles and that the readers often know or have heard of the authors.

Methods

Peer ratings provided by F1000Prime

F1000Prime is a post-publication peer review system of the biomedical literature (papers from medical and biological journals). F1000 Biology was launched in 2002 and F1000 Medicine in 2006. The two services were merged in 2009 and today form the F1000 database. Papers for F1000Prime are selected by a peer-nominated global Faculty of leading scientists and clinicians who then rate them and explain their importance (F1000, 2012). This means that only a restricted set of papers from the medical and biological journals covered is reviewed, and most of the papers are actually not (Kreiman & Maunsell, 2011, Wouters & Costas, 2012).

The Faculty nowadays numbers more than 5,000 experts worldwide, assisted by 5,000 associates, which are organized into more than 40 subjects (which are further subdivided into over 300 sections). On average, 1,500 new recommendations are contributed by the Faculty each month (F1000, 2012). Faculty members can choose and evaluate any paper that interests them; however, the great majority pick papers published within the past month, including advance online papers, meaning that users can be made aware of important papers rapidly (Wets et al., 2003). Although many papers published in popular and high-profile journals (e.g. *Nature*, *New England Journal of Medicine*, *Science*) are evaluated, 85% of the papers selected come from specialized or less well-known journals (Wouters & Costas, 2012). Less than 18 months since Faculty of 1000 was launched, the reaction from scientists has been such that two-thirds of top institutions worldwide already subscribe, and it was the recipient of the Association of Learned and Professional Society Publishers (ALPSP) award for Publishing Innovation in 2002 (http://www.alpsp.org/about.htm) (Wets et al., 2003).

The papers selected for F1000Prime are rated by the members as good, very good, or exceptional, which is equivalent to recommendation scores (rs) of 1, 2, or 3, respectively. Since many papers are not rated by one member alone, but by several, we calculated a mean rs for every paper. In order to categorize the F1000 papers into three quality levels, papers with mean rs < 2 have been categorized as Q1 and papers with mean rs > 2.5 as Q3. Papers with rs in-between are categorized as Q2, then. This is not a categorization of low and high quality because all F1000Prime papers have a very high quality compared to other papers in their field. This is merely a further distinction between high quality papers, as papers with low quality do not get recommended into F1000Prime.

Data sets used from Mendeley and WoS

In January 2014, F1000 provided one of the authors with data on all recommendations (and classifications) made and the bibliographic information for the corresponding papers in their system (n_r=149,227 records, n_p = 114,582 papers). Each of these records with either a PubMed-ID or a DOI was used to retrieve the Mendeley usage statistics via the R (http://www.r-project.org, accessed October 14, 2014) API of Mendeley (https://github.com/Mendeley/mendeley-api-r-example, http://dev.mendeley.com/methods/, both accessed October 14, 2014). An example R script is available at http://dx.doi.org/10.6084/m9.figshare.1335688. In the summer of 2014, a new version of the API was released which we used for this study (Bonasio,

2014). The previous API had some limitations, such as providing only the information of the demographics for the top three categories as a percentage. Another problem (which has not been solved yet) is that most users do not record their country and so only some readership country location information is available (Thelwall & Maflahi, in press). We requested the actual numbers of Mendeley users for each F1000 record (and the result was not truncated after the top three categories). We observed several (probably random) connection problems. Overall, about 99% of the F1000 paper set was found on Mendeley, which implies a rather good coverage of scientific papers on Mendeley (Bornmann & Haunschild, 2015). We recorded a total of 5,885,534 Mendeley reader counts.

For bibliometric analysis in the current study, country information of the authors who published a F1000 paper or published a paper citing a F1000 paper were sought in an in-house database of the Max Planck Society (MPG) based on the WoS and administered by the Max Planck Digital Library (MPDL). Despite different meanings of (citing) authors' and readers' countries, we talk about countries of readers and (citing) authors in the same way in the following sections.

Technical limitations

Only about 17.6% of 5,885,534 Mendeley reader counts (n=1,038,449) provided were available with their country association. For only 1,064 records of the F1000 data set, we found that the sum over all reader's countries was equal to the total number of reader counts. Thus, in the majority of cases (99.3%) some Mendeley readers are missing in our statistic because many readers did not share their location.

In contrast to the Mendeley data (in which the country information is reader-specific), the country information for the (citing) authors is address-specific. If two authors have different addresses, the country information is counted twice. However, if the addresses are identical, they are counted once. This limitation is unavoidable using our current WoS data. A second limitation of the data is that papers with different publication years have been considered without time-normalization in the study. For different publication years, one can expect different numbers of readers and citers: The longer the reader and citation window, the more counts are expectable. Since the counts have not been time-normalized in the study, papers with longer windows will have a greater effect on the results than papers with smaller windows. However, the papers with longer and smaller windows are unsystematically distributed across the different quality levels of the papers. Thus, the missing time-normalization of the data won't influence the investigation of the relationship between the distribution of readers and (citing) authors across countries and quality levels.

Processing and visualization of the data

The Mendeley reader data, as well as the WoS author and citer data, were processed by Perl (http://www.perl.org/) and Gawk (http://awk.info/) scripts. Visualization of the data was carried out using Tableau (http://www.tableausoftware.com/). Plots of country and world maps use the Mercator projection.

Results

The results of the study including all F1000 papers with data from WoS and Mendeley are shown in Figures 1 and 2, as well as Table 1 (all papers). For each country, we calculated the percentage of authors, readers, and citers. In Figure 1, the percentage of authors (red colour), citers (blue colour), and readers (green colour) are visualized for all countries worldwide. Figure 2 shows a

more detailed analysis of Europe as very many circles are overlapping in this region in Figure 1. The left panel of Figure 2 compares readers (green colour) and authors (blue colour) while the right panel compares citers (red colour) and authors (blue colour). The bigger the circle on the maps, the higher the percentage for a country is.

As the results in Figure 1 show most authors, readers, and citers are located in the USA. The results in Table 1 (all papers) point out that 29.2% of all readers, 38.3% of all authors, and 39.9% of all citers come from the USA. The USA is the country with the most readers, authors, and citers – significantly more than any other country. The high percentages of authors and citers point to a high level of research activity in the USA. The population and number of research groups in the USA are significantly higher compared to most other countries. In Table 1 (all papers), the USA is followed by the UK (all papers: readers=10.7%, citers=6.6%, and authors=9.3%). Further countries in the table (Germany, France, Japan, and Canada) show small differences in the percentages compared to the UK (less than 10 percentage points). Despite the rather large number of research groups in Russia and China, it is quite surprising that both do not appear in the top 10 list ordered by the number of Mendeley readers. In fact, we find China on rank 13 and Russia on rank 25, close to Poland and the Czech Republic.

As the results in Table 1 further show, many countries have different percentages of authors, readers, and citers. The US has a similar percentage of authors and citers (see e.g. the numbers for all papers), but the percentage of readers is lower than both other percentages. This result seems to reflect the fact that Mendeley is only one reference manager software among others in the USA. For other countries it is the other way around. For example, while 4.7% of all readers come from Brazil (all papers), less than 1% of all authors and citing authors are working in this country.

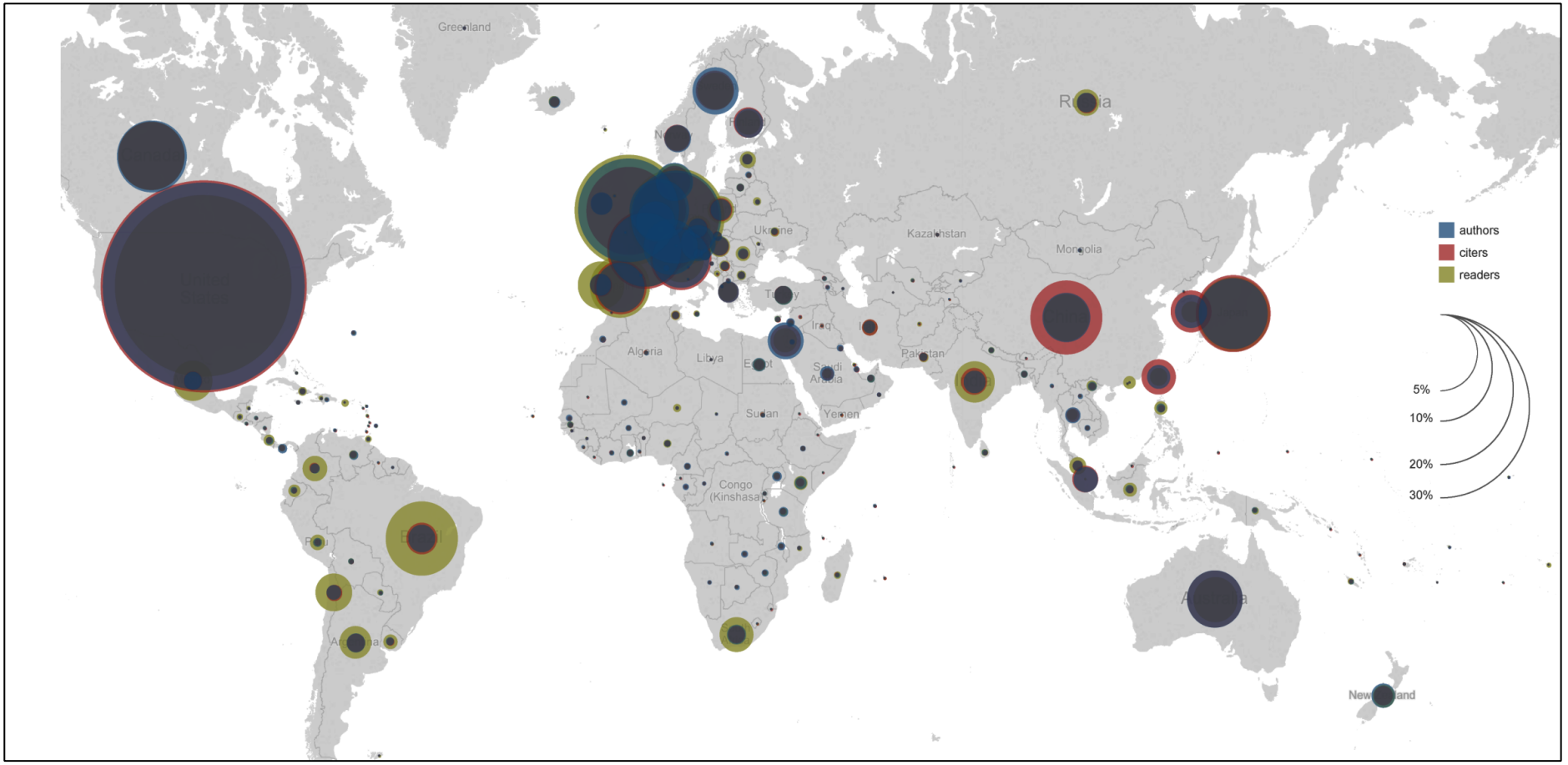


Figure 1. Percentage of authors (blue colour), citers (red colour), and readers (green colour). The circle sizes indicate the share of the country in the amount of readers, citers and authors, respectively. The map is based on all F1000 papers.

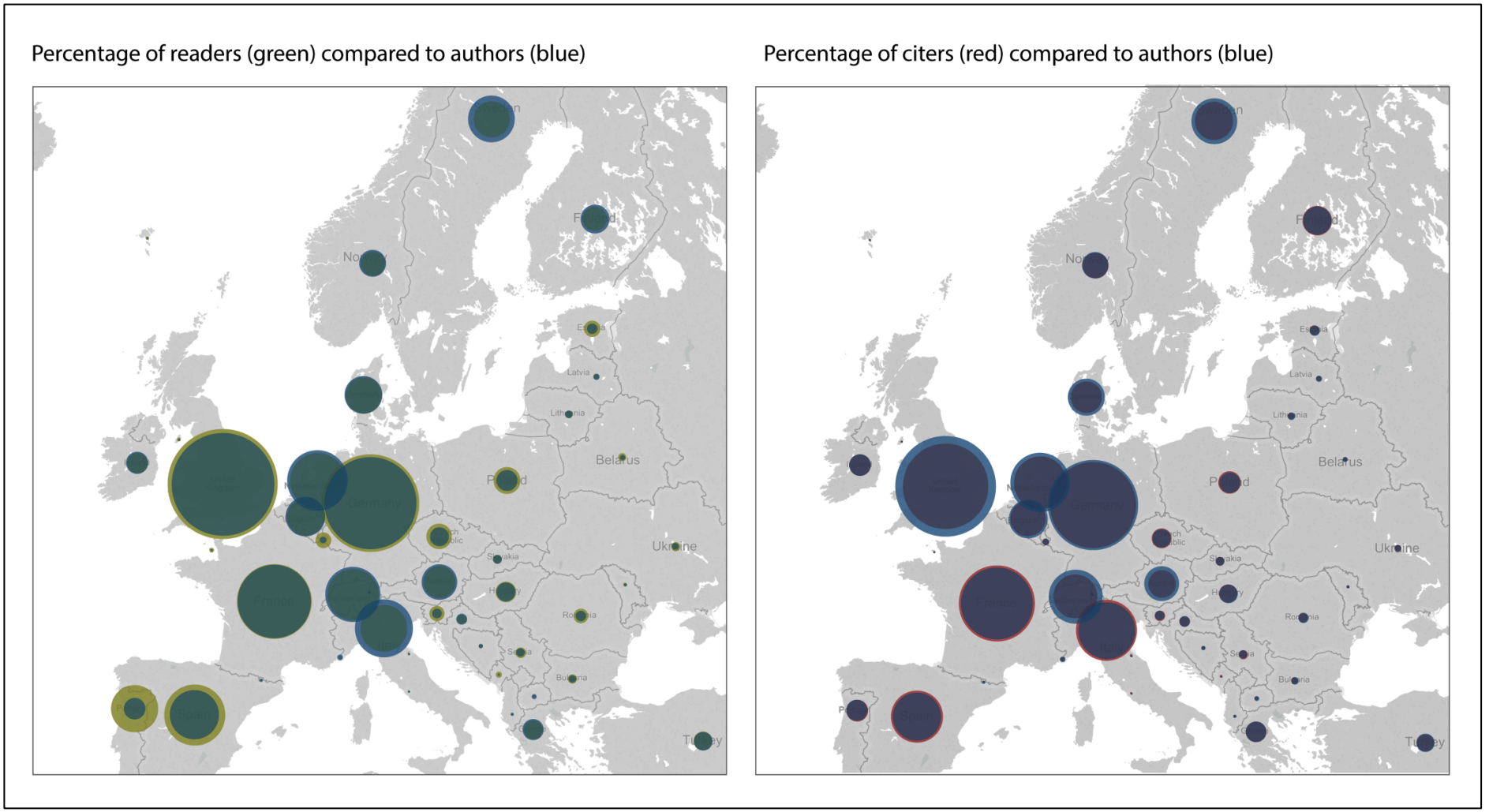


Figure 2. Percentage of readers (green colour) and authors (blue colour) on the left panel, as well as percentages of citers (red colour) and authors (blue colour) visualized on the right panel for European countries. The circle sizes indicate the share of the country in the amount of readers, citers and authors, respectively. The map is based on all F1000 papers.

Table 1. Percentage of authors, citers, and readers from different countries. The percentages are presented for all papers, as well as for papers with Q1 (rs<2), Q2 (rs>=2 and rs<=2.5), and Q3 (rs>2.5) quality. The ten countries are listed with the highest percentage of readers.

All papers	Authors	Citers	Readers	Q1	Authors	Citers	Readers
USA	38.3	39.9	29.2	USA	37.7	39.4	28.7
UK	9.3	6.6	10.7	UK	9.2	6.6	10.7
Germany	7.4	6.8	8.4	Germany	7.4	6.7	8.3
France	4.7	5.2	4.9	Brazil	0.6	0.8	5.0
Japan	4.3	5.1	4.7	France	4.7	5.3	4.9
Brazil	0.6	0.8	4.7	Japan	4.3	5.0	4.5
Canada	4.4	4.0	4.0	Canada	4.5	4.0	4.0
Spain	2.0	2.4	3.2	Spain	2.1	2.5	3.3
Netherlands	3.1	2.5	2.6	Netherlands	3.2	2.5	2.6
Switzerland	2.6	1.7	2.2	Switzerland	2.4	1.6	2.2
Q2	Authors	Citers	Readers	Q3	Authors	Citers	Readers
USA	39.0	40.4	29.4	USA	40.7	41.2	30.6
UK	9.5	6.7	10.7	UK	9.3	6.6	10.7
Germany	7.5	6.9	8.5	Germany	8.0	6.8	8.4
Japan	4.3	5.1	5.0	Japan	4.2	5.4	5.1
France	4.5	5.2	4.8	France	4.6	5.0	4.6
Brazil	0.5	0.7	4.4	Canada	4.2	3.7	4.0
Canada	4.3	3.9	3.9	Brazil	0.6	0.8	4.0
Spain	2.0	2.3	3.1	Spain	1.5	2.2	3.0
Netherlands	3.1	2.4	2.6	Netherlands	2.9	2.3	2.7
Switzerland	2.8	1.7	2.2	Switzerland	2.9	1.7	2.1

This result points out that Brazil rather receives than produces scientific results in the field of biomedical research: Since a low percentage of citing authors reflects a low number of subsequent published papers (following and basing on the F1000Prime papers), this percentage is not only an indicator of reception but also of productivity. Similar results as for Brazil are not only visible on the map in Figure 1 for other south-American countries (such as Argentina or Chile), but also for India and African countries.

From the European countries, Spain and Portugal receive more F1000 papers than they produce (c.f. left panel of Figure 2). Spain is located on rank 8 (see Table 1), and Portugal is located on rank 11. The northern European countries produce more F1000 papers than they cite (c.f. right panel of Figure 2). This is vice versa for most southern European countries.

Table 1 shows the percentage of authors, citers, and readers from different countries not only for all papers, but also for papers with different rs: Q1 (rs < 2), Q2 (2 <= rs <= 2.5), and Q3 (rs > 2.5) section. Comparing the numbers of authors, citers, and readers for different paper quality levels, we see only minor differences for most countries: Brazil shows a somewhat higher amount of readers in the Q1 section (5%) than in the Q3 section (4%), while the percentage of authors and citers does not differ at all between Q3 and Q1 section papers. The USA shows a somewhat higher amount of authors, citers, and readers in the Q3 section (40.7%, 41.2%, and 30.6%, respectively) than in the Q1 section (37.7%, 39.4%, and 28.7%, respectively). The UK shows a nearly constant percentage across quality levels for authors, citers and readers: 9.2%, 6.6%, and 10.7%, respectively for Q1, 9.5%, 6.7%, and 10.7%, respectively for Q2, and 9.3%, 6.6%, and 10.7%, respectively for Q3.

Discussion

By far the highest number of authors, citers, and readers are located in the USA. More F1000 papers are authored, cited, and read in western European countries than in eastern European countries. The amount of F1000 papers authored, cited, and read in China and Russia is small compared to the large number of research groups located there (rank 13 and 25, respectively, according to Mendeley readers). Other reference softwares might be more popular in these countries (or this kind of software is scarcely in use). Traffic data from Alexa.com can be used as an estimate for the Mendeley distribution. The top 5 countries where Mendeley is used seem to be USA (30.4%), India (20.7%), UK (4.3%), Pakistan (3.9%), and Malaysia (3.0%) (http://www.alexa.com/siteinfo/www.mendeley.com, visited on 19 December 2014). Roughly a year earlier, the top 5 countries were somewhat different: USA (16.1%), India (13.2%), Belgium (9.9%), Germany (6.2%), and UK (5.9%) (Thelwall and Maflahi, in press). This relative gain of Mendeley traffic from India, Pakistan, and Malaysia is different from our results, as they do not appear on our top 10 list of Mendeley readers. Within the F1000 readership on Mendeley, India is on rank 15, Malaysia on rank 38, and Pakistan on rank 59. Probably, scientists who use Mendeley in these countries are not that active in the bio-medical research. Belgium, which was in the top 5 list of Mendeley traffic a year ago, is on rank 17 according to our Mendeley readership results of the F1000 paper set.

We find only minor differences in the readership of papers with different quality levels Q1-Q3. The similarities of the results across paper quality levels can be explained with the very high standard of all publications in the F1000Prime set. Also, papers within the Q1 quality section in the F1000 publication set gather a rather high amount of citations (Bornmann 2014). Considering that all papers in the F1000 publication set are of a higher than average quality in the biomedical area, one probably cannot expect a clear difference between quality levels in the Mendeley readership.

Most countries show a quite good balance between consumption and production of F1000 papers. See for example in Table 1, the percentages of Germany are 7.4% authors, 6.8% citers, and 8.4% readers. Although scientists in Germany seem to consume somewhat more of

the literature of the F1000 paper set, the difference between authors (citers) and readers can be neglected, considering the limitations of our study and the (necessary) counting of authors (citers) and readers on unequal footing. In contrast to Germany, the number of readers is significantly higher than the number of authors and citers in some south-American countries (e.g. Brazil, Mexico, Chile, and Argentina) and some European and Asian countries (e.g. Portugal and India).

It is important to keep in mind that we measure authors and citers based on their institutional affiliation and readers on a personal level.

Another problem in the interpretation of the results is that the distribution of the Mendeley software is probably different for each country. Mendeley is free of charge. Thus, one could expect a higher number of Mendeley users in countries with tight research budgets. However, scientists in countries with tight research budgets might not author, cite, or read many publications which got recommended into F1000Prime, as many F1000Prime papers were published in journals with rather high subscription fees.

A third problem in the interpretation of the results is that a rather small number of readers provide their country, as it is not mandatory information. While we found approximately 99% of the F1000 papers at Mendeley, country information were available only for nearly 18% of the reader counts. This is significantly less than the value reported in a previous study done using a much smaller amount of papers (Haustein and Larivière 2014). However, it is reasonable to expect that Mendeley users who do not provide their location are evenly distributed over the world and are reading all quality classes of the F1000 papers.

Acknowledgments

Lutz Bornmann would like to thank Adie Chan, Ros Dignon, and Iain Hrynaszkiewicz from F1000 for providing him with the F1000Prime data set

References

Bar-Ilan, J., Shema, H. & Thelwall M. (2014). Bibliographic References in Web 2.0. In B. Cronin and C. R. Sugimoto (Eds.), *Beyond bibliometrics: harnessing multi-dimensional indicators of performance* (pp. 307-325). Cambridge, MA, USA: MIT Press.

Bonasio, A. (2014). *A look at Mendeley Readership Statistics*. Retrieved October 14, 2014, from http://blog.mendeley.com/academic-features/a-look-at-mendeley-readership-statistics/.

Bornmann, L. (2014). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics*, *8*(4), 935-950.

Bornmann, L. (2015). Alternative metrics in scientometrics: A meta-analysis of research into three altmetrics. *Scientometrics*, 103(3), 1123-1144

Bornmann, L. & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45-80.

Bornmann, L. & Haunschild, R. (2015). Which people use which scientific papers? An evaluation of data from F1000 and Mendeley, *Journal of Informetrics*, 9(3), 477-487

F1000. (2012). What is F1000? Retrieved October 25, from http://f1000.com/about/whatis.

Galloway, L. M., Pease, J. L. & Rauh, A. E. (2013). Introduction to altmetrics for science, technology, engineering, and mathematics (STEM) librarians. *Science & Technology Libraries*, 32(4), 335-345.

Haustein, S. (2014). Readership metrics. In B. Cronin and C. R. Sugimoto (Eds.), *Beyond bibliometrics: harnessing multi-dimensional indicators of performance* (pp. 327-344). Cambridge, MA, USA: MIT Press.

Haustein, S. & Larivièe, V. (2014). A multidimensional analysis of Aslib Proceedings – using everything but the impact factor. *Aslib Journal of Information Management*, *66*(4), 358-380.

Haustein, S. & Peters, I. (2012). Using social bookmarks and tags as alternative indicators of journal content description. *Firstmonday*, *17*(11).

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H. & Terliesner, H. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, 1-19.

Kreiman, G. & Maunsell, J. H. R. (2011). Nine criteria for a measure of scientific output. *Frontiers in Computational Neuroscience*, 5.

Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, *91*(2), 461-471.

Mohammadi, E. & Thelwall, M. (2013). Assessing the Mendeley readership of social science and humanities research. *Proceedings of ISSI 2013 Vienna: 14th International society of scientometrics and informetrics conference*. J. Gorraiz, E. Schiebel, C. Gumpenberger & M. Ho. Vienna, Austria, Austrian Institute of Technology GmbH**:** 200-214.

Mohammadi, E. & Thelwall, M., (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, *65*(8), 1627-1638.

Neylon, C., Willmers, M. & King, T. (2014). Rethinking Impact: Applying Altmetrics to Southern African Research. Ottawa, Canada: International Development Research Centre.

Priem, J. (2014). Altmetrics. In B. Cronin and C. R. Sugimoto (Eds.), *Beyond bibliometrics: harnessing multi-dimensional indicators of performance*. Cambridge, MA, USA: MIT Press.

Priem, J. & Hemminger, B. M. (2010). Scientometrics 2.0: toward new metrics of scholarly impact on the social Web. *First Monday*, *15*(7).

Rodgers, E. P. & Barbrow, S. (2013). A look at altmetrics and its growing significance to research libraries. Ann Arbor, MI, USA, The University of Michigan University Library.

Shema, H., J. Bar-Ilan & Thelwall, M., (2014). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. *Journal of the Association for Information Science and Technology*, *65*(5), 1018-1027.

Sud, P. & Thelwall, M. (in press). Not all international collaboration is beneficial: the Mendeley readership and citation impact of biochemical research collaboration. *Journal of the Association for Information Science and Technology.*

Taylor, M. (2013). Towards a common model of citation: some thoughts on merging altmetrics and bibliometrics. *Research Trends*, 35, 19-22.

Thelwall, M. & Maflahi, N. (in press). Are scholarly articles disproportionately read in their own country? An analysis of Mendeley readers. *Journal of the Association for Information Science and Technology*.

Weller, K. & Peters, I. (2012). Citations in Web 2.0. In A. Tokar, M. Beurskens, S. Keuneke et al. *Science and the Internet* (pp. 209-222). Germany, Düsseldorf: University Press.

Wets, K., Weedon, D. & Velterop, J. (2003). Post-publication filtering and evaluation: Faculty of 1000. *Learned Publishing,* 16(4), 249-258.

Wouters, P. & Costas, R. (2012). *Users, narcissism and control – tracking the impact of scholarly publications in the 21st century*. Utrecht, The Netherlands: SURFfoundation.

Zahedi, Z., Costas, R. & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, 1-23.

Do Mendeley Readership Counts Help to Filter Highly Cited WoS Publications better than average citation impact of journals (JCS)?

Zohreh Zahedi[1,2], Rodrigo Costas[1] and Paul Wouters[1]

z.zahedi.2@cwts.leidenuniv.nl; rcostas@cwts.leidenuniv.nl; p.f.wouters@cwts.leidenuniv.nl
[1]CWTS, Leiden University, P.O. Box 905, Leiden, 2300 AX (The Netherlands)
[2]Department of Knowledge & Information Sciences (KIS), Faculty of Humanities, Persian Gulf University, Bushehr, 7516913817 (Iran)

Abstract
In this study, the 'academic status' of users of scientific publications in Mendeley is explored in order to analyse the usage pattern of Mendeley users in terms of subject fields, citation and readership impact. The main focus of this study is on studying the filtering capacity of Mendeley readership counts compared to journal citation scores in detecting highly cited WoS publications. Main finding suggests a faster reception of Mendeley readerships as compared to citations across 5 major field of science. The higher correlations of scientific users with citations indicate the similarity between reading and citation behaviour among these users. It is confirmed that Mendeley readership counts filter highly cited publications (PPtop 10%) better than journal citation scores in all subject fields and by most of user types. This result reinforces the potential role that Mendeley readerships could play for informing scientific and alternative impacts.

Conference Topic
Altmetrics

Introduction
Mendeley is a popular reference management tool and a rich source of readership metrics for scholarly outputs, used by more than 2.5 million users[1]. This platform collects a wide variety of different metadata[2] for each publication saved by the different types of users in their individual library. Among these metadata, statistics about 'academic status', 'discipline' and 'country' provide useful information on the typologies of users of scientific publications in Mendeley.

Mendeley has different coverage and presence across different fields of science (Zahedi, Costas & Wouters, 2014). A moderate correlation between Mendeley readership and citation counts has been observed for different sets of publications from different fields showing that Mendeley readership counts reflect similar but (perhaps) also other types of impact (Thelwall et al., 2013; Haustein et al., 2013; Zahedi, Costas & Wouters, 2014; Mohammadi & Thelwall, 2014). Also, a weak correlation among number of authors, departments, institutions and countries and readership and citation counts for WoS publications has been observed (Sud & Thelwall, in press; Thelwall & Maflahi, in press). Research on users showed that the majority of Mendeley users per publication are PhDs and students. However, one important limitation with Mendeley data on the analysis of users was the data restriction caused by the reporting of only the three most common user types per publication. Full data on users are necessary in order to properly determine the readership patterns among types of users (Zahedi, Costas & Wouters, 2013 & 2014; Haustein & Larivière, 2014; Mohammadi et al., 2014).

The new Mendeley API provides data on all typologies of readers per publication. This means that 100% of all the users per publication are now fully reported[3]. This study represents one of

[1] http://blog.mendeley.com/start-up-life/mendeley-has-2-5-million-users/
[2] See: http://apidocs.mendeley.com/home/user-specific-methods/user-library-document-details
[3] according to William Gunn in the 1:Am altmetrics conference in London (September 2014) www.altmetricsconference.com/

the first approaches to the analysis of Mendeley readerships based on statistics per publication from all users. We overcome the main limitation of previous studies which were limited to restricted Mendeley users statistics.

In this paper, the usage patterns of the different Mendeley users based on their 'academic status'[4] by fields, citation and readership impact are studied. Also, we analyse the extent to which Mendeley readerships correlate with the number of citations and across 5 major fields of science in the Leiden Ranking (LR). An important focus of this study is on studying the filtering capacity of Mendeley readerships compared to journal citation scores in detecting highly cited publications. Therefore, particular attention will be paid to the extent to which highly cited outputs can be distinguished by these different impact indicators. Similarly, potential differences among Mendeley users in detecting highly cited publications will be also explored. The concrete objectives and research questions of the paper are the following:

O1: To study the general distribution of Mendeley readerships over WoS publications

Q2. What is the distribution of Mendeley readerships across LR fields and by different users?

O3: To study the relationship of Mendeley readerships with bibliometric indicators

Q4. Are there any differences in correlation by different Mendeley users and across LR fields?

O5: To investigate the ability to identify highly cited publications by Mendeley readerships in contrast to journal citation impact indicators

Q6. Which one of these impact indicators can better filter the WoS highly cited publications across LR fields and by different users?

Data and Methodology

For this study, we used a dataset of 1,196,421 Web of Science (WoS) publications from the year 2011 with Digital Object Identifiers (DOI). DOIs were used as the basis to extract readership metrics through the Mendeley REST API in mid-October 2014. The data from Mendeley has been matched with the CWTS in house WoS to add citation data. Citations have been calculated up to 2014.

Although Mendeley has released the full statistics for all the typologies of the users per publications through its API, some Mendeley user statistics are still missing from some publications[5]. These publications were excluded from the analysis due to their unclear reader counts and types. Limiting the dataset to articles and reviews, a final set of 977,067 publications received 12,418,426 total readerships[6] and 6,882,632 total citations. Comparing the ratios of mean citation score per publication (MCS) and mean readerships per publication (MRS), we also find higher MRS (12.7) than MCS (7.04). The actual number of the different types of Mendeley users per publication has been calculated as well as several bibliometrics

[4] These are the different types of users in Mendeley (i.e. PhD students, Professors, Post doc, researchers, Students (under graduates and post graduates), Librarians, Lecturers, Other Professionals and Academic and non-Academic researchers) who have saved publications in their individual libraries. This information allows us to identify users of scientific publications but this information is not free of limitations. For example, it is not clear whether the academic status of the users is updated regularly or how to distinguish users who could belong to more than one category (e.g. a librarian who is also a PhD student).

[5] There are 144,8495 publications with missing readership statistics. These publications have been saved in Mendeley but since their readership counts are missing, they are excluded from the analysis.

[6] We have found some inconsistencies in the counts of readerships. There is a difference between the sum of total readership counts reported by Mendeley (i.e. as they come directly from the readership count provided by Mendeley) and the sum of the individual Mendeley readerships by the different users (calculated by ourselves). (12,418,426 - 12,412,305=6121 differences)

indicators. Precision-recall analysis (Waltman & Costas, 2014) has also been performed, considering 5 major fields of science as represented in the Leiden Ranking (LR)[7].

Analysis and Results

General distribution of Mendeley readerships by major fields of Science and by Mendeley users

Table 1 shows that Biomedical & health sciences (37%) have the highest share of publications with readerships while Mathematics and computer science (8%) have the lowest share. In terms of readership density (i.e. MRS scores) the Life & earth sciences have the highest values (17.5) followed by the Social science & humanities (17), Biomedical & health sciences (14.4) and Natural sciences & engineering (9.7). Mathematics and computer science (9.4) exhibit the lowest readerships density. Also, on average, all fields show higher MRS scores than MCS scores. This could be explained by the relative early publication year (2011) of publications, which could still need some time to get their optimum levels of citations, while in terms of social media, the uptake is normally faster (Haustein et al, 2013), although we still lack information on the obsolescence and time patters of readerships for publications.

Table 1. Mendeley readerships distribution across 5 major fields of science in LR.

LR Main fields of all Publications	P	%	TCS	%	MCS	TRS	%	MRS
Biomedical & health sciences	419,693	37	3617563	44	8,6	6051206	39	14,4
Natural sciences & engineering	322,009	28	2362700	29	7,3	3119704	20	9,7
Life & earth sciences	204,392	18	1469979	18	7,2	3572266	23	17,5
Social sciences & humanities	105,827	9	422046	5	4,0	1795194	12	17,0
Mathematics & computer science	90,813	8	332946	4	3,7	857319	6	9,4
Total		100		100			100	

Total Citation Score (TCS); Total Readership Score (TRS); Mean Citation Score (MCS); Mean Readership Score (MRS)

Figure 1 shows the proportion of readerships by the different types of Mendeley users across the LR fields. Although there are some differences across the fields, in general we find that PhD and students are the most common types of users while Lecturers and Librarian are the least common types of users across all LR fields.

[7] http://www.leidenranking.com/ranking/2013

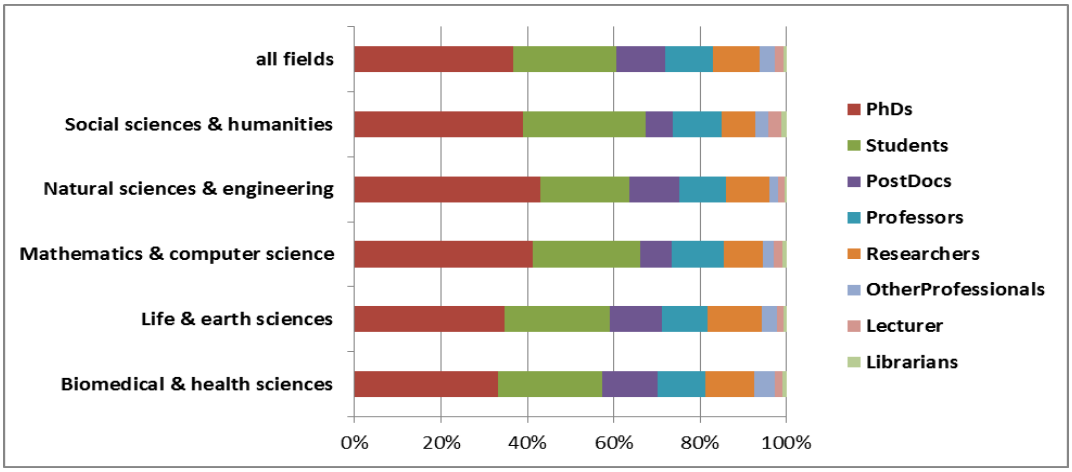


Figure 1. Distribution of Mendeley readerships by the different types of users across LR fields.

Relationship of Mendeley readerships with bibliometric indicators

Spearman correlation analysis among readerships and bibliometric indicators and by the different types of users and across LR Fields has been calculated. The focus here is to explore the extent to which the readerships for the publications saved by the different users in Mendeley are related to their citations and journal indicators. Overall correlation scores among total readerships and bibliometrics indicators are positive and moderate ranging from p=.41 to p=.52 (Table 2).

Table 2. Spearman Correlation analysis of bibliometrics and altmetrics variables.

n=977,067	CS	NCS	JCS	NJCS	RS
CS	1	.93	.57	.43	.52
NCS		1	.40	.46	.50
JCS			1	.75	.44
NJCS				1	.41
RS					1

Citation Score (CS); Normalized Citation Score (NCS); Journal Citation Score (JCS); Normalized Journal Citation Score (NJCS); Readership Score (RS)

Regarding the different types of users, citations have a higher correlation with PhD followed by Students, PostDocs, Researchers, Professors and Other Professionals; however, Librarians and Lecturers exhibit the lowest correlations with citations. These different patterns in terms of correlations among the different types of users might suggest that they have different readership patterns and potentially different readership interests. For example, readership scores for Students, PostDocs, Professors and Researchers correlate most with PhD readership as 'Scientific users', which may indicate their similar scholarly and research usage behaviour. On the other hand, scientific users correlate less with 'other professionals' and Librarians (i.e. suggesting a kind of 'Professional users') and Lecturers as the 'Educational users' (Zahedi, Costas & Wouters, 2013). The latter also correlate most among themselves which may suggest both their similar use of scientific outputs and usage for other purposes than citation such as for self-awareness, teaching and educational or practical and professional purposes (Table3).

Table 3. Spearman Correlation analysis of citation and readerships variables by types of Mendeley users.

n=977,067	CS	PhDs	Students	Post Docs	Professors	Researchers	Other Professionals	Lecturers	Librarians
CS	1	.46	.40	.41	.36	.37	.24	.18	.06
PhDs		1	.58	.49	.48	.47	.25	.27	.08
Students			1	.41	.44	.44	.31	.29	.12
PostDocs				1	.42	.43	.26	.21	.06
Professors					1	.39	.27	.26	.09
Researchers						1	.32	.23	.11
Other Professionals							1	.20	.12
Lecturers								1	.09
Librarians									1

In terms of LR fields, the correlation of citations and readerships is the highest for Social sciences and humanities (p=.61) followed by Natural sciecnes and engineering (p=.59), Life and earth sciences (p=.57), Biomedical and health sciences (p=.55) and the least for Mathematics and computer sciences (p=.45). Regarding the readership by user types and across fields, for most users the highest correlations are in Social sciences and humanities. The lowest correlation with citations is in the field of Mathematics and computer sciences for PhD, Students, PostDocs, Professors and Researchers while for Other Professionals, Lecturers and Librarians the field Natural sciences and engineering displays the lowest correlation with citations (Table 4). This may indicate a relatively stronger use of social media platforms such as Mendeley by scholars in Social science and humanities in their research process than other fields (Rowlands et al., 2011; Tenopir, Volentine & King, 2013).

Table 4. Spearman Correlation analysis of citation and readership by types of Mendeley users across 5 LR Fields.

LR Fields	Total CS and RS	PhD	Student	Post Doc	Professor	Researcher	Other Professional	Lecturer	Librarian
Biomedical & health sciences	.55	.47	.42	.42	.40	.39	.26	.19	.05
Natural sciences &engineering	.59	.51	.43	.39	.35	.33	.17	.18	.04
Life & earth sciences	.57	.53	.46	.43	.40	.39	.24	.22	.06
Mathematics & computer science	.45	.42	.34	.26	.26	.27	.18	.18	.05
Social sciences & humanities	.61	.54	.50	.41	.43	.42	.31	.27	.12

CS (Citation Score); RS (Readership Score)

Analyzing the filtering capacity of highly cited publications by Mendeley readerships

The focus here is to explore the potential use of Mendeley users for filtering highly cited publications compared to journal citation scores. For this purpose, the proportion of top 10% highly cited publications (PPtop 10%)[8] in the sample have been detected. The precision-recall analysis[9] has been performed for all publications in the sample and the 5 LR fields and the different Mendeley users have been explored. Figure 2 shows the general precision-recall analysis of total readership scores and Journal Citation Scores (JCS) for all the publications in the dataset. This figure shows that readerships perform better than JCS in identifying the PPtop 10% most cited publications. The figure indicates that for example a recall of 0.5 (50%) corresponds with a precision of 0.45 (45%) for readership and 0.25 (25%) for journal citation scores in identifying highly cited publications, that is, publications belonging to the top 10% of their field in terms of citations. This means that in order to select half of all highly cited publications we have an error rate of 55% when the selection is made based on readership and an error rate of 75% when the selection is made based on journal citation scores. Since readership outperforms journal citation scores at all levels of recall, we conclude that readership scores identify highly cited publications much better than JCS.

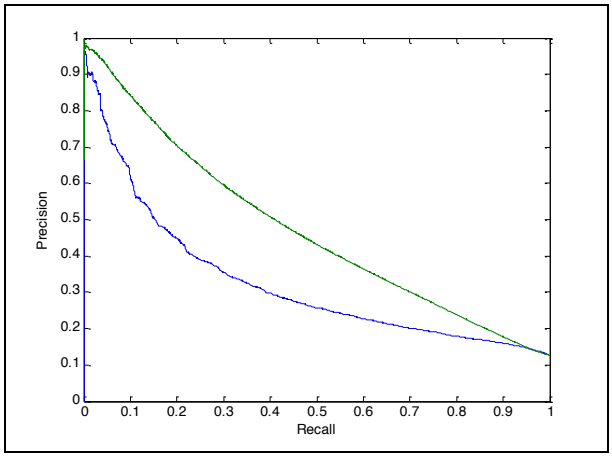
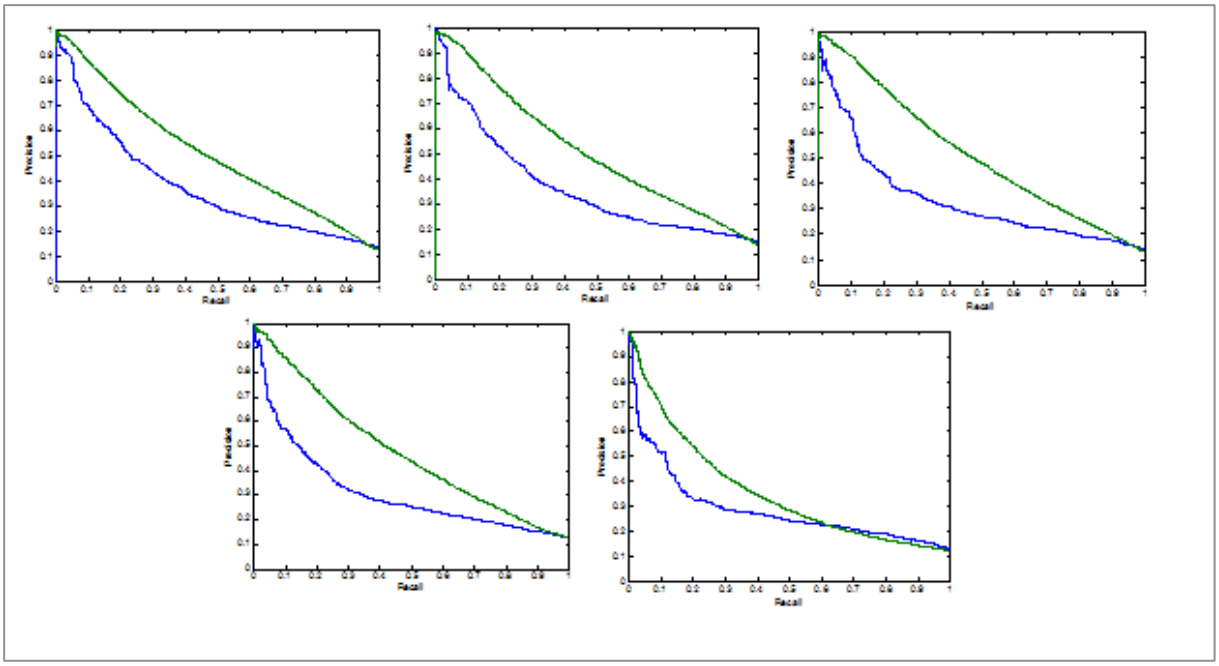


Figure 2. General Precision-recall curves for JCS (blue line) and total readerships (green line) for identifying PPtop10% most highly cited publications.

[8] PP(top 10%) (proportion of top 10% publications). Refers to the proportion of the publications that compared with other publications in the same field and in the same year, belong to the top 10% most frequently cited.

[9] following Waltman & Costas (2014), For a given selection of publications, "precision is defined as the number of highly cited publications in the selection divided by the total number of publications in the selection. Recall is defined as the number of highly cited publications in the selection divided by the total number of highly cited publications".



From left to right: Biomedical & health sciences, Life & earth sciences, Natural Sciences & engineering, Social sciences & humanities, Mathematics & computer science

Figure 3.Precision-recall curves for JCS (blue line) and LR Fields (green line) for identifying PPtop10% most highly cited publications.

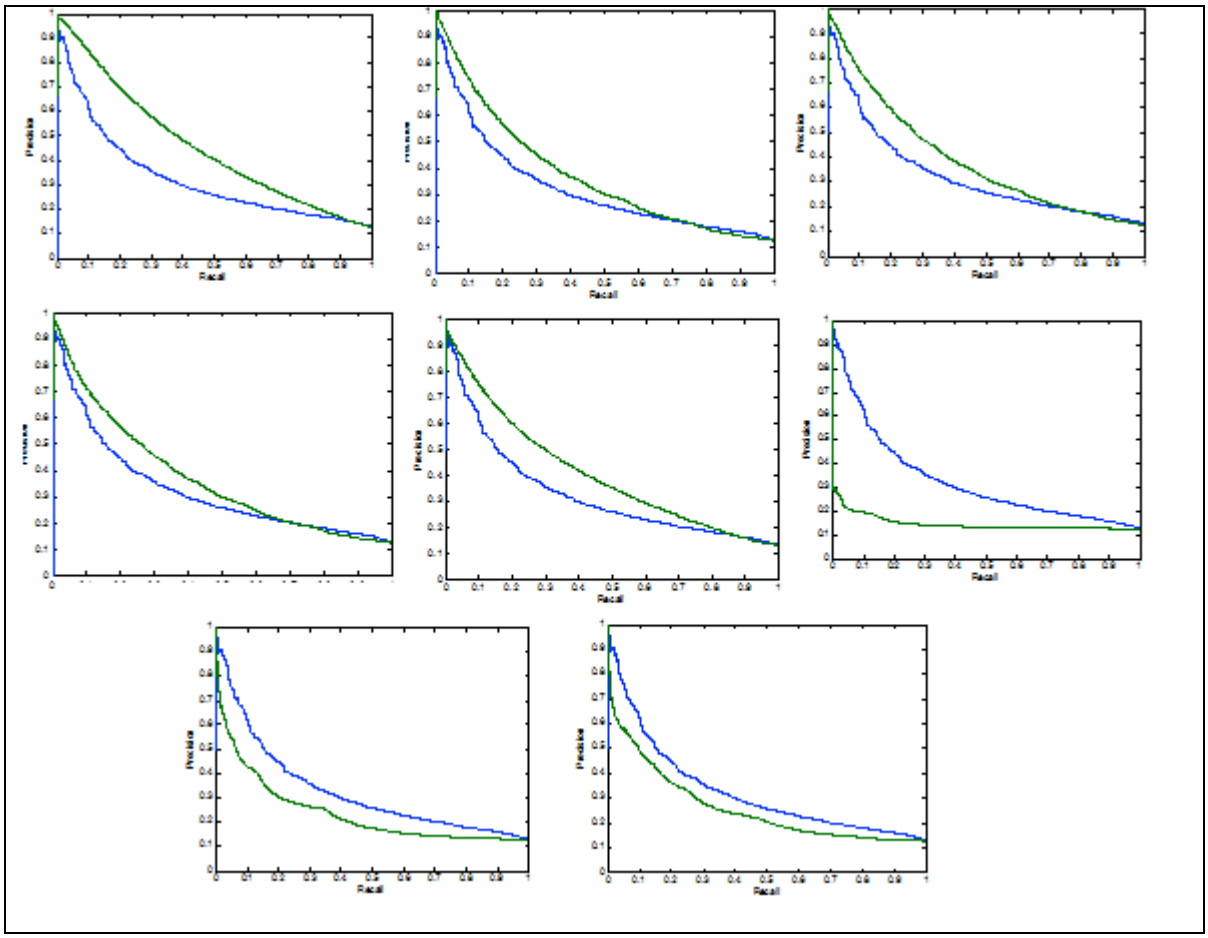
Precision-recall analysis of the different fields of science

The results of the precision-recall analysis for all fields of science again show that readership outperforms JCS scores in filtering highly cited publications. This result supports the idea that Mendeley readership counts filter highly cited publications better than average citation impact of journals (JCS) for all LR fields within our sample. All the figures are similar resembling the general pattern in figure 2 except the figure for Mathematics & computer science, which shows that from recall of 0.6 (60%), the two lines intersect each other and from that point onwards there is a small improvement of JCS over readership scores.

Precision-recall analysis of different types of Mendeley users

The same approach has been done based on the different Mendeley users. Figure 4 shows the results of the precision-recall analysis of readerships scores by the different types of users in Mendeley and Journal Citation Score (JCS). Again, readerships perform better than JCS for most types of users (PhDs, PostDocs, Professors, Researchers and Students vs. Other Professionals, Librarians and Lecturers) in identifying the PPtop10% most highly cited publications within our dataset thus resembling the general pattern in Figure 2. The only exceptions are observed for Librarians, Lecturers and other Professionals where JCS overlaps or outperforms Mendeley readerships. This is in line with the result of the correlation analyse in which these Mendeley user types exhibit less correlations with citations than other types.

Also, regarding the figures for PostDocs, Professors, Researchers and Students, from recall of 0.8 onwards two lines intersect each other and there is a slight improvement of JCS over readerships in the highest level of recall. However, in general, considering readership scores by most types of Mendeley users can help to detect highly cited publications.



From left to right: PhDs, PostDocs, Professors, Researchers, Students, Librarians, Lecturers and Other professionals

Figure 4. Precision-recall curves for JCS (blue line) and type of users readerships (green line) for identifying PPtop10% most highly cited publications.

Main results and discussion

Mendeley is a major multidisciplinary source of readership counts for scholarly publications (Zahedi, Costas & Wouters, 2014) and also it is one of the most promising tools for 'altmetrics' research (Li, Thelwall & Giustini, 2012; Wouters & Costas, 2012). The statistics about the 'Academic Status' of Mendeley users is a valuable source of information to learn more about the academic and non-academic positions of readers of scientific outputs, thus opening the possibility of studying the different types of impact that these different users may entail. Although Mendeley is now reporting the full data per publication, yet more clarity on how Mendeley users are defined is very important, as well as on how the typologies are chosen and updated by the users. For example, the relatively strong correlation between PhDs and Students could suggest that (some) students that become PhD do not update their profiles and therefore they 'read' like PhD students but without updating their 'Academic status' in Mendeley.

The current study has analysed and compared the readership and citation impact of the scholarly publications saved in Mendeley in terms of their types of users and across different LR fields, particularly focusing on the filtering capacity of readership and journal citation impact indicators in identifying highly cited publications. The findings showed that in terms of readership density across the 5 major LR fields, on average, all fields show higher MRS scores than MCS values. This suggests a faster reception of Mendeley readerships as compared to citations and encourages the need to study the temporality and pace of readership

counts. Regarding the types of users, the most common types of users in Mendeley are PhDs and Students, for all LR fields. Correlation analysis shows relatively positive and moderate correlations among the different types of users and citations. The different correlations across users might support the idea that different users could be reading different publications, and thus justifying the use of 'Academic Status' to identify different reading behaviour and typologies of impact. For example, the higher correlations of scientific users with citations, supports their similar reading and citation behaviour vs. other more educational, teaching or professional patterns with lower correlations with citations. This may also be relevant in the analysis of the use of scientific publications in teaching or professional activities. Our results also suggest that readership counts really improve the filtering capacity of highly cited publications over JCS. This is one of the most promising results of this paper, showing the relevance of Mendeley readerships as a relevant filtering tool, something that has not been observed in the previous studies and for other altmetric sources (cf. Costas et al, 2014; Waltman & Costas, 2014). However, it should be taken into account that there are many scholars who don't use Mendeley or any other reference management tools in their scholarly process, so the act of using this type of tools may change in the future. Hence, the use of Mendeley readerships for evaluative purposes still needs careful consideration of its limitations and potential negative effects on the behaviour of individual scholars.

Acknowledgments

The authors are grateful to Erik van Wijk from CWTS for his support on the data collection of Mendeley information for this study. Also, special thanks to Ludo Waltman from CWTS for his fruitful comments on this paper.

References

Costas, R., Zahedi, Z. & Wouters, P. (2014). Do altmetrics correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 1-31. DOI: 10.1002/asi.23309

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2013). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*,101 (2): 1145-1163. DOI: 10.1007/s11192-013-1221-3

Haustein, S. & Larivière, S. (2014). Mendeley as a Source of Readership by Students and Postdocs? Evaluating Article Usage by Academic Status. In *Proceedings of the IATUL Conferences*, Paper 2. http://docs.lib.purdue.edu/iatul/2014/altmetrics/2

Li, X., Thelwall. M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2), 461-471.

Mohammadi, E., Thelwall, M, Larivière, V., & Haustein, S. (2014). Who Reads Research Articles? An Altmetrics Analysis of Mendeley User Categories. *Journal of the Association for Information Sciences and Technology*. Available from:http://www.scit.wlv.ac.uk/~cm1993/papers/WhoReadsResearchArticlesPreprint.pdf

Rowlands, I., Nicholas, D., Russell, B., Canty, N., and Watkinson, A. (2011). Social media use in the research workflow. *Learned Publishing*, 24(3):183-195

Sud, P. & Thelwall, M. (in press). Not all international collaboration is beneficial: The Mendeley readership and citation impact of biochemical research collaboration. *Journal of the Association for Information Science and Technology*.

Tenopir, C., Volentine, R. & King, D.W. (2013).Social media and scholarly reading. *Online Information Review*, 37 (2), 193-216. DOI:10.1108/OIR-04-2012-0062

Thelwall, M,. Haustein, S. Larivière, V. & Sugimoto, C. (2013). Do altmetrics work? Twitter and ten other candidates. *PLOS ONE*, 8(5): e64841. doi:10.1371/journal.pone.0064841

Thelwall, M. & Maflahi, N. (in press). Are scholarly articles disproportionately read in their own country? An analysis of Mendeley readers. *Journal of the American Society for Information Science and Technology*.

Waltman, L., Van Eck, N.J., Van Leeuwen, T.N., Visser, M.S. & Van Raan, A.F.J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1): 37–47.

Waltman, L. & Costas, R. (2014). F1000 Recommendations as a Potential New Data Source for Research Evaluation: A Comparison with Citations. *Journal of the Association for Information Science and Technology*, *65*: 433–445. DOI: 10.1002/asi.23040

Wouters, P. & Costas, R. (2012). Users, Narcissism and control: Tracking the impact of scholarly publications in the 21st century. Utrecht: SURF foundation. Retrieved from: http://www.surffoundation.nl/nl/publicaties/Documents/Users%20narcissism%20and%20control.pdf

Zahedi, Z., Costas, R. & Wouters, P. (2013). What is the impact of the publications read by the different Mendeley users? Could they help to identify alternative types of impact? In PLoS ALM Workshop. 7-9 October. 2013. San Francisco. Available from:

https://openaccess.leidenuniv.nl/handle/1887/23579

http://article-level-metrics.plos.org/files/2013/10/Zahedi.pptx

Zahedi, Z., Costas, R. & Wouters, P. (2014). How well developed are Altmetrics? Cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications". Scientometrics, *101*(2): 1491-1513. DOI:10.1007/s11192-014-1264-0

Zahedi, Z., Costas, R. & Wouters, P. (2014). Assessing the impact of the publications read by the different Mendeley users: Is there any different pattern among users? Proceedings of the IATUL Conferences, Paper 4. Available from :http://docs.lib.purdue.edu/iatul/2014/altmetrics/4

Influence of Study Type on Twitter Activity for Medical Research Papers

Jens Peter Andersen[1] and Stefanie Haustein[2]

[1] *jepea@rn.dk*
Aalborg University Hospital, Medical Library, DK-9000 Aalborg (Denmark)

[2] *stefanie.haustein@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal (Canada)

Abstract
Twitter has been identified as one of the most popular and promising altmetrics data sources, as it possibly reflects a broader use of research articles by the general public. Several factors, such as document age, scientific discipline, number of authors and document type, have been shown to affect the number of tweets received by scientific documents. The particular meaning of tweets mentioning scholarly papers is, however, not entirely understood and their validity as impact indicators debatable. This study contributes to the understanding of factors influencing Twitter popularity of medical papers investigating differences between medical study types. 162,830 documents indexed in Embase to a medical study type have been analysed for the study type specific tweet frequency. Meta-analyses, systematic reviews and clinical trials were found to be tweeted substantially more frequently than other study types, while all basic research received less attention than the average. The findings correspond well with clinical evidence hierarchies. It is suggested that interest from laymen and patients may be a factor in the observed effects.

Conference Topic
Altmetrics

Introduction
In the context of altmetrics, defined as "the study and use of scholarly impact measures based on activity in online tools and environments" (Priem, 2014, p. 266), Twitter has been identified as one of the most interesting and widely-used data sources (Costas, Zahedi, & Wouters, 2014; Thelwall, Haustein, Larivière, & Sugimoto, 2013). Although restricted by brevity—a tweet is limited to 140 characters—Twitter is at the heart of the altmetrics idea to enable a broader scope for impact assessment beyond citation impact. As Twitter is used widely and particularly outside of academia by currently 284 million monthly active users[1], tweets mentioning scientific papers are hoped to capture use by the general public and thus societal impact. Initially suggested as predictors of future citations and thus early indicators of scientific impact (Eysenbach, 2011), more recent large-scale empirical studies suggest that tweets are more likely to reflect online visibility including some social and scientific impact but also self-promotion and buzz (Costas et al., 2014; Haustein, Larivière, Thelwall, Amyot, & Peters, 2014; Haustein, Peters, Sugimoto, Thelwall, & Larivière, 2014). The most tweeted documents seem to attract a lot of online attention rather due to humorous or curious topics than their scientific contributions, often fitting "the usual trilogy of sex, drugs, and rock and roll" (Neylon, 2014, para. 6).
Various, mostly quantitative, studies have shown, with respect to scientific papers, that—after the reference manager Mendeley—Twitter is the altmetrics data source with the second-largest prevalence and it is constantly increasing to currently more than one fifth of 2012 papers being tweeted (Haustein, Costas, & Larivière, 2015). Correlation studies provide evidence that tweets and citations measure different things (for example, Costas et al., 2014;

[1] https://about.twitter.com/company

Haustein, Larivière, et al., 2014; Haustein, Peters, et al., 2014; Priem, Piwowar, & Hemminger, 2012; Thelwall et al., 2013; Zahedi, Costas, & Wouters, 2014). The latest research shows that Spearman correlations with citations for 2012 papers in Web of Science are low at ρ=0.194 for all 1.3 million papers and ρ=0.148 excluding untweeted papers. Beyond the particular differences of Twitter coverage and density between scientific disciplines, research fields and journals reported by various studies (Costas et al., 2014; Haustein, Larivière, et al., 2014; Haustein, Peters, et al., 2014; Zahedi et al., 2014), Haustein et al. (2015) also identified large variations between document types deviating from patterns known for citations. For example, news items and editorial material, which are usually considered non-citable items (Martyn & Gilchrist, 1968), are the most popular types of journal publications on Twitter, showing a tendency of increasing Twitter impact for brief and condensed document types. A study based on a random sample of 270 tweets to scientific papers found that the majority of tweets contained either the paper title or a summary, did not attribute authorship and had a neutral sentiment, while 7% were self-citations (Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013). Other findings suggest that automated diffusion of article links on Twitter plays a role as well (Haustein, Bowman, et al., 2015).

Although these findings provide more evidence that the mechanisms behind tweeting a paper are different from those citing it, the meaning of tweets to scientific papers as well as the role of Twitter in scholarly communication are still unclear, not in the least due to the difficulty to identify 'tweeter motivations' based on 140 characters. This study aims to contribute to a better understanding of tweets as impact metrics by analysing the type of content that is distributed on Twitter. We propose that certain types of articles appeal more to the public than others, for example, because of their potential impact on health issues and everyday life or due to the fact that they are written in a certain way. Previous research has suggested that certain medical study types have a larger citation potential than others (Andersen & Schneider, 2011; Kjaergard & Gluud, 2002; Patsopoulos, Analatos, & Ioannidis, 2005), likely because they are more useful to the research community. In the context of Twitter, medical papers are of particular interest, because, on the one hand, these are particularly relevant to general Twitter users—as opposed to, for example, physics research—and practicing physicians belong to early adopters of social media in their work practice (Berger, 2009). In a survey asking researchers about social media use in research, the uptake by health scientists was, however, slightly below average (Rowlands, Nicholas, Russell, Canty, & Watkinson, 2011).

The aim of this paper is thus to investigate whether there is a connection between different medical study types and the frequency of tweets per article. We hypothesize that some study types are more popular on Twitter due to their attractiveness for a broader audience such as applied medical research relevant to patients as well as meta-analyses summarizing research and condensing results. We will approach this hypothesis by first investigating the potential differences in tweet frequency for a range of medical study types. We argue that logically there should be a connection between the clinical evidence hierarchy (further explained below) and the types of studies patients might consider interesting to discuss or spread on social media, as the highest evidence levels are those which are most likely to affect clinical practice. We therefore expect differences in tweet frequency to be related to evidence levels.

Materials and Methods

Comparing the impact of medical research study types on Twitter requires two pieces of information per research article: a classification of the study type as well as the number of tweets received by each particular paper. Currently no database contains both pieces of information, so that it was necessary to combine data from different sources. For this purpose, the medical study type classifications from the Embase bibliographical database was used,

enriched with metadata from PubMed and Web of Science and then matched to Twitter data from Altmetric.com. The datasets and the matching approach are described in further detail below. Following these descriptions is an account of the specific measurements and statistical tools employed as well as the limitations of this study.

Data collection and matching

Due to Twitter's 140 character limitation, mentions of a scientific paper in tweets are restricted to links to the publisher's homepage or unique document identifiers such as the Digital Object Identifier (DOI) or PubMed ID (PMID). As Twitter only provides access to the most recent tweets[2], it is necessary to constantly query various article identifiers to obtain a database of tweets to scientific papers. Altmetric LLP has been collecting tweets based on multiple document identifiers including the DOI, PMID and the publisher's URL since July 2011 and thus provides a valuable data source for the purposes of our study. To assure reliable and complete Twitter data, we focus our study on papers published 2012. In order to link all tweets to the bibliographic data and study type classification from Embase, the DOI and the PMID are needed.

The study type classifications (see below) for the analysis were retrieved from the Embase bibliographical database. Embase is a major database containing more bibliographical records than PubMed Medline; for example, 24%[3] more for documents published in 2012. It is unclear whether the study type classifications of either database outperforms the other, however, as the indexing of Embase is more exhaustive, we have chosen to use this database for our study. In order to identify relevant papers from Embase (and to be able to perform a citation analysis in the future), *Clinical Medicine* journals were selected from the Web of Science (WoS) based on the National Science Foundation (NSF) journal classification system. The Web of Science also provides bibliographic data and DOIs for the relevant papers, which were used to match Embase study types and tweets from Altmetric.

Embase was queried for the relevant journals using the journal name and various abbreviations as well as the ISSN. Limiting the results to papers published in 2012, the metadata of 593,974 records was retrieved from Embase. In order to obtain the PMID needed to match tweets, PubMed was queried in the same way resulting in 497,619 records. Embase, PubMed and Web of Science were matched using the DOI, PubMed as well as string matches of bibliographic information resulting in 238,560 documents in the final dataset, 94.9% of which with a PMID and 91.1% with a DOI.

The bibliographic metadata was matched to the Altmetric database using the DOI and PMID resulting in 80,116 records with at least one social media event as captured by Altmetric and 74,060 with at least one tweet at the time of data collection in August 2014. This amounts to 31% of the 238,560 being mentioned on Twitter at least once, which corresponds almost exactly to the Twitter coverage of biomedical & health sciences papers found by Haustein, Costas and Larivière (2015). To ensure comparability between tweets published in January and December 2012, we fixed the tweeting window to 18 months (546 days) for each of the tweeted documents, including tweets until 30 June 2013 for papers published on 1 January 2012 and until 30 June 2014 for papers published on 31 December 2012. The day of publication is based on the publication date provided by Altmetric. As this date is not available for all records and is sometimes incorrect, the dataset was further reduced to 52,911 documents, which had an Altmetric publication date in 2012 and not received a tweet before

[2] Twitter's REST API is limited to tweets from the previous week, while the Streaming API provides realtime data only.

[3] For the publication year 2012, Embase contains 1,334,356 records (search: "2012".yr) and PubMed Medline contains 1,072,384 (search: 2012[pdat]).

the publication date. Although these steps lead to an underestimate of the percentage of tweeted papers, they help to reduce biases induced by publication age when comparing the visibility of different medical study types on Twitter.

Medical study type classification

Embase indexes all articles using a controlled vocabulary (the Emtree thesaurus), which contains hierarchically ordered keywords in a classical thesaurus structure. Among these keywords are study type classifications, of which some are directly identifiable as such (e.g. randomised controlled trials), while others require some translation (e.g. "sensitivity and specificity" which is used for diagnostic accuracy studies). The Emtree thesaurus is designed for indexing and retrieval, and there is thus not a given connection between the hierarchical ordering of study type keywords and different levels of research methodology. This is particularly important, as one of the predominant approaches to Western medical research and practice is the so-called evidence based medicine (EBM). One of the cornerstones of EBM is the distinction between study types and their hierarchical ordering based on how much 'evidence' a study is assumed to contribute to the understanding of a given problem (Greenhalgh, 2010). Different hierarchies exist, e.g. the Oxford Centre for Evidence Based Medicine's "Levels of Evidence" (OCEBM Levels of Evidence Working Group, 2011).

Table 1. Medical study type classification system based on Röhrig et al (2009) and OECBM. Classifications with raised numerals have narrower terms, which are not shown here.

	Medical research							
	Primary research						Secondary research	
research_type	A. Basic research		B.Clinical research		C. Epidemiological research		D. Synthesising research	
class	A1. Theoretical	A2. Applied	B1. Experimental	B2. Observational	C1. Experimental	C2. Observational	D1. Meta-analysis	D2. Review
	Method development	Animal study; cell study; genetic engineering/sequencing; biochemistry; material development; genetic studies	Clinical study; phase I-IV	Therapy; prognostic; diagnostic; observational study with drugs; secondary data analysis; case series; case report	Intervention study; field study; group study	Cohort (prospective/historical); case control; cross-sectional; ecological; monitoring, surveillance; Description with registry data		Systematic; narrative
study_type *embase_keyword*	**A1.1 Theoretical study** *Theoretical study* **A1.2 Method development** -	**A2.1 Ex vivo study** *Ex vivo study* **A2.2 In vivo studies** *Animal experiment* **A2.3 In vitro study** *Animal tissue, cells or cell components*[i] *Cell, tissue or organ culture*[ii] *Human tissue, cells or cell components*[iii] **A2.4 Genetic engineering** *Genetic engineering* *and gene technology* *Genetic engineering* *Gene sequence* **A2.5 Biochemistry** *Biochemistry* *Neurochemistry* *Phytochemistry*	**B1.1 Clinical trial** *Clinical trial* *Clinical trial (topic)* *Controlled clinical trial* *Multicenter study* *Phase 1 clinical trial* *Phase 2 clinical trial* *Phase 3 clinical trial* *Phase 4 clinical trial* *Randomized controlled trial*	**B2.1 Case study** *Case report* *Case study* **B2.2 Prognostic study** *Prognosis* **B2.3 Diagnostic study** *Diagnosis* *Diagnostic test* *Sensitivity and specificity* **B2.4 Therapy** - **B2.5 Observational study with drugs** *Observational study AND (major clinical study OR controlled study OR clinical article)*	**C1.1 Intervention study** *Intervention study* **C1.2 Field study** *Field study* **C1.3 Group study** -	**C2.1 Case control study** *Case control study*[iv] **C2.2 Cohort study** *Cohort study* *Longitudinal study* *Retrospective study* *Prospective study* **C2.3 Cross sectional study** *Cross-sectional study* **C2.4 Ecological study** - **C2.5 Monitoring** *Patient monitoring* **C2.6 Surveillance** - **C2.7 Registry study** -	**D1.1 Meta-analysis** *Meta-analysis*	**D2.1 Review** *Review* *Systematic review*

We have chosen to use a particular hierarchy, which allows a classification of study types on their level of research (Röhrig et al., 2009). We have added to the classification of Röhrig et al. (2009) by adding classification codes and the corresponding keywords in Emtree. The resulting system has been validated by two field-experts, and is displayed in Table 1. As can be seen, the classification system allows direct translation between specific Emtree keywords (we have added the broadest terms as well as their relevant narrower terms) and our classification codes on the third level (study_type). The system allows grouping of study

types into classes and research types (levels 2 and 1), thus allowing us to analyse the connection between tweets and the specific study types as well as the broader categories.

Of the entire population of 238,560 records, 162,830 records can be classified using our study type classification system. Of these, 36,595 (22.5%) receive at least one tweet within the fixed 18 months tweet window. Of the remaining 75,730 records without a classification, 16,316 (21.5%) receive at least one tweet. These data delimitations will be used to control for systematic errors in our main dataset (records with classifications). Among those that were classified, 55% had only one classification, 26% had two, 12% had three and the remaining 7% had four or more classifications. References with n classifications are treated as n observations, thus resulting in more than 162,830 observations on either classification level. Some classes in our classification system were not observed at all in the dataset. These classes are omitted in the results section.

Statistical methods and indicators

For each study type classification level we report several statistics for all documents (referred to by $*_A$, e.g. N_A) as well as the subset that has received at least one tweet ($*_T$). The included statistics are number of articles per classification (N), mean tweets per article (μ), the standard deviation from the mean (σ), percentage of articles with at least one tweet (N_T/N_A), and the mean normalised tweets ($\hat{\mu}$) defined as the ratio between μ for a specific classification and μ for the entire population.

As the distributions of tweets for any classification are extremely skewed (see results) similar to citations, the adequacy of the mean as an indicator of average activity is debatable (Calver & Bradley, 2009). However, while the median might be a methodologically more sound choice, the distributions are so extremely skewed that for study type level classification, medians are all 0 when all papers are included and either 1 or 2 if only tweeted papers are included. The corresponding means range from 0.35 to 1.74 and 2.02 to 5.01, providing considerably more information, especially as the scales for the mean are continuous. We therefore use the mean for comparisons, with due care and inclusion of standard deviations and percentage of tweeted articles to provide further information on differences in means. As we have large sample sizes, we expect any major differences in means to be real and not due to chance. However, to test this assumption, all classifications are tested pairwise and against the background population using the independent sample, unpaired Mann-Whitney test.

Limitations

The most obvious error source in this study is the proportion of papers included in the final analysis, compared to the overall population of papers published in 2012. Our background population of 162,830 classified papers only represents 27.4% of the 593,974 records downloaded from Embase. However, it still represents 68.3% of the 238,560 matchable records. This is a fairly high number of papers that could be classified, and if it is possible to improve the matching algorithms, it should also be possible to increase the total number of classified papers comparably. The only systematic error in this regard is the omission of particular documents based on lacking or erroneous DOI's. However, as missing DOI's are also an issue in collecting tweets, this error is not likely to affect the tweet counts with the limitations to tweet-collection that currently exist.

To test if there is a systematic error in the number of tweets per paper, with regard to whether a paper has been classified with a study type or not, we compare the percentage of papers with tweets for classified papers with unclassified papers. For the 162,830 papers with a classification, 36,595 (22.5%) received at least one tweet, while the 75,730 unclassified papers received tweets on 16,316 (21.5%) papers. These values also corroborate findings by Haustein, Costas & Larivière (2015). For the classified papers, mean tweets were 0.67, while

the mean was 0.71 for the unclassified papers. These differences are not random (p = 2.7e-14, using independent two-sample t-test), however, the effect size is also extremely small (Cohen's d = 0.018). We should therefore not consider the lack of study types as confounders for the number of tweets.

While the classification system we have used here was validated by two domain experts, it is only one possible system. Other classifications could have been created, in particular with regard to the translation from Emtree keywords to our classification system. The choices made in this regard will affect the results as presented here. However, when we compare the pairwise scores within a research class, we find high consistency between what could be considered "similar" research types. The only study type, which varies greatly from the other study types in their class is the non-systematic review. This is meaningful, as non-systematic reviews are regarded by medical researchers as much less evidential as their systematic counterparts.

Results

We analysed the classified papers on the three levels present in our classification system: research type, research class and study type. In Tables 2 to 4 we report summary statistics for the three levels, for all papers as well as limited to tweeted papers to determine differences between the share of tweeted papers as well as intensity of (re)use. Results are visualized in Figure 1. In Figures 2 to 4 we provide the results of the pairwise comparison to determine the statistical significance of differences between study types including binary and continuous statistical significance as well as Cohen's d to estimate effect size.

Summary statistics

As can be seen from Tables 2 to 4, there are large differences in the mean tweets per classification, regardless of classification level, although the largest differences are observable in the study types. The differences are clear from the means (μ_A and μ_T), but even more obvious when regarding the relative means ($\hat{\mu}_A$ and $\hat{\mu}_T$). This is also where we find the largest standard deviations, likely due to the smaller N per classification. Meta-analyses and systematic reviews receive considerably more tweets than other study types, which makes the synthesizing research type stand out as well. Overall, a generally increasing interest of the Twitter community can be observed from basic (A) over clinical (B) and epidemiological (C) to synthesizing research (D) papers. Larger variations per research type can be observed for clinical research, where clinical trials are much more tweeted than other study types. In fact, case studies (B2.1) have the lowest mean number of tweets per paper (μ_A), which also reflects in the low mean of observational clinical research (B2) on the research class level. Epidemiological research also performs above average of the entire sample, while basic research (A) consequently performs below, although with somewhat higher scores for genetic engineering (A2.4) than the papers classified as ex vivo (A2.1), in vivo (A2.2) and in vitro (A2.3) studies.

Table 2. Summary statistics for research type.

Research type	N_A	μ_A	σ_A	N_T	N_T/N_A	μ_T	σ_T	$\hat{\mu}_A$	$\hat{\mu}_T$
A. Basic research	130,171	0.434	1.491	25,992	0.200	2.172	2.712	0.642	0.743
B. Clinical research	70,262	0.766	2.699	16,623	0.237	3.238	4.773	1.133	1.108
C. Epidemiological research	43,733	0.963	3.201	12,132	0.277	3.472	5.313	1.425	1.188
D. Synthesising research	38,558	1.005	3.223	10,641	0.276	3.640	5.295	1.486	1.245

Table 3. Summary statistics for research class.

Research class	N_A	μ_A	σ_A	N_T	N_T/N_A	μ_T	σ_T	$\hat{\mu}_A$	$\hat{\mu}_T$
A2. Applied basic research	130,171	0.434	1.491	25,992	0.200	2.172	2.712	0.642	0.743
B1. Experimental clinical research	28,343	1.219	3.495	8,949	0.316	3.860	5.337	1.803	1.321
B2. Observational clinical research	41,919	0.460	1.928	7,674	0.183	2.511	3.894	0.680	0.859
C2. Observational epidemiological research	43,733	0.963	3.201	12,132	0.277	3.472	5.313	1.425	1.188
D1. Meta-analyses	1,883	1.742	4.488	655	0.348	5.009	6.448	2.577	1.714
D2. Reviews	36,675	0.967	3.139	9,986	0.272	3.550	5.199	1.430	1.215

Table 4. Summary statistics for study type.

Study type	N_A	μ_A	σ_A	N_T	N_T/N_A	μ_T	σ_T	$\hat{\mu}_A$	$\hat{\mu}_T$
A2.1. Ex vivo study	1,061	0.425	1.285	223	0.210	2.022	2.155	0.629	0.692
A2.2. In vivo study	52,127	0.437	1.435	10,676	0.205	2.135	2.536	0.647	0.731
A2.3. In vitro study	75,287	0.427	1.519	14,699	0.195	2.190	2.821	0.632	0.749
A2.4. Genetic engineering	1,696	0.606	1.951	394	0.232	2.607	3.345	0.896	0.892
B1.1. Clinical trial	28,343	1.219	3.495	8,949	0.316	3.860	5.337	1.803	1.321
B2.1. Case study	21,788	0.348	1.847	3,204	0.147	2.367	4.292	0.515	0.810
B2.2. Prognostic study	6,618	0.525	1.842	1,407	0.213	2.469	3.341	0.776	0.845
B2.3. Diagnostic study	13,513	0.608	2.081	3,063	0.227	2.682	3.680	0.899	0.917
C2.1. Case control study	2,428	0.975	3.547	664	0.273	3.566	6.065	1.443	1.220
C2.2. Cohort study	34,822	0.943	3.163	9,585	0.275	3.424	5.276	1.394	1.171
C2.3. Cross sectional study	4,891	1.106	3.300	1,440	0.294	3.756	5.201	1.636	1.285
C2.5. Monitoring	1,592	0.956	3.163	443	0.278	3.436	5.242	1.414	1.175
D1.1. Meta-analysis	1,883	1.742	4.488	655	0.348	5.009	6.448	2.577	1.714
D2.1. Review	32,962	0.885	2.909	8,694	0.264	3.354	4.878	1.309	1.147
D2.2. Systematic review	3,713	1.695	4.653	1,292	0.348	4.871	6.839	2.507	1.666

The distributions of tweets per classification are shown in Figure 1, illustrating the highly skewed nature of these distributions, but also the large differences between some categories. The results shown in these boxplots are directly comparable to the summary statistics, and the same classifications stand out as being particularly often tweeted.

From previous research we know that meta-analyses, systematic reviews and clinical trials are also the most highly cited study types (Andersen & Schneider, 2011). However, whether there is a connection between the citedness and tweetedness of medical study types is not obvious from the present data, and will require further research.

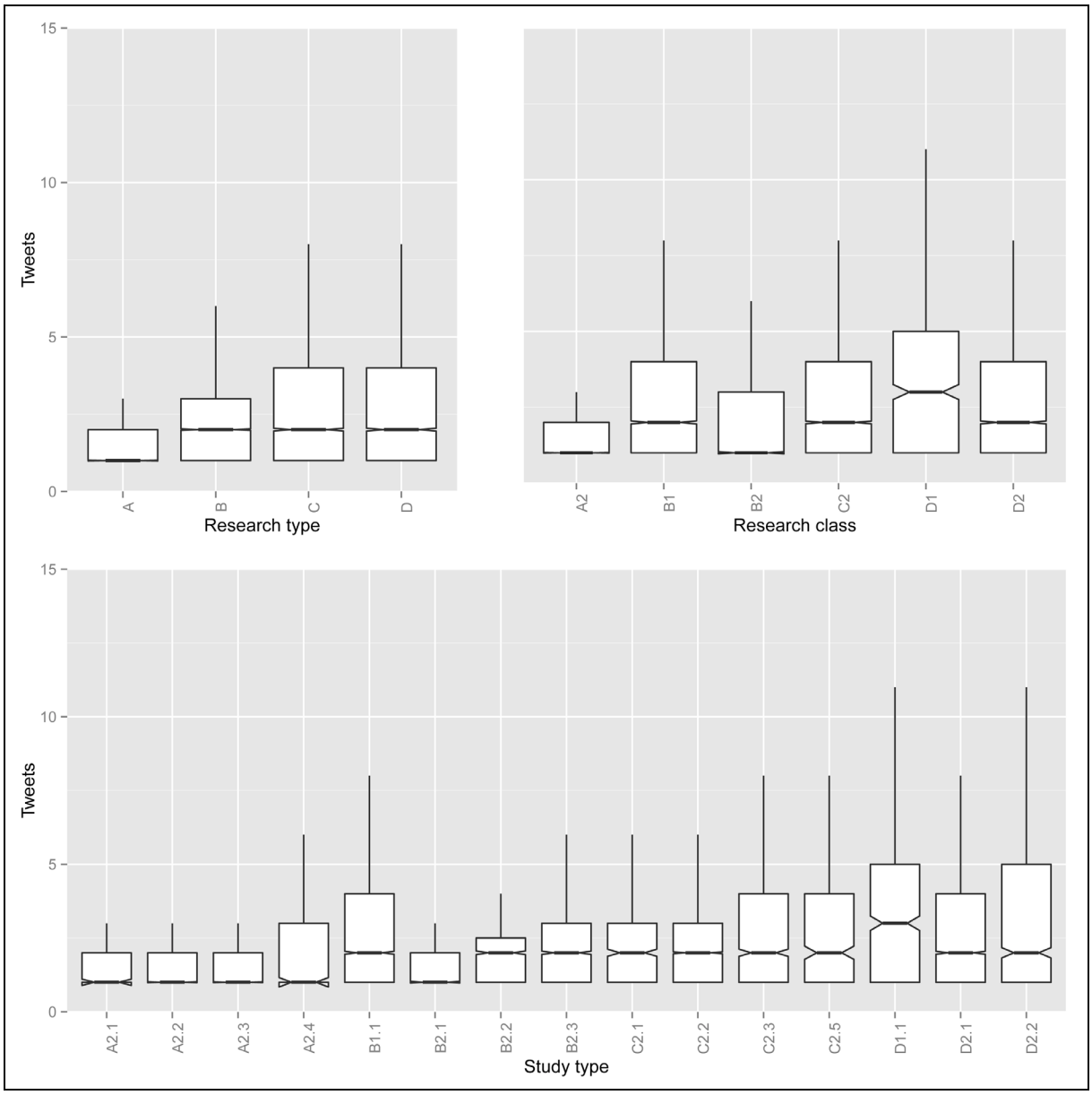


Figure 1. Notched boxplots showing tweet distributions for A) Research type, B) Research class and C) Study type.

Pairwise comparison

In order to analyse the magnitude of differences in classifications further, pairwise comparisons were made on each level. The independent two-sample Mann-Whitney test was used to test whether differences in sample means were due to random effects, and Cohen's *d* was used to estimate the effect size of varying means. There is of course a connection between the *p*-values of the Mann-Whitney tests and Cohen's *d*, to the extent that non-significant differences will also have very small effect sizes, as our sample sizes are quite large. In Figures 2 to 4 these pairwise comparisons are plotted as heatmaps, in which the diagonal and lower half have been omitted. The statistical significance of differences in mean are plotted as both binary maps (*p* below or above 0.05) and as continuous values. On the research type level, basic research stands out the most from the other types, with a lower mean of tweets per paper. For research classes, meta-analyses stand out with very large effect sizes, but overall the effect sizes are somewhat larger on this level than the broader research types. On the study type level, meta-analyses and systematic reviews stand out, but also

clinical trials and epidemiological study types have fairly large effect sizes, compared to other study types.

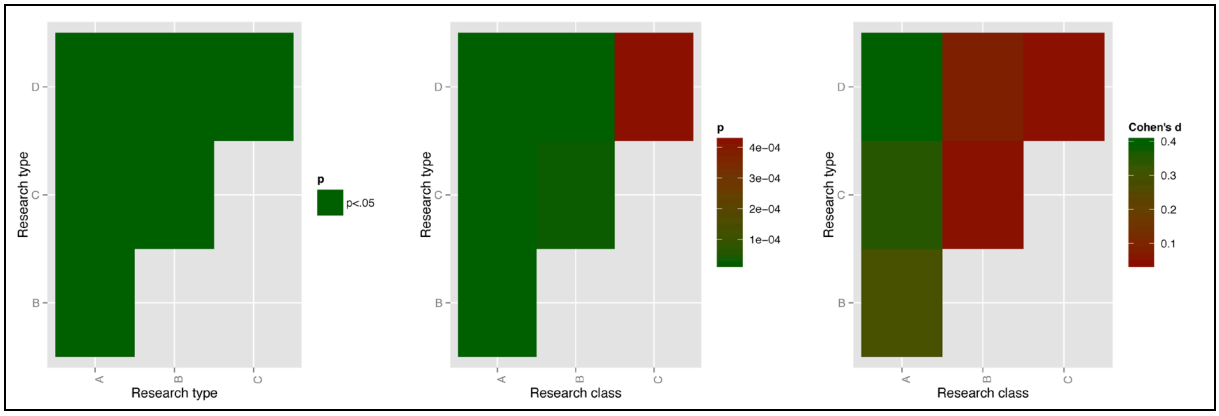


Figure 2. Heatmaps of pairwise comparisons showing A) binary statistical significance, B) continuous statistical significance and C) Cohen's *d* as effect size estimate. All figures are grouped on the research type level.

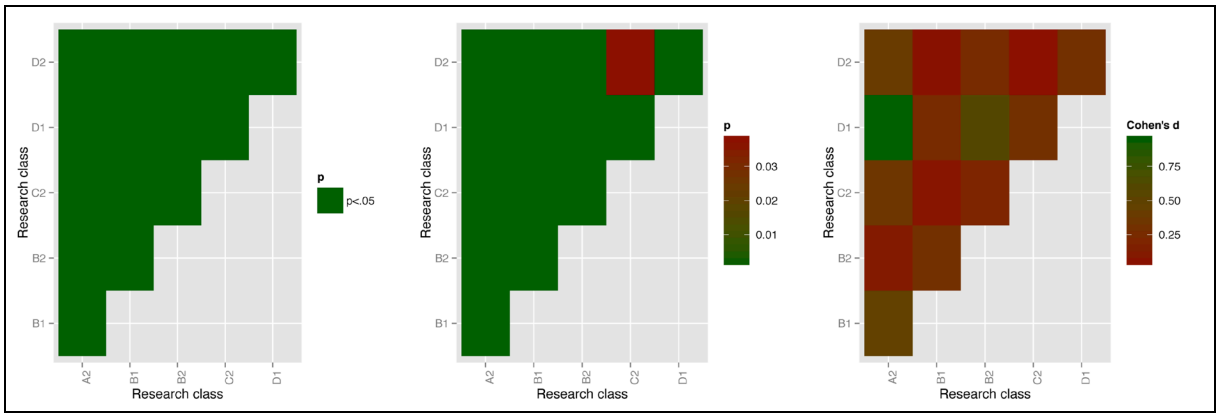


Figure 3. Heatmaps of pairwise comparisons grouped on the research class level. See figure 2 for legend.

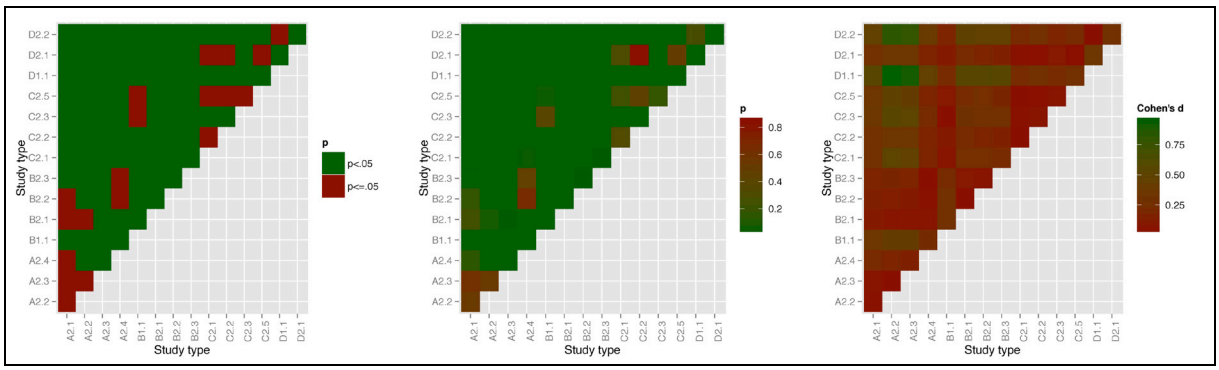


Figure 4. Heatmaps of pairwise comparisons grouped on the study type level. See figure 2 for legend.

Discussion and Outlook

We have analysed the frequency of tweets for medical research papers, distinguished by their specific study type. Our hypothesis was that some study types would be more frequently

tweeted, because they were interesting to a wider audience (e.g., patients and other laymen) than other types. It has not been possible to identify literature on which types of research are actually more useful to laymen, or even which types are most often used. We therefore assume that research, which is close to clinical practise and may contribute to changes in treatments would be more interesting to patients, as they might see a specific benefit to themselves. Based on findings by Haustein, Costas and Larivière (2015) that briefer and condensed document types received more tweets than research articles, we also assumed that synthesising research papers would be more popular on Twitter than basic research.

On the broadest classification level, the results fit well with this assumption, as basic research stands out as the least frequently tweeted research type on average. Basic medical research is also furthest removed from the actual treatment of diseases—so much that some physicians consider it irrelevant to their clinical practise (Andersen, 2013)—which makes them less interesting for the general public of medical laymen and patients active on Twitter. When fine-tuning the analysis to study types, meta-analyses and systematic reviews stand out particularly, followed by clinical trials and epidemiologic study types. This corresponds with typical evidence hierarchies and reflects similar patterns found for citations (Andersen & Schneider, 2011; Kjaergard & Gluud, 2002; Patsopoulos et al., 2005). While this might indicate a relationship between tweets and citations, other studies on a broader level have found this is not the case (Costas et al., 2014; Haustein et al., submitted; Haustein, Larivière, et al., 2014; Zahedi et al., 2014). Other explanations may be that physicians are more likely to tweet about high-evidence studies or that these are also the same types of studies which are most interesting to patients. The latter appears obvious, as high-evidence studies are also more likely to be included in clinical practice guidelines and thus have a greater potential for changing practice. Moreover, results indicating the uptake of social media to be lower among health researchers (Rowlands et al., 2011), while the frequency of tweets per paper in this area is high (Haustein, Peters, et al., 2014), provide some evidence, that the large effect size found for these study types cannot be explained purely by large Twitter-activity from medical researchers. Patients, patient groups and laymen interested in research or other factors may thus play an important role in this observation.

While factors such as entertaining topics may play a role (Neylon, 2014) when looking at the the top per mille most frequently tweeted papers, it is unlikely that all 1,883 meta-analyses, 3,713 systematic reviews and 28,343 clinical trials should have a higher tweet count than other study types due to entertainment value, especially as these are also the most highly regarded study types by the researchers as measured through citations. The mean may of course be affected by single high-scoring studies, however, as can be seen from Figure 1, it is the entire distribution rather than merely the mean, which is increased for these study types. In fact, the maximum tweets per study type is 46 for meta-analyses and 59 for systematic reviews, while it is 65 for two of the basic research study types and 62 for clinical trials. The lowest maximum tweet frequency of a study type is 25 (an in vivo study) and the highest is 67 (a cohort study). It can thus be concluded that medical study types are one of the factors determining popularity of scientific papers on Twitter but they are certainly not the only ones. Apart from factors explored by previous studies and known also from the citation context—such as discipline, publication age, number of authors etc.—Twitter-specific effects should also be investigated. This includes the effect of the number of followers and affordance use as well as the extent to which scientific papers receive tweets due to author and journal self-promotion as well as automated Twitter accounts (Haustein, Bowman, et al., 2015).

Acknowledgements

The authors would like to thank Euan Adie and Altmetric.com for access to their Twitter data. SH acknowledges funding from the Alfred P. Sloan Foundation, grant no. 2014-3-25.

References

Andersen, J. P. (2013). *Conceptualising research quality in medicine for evaluative bibliometrics*. University of Copenhagen. Retrieved June 10, 2015 from http://vbn.aau.dk/files/119316655/JensPeterAndersenThesis.pdf

Andersen, J. P., & Schneider, J. W. (2011). Influence of study design on the citation patterns of Danish, medical research. In *Proceedings of the ISSI 2011 Conference* (pp. 46–53).

Berger, E. (2009). This sentence easily would fit on Twitter: Emergency physicians are learning to "tweet." *Annals of Emergency Medicine*, *54*(2), A23–A25. doi:10.1016/j.annemergmed.2009.06.002

Calver, M. C., & Bradley, J. S. (2009). Should we use the mean citations per paper to summarise a journal's impact or to rank journals in the same field? *Scientometrics*, *81*(3), 611–615. doi:10.1007/s11192-008-2229-y

Costas, R., Zahedi, Z., & Wouters, P. (2014). Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, n/a–n/a. doi:10.1002/asi.23309

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, *13*(4), e123. doi:10.2196/jmir.2012

Greenhalgh, T. (2010). How to read a paper: The basics of evidence-based medicine (4th ed.). Oxford: BMJ Books.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2015). Tweets as impact indicators: Examining the implications of automated bot accounts on Twitter. *Journal of the Association for Information Science and Technology*. Retrieved from http://arxiv.org/abs/1410.4139

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: the effect of document properties and collaboration patterns. *Submitted to PLOS ONE*.

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., & Peters, I. (2014). Tweets vs. Mendeley readers: How do these two social media metrics differ? *It - Information Technology*, *56*(5), 207–215. doi:10.1515/itit-2014-1048

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, *65*(4), 656–669. doi:10.1002/asi.23101

Kjaergard, L. L., & Gluud, C. (2002). Citation bias of hepato-biliary randomized clinical trials. *Journal of Clinical Epidemiology*, *55*(4), 407–10.

Martyn, J., & Gilchrist, A. (1968). *An Evaluation of British Scientific Journals*. London: Aslib.

Neylon, C. (2014). Altmetrics: What are they good for? *PLOS Opens*. Retrieved from http://blogs.plos.org/opens/2014/10/03/altmetrics-what-are-they-good-for/

OCEBM Levels of Evidence Working Group. (2011). *The Oxford 2011 Levels of Evidence*. Oxford: Oxford Centre for Evidence-Based Medicine.

Patsopoulos, N. a, Analatos, A. a, & Ioannidis, J. P. A. (2005). Relative citation impact of various study designs in the health sciences. *JAMA : The Journal of the American Medical Association*, *293*(19), 2362–6. doi:10.1001/jama.293.19.2362

Priem, J. (2014). Altmetrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond bibliometrics: harnessing multi-dimensional indicators of performance* (pp. 263–287). Cambridge, MA: MIT Press.

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. *arXiv*, 1–17. Digital Libraries. doi:http://arxiv.org/abs/1203.4745v1

Röhrig, B., du Prel, J.-B., Wachtlin, D., & Blettner, M. (2009). Types of study in medical research: part 3 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, *106*(15), 262–8. doi:10.3238/arztebl.2009.0262

Rowlands, I., Nicholas, D., Russell, B., Canty, N., & Watkinson, A. (2011). Social media use in the research workflow. *Learned Publishing*, *24*(3), 183–195. doi:10.1087/20110306

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PloS One*, *8*(5), e64841. doi:10.1371/journal.pone.0064841

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., & Haustein, S. (2013). Tweeting links to academic articles. *Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics*, 1–8. Retrieved from http://cybermetrics.cindoc.csic.es/articles/v17i1p1.html

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? cross-disciplinary analysis of the presence of "alternative metrics" in scientific publications. *Scientometrics*. doi:10.1007/s11192-014-1264-0

Is There a Gender Gap in Social Media Metrics?

Adèle Paul-Hus[1], Cassidy R. Sugimoto[2], Stefanie Haustein[1], and Vincent Larivière[3]

[1] *adele.paul-hus@umontreal.ca; stefanie.haustein@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, H3C 3J7 Montreal, Qc. (Canada)

[2] *sugimoto@indiana.edu*
Indiana University Bloomington, School of Informatics and Computing, 1320 East 10th St., 47405 Bloomington, IN (USA)

[3] *vincent.lariviere@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, H3C 3J7 Montreal, Qc. (Canada) and Université du Québec à Montréal, Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Observatoire des Sciences et des Technologies (OST), CP 8888, Succ. Centre-Ville, H3C 3P8 Montreal, Qc. (Canada)

Abstract
The gender gap in science has been the focus of many analyses which have, for the most part, documented lower research productivity and citation impact for papers authored by female researchers. Given the rise of scholarly use of social media to disseminate scientific production and the healthy proportion of women on these sites, further investigation of potential gender disparities in social media metrics are warranted. Comparing event counts from Twitter, blogs, and news with citations, this study examines whether publications with male and female authors differ regarding their visibility on the social web and whether gender disparities can be observed in terms of social media metrics. Findings demonstrate increased gender parity using social media metrics than when considering scientific impact as measured by citations. It is acknowledged that this could be the results of the different impact communities, as the scientific community constituting the citing audience is more male-dominated than the social media environment. The implications for the use of social media metrics as measures of scientific quality are discussed.

Conference Topic
Altmetrics

Introduction
Early Internet use was heavily male-dominated—to the point of being considered a "boy toy" (Morahan-Martin, 1998; Weiser, 2000)—and promises of gender equity in computer-mediated communication were left unrealized (Herring & Stoerger, 2013). However, recent transformations in both the function and functionalities of the Internet have led to increased participation of women, particularly in the use of social networking sites (Kimborough et al., 2013). As of September 2014, slightly more women are using social networking sites than men (Duggan, Ellison, Lampe, Lenhart & Madden, 2015). However, although men and women now both employ social media, the ways in which they use them remain gendered (Correa, Hinsley, de Zuniga, 2010; Koenig, 2015; Muscanell & Guadagno, 2012; Piazza Technologies, 2015).

Twitter—an online social networking service for microblogging—is one of the top websites in the world (Alexa.com). However, despite equality in other social media sites, there appears to be a growing gender disparity in Twitter, with men using the platform at higher rates than women (24 vs 21%) (Duggan et al., 2015). Moreover, the gender gap in Twitter usage has been increasing in the last two years (Duggan & Brenner 2013; Duggan et al., 2015). Gender bias is also reflected by journalism's practices on Twitter, where reporters' tweets severely underrepresent women in quotes (Artwick, 2013). This speaks to women's underrepresentation as authorial voices—that is, voices that can speak as experts and

authority on matters of merit. Given the rise of scholarly use of Twitter (Costas, Zahedi & Wouters, 2014; Haustein, Costas & Larivière, 2015; Holmberg, Bowman, Haustein & Peters, 2014; Pscheida et al., 2013; Rowlands et al., 2011), further investigation of potential gender disparities in scholarly communication and measures of impact from this site are warranted.

Microblogging is not the only web space with demonstrated gender disparities. Given the underrepresentation of women in science (Larivière, Ni, Gingras, Cronin & Sugimoto, 2013; West, Jacquet, King, Correll & Bergstrom, 2013), many studies have sought to examine whether the web might provide a democratizing space for female academics. These studies have shown that men tend to have greater web presence than women (van der Weijden & Calero Medina, 2014) and blog at a greater rate (Puschmann & Mahrt, 2012; Shema, Bar-Ilan & Thelwall, 2012). Bar-Ilan and van der Weijden (2014) recently investigated whether gender specific differences could be found when considering Mendeley (a social bookmarking service) readership counts. Using the gender of one of the co-authors of astrophysics papers—a field where *hyperauthorship* is commonplace (Cronin, 2001), thus making it difficult to distinguish papers attributed to female researchers from male researchers—they showed that the share of papers, to which at least one male contributed were found more often on the platform that those to which at least one women contributed. On the other hand, women attract more profile view in Academia.edu (an academic social networking site) in certain disciplines (Thelwall & Kousha, 2014). Many of these social media sites are associated with less formal ways of discussing and sharing research results with a wider audience (Shema, Bar-Ilan & Thelwall, 2012; 2014). The degree to which this engagement is gender-neutral begs further investigation.

This study builds on these analyses and seeks to examine whether publications with male and female authors differ regarding their visibility on the social web, and whether gender disparities can be observed in terms of social media metrics. Comparing event counts from Twitter, blogs and news with citations, this study aims to answer the following research questions:

- Does the gender gap in scholarly communication observed for publications and citations extend to social media?
- Does the visibility of male and female authored papers differ among Twitter, blogs, and mainstream news media?
- Does the gender gap in social media visibility of scholarly journal articles differ by scientific discipline?

There has been a growing call for researchers to demonstrate social impact (e.g., Force 11, 2011; REF, 2014). Social media metrics have been promoted as a source of such impact measures (Priem, 2014). However, the degree to which gender inequalities exist on such platforms must be investigated prior to wide-scale adoption and use of social media metrics.

Methods

Data were drawn from Thomson Reuters' Web of Science (WoS), which includes the Science Citation Index Expanded, the Social Science Citation Index and the Arts and Humanities Citation Index. These databases index annually documents published in over 12,000 journals across all scholarly disciplines. To determine differences between scientific disciplines, the NSF field classification of journals (National Science Foundation, 2006) was used instead of WoS categories in order to avoid possible double counting of papers by classifying, as the NSF classification assigns each journal to only one specialty.

Only papers published in 2012 were considered, as this year provides the best compromise between the length of the citation window—citations to papers take time to accumulate—and the recent uptake of social media activity (Thelwall, Haustein, Larivière & Sugimoto, 2014). Citations to 2012 papers were counted until the end of 2013, which allows for a citation window of at least one complete year for all papers. Selecting 2012 publications also has the advantage of guaranteeing complete coverage of social media data for the whole year, as Altmetric.com started data collection mid-2011 (Costas, Zahedi & Wouters, 2014).

Altmetric.com was chosen as the data source for social media and mainstream media counts, as it is the most comprehensive source of social media data associated with scientific papers (Robinson-García, Torres-Salinas, Zahedi & Costas, 2014). News items, tweets and scientific blogs entries were selected for the analysis. Mainstream media and news sources captured by Altmetric.com include online mentions of scientific papers in more than 1,000 mainstream media and news outlets such as the Washington Post, Süddeutsche or CNN[1], giving insight on the visibility of a paper among the general public. The audience of Twitter and scientific blogs covered by Altmetric.com may reflect the overlap between the scientific community and the general public as both are widely used outside of academia but also by scholars. These metrics were selected because they represent three different types of social media events and levels of engagement from users, ranging from the one end of the spectrum with an engagement limited to 140 characters on Twitter, to the redaction of whole blog entries or newspaper articles, at the other end. Altmetric.com data includes counts collected up to August 2014. Given the quick uptake of social media-based indicators (excluding Mendeley) reported by Thelwall et al. (2014), we consider that the social media activity window of more than a full year considered in this study is long enough to cover the vast majority of social media activity around papers published in 2012.

The link between WoS papers and the Altmetric.com list of indicators was made using the Digital Object Identifier (DOI). Hence, papers that did not have DOIs were excluded from the analysis. As one might expect, the proportion of papers with DOIs is not distributed evenly across scientific disciplines. While, for most fields, the proportion of journals with publications with a DOI is very high (e.g., above than 70%), a substantial share of journals (30%), particularly in the Social Sciences and Humanities, do not use DOIs (Haustein, Costas & Larivière, 2015). Hence, for papers published in the latter group of journals, results from Altmetric.com are more likely to underestimate their actual online visibility, which represents a limitation of this study (as well as the great majority of social media metrics analyses). Arts and Humanities papers were thus excluded of the analysis because of the low number of papers and of citations. The gender of authors was attributed using the authors' given names, following the method developed in Larivière et al. (2013). The method allowed to assign a gender to the first author of 67.7% (N=696,186) of all 2012 papers that had a DOI (N=1,028,382). The analysis is, thus, based on this dataset of papers, and the gender of the first author is used to categorize the paper as female or male.

The prevalence of social media metrics is measured through intensity, which indicates the mean number of events for papers that show at least one of the particular events (non-zero counts) and coverage, percentage of papers with at least one event. While coverage reflects the probability of a document to be cited or mentioned on the particular platform, the intensity indicate rate aims to measure the frequency or popularity with which documents are (re)used once they are on the platform and remains independent of the coverage and zero values (Haustein, Costas & Larivière, 2015).

The scientific impact of male and female researchers is compared using the average of relative citations (ARC). The ARC provides a field-normalization and thus allows the

[1] http://www.altmetric.com/sources-news.php

comparison of citation impact between the different specialities that have otherwise different citation practices. More specifically, the number of citations received by a given paper is divided by the average number of citations received by articles in the same NSF research specialty published in the same year. An ARC greater than 1 indicates that an article is cited above the world average for the same field, and an ARC below 1 means that it is cited below the world average.

Results

Figure 1 compares the ARC of papers first authored by women and men, respectively, in order to assess whether a gender gap can be found in the dataset of papers used. Figure 1 confirms the widespread gender disparities observed in science (Larivière et al., 2013) in terms of scientific impact. More specifically, in each discipline, papers first authored by male researchers have higher citation impact, with the only exception of Engineering and Technology where papers first authored by female researchers have a slight advantage (ARC value of 1.18 for women and 1.17 for men). Biomedical Research (0.95 for women and 1.11 for men), Professional Fields (1.11 for women and 1.26 for men), Mathematics (1.03 for women and 1.19 for men) and Psychology (0.97 for women and 1.12 for men) show the greatest gender differences regarding citation impact.

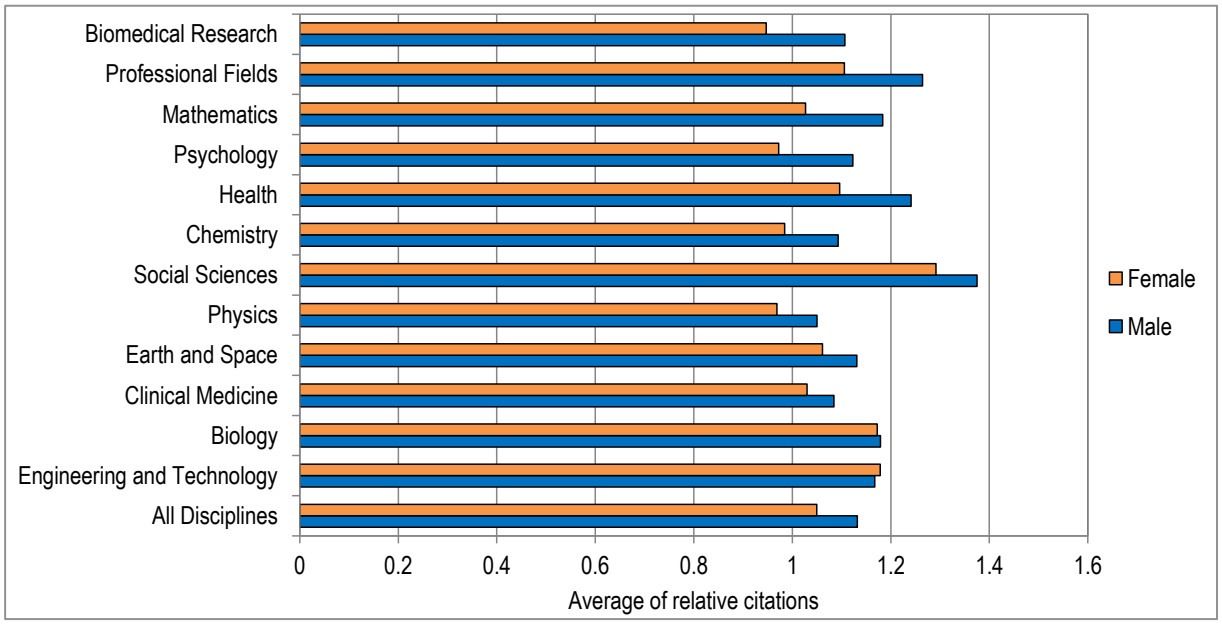


Figure 1. Average of relative citations of papers first authored by female and male researchers, by discipline and ordered by gender gap, 2012.

Figure 2 compares papers first authored by female and male researchers, in terms of intensity of news items (i.e., the mean number of events for all documents with at least one event) and coverage by news items (i.e., the percentage of papers with at least one event). All disciplines taken together, the intensity and the coverage of news items is gender-balanced, with an intensity difference of less than 0.07 event and a coverage difference of less than 1%. Physics (mean number of 1.04 for women and 1.34 for men) and Biomedical Research (1.63 for women, 1.87 for men) are the disciplines showing the strongest gender gap in terms of intensity of news items, in favour of papers first authored by men, corroborating the gender gap found in terms of citation impact (Figure 1). Coverage by news items of papers published in Biomedical Research (1.20% for women, 1.49% for men), Earth and Space (1.17% for women, 1.42% for men), Chemistry (0.59% for women, 0.84% for men) and Psychology

(1.26% for women, 1.50% for men) also confirm the gender gap found in terms of citation impact. However, papers first authored by female researchers in Health (1.32 for women, 1.26 for men), Clinical Medicine (1.39 for women, 1.33 for men) and Professional Fields (1.47 for women, 1.17 for men) have higher mean numbers of news items than that of male researchers while in Biology (0.73% for women, 0.62% for men), Engineering and Technology (0.60% for women, 0.55% for men) and Clinical Medicine (0.67% for women, 0.52% for men) they have a greater coverage.

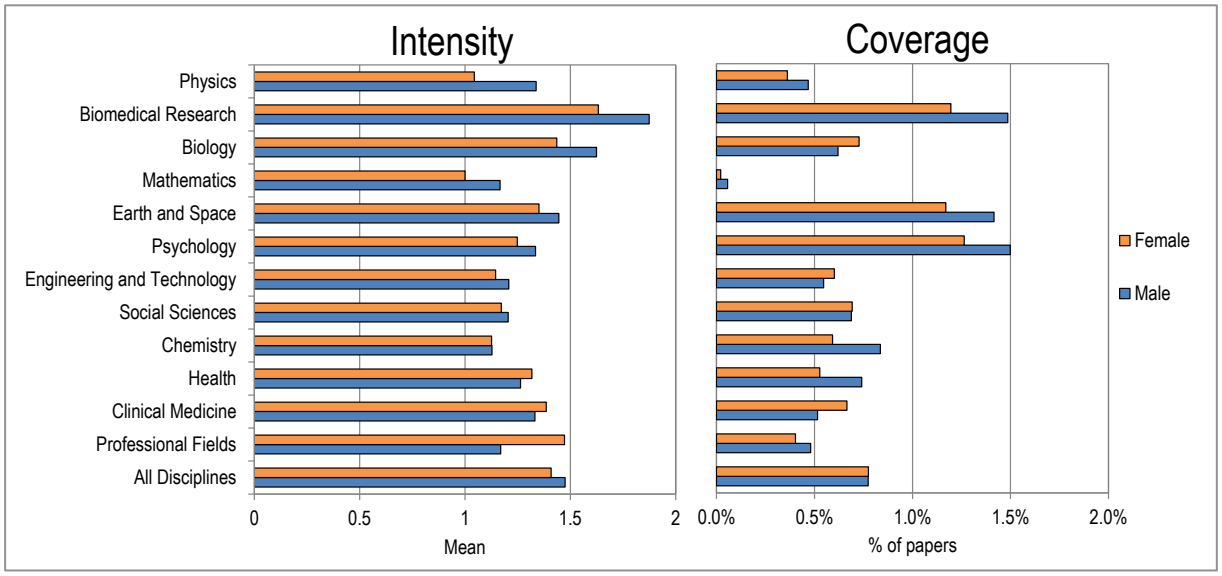


Figure 2. Intensity and coverage of news items of papers first authored by female and male researchers, by discipline, 2012.

Figure 3 provides the average numbers of tweets for all papers with at least one tweet (intensity for non-zero event items) and the percentage of papers with at least one tweet (coverage) by gender. It clearly shows that Twitter is the most popular platform among the three social media and mainstream media metrics analysed here, with an intensity of almost 3 tweets for papers tweeted at least once and coverage of almost 20% of papers (all genders and disciplines taken together). Gender analysis shows that, for all disciplines, papers first authored by female researchers are more intensely tweeted (2.98 tweets for women, 2.94 for men) and have a higher probability of being tweeted than papers first authored by male researchers (21% for women and 18% for men). Consistent with what has been found in terms of citations (Figure 1) and news items (Figure 2), Psychology and Biomedical Research show the highest gap in favour of men in terms of mean numbers of tweets.

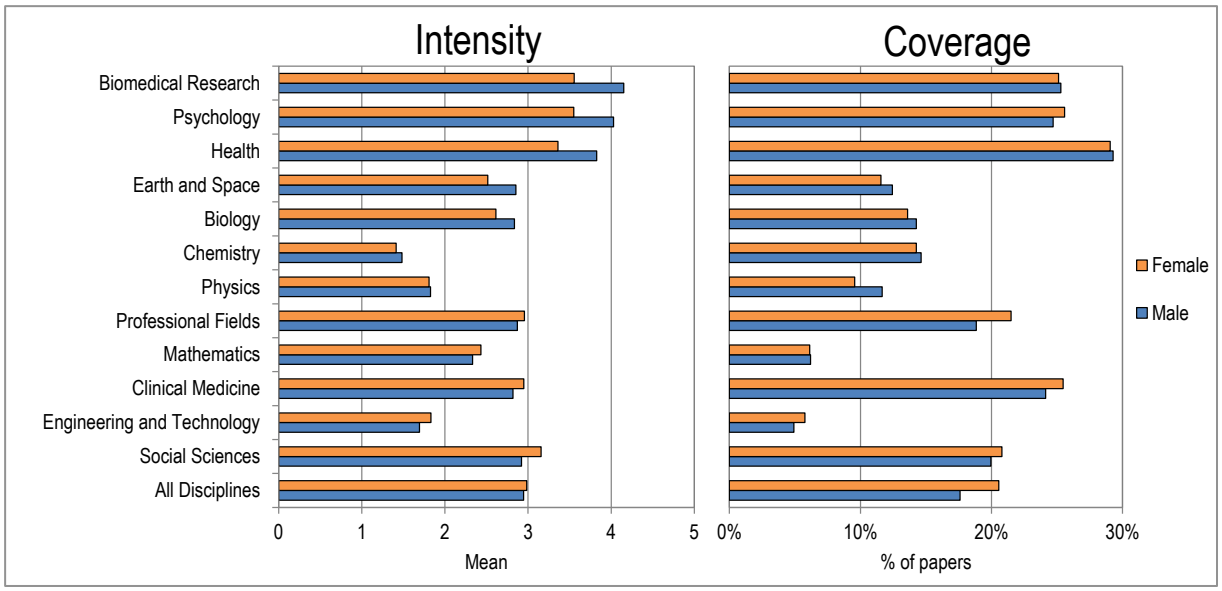


Figure 3. Intensity and coverage of tweets of papers first authored by female and male researchers, by discipline, 2012.

Figure 4 presents intensity and coverage by blog entries of papers first authored by women and men. All disciplines taken together, papers first authored by male researchers show a slightly higher intensity in terms of mean number of blog entries (1.33 for women, 1.40 for men) and higher coverage (1.68% for women, 1.78% for men). As previously shown, Psychology and Biomedical Research present important gender gaps, both in terms of intensity and coverage of blog entries. With respect to intensity, the average of blog entries of papers first authored by female and male researchers are equivalent in Health, Physics and Chemistry and papers authored by women have a slight advantage in Engineering and Technology. Papers authored by female researchers have stronger blog coverage in Clinical Medicine (1.30 % for women, 1.23% for men), Professionals Fields (1.08% for women, 1.02% for men) and Engineering and Technology (0.95% for women, 0.89% for men). However, the extreme gender gap in blog authors—both Puschmann and Mahrt (2012) and Shema, Bar-Ilan and Thelwall (2012) showed that about three quarters of bloggers where male—seems to transfer to the authors cited in blogs as confirmed by the coverage of papers authored by male researchers.

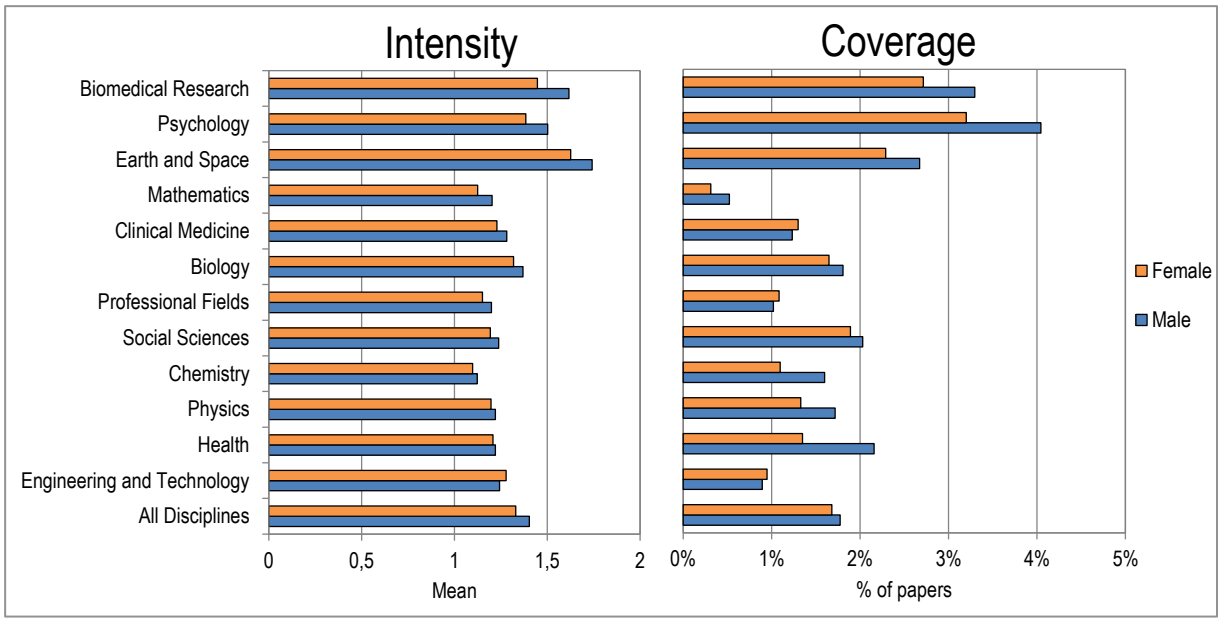


Figure 4. Intensity and coverage of blog entries of papers first authored by female and male researchers, by discipline, 2012.

Discussion and conclusion

Our findings demonstrate a more gender-balanced portrait when considering social media and mainstream media metrics (Figures 2 to 4), than when considering scientific impact as measured by citations (Figure 1). This could be explained by the fact that the impact communities contributing to these metrics are different: the scientific community which constitute the citing audience is more male-dominated than the social media environment (Kimbrough et al., 2013).

However, there is uniformity in the results neither by discipline nor platform. Coverage varied significantly by discipline, as did the mean impact score by gender. Furthermore, gender differences were found when examining microblogging, blogging, and news coverage. This suggests more information is needed before conclusive evidence on gender equality or inequality in social media metrics can be determined.

It could be argued that the diversity of the social media audience gives a broader audience an ability to respond to scholarly communication and therefore these measures of impact are a more honest metric of the absolute value of the work. However, lacking adequate validation of the meaning of social media metrics (Wouters & Costas, 2012), it is perhaps pre-emptive to make such a claim, as many tweets are actually made by bots (Haustein et al., in press). Further research on the nature of highly tweeted research will thus be necessary to assess the underlying mechanisms underneath the observed trends.

Acknowledgments

The authors acknowledge funding from the Alfred P. Sloan Foundation, grant no. 2014-3-25 and would like to thank Euan Adie and Altmetric.com for access to their data.

References

Artwick, C. G. (2014). News sourcing and gender on Twitter. *Journalism*, *15*(8), 1111–1127.

Bar-Ilan, J., & van der Weijden, I. (2014). Altmetric gender bias? Preliminary results. Presented at the 19th International Conference on Science and Technology Indicators, Leiden.

Correa, T., Hinsley, A. W., & de Zuniga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior*, *26*(2), 247–253.

Costas, R., Zahedi, Z., & Wouters, P. (2014). Do “altmetrics” correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, n/a–n/a.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569.

Duggan, M., & Brenner, J. (2013). *The Demographics of Social Media Users — 2012*. Retrieved from http://www.lateledipenelope.it/public/513cbff2daf54.pdf

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015, January). Social Media Update 2014. *Pew Research Center’s Internet & American Life Project*. Retrieved January 19, 2015, from http://www.pewinternet.org/2015/01/09/social-media-update-2014/

Force 11. (2011). *About Force 11*. Retrieved January 20, 2015, from https://www.force11.org/about

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (in press). Tweets as impact indicators: Examining the implications of automated bot accounts on Twitter. *To be published in Journal of the Association for Information Science and Technology*. Retrieved April 15, 2015, from http://arxiv.org/abs/1410.4139

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns. *PLoS ONE*, 10(3), e0120495. doi: 10.1371/journal.pone.0120495

Herring, S. C., & Stoerger, S. (2013). Gender and (A)nonymity in Computer-Mediated Communication. In J. Holmes, M. Meyerhoff, & S. Ehrlich (Eds.), *The Handbook of Language and Gender* (2nd ed., pp. 567–586). Hoboken, NJ: Wiley-Blackwell Publishing. Retrieved from http://info.ils.indiana.edu/~herring/herring.stoerger.pdf

Holmberg, K., Bowman, T. D., Haustein, S., & Peters, I. (2014). Astrophysicists’ Conversational Connections on Twitter. *PLoS ONE*, *9*(8), e106086.

Kimbrough, A. M., Guadagno, R. E., Muscanell, N. L., & Dill, J. (2013). Gender differences in mediated communication: Women connect more than do men. *Computers in Human Behavior*, *29*(3), 896–900.

Koenig, R. (2015, January 6). In STEM Courses, a Gender Gap in Online Class Discussions. *The Chronicle of Higher Education Blogs: Wired Campus*. Retrieved January 7, 2015, from http://chronicle.com/blogs/wiredcampus/in-stem-courses-a-gender-gap-in-online-class-discussions/55399#disqus_thread

Larivière, V., Ni, C. C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*(7479), 211–213.

Morahan-Martin, J. (1998). The Gender Gap in Internet Use: Why Men Use the Internet More Than Women—A Literature Review. *CyberPsychology & Behavior*, *1*(1), 3–10.

Muscanell, N. L., & Guadagno, R. E. (2012). Make new friends or keep the old: Gender and personality differences in social networking use. *Computers in Human Behavior*, *28*(1), 107–112.

National Science Foundation. (2006). Science and Engineering Indicators. Chapter 5: Academic Research and Development. Data and Terminology. Retrieved from http://www.nsf.gov/statistics/seind06/c5/c5s3.htm#sb1

Piazza Technologies. (2015). *STEM Confidence Gap*: *Piazza Blog*. Retrieved January 10, 2015 from http://blog.piazza.com/stem-confidence-gap/

Priem, J. (2014). Altmetrics. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Schorlarly Impact* (pp. 263–288). Cambridge, MA: MIT Press.

Pscheida, D., Albrecht, S., Herbst, S., Minet, C., & Köhler, T. (2013). *Nutzung von Social Media und onlinebasierten Anwendungen in der Wissenschaft. Erste Ergebnisse des Science 2.0-Survey 2013 des Leibniz-Forschungsverbunds „Science 2.0“*. Retrieved from http://www.qucosa.de/fileadmin/data/qucosa/documents/13296/Science20_Datenreport_2013_PDF_A.pdf

Puschmann, C., & Mahrt, M. (2012). Scholarly blogging: A new form of publishing or science journalism 2.0? In A. Tokar, M. Beurskens, S. Keuneke, M. Mahrt, I. Peters, C. Puschmann, & T. van Treeck (Eds.), *Science and the Internet* (pp. 171–182). Düsseldorf: Düsseldorf University Press.

REF. (2014). *REF 2014*. Retrieved from http://www.ref.ac.uk/about/

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: exploring the insides of Altmetric.com. *El Profesional de La Informacion*, *23*(4), 359–366.

Rowlands, I., Nicholas, D., Russell, B., Canty, N., & Watkinson, A. (2011). Social media use in the research workflow. *Learned Publishing*, *24*(3), 183–195.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2012). Research Blogs and the Discussion of Scholarly Information. *PLoS ONE*, *7*(5), e35869.

Shema, H., Bar-Ilan, J., & Thelwall, M. (2014). How is research blogged? A content analysis approach. *Journal of the Association for Information Science and Technology*, n/a–n/a.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE*, *8*(5), e64841.

Thelwall, M., & Kousha, K. (2014). Academia.edu: Social network or Academic Network? *Journal of the Association for Information Science and Technology*, *65*(4), 721–731.

van der Weijden, I., & Calero Medina, C. (2014). *Gender effects on evaluation indicators*. Leiden: CWTS ACUMEN Deliverable.

Weiser, E. B. (2000). Gender Differences in Internet Use Patterns and Internet Application Preferences: A Two-Sample Comparison. *CyberPsychology & Behavior*, *3*(2), 167–178.

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The Role of Gender in Scholarly Authorship. *PLoS ONE*, *8*(7), e66212.

Wouters, P., & Costas, R. (2012). *Users, narcissism and control – tracking the impact of scholarly publications in the 21st century.* SURFfoundation.

PubMed and ArXiv vs. Gold Open Access: Citation, Mendeley, and Twitter Uptake of Academic Articles of Iran

Ashraf Maleki

Malekiashraf@ut.ac.ir
Department of Library and Information Science, University of Tehran, Enqelab sq. Tehran (Iran)

Abstract
Despite contradicting evidence that open access (OA) articles might have greater citation advantage, there is less case studies in developing countries showing whether their global publication availability pattern advantages scientific impact metrics. Also, by addition of altmetrics to the world scientific evaluation system it is less known how different research access channels such as OA publishers, PubMed database and arXiv repository help altmetric indicators. Therefore, this paper investigates the case of WoS publications of Iran (2001-2012) for impact of mentioned publication availability models on citation, Mendeley readership, and tweet counts across four broader disciplines. Findings on 98,453 articles show that gold OA papers (5%) do not benefit significantly more metric counts, except in tweets linking to OA medical publications. Articles in PubMed Central (3%) significantly advantage the three investigated metrics, whereas arXiv preprints (2%) had higher readership advantage only. Different from PubMed publications, tweets to OA medical research were not significantly correlated with citations, suggesting their social impact rather than scientific. Additionally, OA publications are not significantly read by Mendeley users in developing countries, but developed ones, only in life science and biomedicine. Therefore, repository availability appears to be highly impactful in terms of citation and readership, whereas OA publications tend to receive rather high social impact through tweets.

Conference Topic
Altmetric

Introduction
Although traditional citation analysis helps countries to assess academic aspects of research impact and to fund them, so far wider aspects of impact including social and educational influence of research publications have been mainly ignored. However, by developing models of science assessment it seems that there will be better tools to assess influential aspects of research perhaps advantageous for public society rather than academic communities (Bornmann, 2012). Therefore, to improve aspects of wider impact, open access movement encourages researchers to make their research available online using various solutions. The open access (OA) availability of publications was a substantial addition to scholarly communication that enhanced science availability to a wider social audience and the researchers who had no access to subscription-based scientific data sources, especially those in developing countries (Contreras, 2012). With the advent of social networking sites and an access to free and open science, wider audience are now encouraged to publicly distribute science and give feedback about the scientific outputs. Extensive bookmarking of students and academics in research networks such as Mendeley (Mohammadi & Thelwall, 2013; Zahedi, Costas & Wouters, 2014; Haustein & Larivière, 2014) and prevalent reflection of the users' interest in online social networking sites such as Twitter (Haustein et al., 2013; Maleki, 2014) are evidence of wider impact of scientific publications beyond formal citations. Therefore, freely available publications not only advantage more citations (Lawrence, 2001; Gargouri et al., 2010; Laakso & Bjork, 2013), but also there is evidence they benefit from early reflection of impact in online media metrics in a way seemingly different from non-OA. In this respect, many of the top papers with higher altmetric scores in *Altmetric.com* were open access (Van Noorden, 2012). However, in spite of these evidence, there is less case

studies showing whether OA advantage is available for publishing pattern in developing countries, as in this research for Iranian WoS (Web of Science) publications.

The evidence suggests that developing countries have more OA journals than even some distinguished European countries (Bayry, 2013) and institutional repository growth since 2010 (Pinfield et al., 2014), however their journals are less internationally recognized or listed in scientific databases such as PubMed (Bayry, 2013). There are also barriers such as language, lack of knowledge about how OA publishing systems work (Salager-Meyer, 2014), and less funding for the researchers in these countries to contribute in high quality OA journals. Hence, it is less known how availability of their publications advantage citation and altmetric indicators. Therefore currents research aim to test OA impact on formal citations, Mendeley readerships and Twitter mentions (more below) to scholarly publications with Iranian authors, because this country in recent years had a rather noticeable scientific publication growth (e.g. Moin, Mahmoudi & Rezaei, 2005; Brown, 2011).

Furthermore, a fundamental challenge as Moed discussed (2012) is that along with OA journals (gold OA), self-archiving forms of publications (green OA) come a wide variety. There are about 80% of publishers that permit self-archiving (Laakso, 2014) in institutional homepages, subject repositories and web portals that excluding them might decline accuracy of OA advantage analyses (Moed, 2012). Amongst the online repositories, PubMed and arXiv have the highest web presence and impact according to Webometrics ranking (Cybermetrics Lab 2015, see more at http://repositories.webometrics.info), however it is less known how they advantage citations compared to OA journals, which is the subject of current research.

It is necessary to recognize the differences between OA journal and these repositories. PubMed refers to an important search engine for peer-reviewed medical research and has a significant role in research uptake in related fields, whereas arXiv is a preprint repository in *Cornell University* for self-archiving papers even before peer-review, mostly in physical sciences. The gold open access is a widespread solution across disciplines. However, a restricted number of publications in the world currently are published in journals with a free online version, as Harnad estimated gold open access articles about 5% in 2004; and without a considerable change in 2009, this proportion was 5.9% as covered in WoS (Laakso, 2009). However, there were better improvement in green OA reaching to about 12% in 2011 (Björk et al., 2014).

Among altmetric indicators, Mendeley readership and Twitter mentions to articles are known for their prevalent users (Thelwall et al., 2013; Zahedi, Costas & Wouters, 2014). However, evidently the two metrics are different in terms of aspects of impact. Majority of the online users in Mendeley are students (Mohammadi et al., in press; Zahedi et al., 2013; Haustein & Larivière, 2014), but in Twitter are the public audience (Maleki, 2014). They also are different from citation in terms of aspects like statistical distribution pattern (Thelwall & Wilson, in press; Eysenbach, 2011), and incidence, as tweets are fast and immediate (Eysenbach, 2011; Shuai et al., 2012) but Mendeley readerships and citations gradually increase. Also their prevalence is different, as tweets are linking to less publications than Mendeley readerships and citations (Thelwall et al., 2013). Thus, they individually reveal aspects of impact in different ways.

Background Literature

Citation advantage of open access publications

Various studies have reported that OA availability increases citation rate to articles in various fields. The premiere signs of OA citation advantage was reported from conference papers in computer science (Lawrence, 2001). More recently, Gargouri et al. (2010) found both self-selective self-archiving and mandatory self-archiving highly cited. In addition, Laakso and

Bjork (2013) observed that delayed OA policy for 2011 publications with about 78% available within the first year and about 85% within the two year after the publication, increased journal citation rate twice as much as non-OA journals and three times more than immediate OA journals.

In contrast, there are other studies that did not support a citation advantage for OA publications, some of them reviewed in Craig et al. (2007). Amongst more recent evidence Davis did several studies finding no OA citation advantage. He did a randomized control of 11 journals of American Physiological Society, finding no OA advantage after 9-12 month (Davis et al., 2008). His other study included 11 biology and medicine journals among which citations to OA articles fell from 32% in 2003 to 11% in 2007 (Davis, 2011). Gaule and Maystre (2011) also found 17% OA articles in PNAS during 2004 to 2006, where they found no OA diffusion advantage, but rather an author self-selection advantage after adjustment for confounders.

Studies report various evidence that online repositories increase citation advantage of articles, whereas subject repositories are more known to researchers than institutional ones (Cullen & Chawner, 2011). For instance, a study on articles in four math journals deposited in the arXiv indicated 35% more citation on average (Davis & Fromerth, 2007). Wren (2005) also showed that from both OA and non-OA journals with higher Journal Impact Factor (IF) over a third had OA reprints in non-journal websites of which over half had educational domains (.edu), providing a wider access to open research. Furthermore, Jeong and Huh (2014) showed that listing non-OA, non-Medline journals in the open access database of PubMed Central has over years led to an increase in their citation rate and impact factor in comparison with non-OA, non-listed journals.

Wider impact of open access publications

The OA publications were one of the premiere resources of online impact studies of scholarly publications, which revealed aspects of wider impact beyond traditional citations (Kousha & Thelwall, 2006; Vaughan & Shaw, 2007). For instance, Kousha and Thelwall (2006) studied URLs linking to OA publications of library and information science, which were demonstrative of 43% of their formal and 18% informal impact. In another study, Google Scholar unique citation to a sample of articles in 39 WoS OA journals in biology, chemistry, physics and computing was studied finding non-journal Google Scholar citations to OA publications indicator of their wider impact (Kousha & Thelwall, 2008). Other studies revealed usage advantage of online OA publications. Davis (2011) indicated that OA publications had more reader than subscription-based publications but not more citation advantage, for 89% more full-text downloads, 42% more PDF downloads, and 23% more unique visitors.

Only very recently a few studies compared altmetrics across OA publications. Adie (2014) reported that in the *Nature Communication* OA articles attract significantly more Mendeley readers and tweets. Also, Alhoori et al. (2015) displayed that OA papers have 60% more readers and 7% more tweets than non-OA, although non-OA articles were relatively highly covered in both Mendeley and Twitter.

Online Readership Impact assessment in Mendeley

The number of users who bookmarked publications in Mendeley reference sharing site is known as Mendeley readership metric for majority (55%) of users who add papers to their Mendeley libraries for reading or with the intention to read (Mohammadi, Thelwall & Kousha, in press). There is various evidence that Mendeley readerships can be indicative of scientific impact of research and predictor of correlates formal citations (Bar-Ilan, 2012; Thelwall, Haustein, Larivière & Sugimoto, 2013), moderately and weakly in social sciences,

and humanities, respectively (Mohammadi & Thelwall, in press) and strongly in many fields in medical research (Thelwall & Wilson, in press). Wang et al. (2014) reports correlations of Mendeley and citation in a range of 0.36 to 0.61 with 1% significance level in seven PLoS journals and increased html views in correlation with altmetric scores of the articles. A study on arXiv repository examined impact of European astrophysics preprints on Mendeley readerships, finding that 47% of the publications in Scopus are in arXiv, whereas there were more arXiv papers (40%) in Mendeley than Scopus publications (27%) (Bar-Ilan, 2013). Furthermore, Mendeley metric had larger correlation with citations and Journal Impact Factor (IF) than Faculty of 1000 article factors for Genomics and Genetics articles (Li & Thelwall, 2012).

Social Impact Assessment via Twitter mentions

Studies had shown that Twitter is a promising social media to examine social popularity of articles (Thelwall et al., 2013) where tweets linked to about 10% of 1.4 million PubMed articles; and were a fast metric to track comments on arXiv preprints (Shuai et al., 2012). In another study, Wee and Chia (2014) showed that among 20 highly cited WoS articles citations were significantly correlated with altmetric scores in some subject categories including general and internal medicine (Pearson correlation significant in 0.36 level), applied physics (0.39), sociology (0.49), literature (0.62), and music (0.67). The correlation turned out to be significant among articles with highest altmetric scores in multidisciplinary engineering (0.35) and communication (0.31), whilst majority of altmetric scores in various fields coming from Twitter mentions (65% to 89%) rather than Facebook (1% to 11%), news (0 to 19%), and blogs (2% to 11%). Current research is a further exploration into the previous study on Twitter uptake of WoS publications with Iranian authors (Maleki, 2014). The study suggested 5% of publications in 2011-2012 with positive Twitter mentions with the highest uptake was in life science and biomedicine (10%) where links were often created by public society rather than scientific communities (*ibid*).

Research Questions

1. The extent to which are OA, PubMed and arXiv publications by Iranian authors tweeted, read and cited?
2. How do readerships and tweets correlate with formal citations when studies are available through the three above channels across disciplines?
3. Do OA publications advantage more readers in developing countries than developed ones?

Method

As a follow-up study to the previous research on Twitter mentions (Maleki, 2014), the dataset is the same as in the previous research, confined to publications in 2001 to 2012. WoS citations are based on the data available from May 2013 for 98,455 articles with DOIs. Twitter mentions are available according to results in July 2013 through *Altmetric.com* - a subscription based altmetric data provider (see the reasons for choosing *Altmetric.com* in Maleki, 2014); Mendeley readerships are examined via DOI submission to *ImpactStory.org*, another subscription based altmetric data provider which was free at the time of gathering data, in July 2013. *ImpactStory.org* was used because it provided attributes of Mendeley users and because it was different from *Altmetric.com* which provided readers only if papers had social media buzz. However choosing *ImpactStory.org* it was possible to gather a sample of about 30,000 papers rather than all the data.

DOAJ (Directory of Open Access Journals), WOS and Scopus journal datasets are consulted for OA availability of journals and papers OA status is modified based on journals' *Start year* in DOAJ. Data about PubMed archival of the articles was gathered by using DOIs of the publications on the full publication dataset available from PubMed Central. Publications were available via PubMed across four broader research areas for 2,978 papers (3%) the most in life science and biomedicine (2132 papers, 7%). ArXiv preprints of papers were examined using arXiv API, via DOI submission. For this purpose a custom-built program was used to submit 100 DOIs each query to arXiv. The data from arXiv might be not accurate because DOIs are available in arXiv if the authors have provided them for the publications. Results showed that there was overall 489 publication with preprints in arXiv consisting 1.3% of physical science article in 2001 to 2012 and very small proportion in technology (0.1%).

As altmetrics are faster than WoS citations, to learn if tweet and Mendeley uptakes are predictive of later WoS citations the dataset is tested in two time periods. Therefore, an interval is required to be considered for the publications to provide the opportunity to get citations. In case of Twitter, because the reliable and available data is confined to the most recent years (2011 onwards) citations are checked for publications in 2011-2012 in two time intervals after the publication year, the first in July 2013 and the second in December 2014. In Mendeley the data from July 2013 for both recent and older publications could be reliably used, thus the data is compared for recent publications in 2011-2012 and for older publications in 2001-2010. A signed-rank Mann-Whitney test is used to examine differences in medians and means of counts for OA, PubMed and arXiv publication against their counterparts (non-OA, non-PubMed, non-arXiv, respectively) inside each publication period.

A zero inflated negative binomial regressions analysis model is used to assess whether citation, readership and tweet counts dependend on publication access channels. Therefore, articles available via open access journals, PubMed, and arXiv are individually taken as nominal explanatory dummy variables coded as 1, and all the other cases not available in the corresponding availability model coded as 0. The 0 is the reference variable, which is also redundant because OA, PubMed and arXiv are true for minority of the cases. The reason for choosing this model is the overdispersion in the counts or the exceeding variance of the three metric counts from their means.

The analyses were supplemented with users' nationality data on the Mendeley readership counts for the publications. The results are compared across development status of countries for difference in readership of OA, PubMed and arXiv articles in Mendeley. Some articles in Mendeley were recorded with multiple variations, to avoid duplicates the ones with higher readership counts were considered.

Results

The main results of study suggest that out of 98,453 articles in 2001-2012 which had DOIs, 4,772 articles (4.7%) were published in 449 (6%) gold OA journals. There also were 3,043 articles (3%) listed in PubMed Central and 1,489 articles (0.5%) with preprints in arXiv. The articles which were linked by at least one tweet appeared in 1,067 journals, among which there were 116 gold OA journals (11%), 202 journals (19%) with articles indexed in PubMed Central, and 55 journals (5%) with article preprints in arXiv. As mentioned in method a smaller set of publications (35% of all above) were tested for readerships including all articles in 2,522 journals, comprising 273 (11%) gold OA journals, 307 journals (12%) available in PubMed list, and 56 journals (2%) with preprints in arXiv.

The OA journal *PLoS One* with 102 articles all available via PubMed Central had the most articles with tweets (36 papers) and readership counts (83 papers). The following two checked journals with articles available via PubMed with more articles in Mendeley were *Journal of Assisted Reproduction and Genetics* (48 out of 63 papers with readership, and 2 tweeted

papers) and *International Journal of Nanomedicine* (38 out of 47 papers with readership, and 3 tweeted papers). Additionally, the results suggested that tweets link to more articles with preprints in arXiv in the journals *Astrophysics and Space Science* (with 35 tweeted articles and only 20 with preprints in arXiv), *Physical Review D* (27 tweeted articles whereas 75 with preprints in arXiv), and *Physical Review E* (17 tweeted articles, 27 preprints in arXiv) both former journals in astronomy and astrophysics and the latter one in soft-matter physics. However, there were journals with many papers in Mendeley, but poorly available preprints in arXiv; for instance there were 54 articles with readership counts in *International Journal of Theoretical Physics* out of 249 articles whereas only 6 with preprints in arXiv. Other OA journals with numerous articles with both citations and readerships, were *Analytical Science* (84 with readership and 116 with citations out of 118 papers) and *Molecules* (51 articles with readerships and 81 with citations out of 93 and 2 tweeted articles.

Table 1. Spearman correlation between Mendeley readership counts and WoS citations across years in terms of four broader research areas and of OA, PubMed, and arXiv availabilities of articles.

Disciplines / Availability		2012	2011	2010	2009	2008	2001-2007
Life science and biomedicine	OA[a]	.314** 218	.364** 209	.378** 120	.415** 96	.337** 96	.388** 87
	NOA[b]	.236** 1402	.274** 1317	.275** 1020	.296** 803	.339** 609	.302** 1157
	PubMed	.371** 100	.325** 176	.460** 109	.486** 85	.358** 75	.552** 42
	Non-PubMed	.258** 568	.204** 959	.220** 854	.279** 708	.309** 570	.296** 1202
Physical sciences	OA	.159 94	.060 85	.060 76	.194 49	-.016 29	.057 119
	NOA	.293** 838	.229** 816	.237** 691	.275** 539	.282** 470	.167** 1216
	arXiv	.217 35	.291 42	.418* 25	-193 20	-.232 23	-.232 23
	Non-arXiv	.220** 397	.187** 677	.248** 652	.236** 525	.220** 470	.156** 1333
Technology	OA	.160 39	.403 15	-.019 13	-.189 11	-.315 7	.173 19
	NOA	.154** 840	.259** 833	.325** 702	.289** 609	.328** 349	.358** 75
Social sciences and humanities	OA	.304* 52	.188 31	.266 9	.947* 5	.500 3	.293** 4482
	NOA	.363** 56	.259 33	.454* 26	.061 19	.815** 14	.462** 39

Correlation between altmetrics and citations in terms of availability models

Tables 1 and 2 show the correlation between Mendeley readerships and tweets with citations. The readerships of OA articles in life science and biomedicine are appropriately in moderate correlation with citations, and likewise, PubMed publications are correlated, but in stronger levels (correlation coefficients ranging from 0.31 to 0.55). However, the correlations in non-OA and non-PubMed papers are in lower levels (ranging from 0.20 to 0.34) - all correlations are significant in $p < 0.001$. This advantage were not available for the other three broader research areas, where the correlations were significant about non-OA publications rather than OA. The findings suggest that readership of publications with scientific impact have enhanced over years by OA and PubMed availability of life science and biomedicine articles, since older publications are in stronger correlation with citations than newer ones, although they are less numerous.

The figures in Table 2 suggest that there is a weak and significant correlation between tweets and later WoS citations in life science and biomedicine and physical sciences. Different from PubMed articles, tweet to OA publications did not have significant correlation with citations, perhaps for their social impact rather than scientific. On the other hand, correlations between tweets and citations are usually weak and significant after the interval for articles to receive citations in life science and biomedicine (correlations ranging from 0.07 to 0.17 significant in $p < 0.01$) and physical sciences (correlation significant in 0.13, $p < 0.001$). Correlations in all the fields does not show an OA advantage. Instead, there were weak and significant correlation in PubMed and non-OA publications in life science and biomedicine, and non-arXiv and non-OA articles in physical sciences after the interval.

Table 2. Spearman correlation between Twitter mentions and WoS citations in 2011-2012 in terms of four broader research areas and of OA, PubMed, and arXiv availabilities of articles.

Research areas / availability model		2012 Early citation[c]	2012 Later citation[d]	2011 Early citation	2011 Later citation
Life science and biomedicine	OA[a]	.015	.131	.071	.209
		159	159	74	74
	NOA[b]	.072*	.063	.059	.153*
		801	801	256	256
	PubMed	.087	.169*	.002	.143
		200	200	92	92
	Non-PubMed	.049	.034	.056	.147*
		760	760	238	238
Physical sciences	OA	.090	.094	-.178	-.045
		41	41	10	10
	NOA	.074	.130**	.054	.068
		405	405	86	86
	arXiv	-.001	-.009	7	7
		28	28		
	Non-arXiv	.078	.126**	.011	.024
		418	418	89	89
Technology	OA	10	-.048	2	2
			10		
	NOA	-.023	.131	-.017	-.130
		135	135	51	51
Social sciences and humanities	OA	.500	.866	1	1
		3	3		
	NOA	.521**	.345		-.487
		25	25	6	6

[a]. OA: Open Access; [b]. NOA: Non-Open Access; [c]2012 Early citations: citations to 2012 publications in July 2013; [d] citations to 2012 publications in Dec. 2014; Correlation significant at the 0.05 level (*);0.01 level (**).

Metrics dependencies to OA, PubMed and ArXiv publications

As figures in Table 3 show, tweeted gold OA publications (301 papers, 0.8%) are less than non-OA (1,975, 4.4%), whereas in fact more OA articles (11% of all OAs) tend to be tweeted than non-OA (5% of all non-OAs). This happens across the four broader fields with the highest occurrence in life science and biomedicine (15% OA vs. 10% non-OA). Also, findings suggest that tweets tend to link to significantly more PubMed publications in life science and biomedicine (24%), whereas this proportion is higher than tweeted OA publications (15%). The same is observed in physical sciences where arXiv preprints (55%) tend to receive tweets more than OA articles (7%). A Mann-Whitney test suggests that tweets to arXiv (206 tweets to 136 articles) were not significantly more than tweets to publications without arXiv preprint (472 tweets to 406 papers).

Also, tweets to PubMed (1,118 tweets to 293 papers vs. 2,105 tweets to 972 non-PubMed papers), and OA articles in life science and biomedicine (778) are significantly higher than their relative counterparts (i.e. non-PubMed and non-OA, respectively) ($p<0.001$). There were no significant difference between tweets to OA and non-OA in other fields, however. Additionally non-OA publications significantly advantage more citations to tweeted articles in 2011-2012 either in the early stage after publication (3.8 mean tweets to non-OA vs 1.6 tweets to OA) or later stage (10.3 vs. 6). This observations is in line with the correlations above which were significant in cases the publications were non-OA rather than OA in all fields excluding life science and biomedicine.

Table 3. Mean and median tweets and citations to articles with at least one tweet across publication availability models.

		Median Mean					
Source/Publication year		OA	Non-OA	PubMed	Non-PubMed	arXiv	Non-arXiv
Twitter mentions	2011-2012	1 2.9**	1 2.0	2 3.6**	1 1.8	1 1.3	1 1.2
Early citations	Jul. 2013	0 1.6	1 3.8**	1 5.1	1 3.2	1 1.5	1 3.5
Later citations	Dec. 2014	3 6.0	4 10.3**	4 13.1	4 9.1	4 6.7	4 9.7
Total articles	2011-2012	315 (14%)	1975 (84%)	336 (15%)	1954 (85%)	35 (6%)	532 (94%)

*significantly more than its counterpart category (OA vs. non-OA; PubMed vs. non-PubMed; arXiv vs. non-arXiv) in $p < 0.01$ level.**significantly more than its counterpart category in $p < 0.001$ level.

Table 4. Mean and median readerships and citations to articles with at least one Mendeley readership across publication availability models.

		Median Mean					
Source/Publication year		OA	Non-OA	PubMed	Non-PubMed	arXiv	Non-arXiv
Mendeley readers	2001-2010	3 4.3	2 4.1	3 6.9**	1 2.4	5 6.1*	2 3.5
	2011-2012	2 4.2	2 3.3	3 5.8**	2 3.2	4 5.3**	2 2.8
Later citations	2001-2010	5 9.0	5 8.9	4 11.4	4 7.5	9 12.6	6 10.7
Early citations	2011-2012	2 3.0	2 3.3**	1 2.2	1 1.8	1 2.2	1 2.3
Total articles	2001-2010	737 (8%)	8850 (92%)	374 (4%)	9213 (96%)	10 (0.3%)	3211 (99.7%)
	2011-2012	743 (11%)	6079 (89%)	558 (8%)	6264 (92%)	75 (4%)	1758 (96%)

*significantly more than its counterpart category (OA vs. non-OA; PubMed vs. non-PubMed; arXiv vs. non-arXiv) in $p < 0.01$ level.**significantly more than its counterpart category in $p < 0.001$ level.

Table 4 shows proportion of publication with positive Mendeley readership 5% OA (1,480 papers) and 52% non-OA (14,929 papers), with the highest article uptake in life science and biomedicine (9% OA and 61% non-OA) and the least in physical sciences (4% OA and 44% non-OA). Further results show that users tend to read non-OA publications (58%) rather similar to OA (55%) while there is no significant difference in their readership patterns across four broader research areas. However, despite in less papers than OA, PubMed publications (932 papers) tend to have higher readerships (5,566 PubMed vs. 4,675 OA readerships), with

the highest occurring in life science and biomedicine (for 76% PubMed vs. 67% OA papers) ($p < 0.05$). The same was seen in arXiv preprints as their read articles (85 papers) tend to have significantly higher readership counts than non-archive ($p < 0.01$).

The OA publications in the two time periods (8% in 2001-2010 and 11% in 2011-2012) are more than PubMed (4% and 8%) and arXiv (0.3% and 4%). The mean PubMed readerships were significantly more than non-PubMed for the publications in older time period of 2001 to 2010 (6.9 PubMed vs. 2.4 non-PubMed) and for articles in 2011-2012 (5.8 vs. 3.2) ($p < 0.001$). ArXiv preprints in Physical science on average also had higher readerships than non-arXiv in both publication periods (significant in $p < 0.01$ in 2001-2010 and $p < 0.001$ in 2011-2012). There were no significant *citation* advantage for OA, PubMed and arXiv papers with Mendeley readerships, neither in the early nor the later stage after the publication year in none of the four research areas, although non-OA publications in social science and humanities and life science and biomedicine had significantly more readerships than OA.

Table 5 shows results of zero-inflated negative binomial regression analysis. The significance of alpha values in Table 5 identifies overdispersions for the three metrics. Voung statistics being above the critical value of 1.96 approves the overdispersion and the need for the zero inflated method. The estimates of the regression coefficients are shown by the values b and the estimated standard errors are the ratios of the coefficients. Therefore, b values show how much the availability of the articles by various models increases metric counts.

The results in Table 5 suggest that PubMed articles significantly advantage the three metric counts. However, (gold) open access were not significant indicator of neither citations nor the two altmetric counts. In addition, publications with preprints in arXiv had significantly more readership counts only.

Table 5. Zero inflated negative binomial regression analysis for citations, readerships and Twitter mentions by variables of availability channels.

	Citations (2001-2012)		Mendeley Readerships (2001-2012)		Tweets (2011-2012)	
Variables	*b*	*Standard error*	*b*	*Standard error*	*b*	*Standard error*
Open Access	-0.26**	0.02	-0.30**	0.03	-0.39**	0.09
PubMed	0.14**	0.03	0.79**	0.04	0.96**	0.09
ArXiv	-0.64**	0.06	0.37*	0.09	-0.21*	0.09
Constant	2.06**	0.01	1.31**	0.01	0.64**	0.03
Alpha	1.05	0.01**	0.58	0.01**	0.52	0.02**
Vuong Statistics	330.9**		254.9**		64.79**	
Log Likelihood	-200924.8		-39551.92		-3830.57	
Rest Log Liklihood $\chi 2$ (3)	229.2**		579.1**		181.59**	
Publications	98,454		28,758		39,119	

Publication readership across countries development status

An important limitation of statistics about nationality attributes of users is that Mendeley suggests only top three countries with higher readership counts per paper. Based on these data, users were recognized from 141 countries, including 28,966 readerships from developed countries for 16,472 papers and 21,848 readerships from developing countries for 12,699 papers. Median readerships were more in papers with readers from developed countries rather than developing ones (4 vs. 3 readers per paper). The OA life science and biomedicine publications (excluding other field) had significantly more readers in developed countries ($p<0.05$). PubMed publications also had significantly more readerships in developed countries than developing ones ($p<0.001$), whereas there were no such difference about readership of arXiv preprints. In addition, users in developing countries significantly read more non-OA

articles in technology (3,483, mean users = 1.77 vs. 1.68) and physical sciences (3,628, mean users = 1.73 vs. 1.66) ($p<0.001$). All tests were significant in a signed-rank Mann-Whitney test.

Discussions

A main limitation in this research is that it does not include other potential sources of publication availability such as homepages and institutional repositories and social networking sites for self-archiving. Also, a problem may associate with the regression analysis for which the research is very optimistically focused on direct impact of publication access patterns, whereas results might be affected by other correlates of the metrics such as Journal Impact Factor or Immediacy Index. Therefore, designing more complex models for assessment of availability impact might be subject of future studies.

Regarding the first research question results suggest that there are more OA articles (5%) than PubMed listed articles (3%) and arXiv preprints (2%). Also, there are more OA publication with readers (9%) than PubMed (6%) and arXiv (2%), whereas tweets link to relatively more PubMed (15%) papers than OA (14%) and arXiv (6%). Regarding the second question of research there were a significant correlation between tweets and citations to PubMed articles, indicating their scientific impact. However, tweeted OA publications seem to be reflective of social impact rather than scientific since they do not appear correlated with citations neither in early nor later year. In addition, publications in 2012 are more correlated than 2011, suggesting an overtime increasing publication uptake via tweets. A moderately significant and across years decreasing correlation between readerships and citations to OA and PubMed availability of articles in life science and biomedicine (excluding other fields) suggest that older publication had the opportunity to get higher citations.

The mean tweets to both OA (3.3) and PubMed (3.7) life science and biomedicine papers were significantly more than non-OA and non-PubMed, respectively. These publication strategies have obviously enhanced various aspects of research impact. The difference between the mean tweets to arXiv preprints (1.3) and non-arXiv physical science papers (1.2) is statistically significant, however these tweets are very low and does not reflect an aspects of impact, while generally arXiv papers are regularly tweeted for classification and dissemination purposes. The finding from previous study supports this, as papers in physical science are mainly tweeted by subject specific tweeters for classificatory reasons rather than scientific or social impact (Maleki, 2014). In contrast to OA advantage on Twitter mentions of articles (only in life science and biomedicine), Mendeley readerships was not significantly different across gold open access and non-OA publications in the four field.

The regression models for the three metrics also had results in line with the results from previous section. There is a significant citation advantage only for PubMed publications. Both PubMed and arXiv papers advantage Mendeley readerships. The only difference is in tweets where similar to above results show significantly more tweets to PubMed publications, however unlike the above non-OA advantage significantly more tweets than OA, which shows the effect of other hidden variables.

The expected higher readership of OA papers in developing countries failed to be true. A noteworthy result suggests that Iranian OA medical publication readerships by developed countries were significantly higher than developing countries, whereas this connection was vice versa in technology and physical sciences for non-OA articles. This can be connected to development and competitive abilities in research in these areas and/or the distribution of Mendeley users in various fields across countries. In this respect, the inferences need to be made with caution. However, it seems that Iranian medical research tend to get higher uptake by developing countries by appearing in PubMed index.

Conclusions

An important result of the study suggests that PubMed and arXiv strategies of publication availability can enhance the metric counts especially Mendeley readerships. Citations were mainly influenced by PubMed availability of broader field of life science and biomedical research, whereas tweets mainly link by publications available via gold OA journals. Furthermore, nationality of Mendeley readers appear to be informative about publication uptake patterns worldwide. Also, regarding results in this research with the ones from previous study on tweets it seem that Twitter has the potentials to reflect social impact of medical research for which OA availability and PubMed will help. In addition, subject repositories get higher readerships and tweets chance than papers out of them. Future studies might bring more variables associating these metrics for more realistic look at OA advantage in publication and research impact assessment.

Acknowledgment

The author would like to thank Dr. Kayvan Kousha of the Statistical Cybermetrics Research Group, in University of Wolverhampton, for his inspirations and very useful comments.

References

Adie, E. (2014). Attention! A study of open access vs non-open access articles. *Figshare*, doi:http://dx.doi.org/10.6084/m9.figshare.1213690.

Alhoori, H., Ray Choudhury, S., Kanan, T., Fox, E., Furuta, R., & Giles, C. L. (2015). On the Relationship between Open Access and Altmetrics. *iConference 2015 Proceedings*.

Bar-Ilan, J. (2013). Astrophysics publications on arXiv, Scopus and Mendeley: a case study. *Scientometrics*, 1-9.

Bayry, J. (2013). Journals: Open-access boom in developing nations. *Nature, 497*(7447), 40-40.

Björk, B.-C., Laakso, M., Welling, P., & Paetau, P. (2014). Anatomy of green open access. *Journal of the Association for Information Science and Technology, 65*(2), 237-250, doi:10.1002/asi.22963.

Bornmann, L. (2012). Measuring the societal impact of research. *EMBO reports, 13*(8), 673-676.

Brown, M. (2011). China, Turkey and Iran emerge as scientific giants. Wired.co.uk. Accessed 2 April 2014.

Contreras, J. L. (2012). Open access scientific publishing and the developing world. *St. Antony's International Review, 8*(2012), 43-69.

Craig, I., Plume, A., McVeigh, M., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact?A critical review of the literature. *Journal of Informetrics, 1*(3), 239-248, doi:10.1016/j.joi.2007.04.001.

Cullen, R., & Chawner, B. (2011). Institutional repositories, open access, and scholarly communication: a study of conflicting paradigms. *The Journal of Academic Librarianship, 37*(6), 460-470.

Davis, P. M. (2010). Access, readership, citations: A randomized controlled trial of scientific journal publishing. (Vol. 0723, pp. n/a-n/a).

Davis, P. M. (2011). Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *The FASEB Journal, 25*(7), 2129-2134, doi:10.1096/fj.11-183988.

Davis, P. M., & Fromerth, M. J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics, 71*(2), 203-215, doi:10.1007/s11192-007-1661-8.

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. (2008). Open access publishing, article downloads, and citations: randomised controlled trial. [Randomized Controlled Trial Research Support, Non-U.S. Gov't]. *BMJ, 337*, a568, doi:10.1136/bmj.a568.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research, 13*(4).

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., et al. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PloS one, 5*(10), e13636, doi:10.1371/journal.pone.0013636.

Gaulé, P., & Maystre, N. (2011). Getting cited: Does open access help? *Research Policy, 40*(10), 1332-1338, doi:10.1016/j.respol.2011.05.025.

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., et al. (2004). The Access/Impact Problem and the Green and Gold Roads to Open Access. *Serials Review, 30*(4), 310-314, doi:10.1016/j.serrev.2004.09.013.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology, 65*, 656-669, doi:10.1002/asi.23101.

Jeong, G.-H., & Huh, S. (2014). Increase in frequency of citation by SCIE journals of non-Medline journals after listing in an open access full-text database. *Sci Ed, 1*(1), 24-26, doi:10.6087/kcse.2014.1.24.

Kousha, K., & Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. *Scientometrics, 68*(3), 501-517.

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics, 74*(2), 273-294, doi:10.1007/s11192-008-0217-x.

Laakso, M., Welling, P., Bukvova, H., Nyman, L., Björk, B.-C., & Hedlund, T. (2011). The Development of Open Access Journal Publishing from 1993 to 2009. *PloS one, 6*(6), e20961, doi:10.1371/journal.pone.0020961.

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature, 411*(6837), 521-521.

Maleki, A. (2014). *Twitter Users in Science Tweets Linking to Articles: The Case of Web of Science Articles with Iranian Authors.* Paper presented at the American Society for Information Science and Technology, presented at SIG/MET post conference workshop, Seattle, USA,

Moed, H. (2012). Does open access publishing increase citation or download rates. *Research Trends, 28*.

Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015). Who Reads Research Articles? An Altmetrics Analysis of Mendeley User Categories. *Journal of the Association for Information Science and Technology,* doi: 10.1002/asi.23286.

Mohammadi, E., Thelwall, M., & Kousha, K. (in press). Can Mendeley Bookmarks Reflect Readership? A Survey of User Motivations. *Journal of the Association for Information Science and Technology*, doi: 10.1002/asi.23286.

Moin, M., Mahmoudi, M., & Rezaei, N. (2005). Scientific output of Iran at the threshold of the 21st century. *Scientometrics, 62*(2), 239-248.

Mounce, R. (2013). Open access and altmetrics: Distinct but complementary. *Bulletin of The American Society for Information Science and Technology, 39*(4), 14-17.

Pinfield, S., Salter, J., Bath, P. A., Hubbard, B., Millington, P., Anders, J. H. S., et al. (2014). Open-access repositories worldwide, 2005–2012: Past growth, current characteristics, and future possibilities. *Journal of the Association for Information Science and Technology, 65*(12), 2404-2421, doi:10.1002/asi.23131.

Salager-Meyer, F. (2014). Writing and publishing in peripheral scholarly journals: How to enhance the global influence of multilingual scholars? *Journal of English for Academic Purposes,* 13, 78-82.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations. *PloS one,* 7, doi:10.1371/journal.pone.0047523.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web. *PloS one,* 8(5), doi:10.1371/journal.pone.0064841.

Thelwall, M., & Wilson, P. (in press). Mendeley Readership Altmetrics for Medical Articles: An Analysis of 45 Fields. *Journal of the Association for Information Science and Technology*, doi: 10.1002/asi.23501.

Vaughan, L., & Shaw, D. (2007). A new look at evidence of scholarly citation in citation indexes and from web sources. *Scientometrics, 74*(2), 317-330, doi:10.1007/s11192-008-0220-2.

Wang, X., Liu, C., Fang, Z., & Mao, W. (2014). From Attention to Citation, What and How Does Altmetrics Work? *arXiv preprint arXiv:*1409.4269.

Wren, J. D. (2005). Open access and openly accessible: a study of scientific publications shared via the internet. [Research Support, U.S. Gov't, Non-P.H.S.]. *BMJ, 330*(7500), 1128, doi:10.1136/bmj.38422.611736.E0.

Zahedi, Z., Costas, R., & Wouters, P. (2013). What is the impact of the publications read by the different Mendeley users? Could they help to identify alternative types of impact? *Proc. PLoS ALM Workshop.*

Alternative Metrics for Book Impact Assessment: Can *Choice* Reviews be a Useful Source?

Kayvan Kousha[1] and Mike Thelwall[2]

[1] *k.koushal@wlv.ac.uk*

Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY (UK)

[2] *m.thelwall@wlv.ac.uk*

Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY (UK)

Abstract

This article assesses whether academic reviews in *Choice: Current Reviews for Academic Libraries* could be systematically used for indicators of scholarly impact, uptake or educational value for scholarly books. Based on 451 *Choice* book reviews from 2011 across the humanities, social sciences and science, there were significant but low correlations between Choice ratings and citation and non-citation impact metrics. The highest correlations found were with Google Books citations (.350) in science and with WorldCat library holdings counts in the humanities (.304). Books recommended by Choice reviewers for undergraduates were mentioned more often in online course syllabi than were other recommended books. Similarly, books recommended for researchers, faculty members and professionals or graduates tended to receive more Google Books citations than did books recommended for undergraduates. In conclusion, metrics derived from Choice academic book reviews can be used as indicators of different aspects of the value of books but more evidence is needed before they could be used as proxies for peer judgements about individual books.

Conference Topic

Webometrics; Altmetrics

Introduction

Impact assessment in book-based subject areas is more challenging than for article-oriented fields because the major current citation indexes are dominated by academic journal articles, and are therefore inadequate for assessing the research impact of books (Hicks, 1999, Archambault, Vignola-Gagné, Côté, Larivière, & Gingras, 2006, Nederhof, 2006; Huang & Chang, 2008). In recognition of the need to include citations from books (Garfield, 1996), the Thomson Reuters Book Citation Index (BKCI) and Scopus now index selected books, but their coverage seems to be too low to make a difference for impact assessment and they are restricted to just a few publishers and books that are mainly in English (Torres-Salinas et al., 2014). The way that the books are indexed also creates other issues for book impact assessment (Leydesdorff & Felt, 2012; Gorraiz, Purnell, & Glänzel, 2013).

Another important issue is that some academic books, such as textbooks and introductory science books, are primarily written for teaching (Gurung, Landrum, & Daniel, 2012) and other books, such as novels and literary works, may have cultural influence (White, Boell, Yu et al., 2009) or play a public engagement role (Kousha & Thelwall, in press). Moreover, education may be seen as particularly important in the humanities and a core part of its value to society (e.g., Nussbaum, 2012). All of these are unlikely to be reflected by citation counts. Peer review can be used to evaluate the impact of books but it is time-consuming. For instance, in some book-based fields (e.g., history and law) in the 2008 UK Research Assessment Exercise (RAE) reviewers had to assess the research merits of up to 100 books each (Kousha, Thelwall, & Rezaie, 2011). Hence not all of the submitted books may have been examined in detail (Taylor & Walker, 2009). Peer review is also subjective, perhaps

most strongly in the humanities where books are most common. Although critical evaluation is a core skill in the humanities (Small, 2013), it also seems to thrive on controversy and disagreements (Bauerlein, 2002). Moreover, the opinions of reviewers could be more subjective about the teaching or cultural benefits of books than about their research contributions (Weller, 2001).

In response to the weakness of citations for book impact assessment, there have been attempts to assess wider impacts of books (see below), using scholarly book reviews, library holdings statistics, and publisher prestige as well as with altmetrics (Priem, Taraborelli, Groth, & Neylon, 2011). Book reviews are somewhat similar to post-publication reviews for academic articles in systems like Faculty of 1000 (Hunter, 2012; Li & Thelwall, 2012; Mohammadi & Thelwall, 2013; Waltman & Costas, 2014), and both could be useful as additional quality control mechanisms for the critical analysis of published works (Crotty, 2012). The current study explores an alternative source for book impact assessment, *Choice: Current Reviews for Academic Libraries*, which is owned by the American Library Association, and compares it with citation and non-citation metrics. Choice has published reviews of academic books by editors, experts and librarians across different subject areas for about 50 years and is therefore a substantial and successful source of book reviews aimed at librarians making library purchasing decisions. Despite publishing about 7,000 book reviews per year that are relevant to academic libraries, it appears to be an untapped resource in terms of book impact assessment.

Metrics for Book Impact Assessment

Citation Metrics

Web of Science (WoS) and Scopus: Citations to books can be manually extracted from article reference lists (e.g., Cullars, 1998; Krampen, Becker, Wahner & Montada, 2007) or through cited reference searches in WoS (e.g., Bar-Ilan, 2010; Butler & Visser, 2006) or Scopus (Kousha, Thelwall, & Rezaie, 2011), which now includes tens of thousands of books. However, these methods are time-consuming and do not include many citations from books to books. Book to book citations can give different results from article to book citations, especially in book-based fields such as in the humanities and some social sciences (Cronin, Snyder, & Atkins, 1997, Archambault, et al., 2006).

Book Citation Index: The Thomson Reuters Book Citation Index now indexes the references in about 60,000 books and monographs (Book Citation Index, 2014) and is an optional addition to WoS. Nonetheless, only about 3% of BKCI-indexed books are in non-English languages and about 75% of their publishers are from the USA and England (Torres-Salinas et al., 2014). Added to the absence of aggregated citation counts for edited volumes, its use for evaluative purposes would be problematic (Leydesdorff & Felt, 2012; Gorraiz, Purnell, & Glänzel, 2013).

Google Books: Although Google Books (GB) is not a citation index, it can be used to extract citations from digitised books for book impact assessment. GB citations to academic books are more plentiful than citations in traditional citation databases (Scopus and BKCI) in the humanities and in some social sciences but not in science (Kousha & Thelwall, 2009; Kousha, Thelwall, & Rezaie, 2011; Kousha & Thelwall, 2014). For instance, in one study the median number of GB citations was three times higher than the median number of Scopus citations to 1,000 books in the 2008 UK RAE in seven fields (Kousha, Thelwall, & Rezaie, 2011).

Non-Citation Metrics

Book Reviews: Scholarly book reviews are significant academic outputs (Hartley, 2006), especially in some humanities fields, such as history, literature and philosophy (Zuccala &

Van Leeuwen, 2011). One early study found a high correlation (0.620) between the number of reviews in the Book Review Index and the number of library holdings in the OCLC database for 200 novels (Shaw, 1991), suggesting that both indicators may reflect a common factor, such as the popularity of the novels. Another study found that sociology books with more positive reviews tended to attract more citations (Nicolaisen, 2002), although the strength of association between the number of book reviews and citations varies between disciplines (Gorraiz, Gumpenberger, & Purnell 2014). Low but significant Spearman correlations have also been found between the numbers of Amazon book reviews and citation metrics (Kousha & Thelwall, in press).

Libcitations: National or international library holdings statistics can give useful information about potential usage of, or interest in, books (Torres-Salinas & Moed, 2009; White, Boell, Yu et al., 2009). White, Boell, Yu et al. (2009) argued that libcitation statistics could be used as an indication of the cultural benefit of books, especially in the social sciences and humanities. Several follow up studies have found significant, but low, correlations between library holdings statistics and citation metrics for books (Linmans, 2010; Zuccala & Guns, 2013; Kousha & Thelwall, in press), suggesting that library holdings reflect diverse kinds of influence, such as teaching and cultural impacts, that cannot be traced through citations.

Publisher Prestige:

In the absence of credible citation-based indicators for the impact assessment of books, publisher prestige has been proposed as an alternative (Donovan & Butler, 2007). Attempts to estimate the prestige of publishers through surveys of academics have shown that the perception of prestige varies by field (Garand & Giles, 2011; Giménez-Toledo, Tejada-Artigas & Mañana-Rodríguez, 2013). In addition to reputational surveys, BKCI indicators (Torres-Salinas et al., 2012), Scopus citations and matching library holdings data from WorldCat.org (Zuccala, Guns, Cornacchia, & Bod, in press) have also been used to rank academic book publishers.

Syllabus Mentions:

Academics may write textbooks for teaching or monographs that are widely used in teaching rather than, or in addition to, research (Gurung, Landrum, & Daniel, 2012). This kind of teaching contribution may be undervalued or unrewarded (Boyer, 1990; Jenkins, 1995; Healey, 2000) but evidence of inclusion in academic syllabi can reflect some aspects of teaching scholarship success (Albers, 2003; Thompson, 2007). In response, an attempt has been made to capture citations from online course syllabi for WoS-indexed articles across multiple fields, with the results suggesting that online syllabus mentions can be a useful indicator in some social sciences fields (Kousha & Thelwall, 2008).

Research Questions

The following research questions are designed to assess whether ratings and recommendation information in *Choice: Current Reviews for Academic Libraries* could be useful for the impact assessment of academic books.

1. Do Choice book ratings correlate with citation metrics or with other non-citation metrics for books?
2. Are Choice audience recommendations reflected in citation and non-citation metrics? For instance, do books recommended for undergraduates have more syllabus mentions than books recommended for researchers?

Methods

Choice Reviews

The recommendations for 451 book reviews from a free sample issue of *Choice Reviews Online* published in 2011 were extracted from the *Humanities*, *Social & Behavioral Sciences*, and *Science & Technology* categories but omitting reviews for the *Reference* section. The books were selected, with permission of *Choice*, from the collection of free sample reviews. The recommendation levels assigned to Choice reviews (see http://www.ala.org/acrl/choice/about) were converted into a number, from 1 for 'Not recommended' to 5 for 'Essential'.

- *Essential:* A publication of exceptional quality for academic audiences and a core title for academic libraries supporting programs in relevant disciplines.
- *Highly recommended:* A publication of high quality and relevance for academic audiences.
- *Recommended:* A publication containing good content and coverage and suitable for academic audiences.
- *Optional:* A publication that, due to limited value or deficiencies, is marginal for academic audiences.
- *Not recommended:* A poor quality publication or one not suitable for academic audiences.

Choice reviewers include extra information about usefulness for different academic audiences, such as undergraduates, researchers, faculty members and, professionals (Table 1). This information was used for further analyses.

Table 1. Examples of audience recommendations in Choice book reviews.

Audience recommendations	*Examples*
Mainly for undergraduates	*Essential.* Upper-division undergraduates through faculty. *Highly recommended.* Lower-division undergraduates through faculty. *Recommended.* Undergraduate and graduate studies. Optional. Upper-division undergraduates and above.
Mainly for graduates, researchers, professionals and academics	*Essential.* Graduate students, faculty, and professionals. *Highly recommended.* Research libraries and scholars. *Recommended.* All academic and professional audiences. *Optional.* Graduate students, researchers, and faculty.

Google Books Citations

For GB citations, Google Books API searches were used in the previously developed and tested software *Webometric Analyst* (http://lexiurl.wlv.ac.uk, "Books" tab) to extract citations from digitised books indexed by Google Books (for method details see: Kousha & Thelwall, 2014). To locate GB citations in other digitised books, we searched for the first author last name and the first (up to) ten terms of the book title as a phrase search, combined with the publication year.

Lurz "Mindreading animals: The debate over what animals know about other" 2011

For books with three or less words in their titles we added the publisher to the query:

Benford "Performing mixed reality" 2011 "MIT Press"

Syllabus Mentions

For syllabus mentions, an automatic method was used to search for mentions of the 451 books in public online course syllabi indexed by the Bing search engine. *Webometric Analyst* software and a set of rules were used to identify the syllabus mentions in academic websites and to exclude false matches in order to give accurate, although not comprehensive, results. This method was developed to capture academic syllabus mentions for books rather than articles (cf. Kousha & Thelwall, 2008). The first author last name was combined with the book title as a phrase search and either "syllabus" or "course description", with the results of the two combined and false matches automatically filtered out. The automatic syllabus citation extraction method applied in this study seems to give high accuracy (over 90%), although it misses results from non-academic institutions and syllabi stored in password protected databases and systems (see also Kousha & Thelwall, in press).

Barnett "Empire of humanity a history of humanitarianism" "course description"|Barnett "Empire of humanity a history of humanitarianism" "syllabus"

WorldCat Library Holdings

For library holdings, we manually searched for the 451 books in WorldCat online (http://www.worldcat.org) and recorded the number of library holdings for each one.

Mendeley Readers

For Mendeley reader counts, we used the Mendeley API in *Webometric Analyst* with queries combining the last name of the first author, the book title and the publication year for 451 books in the data set (for method details see: Mohammadi & Thelwall, 2014). This returns the number of users of the social reference sharing site Mendeley that have added the book to their personal library.

Amazon.com Reviews

The numbers of customer reviews were automatically extracted from the main Amazon.com URLs for each of the 451 books via *Webometric Analyst* (for method details see: Kousha & Thelwall, 2014 in press).

Sources not used

Not all book impact metrics were collected for the books in the data set. Publisher prestige was not collected because there is not a recognised source of this evidence and it varies by field. WoS/BKCI and Scopus citations were also not collected because Google Books citations have been shown to be superior for book impact assessment in most fields (Kousha & Thelwall, 2009; Kousha, Thelwall, & Rezaie, 2011; Kousha & Thelwall, 2014).

Results

Roughly three-quarters of books with Choice reviews had at least one GB citation (Table 2), and this is higher in the social sciences (80%, median: 3) than in science (68%, median: 2). Moreover, about 45% of the books had one or more academic syllabus mentions and the median number of syllabus mentions is higher in science (1) compared to the humanities (0)

and the social sciences (0). About 30% of the Choice books had at least one Amazon review and all 451 books had at least one WorldCat library holding (median: 394). Nevertheless, only 1.5% of books had at least one Mendeley reader. Follow-up manual investigations with Mendeley searches confirmed that this very low number was not a technical artefact but genuinely reflected the virtual absence of the Choice books from this site. The low Mendeley coverage confirms previous results that, although academic journal articles often have many Mendeley readers (e.g., 78% with one or more readers in the medical sciences, see Thelwall & Wilson, in press), the same is not true for books and monographs (Kousha & Thelwall, in press; see also: Hammarfelt, 2014), suggesting that Mendeley is currently not useful for book impact assessment.

Overall, it seems that GB citations are plentiful enough for book citation impact assessment and academic syllabus mentions, libcitations and Amazon reviews may be common enough to be used to indicate different types of impact, such as teaching, cultural or public interest.

Table 2. Google Books citations, syllabus mentions, libcitation, Amazon reviews and Mendeley reader counts for 451 books with Choice reviews published in 2011 in three broad fields.

Choice subjects	*No. of books*	*Google Books No. (% with GB cites*) median (mean)*	*Syllabus No. (% with syllab.*) median (mean)*	*Libcitation No. (% with holdings*) median (mean)*	*Amazon Rev. No. (% with reviews*) median (mean)*	*Mendeley No. (% with readers*) median (mean)*
Human.	136	474 (69.8%) 2 (3.5)	120 (39.7%) 0 (0.9)	62098 (100%) 356 (456.6)	105 (35.2%) 0 (0.8)	31 (3.7%) 0 (0.2)
Social Sci.	234	1278 (79.9%) 3 (5.5)	349 (45.7%) 0 (1.5)	130018 (100%) 442 (555.6)	951 (34.2%) 0 (4.1)	90 (3.4%) 0 (0.4)
Sci. & Tech	81	367 (67.9%) 2 (4.5)	149 (50.6%) 1 (1.8)	41585 (100%) 391 (513.4)	174 (27.2%) 0 (2.15)	194 (3.7%) 0 (2.4)
Total	451	2119 (74.7%) 2 (4.7)	618 (44.8%) 0 (1.4)	233701 (100%) 394 (518.2)	1230 30.8%) 0 (2.7)	315 (1.5%) 0 (0.7)

*% of books with at least one Google Books citation, academic syllabus mention, WorldCat libcitation, Amazon review and Mendeley reader.

Table 4 compares the metrics between those for books with Choice reviews claiming teaching utility (mainly for undergraduates) and those for books with reviews claiming benefits for graduates, researchers, faculty members and professionals. Books with research or other academic relevance have higher GB citation impact (median 3) than books with benefits for undergraduates (GB median 2). In contrast, books with more teaching utility for undergraduate studies tended to have more academic syllabus mentions (median 1 and 55% with one or more syllabus mentions) than books for academic audiences (median zero and 34% with one or more syllabus mentions). Hence, it seems that Choice reviews are broadly capable of distinguishing between the different types of audiences for books.

Table 3. A comparison of book metrics based on Choice book reviews with different rating recommendation levels.

Recommendation	*No. of books*	*Google Books No. (% with GB cites*) median (mean)*	*Syllabus No. (% with syllab.*) median (mean)*	*Libcitations No. (% with holdings*) median (mean)*	*Amazon Rev. No. (% with reviews*) median (mean)*	*Mendeley No. (% with readers*) median (mean)*
Essential/highly recommended	150	768 (88%) 3 (5.1)	186 (48.6%) 0 (1.2)	85256 (100%) 482.5 (568.4)	440 (40%) 0 (2.9)	51 (5.3%) 0 (0.34)
Other	301	1351 (68.1%) 2 (4.5)	432 (42.8%) 0 (1.4)	148445 (100%) 359 (493.2)	790 (26.2%) 0 (2.6)	264 (2.9%) 0 (0.9)

Table 4. A comparison of book metrics based on Choice recommendations for undergraduates and other academic audiences (graduates, researchers, faculty).

Audience recommendation	*No. of books+*	*Google Books No. (% with GB cites) median (mean)*	*Syllabus No. (% with syllab.) median (mean)*	*Libcitation No. (% with holdings) median (mean)*	*Amazon Rev. No. (% with reviews) median (mean)*	*Mendeley No. (% with readers) median (mean)*
Undergraduates	240	1098 (70.1%) 2 (4.7)	420 (55%) 1 (1.7)	122497 (100%) 394.5 (510.4)	649 (29.6%) 0 (2.7)	267 (5.4%) 0 (1.1)
Graduates, faculty, researchers, profess.	203	1006 (79.8%) 3 (4.9)	197 (34%) 0 (0.9)	108260 (100%) 405 (533.3)	579 (33%) 0 (2.85)	48 (2%) 0 (0.2)

+.Eight books with "Not recommended" Choice reviews were excluded.

There are low but significant positive Spearman correlations between Choice ratings and various citation and non-citation indicators (Table 5). Thus, in general, books with more GB citations, academic syllabus mentions, library holdings or Amazon reviews tended to be recommended more highly by book reviewers. The correlation is highest between Choice ratings and libcitations (0.201). This may reflect academic libraries ordering books based on Choice reviews and recommendations, especially in the United States (About Choice magazine, 2015).

Table 5. Spearman correlations between Choice ratings and other metrics across all fields (n=451).

Metrics	*Choice rating score*	*GB citations*	*Syllabus mentions*	*Libcitations*	*Amazon reviews*
Choice rating score	1	**.142****	**.103***	**.201****	**.141****
GB citations		1	.171**	.189**	.196**
Syllabus mentions			1	.121*	.073
Libcitations				1	.222**
Amazon reviews					1

**. Significant at p=0.01

*. Significant at p=0.05

There are disciplinary differences in the strength of association between Choice ratings and the other metrics (Tables 6-8). The highest correlation is between Choice ratings and GB citations in Science & Technology (0.350), but this correlation is much lower in Social & Behavioural Sciences and in the Humanities category. Hence, it seems that science books with more positive reviews tend to be more cited in other books and so Choice reviews may be a useful indicator for assessing the research contribution of scientific books. This is a surprising finding given that books are not as highly valued in science as in the humanities.

In the Humanities category there is a low and statistically insignificant correlation between Choice ratings and GB citations but this may reflect the weak association between citations and research quality in the humanities more than a lack of correlation between Choice ratings and research value or impact. The higher association between Choice ratings and libcitations (0.304) suggests that books with higher review ratings tend to be more often acquired by academic libraries but that this does not translate into citations. This may represent 'cultural benefits' of humanities books (Belfiore & Upchurch, 2013; White, Boell, Yu et al. 2009) and supports a previous finding that Outstanding Academic Titles in Choice are more likely to be purchased by academic libraries and have slightly higher library usage than non-Choice books (Levine-Clark, & Jobe, 2007). In Humanities there is also a low but significant correlation between Choice ratings and academic syllabus mentions (0.131), suggesting that in some teaching based fields, Choice reviews may reflect the educational merits of books. In Social & Behavioural Sciences, however, there is no relationship between Choice ratings and either citation or non-citation metrics. A possible explanation is that in the social sciences books have very different patterns of scholarly usage in research and teaching and the relationship between the number of book reviews and citations could therefore differ between subject areas (Gorraiz, Gumpenberger, & Purnell 2014).

Table 6. Spearman correlations between Choice rating scores and other metrics in Science & Technology (n=81).

Metrics	*Choice rating score*	*GB citations*	*Syllabus mentions*	*Libcitations*	*Amazon reviews*
Choice rating score	1	**.350****	.090	**.274****	**.297****
GB citations		1	.097	.326**	.250*
Syllabus mentions			1	.196	-.019
Libcitations				1	.028
Amazon reviews					1

Table 7. Spearman correlations between Choice rating scores and other metrics in Humanities (n=136).

Metrics	*Choice rating score*	*GB citations*	*Syllabus mentions*	*Libcitations*	*Amazon reviews*
Choice rating score	1	.144	**.131***	**.304****	.089
GB citations		1	.145	.193*	.170*
Syllabus mentions			1	.045	.025
Libcitations				1	.118
Amazon reviews					1

Table 8. Spearman correlations between Choice rating scores and other metrics in Social & Behavioural Sciences (n=234).

Metrics	*Choice rating score*	*GB citations*	*Syllabus mentions*	*Libcitations*	*Amazon reviews*
Choice rating score	1	.081	.095	.123	.123
GB citations		1	.193**	.127	.179**
Syllabus mentions			1	.117	.116
Libcitations				1	.314**
Amazon reviews					1

Limitations

This study tested only 451 books with Choice reviews from a free issue of *Choice Reviews Online* published in 2011 and a larger data may give different results. The sample of 451 is from the most public part of Choice, its free samples, and so is atypical in that regard. The small sample size was also not enough for a fine grained analysis of individual subject areas and this is an important limitation for the correlation tests because citation practices and educational norms (e.g., typical class sizes and the role of textbooks) can vary substantially between fields in a way that would systematically reduce correlation results when the fields are grouped together. Another limitation is that the data only included GB citations from books to books and so would miss citations from articles to books. Hence, a future study might use cited reference searches in WoS or Scopus order to check whether stronger relationships can be found.

Discussion and Conclusions

This study seems to be the first to assess whether the book reviews in *Choice: Current Reviews for Academic Libraries* reflect the value of books and could be used for indicators of value or impact. The analysis of a small sample of 451 books published in 2011 found weak but often significant relationships with other indicators, suggesting that Choice should be particularly helpful for books that have uses that do not necessarily attract citations.

In answer to the first research question, books that were highly rated in Choice received more GB citations, academic syllabus mentions, libcitations and Amazon reviews than did lower rated books. In answer to the second research question, books recommended for undergraduates (e.g., textbooks) received more academic syllabus mentions, reflecting teaching influence of books, and books recommended for researchers, faculty and professionals received more citations than did books recommended for undergraduates, indicating the ability of Choice reviews to distinguish between the different audiences for books.

The low (but statistically significant) Spearman correlations between Choice ratings and all citation and non-citation indicators suggest that Choice reviews are either somewhat subjective, or (more likely) do not reflect exactly the same aspects of the value of a book (e.g., teaching, research, cultural or social impacts) as any of the other indicators. Hence, the evidence presented here is insufficient to claim that Choice recommendations are reliable indicators of audience or value at the individual book level. Nevertheless, the correlations will be weakened by the broad categories used (e.g., 200 library holdings might be a spectacular success for a monograph on Old Norse but a failure for one on Shakespeare's women). In addition, the correlations will also be weakened by the fact that the other indicators are not direct *measures* of anything (e.g., educational value) but are indirect (not cause-and-effect) reflections and so strong correlations should not be expected. Hence, the low correlations are not evidence that Choice book reviews have little value but probably reflect the complex multifaceted nature of the value of books and the difficulty in finding indicators to effectively reflect those values. In this context, Choice book reviews are a promising new source of post-publication peer review evidence of the value of books. They are a welcome additional source of evidence for the particularly challenging task of book impact assessment and when positive reviews are used for impact assessments of scholarly outputs by evaluators, funders or perhaps even national research assessments (e.g., PBRF, 2013).

References

About Choice magazine, 2015. http://www.ala.org/acrl/choice/about

Albers, C. (2003). Using the syllabus to document the scholarship of teaching. *Teaching Sociology*, *31*(1), 60-72.

Archambault, É., Vignola-Gagné, É., Côté, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, *68*(3), 329-342.

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics*, *82*(3), 495-506.

Bauerlein, M. (2002). Disagreements in the humanities. *Knowledge, Technology & Policy*, *15*(1), 188-195.

Belfiore, E., & Upchurch, A. (Eds.). (2013). *Humanities in the Twenty-first Century: Beyond Utility and Markets*. Basingstoke, UK: Palgrave Macmillan.

Boyer, E. L. (1990). Scholarship revisited. Princeton, NJ: Carnegie Foundation for the Advancement of Teaching. http://184.168.109.199:8080/jspui/bitstream/123456789/2134/1/ED326149.pdf

Butler, L., & Visser, M. (2006). Extending citation analysis to non-source items. *Scientometrics*, *66*(2), 327–343

Cronin, B., Snyder, H. & Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: a study of sociology. *Journal of Documentation*, *53*(3), 263-273.

Crotty, D. (2012). Life after publication - post-publication peer review. *Biochemist*, *34*(4), 26-28. http://www.biochemist.org/bio/03404/0026/034040026.pdf

Cullars, J. (1998). Citation characteristics of English-language monographs in philosophy. *Library & Information Science Research*, *20*(1), 41–68.

Donovan, C., & Butler, L. (2007). Testing novel quantitative indicators of research "quality," esteem and "user engagement:" An economics pilot study. *Research Evaluation*, *16*(4), 231-242.

Garand, J.C., & Giles, M.W. (2011). Ranking scholarly publishers in political science: An alternative approach. *PS: Political Science and Politics*, *44*(2), 375-383.

Garfield, E. (1996). Citation Indexes for Retrieval and Research Evaluation. Consensus Conference on the Theory and Practice of Research Assessment, Capri.

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation*, *22*(1), 64-77.

Gorraiz, J., Gumpenberger, C., & Purnell, P. J. (2014). The power of book reviews: A simple and transparent enhancement approach for book citation indexes. *Scientometrics*, *98*(2), 841-852.

Gorraiz, J., Purnell, P. J., & Glänzel, W. (2013). Opportunities for and limitations of the book citation index. *Journal of the American Society for Information Science and Technology*, *64*(7), 1388-1398.

Gurung, R. A. A., Landrum, R. E., & Daniel, D. B. (2012). Textbook use and learning: A North American perspective. *Psychology Learning and Teaching*, *11*(1), 87-98.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. *Scientometrics*, *101*(2), 1419-1430.

Hartley, J. (2006). Reading and writing book reviews across the disciplines. *Journal of the American Society for Information Science and Technology*, *57*(9), 1194-1207.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, *44*(2), 193−215.

Huang, M., & Chang, Y. (2008). Characteristics of research output in social sciences and humanities: from a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, *59*(11), 1819–1828.

Hunter, J. (2012, August 30). Post-publication peer review: Opening up scientific conversation. *Frontiers in Computational Neuroscience*, doi:10.3389/fncom.2012. 00063

Jenkins, A. (1995). The research assessment exercise, funding and teaching quality. *Quality Assurance in Education*, *3*(2), 4-12.

Kousha, K., & Thelwall, M. (2008). Assessing the impact of disciplinary research on teaching: An automatic analysis of online syllabuses. *Journal of the American Society for Information Science and Technology*, *59*(13), 2060-2069.

Kousha, K., & Thelwall, M. (2009). Google Book Search: Citation analysis for social science and the humanities. *Journal of the American Society for Information Science and Technology*, *60*(8), 1537-1549.

Kousha, K., & Thelwall, M. (2014). An automatic method for extracting citations from Google Books. *Journal of the Association for Information Science and Technology*, *66*(2), 309-320.

Kousha, K., & Thelwall, M. (in press, 2015). Can Amazon.com reviews help to assess the wider impacts of books? *Journal of the Association for Information Science and Technology*. http://www.koosha.tripod.com/AmazonReviewstoAssessBooks-Preprint.pdf

Kousha, K., & Thelwall, M. (in press). An automatic method for assessing the teaching impact of books from online academic syllabi. *Journal of the Association for Information Science and Technology*, http://www.scit.wlv.ac.uk/~cm1993/papers/SyllabiBookCitations.pdf

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62*(11), 2147-2164.

Krampen, G., Becker, R., Wahner, U., & Montada, L. (2007). On the validity of citation counting in science evaluation: Content analyses of references and citations in psychological publications. *Scientometrics*, *71*(2), 191–202.

Levine-Clark, M., & Jobe, M. M. (2007). Do reviews matter? An analysis of usage and holdings of choice-reviewed titles within a consortium. *Journal of Academic Librarianship*, *33*(6), 639-646.

Leydesdorff, L., & Felt, U. (2012). Edited volumes, monographs and book chapters in the Book Citation Index (BKCI) and Science Citation Index (SCI, SoSCI, A&HCI). *Journal of Scientometric Research*, *1*(1), 28–34.

Li, X., & Thelwall, M. (2012). F1000, Mendeley and traditional bibliometric indicators. *17th International Conference on Science and Technology Indicators* (Vol. 3, pp. 1–11). http://sticonference.org/Proceedings/vol2/Li_F1000_541.pdf

Linmans, A. J. M. (2010). Why with bibliometrics the humanities does not need to be the weakest link. Indicators for research evaluation based on citations, library bindings and productivity measures. *Scientometrics*, *83*(2), 337–354.

Mohammadi, E. & Thelwall, M. (2013). Assessing non-standard article impact using F1000 labels. *Scientometrics*, *97*(2), 383-395.

Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, *65*(8), 1627-1638.

Nederhof, A. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*(1), 81−100.

Nicolaisen, J. (2002). The scholarliness of published peer reviews: A bibliometric study of book reviews in selected social science fields. *Research Evaluation*, *11*(3), 129-140.

Nussbaum, M. C. (2012). *Not for profit: Why democracy needs the humanities*. Princeton, NJ: Princeton University Press.

PBRF. (2013). Performance-Based Research Fund Quality Evaluation Guidelines 2012. http://www.tec.govt.nz/Documents/Publications/PBRF-Quality-Evaluation-Guidelines-2012.pdf

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2011). altmetrics: a manifesto. http://altmetrics.org/manifesto.

Shaw, D. (1991). An analysis of the relationship between book reviews and fiction holdings in OCLC. *Library and Information Science Research*, *13*(2), 147-154.

Small, H. (2013). *The Value of the Humanities*. Oxford: Oxford University Press.

Taylor, J. & Walker, I. (2009). Peer assessment of research: How many publications per staff? Lancaster University Management School, Working Paper 2009/035. http://eprints.lancs.ac.uk/31757/1/006236.pdf

Thelwall, M. & Wilson, P. (in press). Mendeley readership Altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology*. http://www.scit.wlv.ac.uk/~cm1993/papers/MendeleyInScienceAltmetricsPreprint.pdf

Thompson, B. (2007). The syllabus as a communication document: Constructing and presenting the syllabus. *Communication Education*, *56*(1), 54-71.

Torres-Salinas, D. & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics*, *3*(1), 9-26.

Torres-Salinas, D., Robinson-García, N., Campanario, J. M., & López-Cózar, E. D. (2014). Coverage, field specialisation and the impact of scientific publishers indexed in the book citation index. *Online Information Review*, *38*(1), 24-42.

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E., & Delgado López-Cózar, E. (2012). Towards a 'Book Publishers Citation Reports'. First approach using the 'Book Citation Index'. *Revista Española de Documentación Científica*, *35*(4), 615-620.

Torres-Salinas, D., Rodríguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J., & García, J. A. (2013). Mapping citation patterns of book chapters in the Book Citation Index. *Journal of Informetrics*, *7*(2), 412-424.

Waltman, L. & Costas, R. (2014). F1000 Recommendations as a potential new data source for research evaluation: a comparison with citations. *Journal of the Association for Information Science and Technology*, *65*(3), 433-445.

Weller, A. C. (2001). *Editorial Peer Review: Its Strengths and Weaknesses*. Medford, N.J: Information Today.

White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, *60*(6), 1083-1096.

Zuccala, A., & Guns, R. (2013). Comparing book citations in humanities journals to library holdings: Scholarly use versus "perceived cultural benefit" (RIP). In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. Moed (Eds.) *Proceedings of 14th International Conference of the International Society for Scientometrics and Informetrics* (pp.353–360). Vienna, Austria: AIT Austrian Institute of Technology GmbH Vienna.

Zuccala, A. & Van Leeuwen, T. (2011). Book reviews in humanities research evaluations. *Journal of the American Society for Information Science and Technology*, *62*(10), 1979–1991.

Zuccala, A., Guns, R., Cornacchia, R., & Bod, R. (in press). Can we rank scholarly book publishers? A bibliometric experiment with the field of history. *Journal of the Association for Information Science and Technology*. http://www.illc.uva.nl/evaluating-humanities/RankingPublishers(Preprint_2014).pdf

A Longitudinal Analysis of Search Engine Index Size

Antal van den Bosch[1], Toine Bogers[2], and Maurice de Kunder[3]

[1] *a.vandenbosch@let.ru.nl*
Centre for Language studies, Radboud University, PO Box 9103, 6500 HD Nijmegen (The Netherlands)

[2] *toine@hum.aau.dk*
Department of Communication, Aalborg University Copenhagen, A.C. Meyers Vænge 15, 2450 Copenhagen (Denmark)

[3] *maurice@dekunder.nl*
De Kunder Internet Media, Toernooiveld 100, 6525 EC, Nijmegen (The Netherlands)

Abstract
One of the determining factors of the quality of Web search engines is the size of their index. In addition to its influence on search result quality, the size of the indexed Web can also tell us something about which parts of the WWW are directly accessible to the everyday user. We propose a novel method of estimating the size of a Web search engine's index by extrapolating from document frequencies of words observed in a large static corpus of Web pages. In addition, we provide a unique longitudinal perspective on the size of Google and Bing's indexes over a nine-year period, from March 2006 until January 2015. We find that index size estimates of these two search engines tend to vary dramatically over time, with Google generally possessing a larger index than Bing. This result raises doubts about the reliability of previous one-off estimates of the size of the indexed Web. We find that much, if not all of this variability can be explained by changes in the indexing and ranking infrastructure of Google and Bing. This casts further doubt on whether Web search engines can be used reliably for cross-sectional webometric studies.

Conference Topic
Webometrics

Introduction
Webometrics (or cybermetrics) is commonly defined as the study of the content, structure, and technologies of the World Wide Web (WWW) using primarily quantitative methods. Since its original conception in 1997 by Almind & Ingwersen, researchers in the field have studied aspects such as the link structure of the WWW, credibility of Web pages, Web citation analysis, the demographics of its users, and search engines (Thelwall, 2009). The size of the WWW, another popular object of study, has typically been hard to estimate, because only a subset of all Web pages is accessible through search engines or by using Web crawling software. Studies that attempt to estimate the size of the WWW tend to focus on the surface Web—the part indexed by Web search engines—and often only at a specific point in time.

In the early days of search engines, having the biggest index size provided search engines with a competitive advantage, but a changing focus on other aspects of search result quality, such as recency and personalization, has diminished the importance of index size in recent years. Nevertheless, the size of a search engine's index is important for the quality of Web search engines, as argued by Lewandowski and Höchstötter (2008). In addition, knowledge of the size of the indexed Web is important for webometrics in general, as it gives us a ceiling estimate of the size of the WWW that is accessible by the average Internet user.

The importance of index sizes in the early days of Web search resulted in several estimation methods, most of which used the overlap between different Web search engines to estimate the size of the indexed Web as a whole. Bharat and Broder (1998) used an overlap-based method to estimate the size of the WWW at around 200 million pages. Lawrence & Giles

(1998, 1999) produced higher estimates of 320 and 800 million pages in 1998 and 1999 using a similar method, and Gulli and Signorini (2005) updated these estimates to 11.5 billion pages. The last decade has seen little work on index size estimation, but a general problem with all of the related work so far is that all the analyses have been cross-sectional. There has been no analysis of index size on a longer time scale that sheds light on the robustness of the different estimation methods. The handful of studies that have taken a longer-term perspective have typically focused on Web page persistence (Koehler, 2004) or academic link structure (Payne & Thelwall, 2008), but never search engine index size.

In this paper we present a novel method of estimating the size of a Web search engine's index by extrapolating from document frequencies of words observed in a large static corpus of Web pages. In addition, we provide a unique longitudinal perspective on our estimation method by applying it to estimate the size of Google and Bing's[1] indexes over a period of close to nine years, from March 2006 until January 2015.[2]

We find that index size estimates of these two search engines tend to vary wildly over time, with Google generally possessing a larger index than Bing. This considerable variability has been noted in earlier work (e.g., Rousseau, 1999; Payne & Thelwall, 2008), which raises doubts about the reliability of previous one-off estimates of the size of the indexed Web. In our analysis, we find that much of this variability can be explained by changes in the indexing and ranking infrastructure of Google and Bing. This casts further doubt on whether Web search engines can be used reliably for one-off Webometric studies, confirming similar sentiments expressed by, for instance, Payne and Thelwall (2008), and Thelwall (2012).

The remainder of this paper is organized as follows. The next section contains a review of related work in webometrics and on estimating the size of the indexed WWW. We then explain our estimation method in more detail, followed by the results of our estimation method and an analysis of the variability we uncover. We then discuss our findings and draw our conclusions.

Related work

Since its inception, researchers have studied many different aspects of the Web. This section provides a brief overview of some of the key studies on measuring different properties of Web search engines and the WWW, in particular work on estimating their size.

Measuring the Web

Over the past two decades many aspects of the WWW have been studied, such as the link structure of the Web that emerges from the hyperlinks connecting individual Web pages. Broder et al. (2000) were among the first to map the link structure of the WWW. They showed that the Web graph can be visualized as a bow-tie structure with 90% of all pages being a part of the largest strongly connected component, which was confirmed in 2005 by Hirate et al. (2008). Payne and Thelwall (2008) performed a longitudinal analysis of hyperlinks on academic Web sites in the UK, Australia and New Zealand over a six-year period. They found that the inlink and outlink counts were relatively stable over time, albeit with large fluctuations at the individual university level. As a result, they concluded that such variability could create problems for the replicability and comparability of webometrics research. Other related work on analyzing the link structure of the Web includes Kleinberg et al. (1999) and Björneborn (2004).

[1] Formerly known as Microsoft Live Search until May 28, 2009.

[2] Recent daily estimates produced by our method can be accessed through http://www.worldwidewebsize.com/. The time series data displayed in Figure 1 are available online at http://toinebogers.com/?page_id=757.

Web search engines are an essential part of navigating the WWW and as a result have received much attention. Many different aspects of Web search have been investigated, such as ranking algorithms, evaluation, user behavior, and ethical and cultural perspectives. Bar-Ilan (2004) and Zimmer (2010) provide clear, multi-disciplinary overviews of the most important work on these aspects.

From a webometric perspective the hit counts, search engine rankings, and the persistence of the indexed URLs are highly relevant for the validity and reliability of webometric research using Web search engines. Rousseau (1999) was among the first to investigate the stability of search engine results by tracking the *hit counts*—the number of results indicated for a query—for three single-word search terms in Altavista and NorthernLight over a 12-week period in 1998. Altavista exhibited great variability over a longer time period, even with only three anecdotal query words. Rousseau attributed this to changes in Altavista's infrastructure with the launch of a new version in 1998. Thelwall (2008) also performed a cross-sectional, quantitative comparison of the hit counts and search engine results of Google, Yahoo!, and Live Search. He extracted 1,587 single-word queries from English-language blogs "based purely on word frequency criteria" (Thelwall, 2008, p. 1704), found strong correlations between the hit count estimates of all three search engines, and recommended using Google for obtaining accurate hit count estimates. Uyar (2009) extended Thelwall's work by including multi-word queries. He found that the number of words in the query significantly affects the accuracy of hit counts, with single-word queries providing nearly double the hit count accuracy as compared to multi-word queries. Finally, Thelwall and Sud (2012) investigated the usefulness of the Bing Search API 2.0 for performing webometric research. They examined, among other things, the hit count estimates and found that these can vary by up to 50% and should therefore be used with caution in webometric research.

Bar-Ilan et al. (2006) compared the rankings of three different Web search engines over a three-week period. They observed that the overlap in result lists for textual queries was much higher than for image queries, where the result lists of the different search engines showed almost no overlap. Spink et al. (2006) investigated the overlap between three major Web search engines based on the first results pages and found that 85% of all returned top 10 results are unique to that search engine.

The issue of Web page persistence in search engine indexes—how long does a Web page remain indexed and available—was first examined by Bar-Ilan (1999) for a single case-study query during a five-month period in 1998. She found that for some search engines up to 60% of the results had disappeared from the index at the end of the period. She hypothesized that the distributed nature of search engines may cause different results to be served up from different index shards at different points in time. Koehler (2004) reported on the results of a six-year longitudinal study on Web page persistence. He also provided an overview of different longitudinal studies on the topic and concluded, based on the relatively small number of studies that exist, that Web pages are not a particularly persistent medium, although there are meaningful differences between navigation and content pages.

Index size estimation

In the last two decades, various attempts have been directed at estimating the size of the indexed Web. Some approaches focus on estimating the index size of a single search engine directly, while a majority focuses on estimating the overlap to indirectly estimate the size of the total indexed Web.

Highly influential work on estimating index size was done by Bharat & Broder (1998), who calculated the relative sizes of search engines by selecting a random set of pages from one engine, and checking whether each page was indexed by another engine. They used 35,000 randomly generated queries of 6 to 8 words selected at random from a Web-based lexicon and

sent these queries to four search engines. One of every top-100 results pages was randomly selected, after which they calculated the relative sizes and overlaps of search engines by selecting this random set of pages from one engine, and checking whether the page was indexed by another engine. By combining their method with self-reported index sizes from the commercial search engines, they estimated the size of the WWW to be around 200 million pages. Gulli et al. (2005) extended the work of Bharat and Broder by increasing the number of submitted queries by an order of magnitude, and using 75 different languages. They calculated the overlap between Google, Yahoo!, MSN Live, and Ask.com, and updated the previous estimates to 11.5 billion pages in January 2005. Most approaches that use the work of Bharat and Broder as a starting point focus on improving the sampling of random Web pages, which can be problematic because not every page has the same probability of being sampled using Bharat and Broder's approach. Several researchers have proposed methods of near-uniform sampling that attempt to compensate for this ranking bias, such as Henzinger et al. (2000), Anagnostopoulos et al. (2006), and Bar-Yossef and Gurevich (2006, 2011).

Lawrence and Giles (1998) estimated the indexed overlap of six different search engines. They captured the queries issued by the employees of their own research institute and issued them to all six engines. The overlap among search engines was calculated on the aggregated result sets, after which they used publicly available size figures from the search engines to estimate the size of the indexed Web to be 320 million pages. Lawrence and Giles updated their previous estimates to 800 million Web pages in July 1999. Dobra et al. (2004) used statistical population estimation methods to improve upon the original 1998 estimate of Lawrence and Giles. They estimated that Lawrence and Giles were off by a factor of two and that the Web contained around 788 million Web pages in 1998. Khelghati et al. (2012) compared several of the aforementioned estimation methods as well as some proposed modifications to these methods. They found that a modified version of the approach proposed by Bar-Yossef et al. (2011) provided the best performance.

Estimating the Size of a Search Engine Index through Extrapolation

On the basis of a textual corpus that is fully available, both the number of documents and the term and document frequencies of individual terms can be counted. In the context of Web search engines, however, we only have reported hit counts (or document counts), and we are usually not informed about the total number of indexed documents. Since it is the latter we are interested in, we want to estimate the number of documents indexed by a search engine indirectly from the reported document counts.

We can base such estimates on a training corpus for which we have full information on document frequencies of words and on the total number of documents. From the training corpus we can extrapolate a size estimation of any other corpus for which document counts are given. Suppose that, for example, we collect a training corpus T of 500,000 web pages, i.e. $|T| = 500{,}000$. For all words w occurring on these pages we can count the number of documents they occur in, or their document count, $d_T(w)$. A frequent word such as *are* may occur in 250,000 of the documents, i.e., it occurs in about one out of every two documents; $d_T(are) = 250{,}000$. Now if the same word *are* is reported to occur in 1 million documents in another corpus C, i.e., its document count $d_C(are) = 1{,}000{,}000$, we can estimate by extrapolation that this corpus will contain about $|C| = \frac{d_C(are) \times |T|}{d_T(are)}$, i.e., 2 million documents.

There are two crucial requirements that would make this extrapolation sound. First, the training corpus would need to be representative of the corpus we want to estimate the size of. Second, the selection of words[3] that we use as the basis for extrapolation will need to be such

[3] We base our estimates on words rather than on multi-word queries based on the findings of Uyar (2009).

that the extrapolations based on their frequencies are statistically sound. We should not base our estimates on a small selection of words, or even a single word, as frequencies of both high-frequency and low-frequency words may differ significantly among corpora. Following the most basic statistical guidelines, it would be better to repeat this estimation for several words, e.g., twenty times, and average over the extrapolations.

A random selection of word types is likely to produce a selection with relatively low frequencies, as Zipf's second law predicts (Zipf, 1995). A well-known issue in corpus linguistics is that when any two corpora are different in genre or domain, very large differences are likely to occur in the two corpora's word frequencies and document frequencies, especially in the lower frequency bands of the term distributions. It is not uncommon that half of the word types in a corpus occur only once; many of these terms will not occur in another disjoint corpus, even if it is of the same type. This implies that extrapolations should not be based on a random selection of terms, many of which will have a low frequency of occurrence.

The selection of words should sample several high-frequency words but preferably also several other words with frequencies spread across the document frequency bands.

It should be noted that Zipf's law concerns word frequencies, not document frequencies. Words with a higher frequency tend to recur more than once in single documents. The higher the frequency of a word, the more its document frequency will be lower than its word frequency. A ceiling effect thus occurs with the most frequent words if the corpus contains documents of sufficient size: they tend to occur in nearly all documents, making their document frequencies about the same and approaching the actual number of documents in the corpus, while at the same time their word token frequencies still differ to the degree predicted by Zipf's law (Zipf, 1995). This fact is not problematic for our estimation goal, but it should be noted that this hinges on the assumption that the training corpus and the new corpus of which the frequencies are unknown, contain documents of about the same average size.

As our purpose is to estimate the size of a Web search engine's index, we must make sure that our training corpus is representative of the web, containing documents with a representative average size. This is quite an ambitious goal. We chose to generate a randomly filtered selection of 531,624 web pages from the DMOZ[4] web directory. We made this selection in the spring of 2006. To arrive at this selection, first a random selection was made of 761,817 DMOZ URLs, which were crawled. Besides non-existing pages, we also filtered out pages with frames, server redirects beyond two levels, and client redirects. In total, the DMOZ selection of 531,624 documents contains 254,094,395 word tokens (4,395,017 unique word types); the average DMOZ document contains 478 words.

We then selected a sequence of DMOZ words by their frequency rank, starting with the most frequent word, and selecting an exponential series where we increase the selection rank number with a low exponent, viz. 1.6. We ended up with a selection of the following 28 words, the first nine being high-frequency function words and auxiliary verbs: *and*, *of*, *to*, *for*, *on*, *are*, *was*, *can*, *do*, *people*, *very*, *show*, *photo*, *headlines*, *william*, *basketball*, *spread*, *nfl*, *preliminary*, *definite*, *psychologists*, *vielfalt*, *illini*, *chèque*, *accordée*, *reticular*, *rectificació*. The DMOZ directory is multilingual, but English dominates. It is not surprising that the tail of this list contains words from different languages.

Our estimation method then consists of retrieving document counts for all 28 words from the search engine we wish to estimate the number of documents for, obtaining an extrapolated estimate for each word, and averaging (taking a mean) over the 28 estimations. If a word is not reported to occur in any document (which hardly happens), it is not included in the average.

[4] DMOZ is also called the Open Directory Project, http://www.dmoz.org/.

To stress-test the assumption that the DMOZ document frequencies of our 28 words yield sensible estimates of corpus size, we estimated the size of a range of corpora: the New York Times part of the English Gigaword corpus[5] (newspaper articles), the Reuters RCV1 corpus[6] (newswire articles), the English Wikipedia[7] (encyclopedic articles, excluding pages that redirect or disambiguate), and a held-out sample of random DMOZ pages. Table 1 provides an overview of the estimations on these widely different corpora. The size of the New York Times corpus is overestimated by a large margin of 126%, while the sizes of the other three corpora are underestimated. The size of the DMOZ sample—not overlapping with the training set, but drawn from the same source—is relatively accurately estimated with a small underestimation of 1.3%. Larger underestimations, for Reuters RCV1 and Wikipedia, may be explained by the fact that these corpora have shorter documents on average.

The standard deviations in Table 1, computed over the 28 words, indicate that the different estimates are dispersed over quite a large range. There seems to be no correlation with the size of the difference between the actual and the estimated number of documents. Yet, the best estimate, for the small DMOZ held-out sample (–1.3% error), coincides with the smallest standard deviation.

Table 1. Real versus estimated numbers (with standard deviations) of documents on four textual corpora, based on the DMOZ training corpus statistics: two news resources (top two) and two collections of web pages (bottom two). The second and third column provides the mean and median number of words per document.

	Words per document					
Corpus	*Mean*	*Median*	*# Documents*	*Estimate*	*St. dev.*	*Difference*
New York Times '94-'01	837	794	1,234,426	2,789,696	1,821,823	+126%
Reuters RCV1	295	229	453,844	422,271	409,648	– 7.0%
Wikipedia	447	210	2,112,923	2,024,792	1,385,105	– 4.2%
DMOZ test sample	477	309	19,966	19,699	5,839	– 1.3%

After having designed this experiment in March 2006, we started to run it on a daily basis on March 13, 2006, and have continued to do so. Each day we sent the 28 DMOZ words as queries to two search engines: Bing[i] and Google[8]. We retrieve the reported number of indexed pages on which each word occurs as it is returned by the web interface of both search engines, not their APIs. This number is typically rounded: it retains three or four significant numbers, the rest being padded by zeroes. For each word we use the reported document count to extrapolate an estimate of the search engine's size, and average over the extrapolations of all words. The web interfaces to the search engines have gone through some changes, and the time required to adapt to these changes sometimes caused lags of a number of days in our measurements. For Google 3,027 data points were logged, which is 93.6% of the 3,235 days between March 13, 2006 and January 20, 2015. For Bing, this percentage is 92.8% (3,002 data points).

Results

Figure 1 displays the estimated sizes of the Google and Bing indices between March 2006 and January 2015. For visualization purposes and to avoid clutter, the numbers are unweighted

[5] https://catalog.ldc.upenn.edu/LDC2003T05.
[6] http://trec.nist.gov/data/reuters/reuters.html.
[7] Downloaded on October 28, 2007.
[8] We also sent the same 28 words to two other search engines that were discontinued at some point after 2006.

running averages of 31 days, taking 15 days before and after each focus day as a window. The final point in our measurements is January 20, 2015; hence the last point in this graph is January 5, 2015. Rather than a linear, monotonic development we observe a rather varying landscape, with Google usually yielding the larger estimates. The largest peak in the Google index estimates is about 49.4 billion documents, measured in mid-December 2011. Occasionally, estimates are as low as under 2 billion pages (e.g. 1.96 billion pages in the Google index on November 24, 2014), but such troughs in the graph are usually short-lived, and followed by a return to high numbers (e.g., to 45.7 billion pages in the Google index on January 5, 2015).

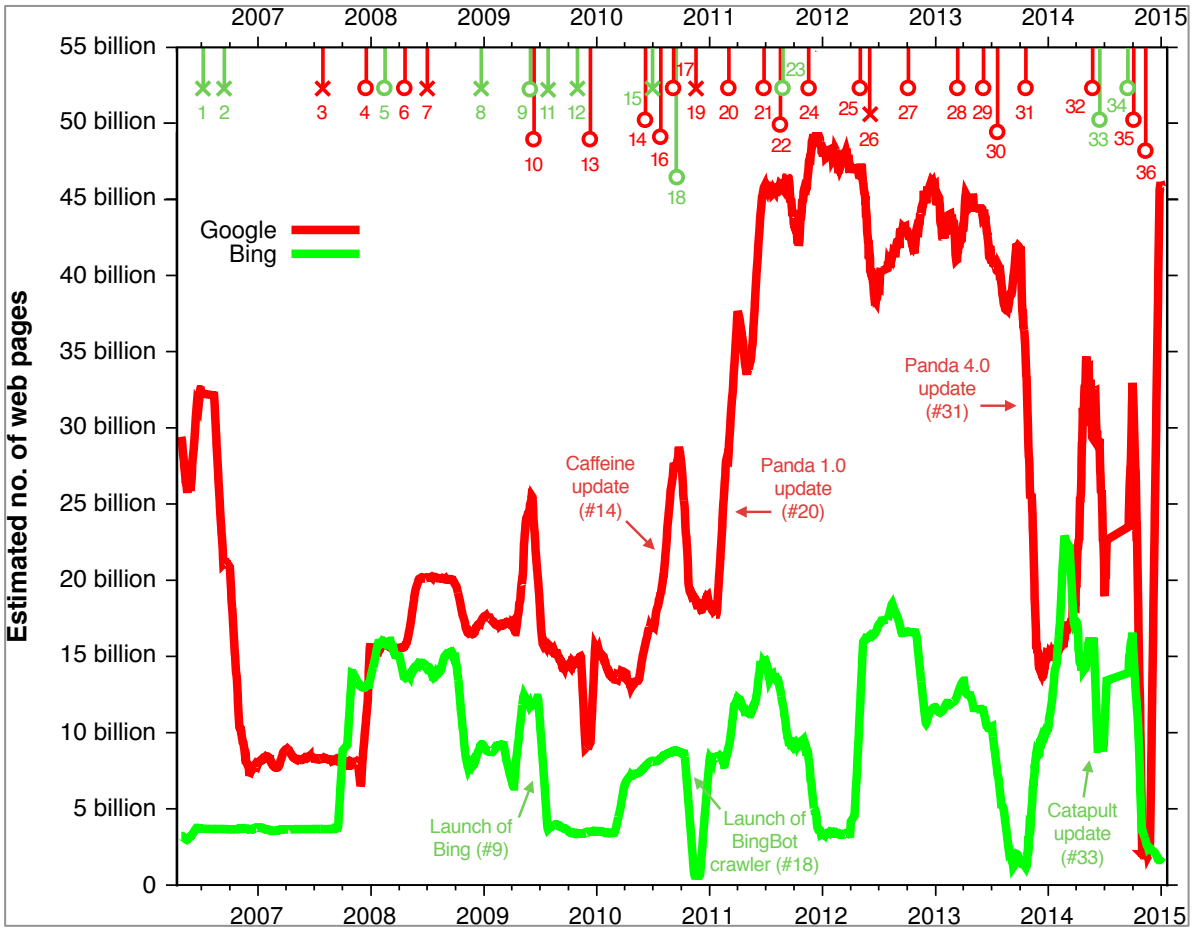


Figure 1. Estimated size of the Google and Bing indexes from March 2006 to January 2015. The lines connect the unweighted running daily averages of 31 days. The colored, numbered markers at the top represent reported changes in Google and Bing's infrastructure. The colors of the markers correspond to the color of the search engine curve they related to; for example, red markers signal changes in Google's infrastructure (the red curve). Events that line up with a spike are marked with an 'O', other events are marked with an '✕'.

Extrinsic variability

The variability observed in Figure 1 is not surprising given the fact that the indexing and ranking architectures of Web search engines are updated and upgraded frequently. According to Matt Cutts[9], Google makes "roughly 500 changes to our search algorithm in a typical year", and this is likely the same for Bing. While most of these updates are not publicized,

[9] http://googleblog.blogspot.com/2011/11/ten-algorithm-changes-on-inside-search.html.

some of the major changes that Google and Bing make to their architectures are announced on their official blogs. To examine which spikes in Figure 1 can be attributed to publicly announced architecture changes, we went through all blog posts on the Google Webmaster Central Blog[10], the Google Official Blog[11], the Bing Blog[12], and Search Engine Watch[13] for reported changes to their infrastructure. This resulted in a total of 36 announcements related to changes in the indexing or ranking architecture of Google and Bing[14]. The colored, numbered markers at the top of Figure 1 show how these reported changes are distributed over time.

For Google 20 out of the 24 reported changes appear to correspond to sudden spikes in the estimated index size, and for Bing 6 out of 12 reported changes match up with estimation spikes. This strongly supports the idea that much of the variability can be attributed to such changes. Examples include the launch of Bing on May 28, 2009 (event #9), the launch of Google's search index Caffeine on June 8, 2010 (event #14), the launch of the BingBot crawler (event #18), and the launches of Google Panda updates, and Bing's Catapult update (events #20, #31, and #33).

Of course not all spikes can be explained by reported events. For example, the spike in Bing's index size in October 2014 does not match up with any publicly announced changes in their architecture, although it is a likely explanation for such a significant change. In addition, some changes to search engine architectures are rolled out gradually and would therefore not translate to spikes in the estimated size. However, much of the variation in hit counts, and therefore estimated index size, appears to be caused by changes in the search engine architecture—something already suggested by Rousseau in his 1999 study.

Discussion and Conclusions

In this paper we presented a method for estimating the size of a Web search engine's index. Based on the hit counts reported by two search engines, Google and Bing, for a set of 28 words, the size of the index of each engine is extrapolated. We repeated this procedure and performed it once per day, starting in March 2006. The results do not show a steady, monotonic growth, but rather a highly variable estimated index size. The larger estimated index of the two, the one from Google, attains high peaks of close to 50 billion web pages, but occasionally drops to small indices of 2 billion pages as well. Are we measuring the extrinsic variability of the indices, or an intrinsic variability of our method? Our method is fixed: the same 28 words are sent to both search engines on every day. The frequencies of our test words are unlikely to change dramatically in a corpus as big as a crawl of the indexed Web; especially the document counts for our high-frequent words in our list should approximate (or at least be in the same order of magnitude as) the total number of documents in the index. We therefore believe that the variability we measure is largely, if not entirely attributable to the variability of the index of Google and Bing. In other words, what we are measuring is the genuine extrinsic variability of the indices, caused by changes (e.g., updates, upgrades, overhauls) of the indices. In Figure 1 we highlighted several publicly announced changes to both search engines' indices, many of which co-occur with drastic changes in index size as estimated by our method (20 out of the 24 reported changes in the Google index, and 6 out of 12 changes in Bing's index).

This variability, noted earlier also by Rousseau (1999), Bar-Ilan (1999), and Payne and Thelwall (2008), should be a cause for concern for any non-longitudinal study that adopts

[10] http://googlewebmastercentral.blogspot.com/.
[11] http://googleblog.blogspot.com/.
[12] http://blogs.bing.com/.
[13] http://searchenginewatch.com.
[14] A complete, numbered list of these events can be found at http://toinebogers.com/?page_id=757.

reported hit counts. It has been pointed out that "Googleology is bad science" (Kilgariff, 2007), meaning that commercial search engines exhibit variations in their functioning that do not naturally link to the corpus they claim to index. Indeed, it is highly unlikely that the real indexable Web suddenly increased from 20 to 30 billion pages in a matter of weeks in October 2014; yet, both the Bing and Google indices report a peak in that period. It is important to note, however, that the observed instability of hit counts does not automatically imply that measuring other properties of search engines for use in webometric research, such as result rankings or link structure, suffer from the same problem.

Our estimates do not show a monotonic growth of Web search engines' indices, which was one of the hypothesized outcomes at the onset of this study in 2006. The results could be taken to indicate that the indexed Web is not growing steadily the way it did in the late 1990s. They may even be taken to indicate the indexed Web is not growing at all. Part of this may relate to the growth of the unindexed Deep Web, and a move of certain content from the indexed to the Deep Web.

The unique perspective of our study is its longitude. Already in 1999, Rousseau remarked that collecting time series estimates should be an essential part of Internet research. The nine-year view visualized in Figure 1 shows that our estimation is highly variable. It is likely that other estimation approaches, e.g. using link structure or result rankings, would show similar variance if they were carried out longitudinally. Future work should include comparing the different estimation methods over time periods, at least of a few years. The sustainability of this experiment is non-trivial and should be planned carefully, including a continuous monitoring of the proper functioning. The scripts that ran our experiment for nearly nine years, and are still running, had to be adapted to changes in the web interfaces of Google and Bing repeatedly. The time required for adapting the scripts after the detection of a change caused the loss of 6-7% of all possible daily measurements.

Our approach, but also the different approaches discussed in the section on related research introduce different kinds of biases. We list here a number of possible biases and how they apply to our own approach:

Query bias. According to Bharat and Broder (1998), large, content-rich documents have a better chance of matching a query. Since our method of absolute size estimation relies on the hit counts returned by the search engines, it does not suffer from this bias, as the result pages themselves are not used.

Estimation bias. Our approach relies on search engines accurately reporting the genuine document frequencies of all query terms. However, modern search engines tend to not report the actual frequency, but instead estimate these counts, for several reasons. One such reason is their use of federated indices: a search engine's index is too large to be stored on one single server, so the index is typically divided over many different servers. Update lag or heavy load of some servers might prevent a search engine from being able to report accurate, up-to-date term counts. Another reason for inaccurate counts is that modern search engines tend to use document-at-a-time (DAAT) processing instead of term-at-a-time (TAAT) processing (Turtle & Flood, 1995). In TAAT processing the entire postings list is traversed for each query term in its entirety, disregarding relevant documents with each new trip down the postings list. In contrast, DAAT processing the postings list is traversed one document at a time for all query terms in parallel. As soon as a fixed number of relevant documents—say 1,000—are found, the traversal is stopped and the resulting relevant documents are returned to the user. The postings list is statically ranked before traversal (using measures such as PageRank) to ensure high quality relevant documents. Since DAAT ensures that, usually, the entire postings list does not have to be

traversed, the term frequency counts tend to be incomplete. Therefore, the term frequencies are typically estimated from the section of the postings list that was traversed.

Malicious bias. According to Bharat and Broder (1998, p. 384), "a search engine might rarely or never serve pages that other engines have, thus completely sabotaging our approach". This unlikely scenario is not likely to influence our approach negatively. However, if search engines were to maliciously inflate the query term counts, this would seriously influence our method of estimating the absolute index sizes.

Domain bias. By using text corpora from a different domain to estimate the absolute index sizes, a domain bias can be introduced. Because of different terminology, term statistics collected from a corpus of newswire, for instance, would not be applicable for estimating term statistics in a corpus of plays by William Shakespeare or corpus of Web pages. We used a corpus of Web pages based on DMOZ, which should reduce the domain bias considerably. However, in general the pages that are added to DMOZ are of high quality, and are likely to have a higher-than-average PageRank, which might introduce some differences between our statistics and the ideal statistics.

Cut-off bias. Some search engines typically do not index all of the content of all web pages they crawl. Since representative information is often at the top of a page, partial indexing does not have adverse effect on search engine performance. However, this cut-off bias could affect our term estimation approach, since our training corpus contains the full texts for each document. Estimating term statistics from, say, the top 5 KB of a document can have a different effect than estimating the statistics from the entire document. Unfortunately, it is impractical to figure out what cut-off point the investigated search engines use so as to replicate this effect on our training corpus.

Quality bias. DMOZ represents a selection of exemplary, manually selected web pages, while it is obvious that the web at large is not of the same average quality. Herein lies a bias of our approach. Some aspects of the less representative parts of the web have been identified in other work. According to Fetterly et al. (2005), around 33% of all Web pages are duplicates of one another. In addition, in the past about 8% of the WWW was made up of spam pages (Fetterly et al., 2004). If this is all still the case, this would imply that over 40% of the Web does not show the quality nor the variation present in the DMOZ training corpus.

Language bias. Our selection of words from DMOZ are evenly spread over the frequency continuum and show that DMOZ is biased towards the English language, perhaps more than the World Wide Web at large. A bias towards English may imply an underestimation of the number of pages in other languages, such as Mandarin or Spanish.

This exploratory study opens up at least the following avenue for future research that we intend to pursue. We have tacitly assumed that a random selection of DMOZ pages represents "all languages". With the proper language identification tools, by which we can identify a proper DMOZ subset of pages in a particular language, our method allows to focus on that language. This may well produce an estimate of the number of pages available on the Web in that language. Estimations for Dutch produce numbers close to two billion Web pages[ii]. Knowing how much data is available for a particular language, based on a seed corpus, is relevant background information for language engineering research and development that uses the web as a corpus (Kilgariff & Grefenstette, 2003).

References

Almind, T.C. & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. *Journal of Documentation*, *53*, 404–426.

Anagnostopoulos, A., Broder, A. & Carmel, D. (2006). Sampling search-engine results. In *Proceedings of WWW '06*, (pp. 397–429). New York, NY, USA: ACM Press.

Bar-Ilan, J. (1999). Search engine results over time: a case study on search engine stability. *Cybermetrics*, *2*, 1.

Bar-Ilan, J. (2004). The use of Web search engines in information science research. *Annual Review of Information Science and Technology*, *38*, 231–288.

Bar-Ilan, J., Mat-Hassan, M. & Levene, M. (2006). Methods for comparing rankings of search engine results. *Computer Networks*, *50*, 1448–1463.

Bar-Yossef, Z. & Gurevich, M. (2006). Random sampling from a search engine's index. In *Proceedings of WWW '06* (pp. 367–376). New York, NY, USA: ACM Press.

Bar-Yossef, Z. & Gurevich, M. (2011). Efficient search engine measurements. *ACM Transactions on the Web*, *5*, 1–48.

Bharat, K. & Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of WWW '98* (pp. 379–388). New York, NY, USA: ACM Press.

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. PhD thesis, Royal School of Library and Information Science.

Dobra, A. & Fienberg, S.E. (2004). How large is the World Wide Web? In *Web Dynamics* (pp. 23–43). Berlin: Springer.

Fetterly, D., Manasse, M. & Najork, M. (2005). Detecting phrase-level duplication on the World Wide Web. In *Proceedings of SIGIR '05* (pp. 170–177). New York, NY, USA: ACM Press.

Fetterly, D., Manasse, M. & Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam Web pages. In *Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004* (pp. 1–6).

Gulli, A. & Signorini, A. (2005). The indexable Web is more than 11.5 billion pages. In *Proceedings of WWW '05* (pp. 902–903). New York, NY, USA: ACM Press.

Henzinger, M., Heydon, A., Mitzenmacher, M. & Najork, M. (2000). On near-uniform URL sampling. *Computer Networks*, *33*, 295–308.

Hirate, Y., Kato, S. & Yamana, H. (2008). Web structure in 2005. In Aiello, W., Broder, A., Janssen, J. & Milios, E. (Eds.) *Algorithms and Models for the Web-Graph*, vol. 4936, Lecture Notes in Computer Science, (pp. 36–46). Berlin: Springer.

Khelghati, M., Hiemstra, D. & Van Keulen, M. (2012). Size estimation of non-cooperative data collections. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services* (pp. 239–246). New York, NY, USA: ACM Press.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on Web as corpus. *Computational Linguistics*, *29*.

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, *33*, 147–151.

Kleinberg, J.M., Kumari, R., Raghavan, P., Rajagopalan, S. & Tomkins, A.S. (1999). The Web as a Graph: Measurements, Models, and Methods. In *COCOON '99: Proceedings of the 5th Annual International Conference on Computing and Combinatorics* (pp. 1–17). Berlin: Springer.

Koehler, W. (2004). A longitudinal study of Web pages continued: a report after six years. *Information Research*, 9. http://www.informationr.net/ir/9-2/paper174.html.

Lawrence, S. & Giles, C.L. (1998). Searching the World Wide Web. *Science*, *280*, 98–100.

Lawrence, S. & Giles, C.L. (1999). Accessibility of Information on the Web. *Nature*, *400*, 107–109.

Lewandowski, D. & Höchstötter, N. (2008). Web searching: a quality measurement perspective. In Spink, A. & Zimmer, M. (Eds.) *Web Search*, vol. 14, Information Science and Knowledge Management (pp. 309–340). Berlin: Springer.

Payne, N. & Thelwall, M. (2008). Longitudinal trends in academic Web links. *Journal of Information Science*, *34*, 3–14.

Rousseau, R. (1999). Daily time Series of common single word searches in AltaVista and NorthernLight. *Cybermetrics*, *2*, 1.

Spink, A., Jansen, B.J., Kathuria, V. & Koshman, S. (2006). Overlap among major Web search engines. *Internet Research*, *16*, 419–426.

Thelwall, M. (2008). Quantitative comparisons of search engine results. *Journal of the American Society for Information Science and Technology*, *59*, 1702–1710.

Thelwall, M. (2009). Introduction to Webometrics: Quantitative Web Research for the Social Sciences. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1, pp. 1–116.

Thelwall, M. & Sud, P. (2012). Webometric research with the Bing search API 2.0. *Journal of Informetrics*, *6*, 44–52.

Turtle, H. & Flood, J. (1995). Query evaluation: Strategies and optimizations. *Information Processing & Management*, *31*, 831–850.

Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, *35*, 469–480.

Zimmer, M. (2010). Web Search Studies: Multidisciplinary Perspectives on Web Search Engines. In Hunsinger, J., Klastrup, L. & Allen, M. (Eds.) *International Handbook of Internet Research* (pp. 507–521). Berlin: Springer.

Zipf, G.K. (1949). *Human Behaviour and the Principle of Least Effort*; Reading, MA: Addison-Wesley.

Online Attention of Universities in Finland: Are the Bigger Universities Bigger Online too?

Kim Holmberg[1]

[1] *kim.j.holmberg@utu.fi*
Research Unit for the Sociology of Education, University of Turku, 20014 Turku (Finland)

Abstract
As universities have entered a time of increased demand for public outreach and measurable impact, the universities are also exploring social media for student recruitment and science communication. Because many of the popular social media sites are free to use they could provide more democratic channels for organizational communication and marketing efforts. This research in progress investigates the social media presences of 14 universities in Finland and studies whether the offline performances of the universities are reflected in social media. The results suggest that while the RG score from ResearchGate and the Google Trends score for relative search volume correlate well with both productivity of the universities and university rankings, some of the other social media sites do not reflect the institutional characteristics as well. This is assumed to be a result of different types of usage and different purposes of the different social media sites.

Conference Topics
Webometrics; Altmetrics; Country-level studies

Introduction
Universities have entered a time of increased demand for public outreach and measurable impact. While competing for students the Humboldtian research universities try their best to conduct high quality research for the benefit of the society and to create a foundation for the research based education. At the same time social media has become mainstream in organizational communication (e.g., Badea, 2014; Huang, Baptista & Galliers, 2013; Lovejoy & Saxton, 2012). Organizations use social media for various purposes, both internally and externally, and for universities social media would seem to be an especially efficient tool for public outreach and for recruiting students. Social media are particularly efficient for sharing information through the online social networks, an aspect that would allow universities to efficiently reach their audiences. As the most popular social media sites are free to use, they may provide a more democratic way for universities to reach out to the various audiences and interest groups. This research in progress investigates whether this is true in the case of 14 Finnish universities: are smaller universities taking full advantage of the more democratic ways of communication or are the bigger universities with more resources also "bigger" in social media?

Literature review
Forkosh-Baruch and Hershkovitz (2012) investigated the use of social media sites Twitter and Facebook for scholarly purposes among higher education institutes in Israel. Their findings showed how the social media sites were extensively used for sharing academic or professional news. The authors suggest that use of these social media sites could therefore promote knowledge sharing and informal learning. Based on a content analysis of the messages shared in social media by the group of Israeli HEIs, the authors also discovered that the social media usage patterns followed similar offline usage patterns. The similar patterns here being the perception that colleges are more open and social, while universities tend to focus more on research and involvement in the research community; characteristics that were discovered in the content of the analyzed social media messages. Because of this lack of socializing and

interactivity among the universities, the authors conclude that "*the potential of SNS [social networking sites] as means of sharing academic knowledge in higher education institutes in Israel has not been actualized yet, but is indeed being explored by these organizations…*" With this the authors emphasize the importance of interactivity and audience involvement in organizational communication in social media.

In addition to social media visibility, interest towards universities, as measured by search volume on Google Trends, has also been discovered to have a connection with academic reputation (Vaughan & Romero-Frías, 2014). Vaughan and Romero-Frías (2014) used Google Trends to collect the relative search volume of the top 50 universities in the QS ranking from the US and the 56 Spanish universities included in the ARWU ranking. Their findings indicate that highly ranked universities attracted also more attention, as measured by search volume. In Google Trends the results can also be focused on searches within specific countries; one could for instance look up the search volume for "Kate Upton" in the UK or "Justin Bieber" in Norway. Vaughan and Romero-Frías (2014) discovered that while a great amount of searches for the US universities came from outside the US, only a few searches for the Spanish universities came outside of Spain, which according to the authors also reflects the international positions of the two sets of universities. As searches in English in general and for universities in English in particular may be assumed to be relatively low in non-English speaking countries, it may not make sense in all cases to focus on the country-level search volume in English. For instance in the case of Finnish universities we can assume that searches for them from Finland would mainly use their Finnish names, while the volume of searches in English would mainly reflect the international attention and interest.

Thelwall and Kousha (in press) took another approach to study universities' online presences and investigated whether the usage of ResearchGate and the publications uploaded to it by researchers has a connection with the "academic hierarchies" of different university rankings. ResearchGate is a scholarly social networking site where scholars can create their own profile pages and upload their publications to it, network with other researchers, and find possibly relevant and interesting publications, based partly on their own interests (as indicated on their profile pages) and partly on the interests of those in their social network. Based on researchers' activity on the site and their publications (both number of publications and the journal impact factor of the journals where the papers have been published in) ResearchGate calculates RG scores as a measure of individual researchers' "scientific reputation". The exact formula with which the RG score is calculated is, however, not revealed by ResearchGate. This approach can also be criticized because use of journal impact factors to evaluate or rank individual researchers has increasingly been criticized and condemned (e.g., DORA, 2013). Collectively the RG scores for researchers from a specific institution can give an institutional RG score, supposedly indicating institutional reputation. This is the score that Thelwall and Kousha (in press) used to compare to different university rankings. Their findings showed a moderate correlation between the rankings on ResearchGate and the other university rankings (The Higher Education ranking, QS world university rankings, Academic Ranking of World Universities, CWTS Leiden ranking, and the ranking on Webometrics.info). Because the rankings on ResearchGate are based on researchers' activities on the site and their research work, the findings by Thelwall and Kousha (in press) suggest that the usage of ResearchGate "broadly reflects traditional academic capital."

The current university rankings do place somewhat different weight on different things. For instance the ranking provided by the Webometrics.info measure online visibility, presence and impact, weighting most on visibility as measured by hyperlinks, while the other rankings use more traditional measures of research productivity and impact, i.e. publications and citations, and give them different weights (Aguillo, Bar-Ilan, Levene & Ortega, 2010). Still the different university rankings tend to give similar results, which would suggest that

universities performing well in one area also perform well on other areas. In other words, a university that is performing well when assessed with publications and citations seems to also perform well online. But whether this is reflected to the universities usage of social media and the attention they receive there is unclear. Attention and visibility in social media, as measured with various social media metrics, has been suggested to be a potential indicator of research impact (e.g., Bollen, Van De Sompel, Hagberg & Chute, 2009; Priem & Hemminger, 2010; Lin & Fenner, 2013). These new social media metrics, the so called altmetrics, could potentially give a more nuanced view of the attention towards research outputs. It has also been suggested that altmetrics could provide indicators for the societal impact of research (Bornmann, 2014) or provide knowledge about the interest towards research from a wider audience outside academia (Haustein, 2014). Although not yet extensively studied, altmetrics may also be able to provide country-level indicators of research impact, as Alhoori et al. (2014) have discovered significant correlations between bibliometric data and some altmetrics when aggregated to the country-level.

The research in progress presented here investigates the social media presence of 14 universities in Finland and with that opens research for institutional altmetrics.

Data and methods

The 14 universities in Finland all have online presences in social media. All have profiles, pages or groups on the most popular social media sites Facebook, Twitter, YouTube and LinkedIn, and some also have accounts on Instagram, Flickr or Pinterest. These are usually linked to from the university's webpage. The goal of this research is to 1) study how universities are using social media, 2) how much attention they have attracted, and 3) whether this attention is connected to other offline descriptive metrics about the universities' resources and performance.

Descriptive statistics were manually collected by visiting the universities' official social media profiles, as linked to from the universities' websites. The data consists of the number of tweets, followers and following on Twitter, "likes" on Facebook, subscriptions to and views on the universities' YouTube channel, followers on LinkedIn, and the universities RG score on ResearchGate. In addition to this universities' relative search volumes, as indicated by Google Trends, were retrieved. As the Google Trends score is a score relative to the search volume of the other words searched at the same time (maximum of five different terms compared in one search), we retrieved the scores for the universities' names in English by keeping the two universities with the highest scores included in the search for reference. This way all the scores were relative to those universities with the biggest search volume. The descriptive data about the universities and their performance were retrieved from the report of the State of Scientific Research in Finland, commissioned by the Academy of Finland (http://www.aka.fi/en-GB/A/Decisions-and-impacts/The-state-of-scientific-research-in-Finland/). This performance data consists of variables from 2012; the number of PhDs awarded, total person-years of the teaching and research staff, research funding, and number of publications. In addition to these the rankings of the Finnish universities were retrieved from the following university rankings; CWTS Leiden, ARWU, QS, THE, and Webometrics.info. Only Webometrics.info could provide the rankings for all but one of the 14 universities: the ranking of the fairly new University of the Arts (the former Academy of Fine Arts, Sibelius Academy and Theatre Academy merged to the University of the Arts in 2013). Nine of the 14 universities were found on QS ranking, seven on the CWTS ranking and on THE ranking, and five on the ARWU ranking. Only rankings from Webometrics.info and the QS were used in further analysis.

Spearman rank correlations between the social media metrics and offline data about the universities' performance were investigated to discover whether social media usage would

follow the academic capital at these universities. In addition to this, connections between the social media metrics and university rankings were also tested to see whether the universities reputation and performance was reflected in social media attention and usage.

Results

The different offline university specific metrics are clearly associated, showing how number of students and faculty, funding and publications are all very tightly connected (Table 1). This naturally means that universities with more funding have bigger faculty, more students and produce more publications. As some of these metrics are also used for university rankings it is only natural that the rankings correlate well with these (0.830, n=13, between publications and Webometrics.info; 0.867, n=9, between publications and QS ranking, both Spearman rank correlations significant at level 0.05). The universities that were omitted from the analysis due to non-existent data on Webometrics.info and QS were the universities with the least publications, a probable explanation why they were not covered by the university rankings.

Table 1. Spearman rank correlations between the social media metrics and offline metrics of the 14 universities in Finland. Correlations in bold are significant at the 0.05 level, two-tailed. (RG = RG score; GT = Google Trends score; Tw = Tweets in Twitter; Tw.a = Followers on Twitter; Tw.b = Following on Twitter; FB = Facebook likes; YTs = YouTube subscriptions; YTv = YouTube views; LI = LinkedIn followers; Phd. = PhDs awarded in 2012; Fa. = Faculty in 2012; Fu. = Research funding in 2012; Pu. = Peer-reviewed publications in 2012).

	RG	GT	Tw	Tw.a	Tw.b	FB	YTs	YTv	LI	PhD.	Fa.	Fu.	Pu.
RG	**1**	**0,679**	0,473	0,367	0,046	0,389	0,337	0,204	0,385	**0,923**	**0,938**	**0,952**	**0,969**
GT		**1**	0,444	0,435	0,251	0,266	0,316	0,342	0,160	**0,750**	**0,690**	**0,648**	**0,746**
Tw			**1**	**0,776**	0,516	0,345	**0,579**	**0,587**	**0,618**	**0,670**	**0,604**	0,534	**0,543**
Tw.a				**1**	0,499	0,059	**0,557**	**0,613**	**0,749**	**0,551**	0,468	0,393	0,420
Tw.b					**1**	-0,099	0,233	0,314	0,196	0,192	0,143	0,064	0,116
FB						**1**	0,260	0,015	-0,178	0,463	**0,574**	**0,604**	0,389
YTs							**1**	**0,871**	**0,700**	0,397	0,414	0,392	0,317
YTv								**1**	**0,754**	0,333	0,266	0,231	0,284
LI									**1**	0,423	0,349	0,323	0,380
PhD.										**1**	**0,974**	**0,949**	**0,960**
Fa.											**1**	**0,987**	**0,947**
Fu.												**1**	**0,943**
Pu.													**1**

Overall the number of tweets and Facebook 'likes' correlated moderately with the performance metrics of universities (Table 1), with tweets giving somewhat higher correlations on average than Facebook. While the number of followers on Twitter had some connection to the offline metrics, the number of followed accounts only had a very weak connection. This suggests that larger universities are not necessarily more active on Twitter, but that they still generate more attention.

Our findings indicate that research productivity (and the other offline metrics), as measured by the number of peer-reviewed publication from 2012, did correlate almost perfectly with the RG score on ResearchGate (0.969 Spearman, significant at the 0.05 level). The RG score did, however, not correlate well with many of the other social media metrics. Search volume on Google Trends also correlated well with the offline metrics, with the Spearman rank correlation between Google Trends score and number of publications being 0.746, significant

at 0.05 level. The relationships of these two cases are illustrated in figures 1 and 2. In both cases the University of Helsinki, the largest university in Finland, appear as an outlier due to its size. In figure 2 we can see a bit more scattering and how the University of Jyväskylä, and to some extent University of Eastern Finland and Aalto University, although not having exceptionally many publications still have managed to attract significant interest as measured by search volume on Google.

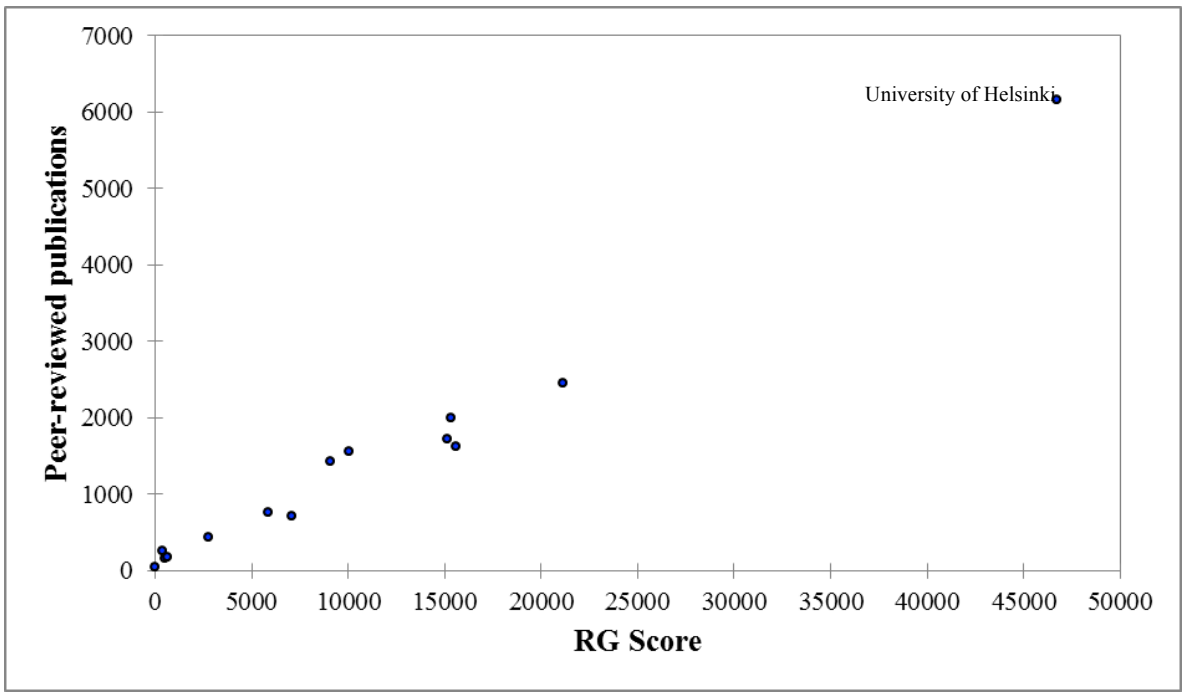


Figure 1. Correlation between the RG score (from ResearchGate) and the number of peer reviewed publications in 2012 at the Finnish universities (0.969 Spearman).

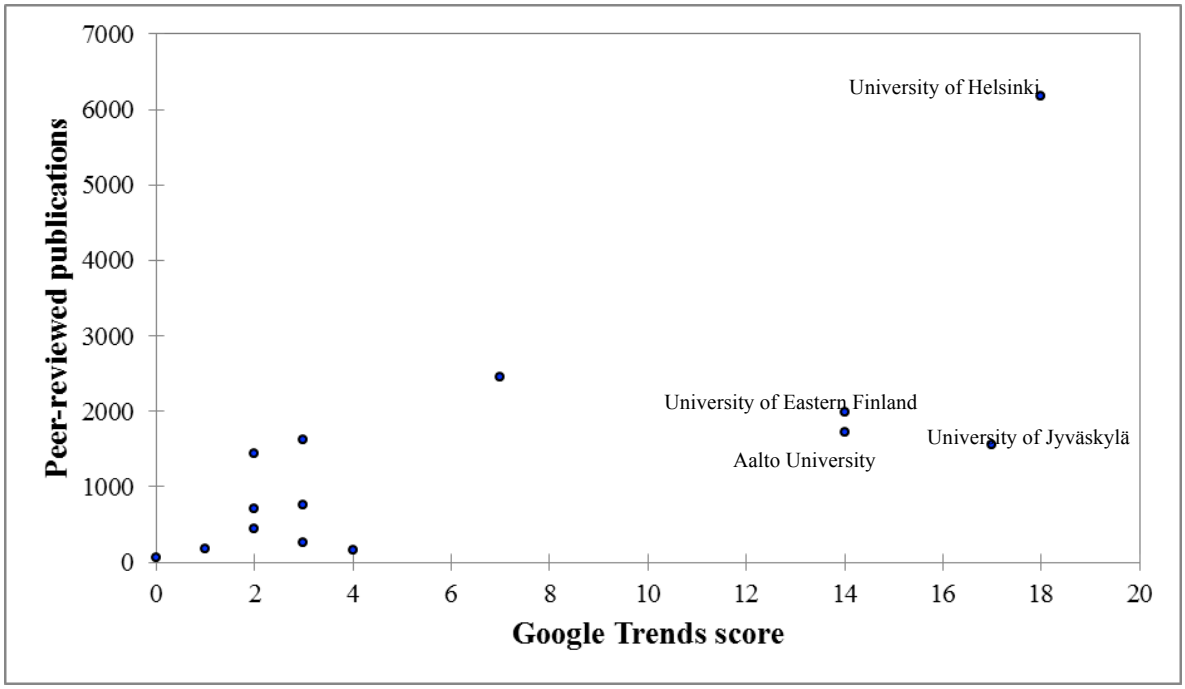


Figure 2. Correlation between the search volume as measured by Google Trends and the number of peer reviewed publications in 2012 at the Finnish universities (0.746 Spearman).

Discussion and conclusions

We set out to investigate the social media presences of 14 universities from Finland and the attention they have received in social media. Our results show that while in many cases the larger and more productive universities are also more active or receive more attention in social media; this is not always the case (Table 1). This suggests that the smaller universities, at least in this small sample, are benefitting from the more democratic channels of social media. Our findings also suggest, in line with the findings by Thelwall and Kousha (in press), that the institutional RG scores and the RG scores for individual researchers on ResearchGate, may be a promising source for altmetrics at institutional and possibly even country level. Due to the uncertainty of how the RG score exactly is calculated and because of the use of journal impact factors in that calculation more research into the topic is clearly needed.

The next step of this research in progress will be a content analysis of the universities social media accounts. This will provide new knowledge about how the universities are represented in social media, for what purposes they use social media, and how attention in social media is created. This will provide important background information for institutional altmetrics.

References

Aguillo, I., Bar-Ilan, J., Levene, M. & Ortega, J.L. (2010). Comparing university rankings. *Scientometrics*, *85*, 243-256. DOI: 10.1007/s11192-010-0190-z

Alhoori, H., Furuta, R., Tabet, M., Samaka, M. & Fox, E.F. (2014). Altmetrics for country-level research assessment. *Lecture Notes in Computer Science*, *8839*, 59-64. DOI: 10.1007/978-3-319-12823-8_7

Badea, M. (2014). Social media and organizational communication. *Procedia Social and Behavioral Sciences*, *149*, 70-75. DOI: 10.1016/j.sbspro.2014.08.192

Bollen, J., Van De Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures, *PLoS One*, *4*(6). DOI: 10.1371/journal.pone.0006022.

Bornmann, L. (2014). Validity of altmetrics data for measuring societal impact: a study using data from Altmetric and F1000 Prime. *Journal of Informetrics, 8*(4), 935-950. DOI: 10.1016/j.joi.2014.09.007

DORA (2013). San Francisco Declaration on Research Assessment. Retrieved on April 15, 2015, from http://www.ascb.org/dora-old/files/SFDeclarationFINAL.pdf.

Forkosh-Baruch, A. & Hershkovitz, A. (2012). A case study of Israeli higher-education institutes sharing scholalrly information with the community via social networks. *Internet and Higher Education*, *15*, 58-68. DOI: 10.1016/j.iheduc.2011.08.003

Haustein, S., Peters, I., Sugimoto, C.R., Thelwall, M. & Lariviére, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, *65*(4), 656-669. DOI: 10.1002/asi.23101

Huang, J., Baptista, J. & Galliers, R.D. (2013). Reconceptualizing rhetorical practices in organizations: The impact of social media on internal communications. *Information & Management*, *50*(2-3), 112-124. DOI: 10.1016/j.im.2012.11.003

Lin, J. & Fenner, M. (2013). The many faces of article-level metrics. *Bulletin of the Association for Information Science and Technology*, *39*(4). DOI: 10.1002/bult.2013.1720390409

Lovejoy, K. & Saxton, G.D. (2012). Information, community, and action: How nonprofit organizations use social media. *Journal of Computer-Mediated Communication*, *17*(3), 337-353. DOI: 10.1111/j.1083-6101.2012.01576.x

Priem, J., & Hemminger, B. M. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. First Monday, vol. 15, no. 7. Retrieved on January 20, 2015 from http://firstmonday.org/article/viewArticle/2874/2570.

Thelwall, M. & Kousha, K. (in press). ResearchGate: disseminating, communicating and measuring scholarship? *Journal of the Association for Information Science and Technology*. DOI: 10.1002/asi.23236.

Ranking Journals Using Altmetrics

Tamar V. Loach[1,2] and Tim S Evans[2]

[1] *t.loach @ digital-science.com*
Digital Science, 2 Trematon Walk, Wharfdale Road, London, N1 9FN (U.K.)

[2] *{t.loach, t.evans} @ imperial.ac.uk*
Imperial College London, Centre for Complexity Science, South Kensington Campus, London SW7 2AZ (U.K.)

Abstract
The rank of a journal based on simple citation information is a popular measure. The simplicity and availability of rankings such as Impact Factor, Eigenfactor and SciMago Journal Rank based on trusted commercial sources ensures their widespread use for many important tasks despite the well-known limitations of such rankings. In this paper we look at an alternative approach based on information on papers from social and mainstream media sources. Our data comes from altmetric.com who identify mentions of individual academic papers in sources such as Twitter, Facebook, blogs and news outlets. We consider several different methods to produce a ranking of journals from such data. We show that most (but not all) schemes produce results, which are roughly similar, suggesting that there is a basic consistency between social media based approaches and traditional citation based methods. Most ranking schemes applied to one data set produce relatively little variation and we suggest this provides a measure of the uncertainty in any journal rating. The differences we find between data sources also shows they are capturing different aspects of journal impact. We conclude a small number of such ratings will provide the best information on journal impact.

Conference Topic
Altmetrics

The background and purpose of the study
Journal metrics, such as the Thomson Reuters Journal Impact Factor, were originally developed in response to a publisher need to demonstrate the academic attention accorded to research journals. Over the intervening 50 years since Garfield's work in the field, the Impact Factor and other metrics, such as Eigenfactor (Bergstrom, 2007), have been used and misused in a variety of contexts in academia. An oft-discussed perception is that a journal-level metric is a good proxy for the quality of the articles contained in a journal.

In the evaluation and bibliometrics communities citation counting is generally understood not to be an appropriate proxy for quality but rather a measure of attention. The type of attention being measured in this case is quite specific and has particular properties. What is being measured is the attention to a paper of peers in related fields. The bar for registration of this attention is relatively high – the researcher or researchers making the citation must deem the target article to be of sufficient value that they include a citation in a work of their own that in turn is deemed publishable (e.g. see Archambault & Lariviére, 2009, and references therein). The timescale associated with citations is also long – typically being limited by the review and publication process associated with particular fields. Additionally, it is accepted that journal-level metrics say little regarding the merit of particular articles in the journal since journal-level metrics are often calculated based on thousands of articles and are often biased by the performance of the tails of the distribution of citations. These realisations have led to the recent growth in popularity of article-level metrics or altmetrics.

Altmetrics have broadened the range of types of attention that we can measure and track for scholarly articles. Mostly based in social and traditional media citations, the altmetric landscape is one that is constantly changing with the introduction of different data sources all the time. While, one the one hand, altmetrics suffer from all the unevenness of traditional citations, they occur over different timescales, which provides us with a more nuanced view

of the lifecycle of a scholarly work. Aggregating alternative metrics at a journal level will complement Journal Impact Factor, giving us new insights into different facets of attention.

Traditional citation-based metrics are difficult to calculate since they are based on the bibliometric journal databases, such as Thomson Reuters' Web of Science. Conversely, Altmetrics are conglomerates of disparate sources of references to research output derived from non-traditional sources, primarily modern electronic sources characterised by fast response times (see Bornmann, 2014, for a recent overview). The lack of any systematic peer review is another characteristic of most altmetric data. The open and electronic nature of much altmetric data offers the prospect of alternative paper and journal metrics, which may be more accessible to stakeholders. The rapid response of such data to innovations suggests such metrics might offer improvements over metrics based on slower traditional sources.

This paper considers a number of approaches to the aggregation of altmetric data in order to create a robust journal-level metric that complements the existing citation-based metrics already in use across the academic community. The aim is not to create a contender for a single metric to quantify journal output but instead to create a useful measure that gives "the user" a sense of the non-citation attention that a journal attracts in the same way that Journal Impact Factor, Eigenfactor and other related metrics give this sense for citation attention.

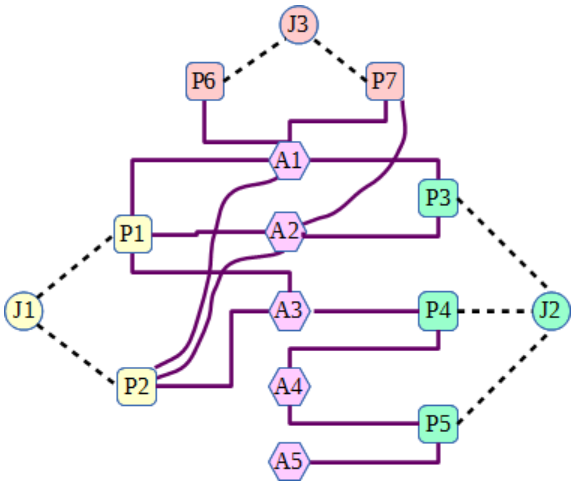


Figure 1. The relationships recorded in our altmetric.com data. The raw data illustrated here contains fifteen "mentions" (solid lines) by five "authors" (hexagons A1 to A5) of seven papers (squares P1 to P7). We also know the journal (circles), which published a paper (dashed lines).

Data Sources

In this paper we use the 2013 IF (Impact Factor) and EF (Eigenfactor) as examples of traditional sources of journal ratings. Our altmetric data comes from 20 months of data from altmetric.com, a commercial company. For each mention about a paper we had the journal in which it was published, the source (twitter, Facebook, etc.) and the account (here termed an 'author'), as shown in Figure 1. In our case, a 'paper' has to be an article coming from a known journal. A single 'author' for us is a single account (e.g. one twitter account) or a single source (a news outlet such as a newspaper). In some cases several different authors may be responsible for one site or one author could provide information to many different sites or accounts (a twitter account, a facebook account, a blog, etc) but in our data such an author appears as many distinct authors.

Methods

The simplest type of journal altmetric is one based on basic counts where each mention of a paper in a journal adds one to that journal's count. We collected counts for social media '*sbc*',

non-social media '*nsbc*' (e.g. downloads) and combined scores '*bc*' (for blind count i.e. with no weighting for different sources). We also obtained the current journal rating produced by altmetric.com (denoted '*ca*'), which is a weighted count rating in which different sources are given different weights (blogs and news sources get highest weighting).

Network Definitions

A criticism of simple count based methods, such as Impact Factor or our altmetric counts discussed above, is that some citations or some altmetric authors are more important than others. Eigenfactor is an illustration of a response to these criticisms in the realm of traditional data (Bergstrom, 2007), as it uses a network based view to arrive at a PageRank style measure. We will also turn to a network-based view in order to look at a wide range of measures, which probe the relationships between journals on a much larger scale.

There are many possible network representations of our data. In this paper we will focus only on networks in which the nodes represent journals. The central idea in our construction of the relationship between two journals is that we only want to consider activity from authors who mention both journals because only these authors are making an implicit comparison between journals. The activity of each author is used to define their own "field of interest" in a self-consistent manner and so the activity of authors is used to make comparisons between journals in the same field as defined by each author's interests. This ensures that at a fundamental level we avoid the much discussed problem of making comparisons between papers or journals from different fields. An author only interested in medical issues will only contribute to the evaluation of Nature, Science and so forth in terms of their interest in these multidisciplinary journals relative to Cell or other specialised journals.

A useful analogy here is that each journal is a team and an author who mentions articles published in two journals represents one game between these journals – our pairwise comparison. The score in each game is the number of mentions so in comparing two journals *j* and *l*, the score for journal *j* from the game represented by author *a* is recoded as the entry J_{ja}.in a rectangular matrix. In Figure 1 the game between J1 and J2 represented by author A2 has the result 2-1, a 'win' for journal 1 over journal J2 suggesting that we should rate journal J1 more highly than journal J2 given the activity of this one author.

We shall consider three different ways of quantifying the journal relationships, the network edges. Our first approach gives us an adjacency matrix *S* where the entry S_{jl} gives the weight of the edge from journal *j* to journal *l*, and this is given by $S_{jl} = \frac{1}{|A_{jl}|}\sum_{a\in A_{jl}} J_{ja}$, where $A_{jl} = \{a|J_{ja} > 0, J_{la} > 0\}$. Here *j* and *l* represent different journals and *a* is one author. J_{ja} is a matrix, which is equal to the number of papers mentioned by author *a* which were published in journal *j*. The expression for S_{jl} is counting the number of times papers published in journal *j* are mentioned by authors who also mention papers in journal *l,* with the total normalised by the number of such authors. Note that this defines a sparse, weighted and directed network. In our conventions if journal *j* is better than journal *l* we will have $S_{jl} > S_{lj}$.

Our second definition gives us an adjacency matrix *P* where $P_{jl} = \frac{1}{|A_{jl}|}\sum_{a\in A_{jl}} \theta(J_{ja} - J_{la})$. Here $\theta(x) = 1$ if $x > 0$ otherwise this function gives 0. This definition counts how many authors mention more papers in journal *j* than they do papers in journal *l*., normalising again by the number of authors who are able to make this pairwise comparison. Again $P_{jl} > P_{lj}$ if journal *j* is better than journal *l*.

Finally we define an adjacency matrix *Q* where $Q_{jl} = \frac{1}{|A_{jl}|}\sum_{a\in A_{jl}} \Theta(J_{ja} - J_{la})$. Here $\Theta(x) = 1$ if $x > 0$, $\Theta(0) = 0.5$ while for negative values this function gives 0. This definition counts how many authors mention more papers in journal *j* than they do papers in journal *l*

but when this is balanced gives an equal weighting to both side. This definition has the useful property that $Q_{jl} + Q_{lj} = 1$ (not generally true for matrix P).

Network Measures

Once we have our network with journals as nodes, we need to find ways to use this structure to define which nodes are the most important. Measures which quantify the importance of a node are known as centrality measures in social network analysis. Unfortunately, many standard measures do not take into account the weights or directions of edges, both of which carry crucial information in our case. We used two well-known network centrality measures to illustrate our approach: PageRank and HITS (e.g. see Langville & Meyer, 2012). Both may be cast as eigenvector problems and there are fast algorithms for large networks which are readily available. We apply these two methods to all three networks, giving six different ratings e.g. 'qpr' indicates a PageRank rating derived from a Q matrix while 'ph' indicates a HITs rating derived using a P matrix.

We also tried a different type of measure known as Points Spread Rating (denoted 'psr') (p.117-120, Langville & Meyer, 2012) where the rating r_j for journal j is $r_j = \sum_l (S_{jl} - S_{lj}) / n_j$, (similarly for the P and Q matrices) and n_j is the number of journals. This expression ensures that the differences $(r_j - r_l)$ in the rating of any two journals j and l are as close as possible to the actual differences in the number of average mentions of papers.

Comparing Ratings

Once we have obtained different ratings, the final task is to make a comparison. The simplest approach is to make a qualitative comparison of the top ranked journals in each case. For a more quantitative approach we used standard methods of multivariate statistics. First we found a correlation matrix whose entries express the similarity of two rating methods: the Pearson correlation matrix based on the numerical values of the ratings obtained, Spearman's matrix which based on the ranking of journals, and finally Kendall's tau. These were analysed using principle component analysis or hierarchical clustering methods.

Findings

In terms of the altmetric data we found typical fat-tailed distributions, both for the number of mentions of a paper from different sources and in terms of the number of mentions put out by a single author. Some sources, such as twitter, are significantly larger than others.

When comparing different journal rating schemes, some results were found only with Spearman and Kendall tau correlation measures (which are based on the ranks of journals). The Pearson measure (based on actual rating values gave slightly different results in some cases. However in most cases there good agreement. Some typical results are shown in Figure 2 and numbers for ranking schemes in the following text refer to the labels in Figure 2.

The variation between different rating schemes for the same altmetric data source gives relatively little variation, roughly on the same scale as the difference we find between IF and EF. The four different methods shown for ratings based on Facebook mentions (6,12,16,19) are a typical example. Clearly our Points Spread Rating scheme (psr, 21,22,23) and our simple counts of non-social media mentions (nsbc, 6) produces outliers.

Some sources, such as Facebook and News, were also noticeably different from IF and EF, but the difference was much smaller than that found with the psr rating. One source, which gave ratings well correlated with IF and EF was blogs (8, 11, 15, 18).

Likewise, most of our simple count based ratings were just as close to IF (3) or EF (5) as these two rating schemes were to each other. This includes our unweighted count of all mentions (bc, 1), the number of times papers are mentioned (pc, 7), counts of just social

media mentions (sbc, 14), and in particular the more sophisticated weighted journal ranking produced by altmetric.com (ca, 2).

Most of our work focused on statistics for the whole collection. A look at the top journals, see Table 1, confirmed that at an individual level our new altmetric network ratings were giving sensible results, but with variations which indicate the uncertainty in such rankings.

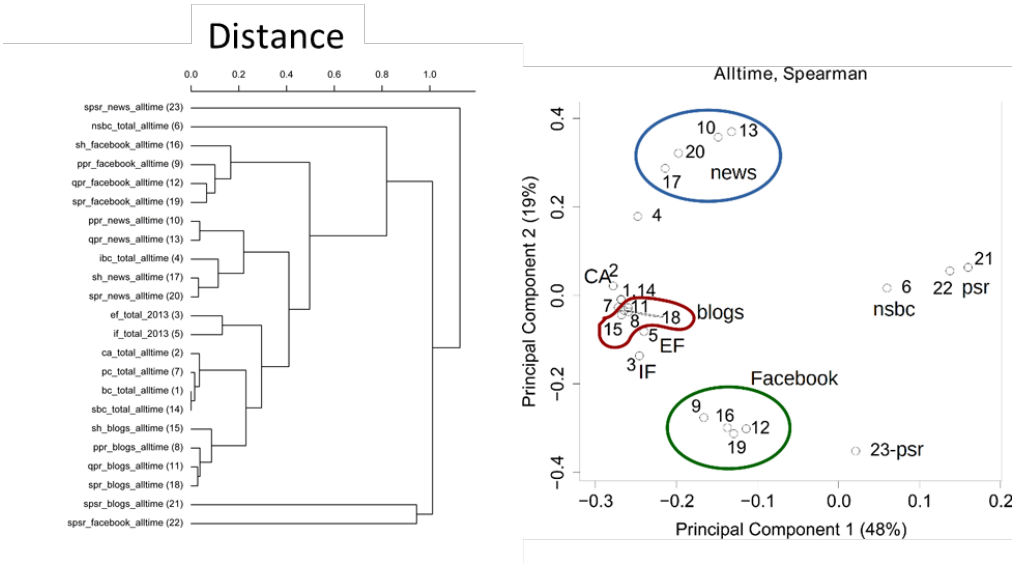


Figure 2. A comparison of some of the different ranking schemes using a Spearman correlation matrix. On the left a dendrogram and on the right a scatter plot using the first two principle components of PCA. For clarity, only a limited subset of our ratings were used in these plots.

Discussion

Given our differences between ranking based comparisons (Spearman and Kendall Tau) and results based on Pearson correlation matrices, this suggests that ratings are dominated by the measurement of the few journals, which have most of the mentions (fat tails). This is one reason we favour Spearman correlation matrices in Figure 2 and would suggest this makes sense in most journal ranking contexts.

Our Points Spread Rating scheme (psr, 21, 22, 23) seems to be reflecting very different patterns in the data from those found using other approaches. Given that the other approaches include Impact Factor, widely accepted as a measure of journal attention, we think it is hard to see a role for PSR to rank journals. Likewise, the simple blind counts of non-social media mentions (nsbc, 6) does not appear to be useful.

The remaining different altmetric sources and rating methods do show enough similarity to suggest that they are all an acceptable measure of journal importance. At the same time there are some interesting differences indicating that our altmetric based schemes are capturing different features of the impact of journals. At the very least this diversity will indicate the level of uncertainty in rating schemes. Two possible reasons for the close correlation of blogs and IF are as follows. Perhaps papers in high IF journals are of intrinsic interest to blog writers. Alternatively blog authors may read a limited number of journals but these tend to be those with high IF. Probably both factors are important, each reinforcing the other to produce the strong correlation we find.

Another interesting feature is that most of our simple count based ratings, which are not normalised by the number of articles per journal, are also well correlated with IF (3) which does use normalised counts. This can be explained if there is a correlation between the number of papers in a journal and its impact, something we can see in of count of number of papers (pc, 7). We will be looking at normalised altmetric counts in the future but it appears

normalisation may not be essential. In particular, we note the altmetric.com journal rating (ca, 2) is well correlated and so provides a good handle on the impact of journals.

Table 1. Top ten journals based on various network based altmetric measures.

Rank	*Q, HITS, Blogs*	*Q, HITS, News*	*S, PageRank, Google+*
1	Nature	Nature	Nature
2	PNAS	PNAS	PLoS ONE
3	Science	PLoS ONE	Science
4	PLoS ONE	Science	PNAS
5	New England J. of Med.	New England J. of Med.	New England J. of Med.
6	British Medical J.-C.R.Ed.	British Medical J.-C.R.Ed.	British Medical J.-C.R.Ed.
7	The Lancet (British Ed.)	Nature Communications	Scientific Reports
8	JAMA	JAMA	JAMA
9	Proc. Royal Soc. B:	The Lancet (British Ed.)	The Lancet (British Ed.)
10	Current Biology	Pediatrics	PLoS Biology

The fact that we tried many different rating methods and that (with the exception of psr based measures) they showed variations on scales no bigger than those found between IF and EF, suggests that no one method is optimal in any sense. However we can use such a suite of metrics to get a handle on the uncertainty associated with any measure. This would be of great utility for users and a contrast to the three decimal point 'accuracy' associated with IF results.

Conclusions

We have shown how to use altmetric data to provide a reasonable journal ranking. Most types of altmetric data appear to give useful information in the sense that the correlation with IF is acceptable. At the same time altmetric data can be sufficiently different that it might reflect different types of impact. Our results suggest that different rating methods can provide a measure of the uncertainty of any journal ranking. Confirming these patterns over longer periods and producing a better understanding of the social reasons for the patterns we have found are future directions for our work. It would also be interesting to compare our results with journal attention measures derived from journal usage patterns, see for example Bollen et al 2009, an aspect not included in our data.

Acknowledgments

We would like to thank Euan Adie (altmetric.com) for providing us with the altmetric data, and for useful discussions along with Jonathan Adams and Daniel Hook (Digital Science).

References

Archambault, É., & Lariviére, V. (2009). History of the journal impact factor: Contingencies and consequences. Scientometrics, *79*(3), 635–649.

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, *68*, 314-316.

Bollen, J., Van de Sompel, H., Hagberg, A. & Chute, R. (2009). A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE, 4*, e6022.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, *8*, 895-903.

Langville, A. N. & Meyer, C. D. (2012) *Who's #1?: The science of rating and ranking*, Princeton: Princeton University Press.

Who Tweets about Science?

Andrew Tsou, Tim Bowman, Ali Ghazinejad, and Cassidy Sugimoto

atsou@umail.iu.edu, ali.ghazinejad@gmail.com, sugimoto@indiana.edu, tim.bowman@gmail.com
School of Informatics and Computing, Indiana University Bloomington (Bloomington, IN, USA)

Abstract
Twitter is currently one of the primary venues for online information dissemination. Although its detractors portray it as nothing more than an exercise in narcissism and banality, Twitter is also used to share news stories and other information that may be of interest to a person's followers. The current study sampled tweeters who had tweeted at least one link to an article in one of four leading journals, with a focus on studying who, precisely, these tweeters were. The results showed that approximately 76% of the sampled accounts were maintained by individuals (rather than organizations), 67% of these accounts were maintained by a single man, and 34.4% of the individuals were identified as possessing a Ph.D, suggesting that the population of Twitter users who tweet links to academic articles does not reflect the demographics of the general public. In addition, the vast majority of students and academics were associated with some form of science, indicating that interest in scientific journals is limited to individuals in related fields of study.

Conference Topic
Altmetrics

Introduction
Twitter is currently one of the primary venues for online information dissemination. Nearly a quarter of adult Internet-users take advantage of Twitter (Pew Research, 2014), and according to Alexa (2015), as of January 22, 2015, Twitter is ranked as the 8th most visited site on the Web (and the 7th most visited in the United States). Although its detractors portray it as nothing more than an exercise in narcissism and banality, Twitter is also used to share news stories and other information that may be of interest to a person's followers. Amidst much vapidity can be found discussions or links of genuine merit, and indeed, it has been found that "academic articles are now frequently tweeted and so Twitter seems to be a useful tool for scholars to use to help keep up with publications and discussions in their fields" (Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013, p. 1). Previous research has discussed the content of such tweets, their sentiments (Thelwall, Tsou, Weingart, Holmberg, & Haustein, 2013), tweeting behaviour across venues and disciplines (Haustein, Peters, Sugimoto, Thelwall, & Larivière, 2014), the use of Twitter for altmetrics (Thelwall, Haustein, Larivière, & Sugimoto, 2013), and the effect that automated bots have on the legitimacy of using tweets to assess academic impact (Haustein et al., 2014). However, the demographics of tweeters who post links to academic articles have not yet been investigated. This study proposes to address this gap.

Methods

Sampling frame.

The initial sampling frame was a list of individuals who had provided a link to an academic article in a tweet. These tweets were gathered by running a Twitter query approximately every hour from March 17, 2012 to March 17, 2013 for each of a number of URLs of journals (Table 1). The journals were selected as leading journals that were widely tweeted (based on a manual examination of the data) and had a simple URL format for articles that could be collected by a query. Collecting tweets in this way was a practical step because many people link to articles if they mention them and it is easy to search for articles by part of URL. In

each case article URLs had a common starting text, such as a domain name, and queries for this common part matched all articles in the site. Although Twitter shortens almost all URLs in tweets, it is possible to use URL-based queries because Twitter search returns matches for the original URLs rather than the shortened versions.

Table 1. Queries for links to academic articles in Twitter.

Source	Twitter query
Nature journal	"go.nature.com"
PLOS ONE journal	"plosone.org/article"
PNAS journal	"pnas.org/content"
Science journal	"scim.ag"

This method does not retrieve all tweets of academic articles published in the selected journals. In particular, it does not capture links to copies of the articles elsewhere (e.g., self-archived preprints) and does not capture articles mentioned by name rather than by link. Also, Twitter does not guarantee comprehensive matches to all searches so it is likely that not all URLs matching the above set of queries were found. Some data was also lost due to power cuts and an enforced shutdown at Wolverhampton in December 2012. However, this provides an authoritative list of scholarly tweets.

Sample

From this sampling frame, a list of all unique twitter accounts was generated. From this list, a sample of 500 unique tweeters for each journal was randomly selected. Duplicate accounts were removed and replaced so that the sample represented 2,000 unique accounts (this was necessary as some accounts tweeted articles from more than one journal).

Survey

The initial plan was to directly survey the journal tweeters and, accordingly, a survey was set up in Qualtrics and a separate DID Cascades Twitter account was established for the purpose of tweeting a link to the survey to all 2000 account. We set up an automated system to send out invitations to the survey to the identified twitter handles in batches small enough to not violate Twitter's mass tweeting policies. However, even working within these parameters, our account was suspended immediately upon our first batch of survey invitations. We mention this failure here as it is relevant to conducting research in this environment. Although some modes of inquiry (e.g., large-scale survey research) may be more appropriate for answering certain questions, they are untenable due to the current affordances of the platform. These limitations should be taken into consideration for future analyses.

Codebook construction

Given that obtrusive research was not possible, we turned to unobtrusive measures (i.e., content analysis) to analyse the identities of those who tweet about science. The codebook was developed inductively through several iterative explorations with four researchers. Variables such as gender, academic affiliation, and (in the case of non-individuals) organization type were collected. Iterative coding led to refining of the initial categories (e.g., the "Finance" category originally proposed was expanded to "Business/Finance", "Freelance" was incorporated into the coding due to the high frequency of this position, and "Non-profit" was added in the organizational category).

One of the initial desires was to be able to tag those who were "affiliated with science." This was intended to distinguish between the "layperson" and the "scientists". This seemingly

simple distinction proved to be overwhelmingly difficult to code unobtrusively. Those who explicitly identified with academic institutions and were readily associated with science departments within those institutions were easy to identify. However, many of the non-academics were also affiliated with science in some form (e.g., government positions in science and technology). This also led to the issue of determining what constitutes science (e.g., are humanists, entrepreneurs, and technologists scientists?). This was equally difficult for organizations. For example, an online consumer or financial corporation might not have science as the main objective, but have an arm of the organization that conducts research. This question was further complicated by false negatives—that is, instances where we could not provide evidence that the individual was associated with science, but also could not provide evidence that they were not.

The issue of false positives and false negatives on other questions was addressed by adding an "unknown" option in addition to "yes/no" options. For example, one question asked whether the individual was a student. As it was frequently impossible to definitively state whether or not an individual was not a student (i.e., the lack of information regarding a person's reenrolment in a university would not, in itself, extinguish the possibility of their academic involvement at the student level). However a "no" option remained available for those situations in which it could be ascertained with a high degree of certainty that the individual was not or no longer a student (e.g., from a detailed LinkedIn profile or online curriculum vita).

Coding

Initial coding began in May 2013 and was completed on December 15, 2013. Coding was done by two coders for whom a high interrater reliability was ascertained. The twitter handles were used as the initial point of departure for the search. Coders determined what they could from the information provided in the short biographical information on twitter. If a url was provided on twitter, this was followed. Google searches were also employed, using as a seed the person's first name and/or twitter handle and limiting searches to the first three pages of results. Where there was a dispute between sources, the more contemporary source was used.

The first coding variable asked the coder to distinguish whether the account was held by an individual or an organization. Although most accounts are technically managed by a single person, a distinction was made between people who represented themselves and people who represented a company or organization. If a person simply affiliated with an organization, they were still coded as an individual.

Research centers at universities were coded as university. Research centers outside of a university setting were coded as non-profits. Although universities could be considered "government" or "non-profit" (and in some rare cases a corporation), all academic institutions were coded as universities.

Results

Approximately three-quarters of the sampled accounts could be identified as belonging to individuals (n=1520), while slightly under 23% belonged to organizations (n=459) (Figure 1). Of the accounts belonging to people, the majority were associated with a male tweeter (Figure 2). Nearly 12% of the individuals were identified as students (either undergraduate, master's, or doctoral). Of the students, 67.2% were doctoral students or candidates. It should be noted that, for some codes, a failure to mark a quality as "present" does not necessarily indicate that the reverse is true. For example, it is likely not the case that 88.2% of the individuals are *not* students; rather, all that we can say is that we were able to identify 11.8% of the individuals as students.

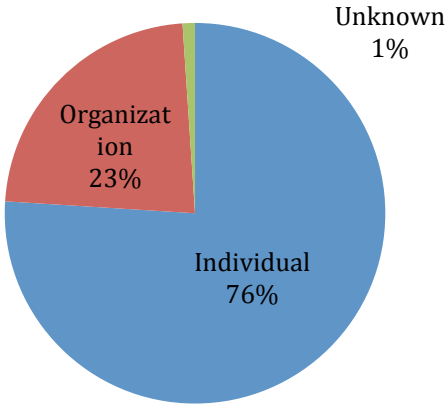


Figure 1. Twitter accounts by type.

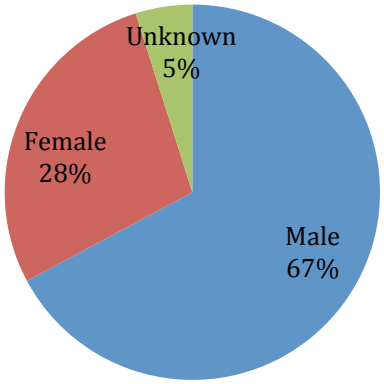


Figure 2. Individual accounts by gender.

In terms of the entire population of individuals, 34.4% were identified as possessing a Ph.D (this discounts the students who were working towards a Ph.D), suggesting that the population of Twitter users who tweet links to academic articles does not reflect the demographics of the general public. STEM fields were dominant both within the group of users identified as students and within the group of users identified as working in academe.

In terms of the students, 52.4% were affiliated with general science, 15.1% were associated with health/medical study, and 10.8% were associated with technology/engineering. In terms of the academics, 62% were associated with general science, 10.4% were affiliated with health/medical study, 8.1% were associated with the social sciences, and 7.5% were affiliated with technology/engineering (Figure 3).

Of the organizations, 41.6% were identified as non-profits, 29.2% were identified as corporations, and 13.1% were identified as universities. 18.9% were classified as news/media/outreach institutions (note that this was considered a non-exclusive category independent of the earlier classifications).

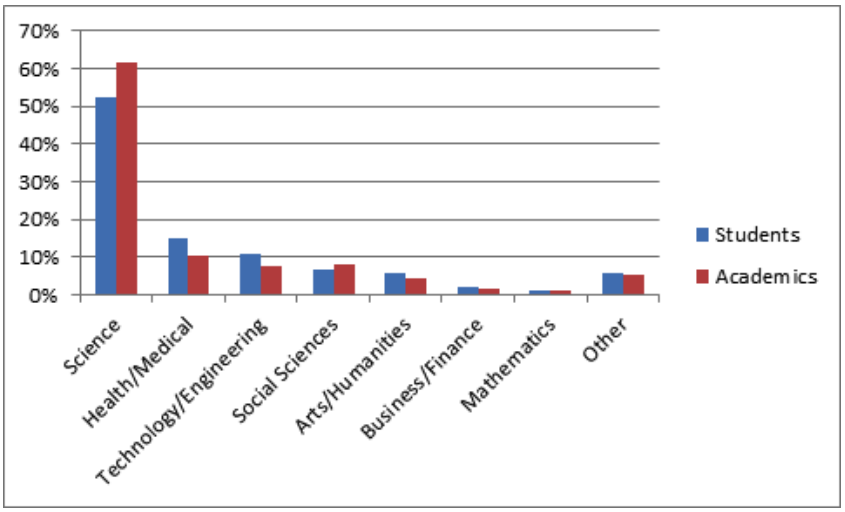


Figure 3. Proportion of twitter accounts by disciplinary domain.

Discussion and Conclusion

The demographics of the individual tweeters did not reflect the general population of Twitter users. Whereas women are overall slightly more likely to take advantage of social networking sites than men are (Kimbrough, Guadagno, Muscanell, & Dill, 2013; Pew Research, 2014), men use Twitter slightly more (24% of male Internet users, compared to 21% of female Internet users). Our study was much more male-baised, with nearly 70% of individual accounts maintained by men. This percentage is in keeping with male to female ratios found in the scientific workforce and scholarly publishing (Larivière et al., 2013).

A growing body of literature seeks to validate social media metrics, or "altmetrics" as valid forms of the social (i.e., public) impact of scholarly research. However, this research indicates that a large portion (i.e., nearly half) of those who tweet about science already have a doctoral degree or are in pursuit of one. This proportion far exceeds the 1% of the US population, for instance, holding a doctoral degree (Petersons, 2014). This suggests caution when utilizing social media metrics as an indication of the value of the work for the public. Rather, this emphasizes the strong use of these tools for dissemination and discussion of scholarship *among scholars*. Acknowledgement of the scholarly context of social media metrics must be taken into account in evaluative uses of these metrics.

Limitations

The study only considered journals that were frequently tweeted. It is possibly that the demographics of users who tweet articles from less popular journals might differ from those of tweeters who share links to the highest echelon of scientific journals. In addition, the information that could be gathered about the tweeters was limited to what was readily available online. Accordingly, the percentages generated by the study represent conservative estimates rather than absolute figures.

Future research might consider a wider variety of journals, as well as employing other methods to ascertain tweeter demographics (e.g., studying the users' tweets in an attempt to ascertain gender, academic affiliation, etc. for those users for whom such information was not publicly available). In addition, it is theoretically possible to directly survey the tweeters who shared links to academic articles, although such an approach would likely rely on publicly available contact information (primarily e-mail addresses), and would most likely face the same issues that were encountered in this study.

Acknowledgments

Funding for this project was provided by the Alfred P. Sloan Foundation ("Understanding the use and meaning of social media metrics in scholarly communication") and the National Science Foundation (grant SMA-1208804) as part of the Digging into Data initiative. We would like to thank Mike Thelwall for providing data for this project.

References

Alexa (2015). Alexa top 500 global sites. Retrieved from http://www.alexa.com/topsites.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2014). Tweets as impact indicators: Examining the implications of automated bot accounts on Twitter. arXiv preprint arXiv:1410.4139.

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. Journal of the Association for Information Science and Technology, *65*(4), 656-669.

Kimbrough, A. M., Guadagno, R. E., Muscanell, N. L., & Dill, J. (2013). Gender differences in mediated communication: Women connect more than do men. Computers in Human Behavior, *29*(3), 896 900. doi:10.1016/j.chb.2012.12.005

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C.R. (2013). Global gender disparities in science. *Nature, 504*(7479), 211-213.

Petersons (2014). Ph.D. programs are rigorous educational experiences. Retrieved from http://www.petersons.com/graduate-schools/phd-programs-rigorous-educational.aspx

Pew Research Center (2014). Social media update 2014. Retrieved from http://www.pewinternet.org/files/2015/01/PI_SocialMediaUpdate20144.pdf

Pew Research Center (2014). Social networking fact sheet. Retrieved from http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. PloS one, *8*(5), e64841.

Thelwall, M., Tsou, A., Weingart, S., Holmberg, K., & Haustein, S. (2013). Tweeting links to academic articles. Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics, (17), 1-8.

Classifying altmetrics by level of impact

Kim Holmberg

kim.j.holmberg@utu.fi
Research Unit for the Sociology of Education, University of Turku, 20014 Turku (Finland)

Introduction

In the light of current knowledge we can conclude that altmetrics do not present an alternative for traditional citation-based analysis of research impact (e.g., Haustein *et al*., 2014). Altmetrics have instead the potential to show some other aspects of research activities and provide a more nuanced view of the impact research has made on various audiences (Liu & Adie, 2013; Piwowar, 2013). Altmetrics come in many forms and from many different sources, all of which can represent different aspects of the online activity or of the different levels of impact that various research products have made on different audiences. What exactly the different altmetrics represent we do not yet know, but the greatest advantage of altmetrics may be exactly in this diversity.

Aggregating all altmetrics to a single indicator would remove this advantage. With aggregation of different altmetrics we are just creating another impact factor, another indicator that in the worst case is used for something that it is neither designed for nor capable of indicating. However, because of the wide variety of different sources for altmetrics, some form of aggregation or classification is needed and different types of classifications are already used by some service providers. Here we present another approach, one based on the level of impact. With this we hope to stimulate further discussion about the actual meaning of altmetrics.

Diversity of altmetrics

The diversity of altmetrics has two interesting dimensions; the diversity of people creating the altmetrics, and the diversity of the impact they indicate. In any research assessment what we want to measure is value or quality of research. Quality is of course very subjective and difficult to quantify. Because we cannot evaluate quality directly, particularly not at large scale, we use volume of impact as a proxy for value (i.e. number of citations or more recently number of online mentions).

The different data sources and different data types collected from the mentions of research products in various social media sites can represent a wide spectrum of different levels of impact. For instance, while a tweet does not necessarily hold any indication of impact other than awareness, a blog entry or a Wikipedia citation reflect some level of influence or impact. The people creating the altmetrics then again range from researchers and practitioners to the public.

Aggregating altmetrics

In social media analytics the mentions of brands and products in various social media are often placed and grouped together on a spectrum according to level of engagement, ranging from visibility to influence and finally reaching engagement as the most desired level of reaction. In the context of altmetrics, Piwowar and Priem (2013) write about the different "flavours" of impact that altmetrics could potentially reflect, referring to the diversity of altmetrics and possibility to group similar metrics into these "flavours". This is in line with the ideas presented at PLoS too, with different sources and different timings of altmetrics reflecting engagement from different audiences and possibly also that of different purposes for the engagement (Lin & Fenner, 2013).

This approach has already been taken by some of the altmetrics service providers as they group the data collected from various sources into what reflects different types of activities. PLoS for instance groups the metrics they use into views, saves, mentions, and citations. These do roughly translate to what we can assume to be different levels of impact, reflecting the variety of actions and interactions that one can have with the research products. Saving a research product suggests that the research product have made a bigger impact than just viewing it suggests, mentioning it suggests additionally increased level of impact, and citing it suggests what could perhaps be considered as the ultimate level of impact, at least when the goal is to investigate scientific impact.

Aggregation by the level of impact

Indicators of impact come in many diverse forms on the web and in social media and the different social media sites and the different activities within them can provide various metrics of different levels of impact. A potential approach to aggregating altmetrics would be to use these different levels of impact as they are and to not try to combine them according to source or type of activity they represent.

When the metrics indicate low impact we cannot really be sure whether the research has made any impact at all as evidence of it is usually not clear; a page view, clicking on a tweet button next to the article, or sharing a research article on Facebook, all indicate that the user has seen what they are sharing but nothing indicates that it has made any

impact on them, that they would have been influenced by it, or that they would have changed their behaviour because of it. Metrics indicating a medium level of impact would already come attached with at least some information that the research has made an impact, that it has in some way influenced the user. Whether the research product has been mentioned somewhere online or been bookmarked with the intent to use it later, the metrics generated from the activities at this level suggest that the users have been influenced some way, that the research has made at least some impact. Metrics indicating a high level of impact usually come attached with some additional, perhaps more qualitative data that we can use to investigate how the research has influenced the user and confirm what kind of impact it has made. A rough classification of different types of altmetrics that indicate different levels of impact could follow the one presented in Table 1. Besides impact, we can also measure reach with altmetrics; how many people have become aware of the research and how many of them have been influenced by it in some way.

Table 1. Levels of impact.

	Altmetrics		
Level of impact	Low	Medium	High
Reach	High	Medium	Low
Example activities	Awareness, visibility	Influence, interaction	Usage
Example metrics	Tweets, 'likes', shares, …	Mentions, downloads, bookmarks, …	Blog posts, …

More research is needed and both quantitative and qualitative methods are needed to confirm what level of impact different types of actions in different social media reflect and how they relate to each other.

Benefits of the proposed approach

Focusing future research on the level of impact has a couple of benefits compared to other approaches. First of all, impact is what we want to measure, hence grouping different metrics based on the level of impact they reflect makes sense. Second, using all the unique metrics (e.g., tweets, retweets, blog mentions, link in blogroll, Facebook shares, "likes", and mentions) would create a massive number of different metrics that would be difficult to a) keep track of, b) present, and c) control. Third, aggregating the different metrics by type of activity they represent may not give an accurate picture of the impact they represent, as similar types of activities on for instance different social media sites may be reflecting different levels of impact and/or different types of users. And fourth, aggregating all the metrics into a single indicator would just be creating another impact factor, but this time from a much wider diversity of different metrics indicating different aspects and which probably should not be aggregated at all because of that. And finally, focusing on the different indicators for different levels of impact instead of some specific sites would not be such a vulnerable approach relying on the continued existence and goodwill of the social media sites to allow access to their data.

Conclusions

We propose the classification of altmetrics based on the level of impact reflected by the specific altmetrics. This approach would have some clear benefits compared to aggregations based on activity or source of altmetrics. More research is, however, needed to establish the different levels. The key challenges for future altmetric research are a) identifying the groups of people that create different altmetrics, and b) mapping the different levels of impact the different metrics reflect. This line of research would bring us again one step closer to fully understand what altmetrics indicate, and with that, the meaning of altmetrics. It is nevertheless important to recognize that the true meaning of any altmetrics lies in the stories behind the numbers. Hence it is important that any altmetrics are presented together with the accompanied stories to give the full context in which they have been generated.

References

Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M., & Larivière, V. (2014). Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the American Society for Information Science and Technology*, *65*(4), 656-669.

Lin, J. & Fenner, M. (2013). Altmetrics in evolution: defining and redefining the ontology of article-level metrics. *Information Standards Quarterly*, *25*(2), 20-26. http://dx.doi.org/10.3789/isqv25no2.2013.04.

Liu, J. & Adie, E. (2013). New perspectives on article-level metrics: developing ways to assess research uptake and impact online. *Insights: the UKSG Journal*, *26*(2), 153-158.

Piwowar, H. (2013). Altmetrics: what, why and where? *Bulletin of the Association for Information Science and Technology*, 39(4).

Piwowar, H. & Priem, J. (2013). The power of altmetrics on a CV. *Bulletin of the Association for Information Science and Technology*, 39(4).

Characterizing In-text Citations using N-gram Distributions

Marc Bertin[1] and Iana Atanassova[2]

[1] bertin.marc@gmail.com
Centre Interuniversitaire de Rercherche sur la Science et la Technologie (CIRST),
Université du Québec à Montréal (UQAM), Canada

[2] iana.atanassova@univ-fcomte.fr
Centre Tesniere, University of Franche-Comte, France

Introduction

This article focuses on a Natural Language Processing (NLP) approach for the analysis of citation functions in scientific papers. Bibliometric studies traditionally rely on citation metadata and count the number of times a publication has been cited. However, some recent studies rely also on full text processing on papers, e.g. (Boyack et al., 2013), (Bertin et al., 2013, 2014). The full text content of papers and more specifically the sentences containing citations provide valuable information on the functions of citations that can be exploited through NLP. To study citation acts, we need to consider full text papers and their rhetorical structure.

The main question that we want to answer here is whether the most frequent citation patterns are correlated to the rhetorical structure of scientific papers. We investigate the properties of the linguistic patterns that appear in citation contexts. For this, we study the distribution of n-gram classes containing verb forms, and we show the existence of three different types of distributions according to the rhetorical structure.

Method

By analyzing a large corpus of articles, we propose a quantitative study of the linguistic patterns around in-text citations. Some words or sets of words in n-grams are more frequent than others (Cavnar & Trenkle, 1994), and this idea is consistent with Zipf's Law (Zipf, 1949). The difficulty is that the calculation of n-grams in contexts results in a combinatorial explosion. We propose several filters to reduce the number of patterns.

The rhetorical structure of scientific papers is typically organized around a standardized pattern, known as the IMRaD structure (Introduction, Methods, Results and Discussion). We identify the four main section types of this structure by analysing section titles. Then, we consider the set of sentences containing citations and belonging to each section type.

We represent citation contexts by using sequences of words of length n called n-grams where $2<n<=5$. In our approach we consider only n-grams within sentence boundaries because sentences are natural building blocks of the text. For each n-gram we observe its frequencies in the four section types of the IMRaD structure. For our study, we select only the n-grams that contain at least one verb form. In this way, the number of n-grams to process is much smaller and we eliminate word patterns containing only nominal groups like: “In this paper”, “the present article”, “the result of” etc. for 3-grams.

Dataset

We performed an automatic analysis of the seven peer-reviewed academic journals published in Open Access by the Public Library of Science (PLOS). The corpus contains about 85,660 research articles. Most of the articles are in the biomedical domain, but the corpus covers all fields of Human and Natural Sciences, as the publisher’s main journal, PLOS ONE, is multidisciplinary. Around 98% of the articles in the corpus follow the IMRaD structure, which is imposed by editorial requirements.

Results

We select the most frequent verb forms in order to construct n-gram classes from in-text citation contexts. This data will be used to obtain a first typology of the distribution of n-grams depending on the rhetorical structure of articles.

The following figures present distributions of n-grams classes for the IMRaD sections. We can distinguish between three different type of classes, and we give one example of each. The horizontal axis presents the text progression of the section from 0% to 100%. The vertical axis gives the percentage of occurrences of each class relative to its occurrences in citation contexts in the entire article.

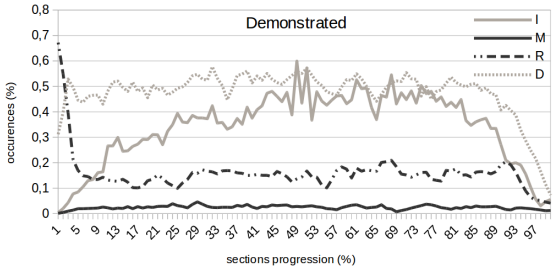


Figure 1. Demonstrated.

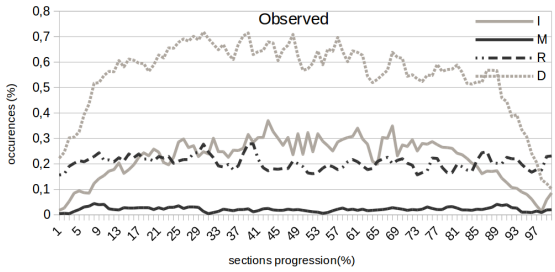


Figure 2. Observed.

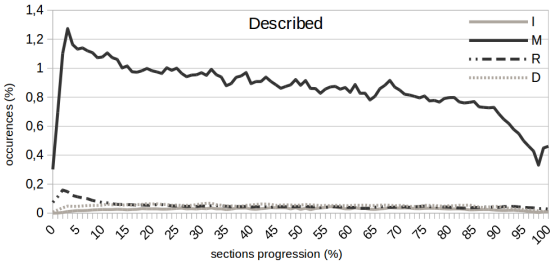


Figure 3. Described.

Discussion

Figure 1 shows the first class, which includes n-grams containing the verb *Demonstrated***.** These n-grams appear with roughly equivalent frequencies in the sections Results and Discussion, but, at the same time the Methods section contains much lower frequencies of these patterns.

Figure 2 shows the second class type, which includes n-grams with the verb *Observed***.** We can observe another type of distribution, with relatively very high frequencies in the Discussion section.

Figure 3 shows the distribution of n-grams with the verb *Described*. We can observe that the structure of the Methods section is unique, as the class *Described* is present with a very high frequency in this section and especially at the beginning of the section. Moreover, Figures 1 and 2 show that on the distributions for the other classes, the Methods section contains relatively few occurrences. In other words, the class *Described* is characteristic of the Methods section, where it appears with very high frequency, and it is very rare in all the other sections. The Methods section displays very low frequencies for all classes except *Described*.

These results imply that each section, depending on its nature, authorizes more or less easily the usage of specific patterns containing verbs. The Methods section is rather closed in nature, where we find a very small number of high frequency verbs. At the same time, the Discussion section is open to different forms and allows a larger number of variations in terms of the linguistic means that authors use in citation contexts.

Conclusion

The purpose of this study is to demonstrate the existence of frequent n-gram patterns in citation contexts and their strong relation with the rhetorical structure of scientific articles. Studying the n-gram classes containing verb forms, we show the existence of three different types of distributions according to the rhetorical structure. From our point of view, the problem of the automatic annotation of citation contexts is strongly related to identifying significant surface patterns for the annotation process.

Acknowledgments

We thank Benoit Macaluso of the Observatoire des Sciences et des Technologies (OST), Montreal, Canada, for harvesting and providing the PLOS data set.

References

Bertin, M., & Atanassova, I. (2014). A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. *Proceedings of the First Workshop on Bibliometric-enhanced Information Retrieval co-located with 36th European Conference on Information Retrieval* (ECIR 2014). Amsterdam, The Netherlands.

Bertin, M., Atanassova, I., Larivière, V., & Gingras, Y. (2013). The Distribution of References in Scientific Papers: an Analysis the IMRaD Structure. *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics* (pp. 591–603). Vienna.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology, 64*(9),1759–1767.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. Ann Arbor MI, 48113(2), 161–175.

Small, H. (1982). Citation context analysis. *Progress in Communication Sciences, 3*, 287–310.

Zipf, G. K. (1949). Human behavior and the principle of least effort.

[illegible]

[illegible]

[1]breeze7zhou@163.com

[2] zhangcz@njust.edu.cn.

[illegible]

Introduction

[illegible]

Methodology

Framework

[illegible]

Data

[illegible]

[illegible]

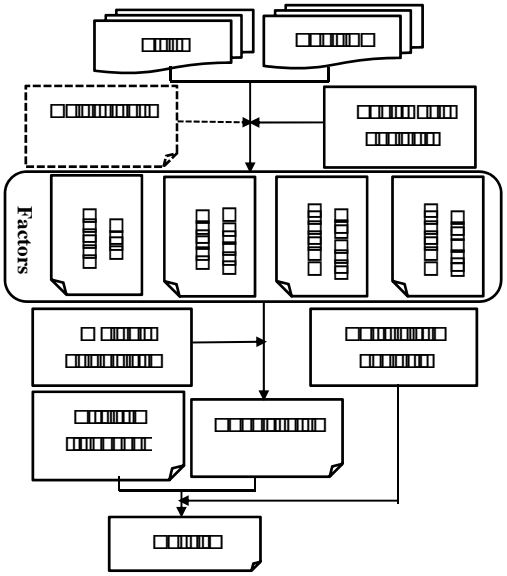


Figure 1. Frameworks ofcorrelation analysis.

Factor calculations

Calculations of numbers of positive reviews and negative reviews

[illegible]

Calculations of aspect values and star values

[illegible] sp_{ij} [illegible] s_i [illegible] r_j[illegible]

* *Corresponding author: Chengzhi Zhang, Tel: +86-25-84315963.*

[illegible]

Table 1.Calculations of book scores.

aspect values	$VAB_{it} = \sum_{j=1}^{N} sp_{ij} \Big/ \sum_{j=1}^{N} \lvert sp_{ij} \rvert$ $i = 1,2,\ldots,10, t = 1,2,\ldots M$
	$VAB'_{it} = \sum_{j=1}^{N} (sp_{ij} * h_j) \Big/ \sum_{j=1}^{N} \lvert sp_{ij} \rvert$
star values	$VSB_{jt} = \sum_{j=1}^{N} star_j / \mathrm{N}$
	$VSB'_{jt} = \sum_{j=1}^{N} (star_j * h_j) / \mathrm{N}$

[illegible] VAB_{it} [illegible] s_i [illegible] b_t [illegible] $VAB_{it}{}'$ [illegible] N [illegible] s_i [illegible] b_t [illegible] i [illegible] M [illegible] h_j [illegible] r_j [illegible] VSB_{jt} [illegible] r_j [illegible] b_t [illegible] VSB'_{jt} [illegible] $star_j$ [illegible] r_j [illegible] N [illegible] b_t [illegible]

Calculations of book scores

[illegible]

Table 2.Calculations of book scores.

Steps	Formulas
[illegible]	$p_{ij} = \frac{v_{ij}}{\sum_{i=1}^{N} v_{ij}}$ $i = 1,2,\ldots,N, j = 1,2,\ldots,m$
[illegible]	$e_j = -\frac{1}{\ln(n)} \sum_{i=1}^{N} p_{ij} \ln(p_{ij})$
[illegible]	$w_j = 1 - e_j / m - \sum_{j=1}^{m} e_j$
[illegible]	$SB_i = \sum_{j=1}^{m} p_{ij} * w_j$

[illegible] p_{ij} [illegible] b_i [illegible] f_j [illegible] v_{ij} [illegible] b_i [illegible] f_j [illegible] N [illegible] m [illegible] e_j [illegible] f_j [illegible] w_j [illegible] f_j [illegible] SB_i [illegible] b_i [illegible]

Experimental result analysis

[illegible]

Table 3. Results of correlation analysis.

Domains	Without H.	With H.
[illegible]	0.383*	0.378*
[illegible]	0.401**	0.417**
[illegible]	[illegible]	0.240*

Conclusions

[illegible] **WH (with helpfulness) conclusion** [illegible] **OC (overall correlation) conclusion** [illegible]

Acknowledgments

[illegible]

References

[illegible] *A holistic lexicon-based approach to opinion mining*. [illegible]

[illegible] *Intelligent Systems and their Applications, IEEE, 13* [illegible]

[illegible] *Power System Technology, 33* [illegible]

[illegible] *Introduction to modern information retrieval*.

[illegible] *A report on the academic influence of Chinese humanity and social science books: China Social Science Press* [illegible]

[illegible] *Mis Quarterly, 38* [illegible]

[redacted]

[redacted]

[1] natalie.friedrich@hhu.de, stock@phil.hhu.de
[redacted]

[2] stefanie.haustein@umontreal.ca, timothy.bowman@umontreal.ca
[redacted]

Introduction

[redacted]

Dataset and Methods

Tweets and research articles

[redacted]

Sentiment tools

[redacted] s_1 [redacted] s_2 [redacted]

[redacted]

Intellectual coding of sentiments

[redacted] s_i [redacted] s_0 [redacted]

Cleaning tweets

[redacted] t_0 [redacted]

Preliminary Results

Table 1. Intellectual (s_0) and automated (s_1, s'_1, s_2) sentiment detection for 1,000 tweets.

		Sentiments (%)			Agreement w/ s_0	
		+	–	*n*	%	*K*
	s_i				*n/a*	
	s_0				*n/a*	
t_0	s_1					
	s_2					
t_a	s_1					
	s'_1					
	s_2					

Discussion and Future Work

References

Journal of the Association for Information Science and Technology

PLoS ONE 8(5)

Cybermetrics: International Journal of Scientometrics, Informetrics and Bibliometrics

Journal oft he American Society for Information Science und Technology, 61

[redacted]

[redacted]

[3] malekiashraf@ut.ac.ir

[redacted]

Introduction

[redacted]

Method

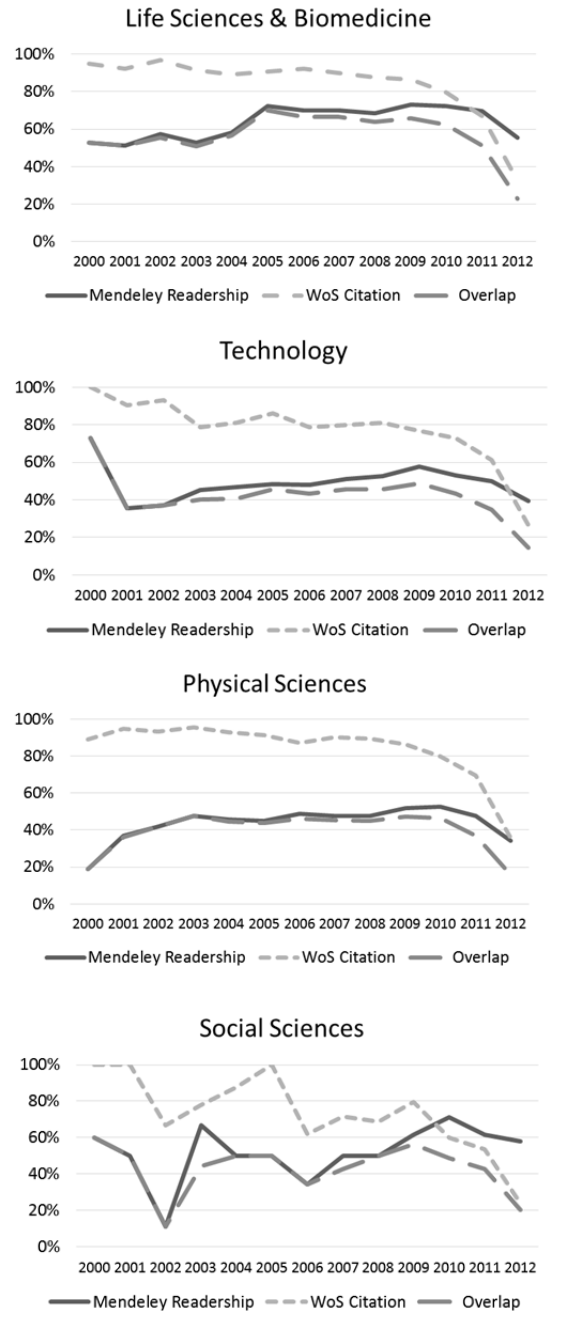
[redacted] *ImpactStory.org* [redacted]

Results

[redacted]

Career stages and nationalities of Mendeley users

[redacted]



Figures 1-4. Trend of relative proportion of publication uptake via formal WoS citations, Mendeley readerships and both of them (overlap) across four broader research areas- Y-axis shows percent of publications in each year.

Discussions and Conclusions

Acknowledgments

References

Proceedings of the IATUL Conferences

Scientometrics 91

Journal of the Association for Information Science and Technology 65

Journal of the Association for Information Science and Technology

Journal of the Association for Information Science and Technology

[illegible]

[illegible]

[1] ronaldfa@gmail.com
[illegible]

[2] tiago.murakami@dt.sibi.usp.br, [3] jan.lara@sibi.usp.br [4] sifausto@usp.br
[illegible]

Introduction

[illegible]

Methods

[illegible]

Results

Table 1. Mentions per year.

Year	Articles	Mentions	(%)
1999	6	22	1,89
2000	23	30	2,58
2001	26	29	2,49
2002	29	30	2,58
2003	27	23	1,98
2004	29	109	9,36
2005	24	31	2,66
2006	27	56	4,81
2007	26	85	7,30
2008	31	77	6,62
2009	34	68	5,84
2010	34	96	8,25
2011	39	112	9,62
2012	43	119	10,22
2013	32	79	6,79
2014	11	198	17,01
Total	**441**	**1164**	**100**

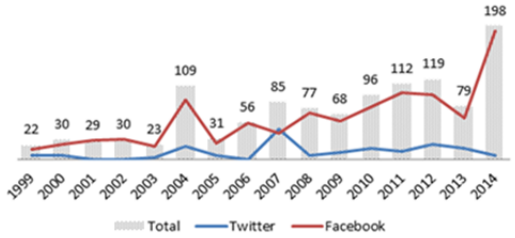


Figure 1. Mentions by Social Media.

Discussion

Conclusions

References

Open Access and Scholarly Communications Indicators in Latin America

[illegible]

[illegible]

tdbowman@indiana.edu

[illegible]

Introduction

[illegible]

Dataset and Methods

Survey of Professors

[illegible]

Tweet Collection

[illegible]

Research Article Collection

[illegible]

Results

Comparison of Survey Results

[illegible]

Twitter Use Type

Tweet Analysis

Tweet and Publication Activity Comparison

Discussion and Future Work

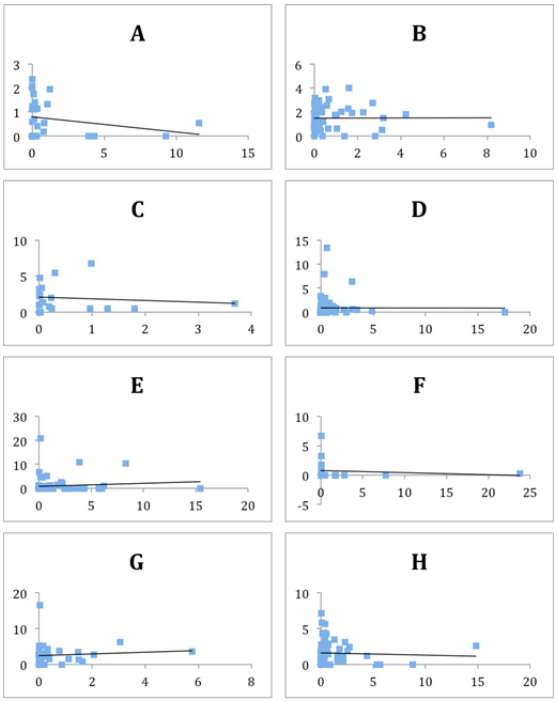


Figure 1. Average citation impact [y-axis] and average mean tweets-per-day [x-axis] for 395 professors in Anthropology [A], Biology [B], Chemistry [C], Computer Science [D], English [E], Philosophy [F], Physics [G], & Sociology [H].

References

Perceiving, Acting, and Knowing: Toward an Ecological Psychology

Journal of Computer Assisted Learning,

[1] jpqiu@whu.edu.cn

[2] yuhouq@yeah.net

Introduction

Exploring the meaning of impact

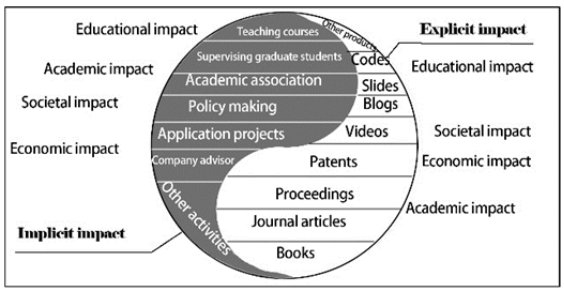


Figure 1. The composition of scholars' impact.

Impact Generation Model

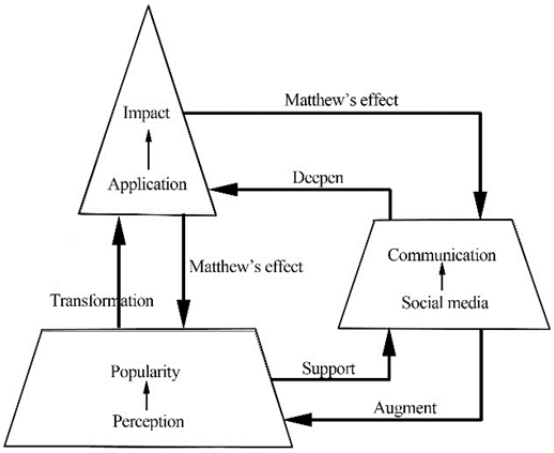


Figure 2. Impact generation model.

Stratifying altmetrics indicators

An economic analysis of level of engagement phenomenon

Stratification of altmetric indicators

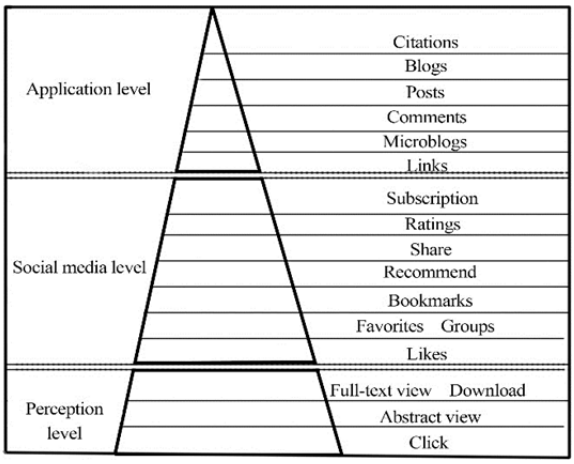


Figure 3. Stratification of altmetrics indicators in the pyramid form.

Merits of the stratification

Acknowledgments

References

Open Repositories.

ASIST, 39

Nature, 493

Nature

ASIST, 47

ISSI 2015

CITATION AND CO-CITATION ANALYSIS

Citation Type Analysis for Social Science Literature in Taiwan

Ming-yueh Tsay

mytsay@nccu.edu.tw
Graduate Institute of Library, Information and Archival Studies, National ChengChi University (Taiwan)

Abstract
Through citation analysis, this study explored the distribution of document type, language and publication year for citations in social science journals. Samples were research articles published in 2010 from first-rank journals, as assessed by the Department of Humanities and Social Sciences, National Science Council and indexed in Taiwan Social Sciences Citation Index (TSSCI). The section in which citations appeared, namely introductions, methodologies, results, and conclusions, were also examined. Conclusions and suggestions are made based on the research results and interdisciplinary comparisons. For social science studies in Taiwan, the major findings are as follows: 1. Journals and books were the most cited materials, and English was the language of most citations. 2. Social scientists in Taiwan tended to cite materials published within 10 years with a citing half-life of approximately 11 years. 3. The ratio of articles following the IMRAD format was high in Taiwan social science journals. 4. Citations in these social science journals occurred most frequently in the introduction section, while they occurred least frequently in the conclusions. 5. Social scientists mostly cite to set the stage for their current studies. 6. The citation type is highly related to the citation location.

Conference Topic
Citation and co-citation analysis

Introduction
Since the social sciences are associated with human society, its patterns, where it goes and how it works, it can enrich the values and contents of our lives. In contrast, the "hard sciences" have been the focus of attention with the rapid growth of technology grew, and the social sciences have received less attention. This has led to a lack of balance between academic and technical research in many developing countries. To gain attention and support from governments and the public, social scientists need to promote their research outcomes and impacts much more effectively via the presentation and communication of their scholarly articles.

A research article may consist of body text and references; the former is the citing article, and the latter are cited articles. Relations between the citing and the citied may explain the interaction, development and communication among disciplines, and can reveal current research interests and future trends. Citations have multiple roles and unique functions in scholarly communication; for example, a cited article may present broader research contents, explain methods applied in a research or provide information and discussion that support a specific perspective.

The importance of journal articles for scholarly communication and academic assessment motivates the present study on Taiwanese social science journal articles to explore and compare their characteristics and types of citations via methods of bibliometric and citation analysis. The research outcomes may improve the knowledge of citation, and serve as reference for future empirical researches for the social science studies in Taiwan.

Other Citation Studies
Citations have been studied using context or content analysis, whereby the analysis determines the citation type based on the surrounding text. Frost (1979) mentioned the complexity of citation function and that the classification of citation function and proper schemes for classification received little attention in citation studies. To explore the nature of

citation use, some various schemes of classification for different disciplines have been developed to explain the functions of citations and the relations between body text and citations.

In Moravcsik and Murugesan's (1975) study physics citations fall into the "applied/used" category with 60% and 40% of the citations being general acknowledgements. In the study of Voos and Dagaev (1976), inspected the locations of each citation in sample articles and found that articles of biology and medicine were mostly cited within two to three years after their publication, and were cited the most in the introduction section, and next in the discussion section.

Peritz (1983) selected a variety of social science journals in which the basic methodologies of empirical social research were used and analyzed into the categories of a citation classification scheme. That study revealed that generally, "setting the stage for the present study" citations rank first. To carry out the reliability citation classification scheme, Peritz further investigated the association between classification and location and found that the marginal frequencies of the location introduction, methods and discussion were fairly close to the frequencies of the classification categories of setting the stage, methodology, and comparison and argument, respectively.

More recently, Harwood (2008) interviewed six informants who were computer scientists and six who were sociologists on the functions of citations in their writing. His findings reveal that position, supporting, and credit are relatively frequent across both disciplines, although the engaging function is far more frequent in the sociology texts.

Case and Miller (2011) investigated the citation practice of a group of citing authors with an interest in bibliometric or scientometric research, finding that the most popular reason was "this reference is a 'concept marker'," which distantly followed by "reviews prior work in the area" and other reasons.

The above literature survey shows there have been many studies investigating citation category and citation practices, which are likely to vary from discipline to discipline. This motivates the present study to further explore the citation type of articles cited in the social science journals published in Taiwan.

Research Method and Limitation

The journals selected in this research were six first-ranked journals indexed in the Taiwan Social Sciences Citation Index (TSSCI) in the disciplines of sociology, education, psychology, political science, economics and management. In this study, it is assumed that the first ranked journal of each discipline may represent the research characteristics of that discipline.

Articles published in 2010 and following the IMRAD format were selected as research samples, though articles published earlier than 2010 were also collected if there were insufficient samples. The titles of journals and number of articles selected for the six disciplines were: sociology, *Taiwanese Journal of Sociology*, 15 articles (2008-2010); education, *Bulletin of Educational Psychology*, 31 articles (2010); psychology, *Chinese Journal of Psychology*, 16 articles (2010); political science, *Taiwan Political Science Review*, 16 articles (2008-2010); economics, *Academia Economic Papers*, 13 articles (2010); management, *Journal of Management*, 25 articles (2010).

In the present study, if introductions and literature reviews were in two different sections, they were considered as an introduction in combination; if results and discussion were in one section, they would be categorized as result. Citations were categorized, on the basis of the classification scheme proposed by Peritz (1983), which requires little subjective judgment and is easy to carry out even without in-depth knowledge of the subject field.

Full texts and references of all 116 research articles were downloaded from online databases

or photocopied from printed journal and processed with Excel (Microsoft, U.S.) into bibliographical files. Employing bibliometric techniques and citation analysis, this study explored article type of journal, language of citation, citation years, document types of citations, citation types and locations of citations, the relations between citation type and location, and comparison among six disciplines of social sciences in Taiwan.

This study conducted purposive sampling to acquire journal articles for citation analysis, whose results might thus be limited indeed and less representative for each or the whole of humanities disciplines. Nevertheless, the current study aims to distinguish the meaningful characteristics of article structures, citation locations, and citation types of the six social science disciplines of Taiwan; also the method of "citation content analysis" used in this study to explore the nature of citation types is qualitative and justified by the attempt to interpret the existing phenomena. In the above senses, purposive sampling and unequal sample size seemed to be acceptable limitations.

Results

In this study, citation characteristics and locations in body texts are discussed according to article type of journal, language of citation, year of the highest citation and citation half-life, document type of citation, citation location and citation type.

Article Type of Journal

Papers published in social science journals in Taiwan are mainly divided into research articles and review articles. In general, research articles comply with the IMRAD format. The ratio of articles following the IMRAD format was high in the social sciences. Table 1 demonstrates review and research articles, both appeared in the disciplines of political science and sociology, while journals in the fields of psychology, education, economics and management preferred research articles.

Table 1 shows that, among the six disciplines, education, economics and management composed completely (100%) of research articles that follow the format of IMRAD.

Table 1. Article types in social science journals of Taiwan.

Discipline (Journal name)	*Papers*	*English article*	*Chinese review article*	*Chinese research article*	*% of Chinese Research article*
Political Science (*Taiwan Political Science Review*)*	30	1	13	16	55.2%
Sociology (*Taiwanese Journal of Sociology*)*	25	0	10	15	60.0%
Education (*Bulletin of Educational Psychology*)	32	1	0	31	100.0%
Psychology (*Chinese Journal of Psychology*)	23	4	3	16	84.2%
Economics (*Academia Economic Papers*)	18	5	0	13	100.0%
Management (*Journal of Management*)	30	5	0	25	100.0%
Total	158	16	26	116	81.7%

*Semi-annual journal. Sample articles of these journals were dated back to 2008 from 2010; samples of other journals were articles published in 2010.

Language of Citation

Materials in Chinese and English were the major source of references cited in social science articles, with the former accounting for 21.5% and the latter 78% of the total references collected. Most of the references in economics (93.5%) and management (92.1%) were English papers, while Chinese articles were infrequently used in both disciplines. Domestic research articles and reference materials, however, were used quite often by scholars of sociology (30.7%) and political science (43.4%).

Year of the Highest Citation and Citing Half-Life

Table 2 reveals the year of highest citation, citation age and citing half-life of articles in sample journals. Citing half-life refers to the time span from the current year to the year whose accumulated number of citations accounts for 50% of total citations in the journal. For example, the citing half-life of *Chinese Journal of Psychology* shown in Table 2 was 11, indicating that half of its citations were younger than 11 years as the citing articles being published. The time span of citation half-life reflects the currency of cited materials: the longer the citing half-life, the older the cited materials, and vice versa.

Table 2. Distribution of year of the highest citation, citation age of the highest citation and citing half-life.

Discipline (Journal name)	*Year of the highest citation*	*Citation Age*	*Citing half-life*
Political Science (*Taiwan Political Science Review*)*	2007	4	10.6
Sociology (*Taiwanese Journal of Sociology*)*	2006	5	11.5
Education (*Bulletin of Educational Psychology*)	2005	6	11.2
Psychology (*Chinese Journal of Psychology*)	2006	5	11.0
Economics (*Academia Economic Papers*)	2007	4	11.4
Management (*Journal of Management*)	2004	7	11.2
Average	2006	5.2	11.2

Based on the year of highest citations, the number of citations earlier than 2004 is decreasing for earlier articles. In other words, the older the articles were, the fewer citations they received. In general, for articles published in 2010 the peak of citations fell between 2004 and 2007 that suggests citations that received from the sample journals reached a peak after four to seven years of its publication, five years in most cases. A large number of citations came from articles published in the recent several years, indicating that social scientists have a tendency to cite the most recent articles. In the social science fields, scholars tended to cite materials with a citing half-life of approximately 11 years. For social scientists in most disciplines, 50% of their research needs could be satisfied by articles published after 2000, and the tendency to cite the most recent articles indicates the social science research depends on more current literature.

Document Type of Citation

In the six top journals selected as samples in this study, there were 116 Chinese articles following the IMRAD format, citing 6,063 references to the bibliographic files built by this study. According to the bibliographic data collected, journals and books were the most

frequently cited, accounting for 88% (journals 65% and books 23%) of all types of cited materials. The uses of journals and of books in economics were quite different, with the highest interval over 76%, in which journals accounted for 82% of cited materials while books accounted for 6%. The second-highest difference between citations of journals and of books was in management, where journals accounted for 80% of the citations, which was 67% higher than books. For other disciplines, such as social science (journals 53% vs. books 35%) and political science (journals 49% vs. books 34%), the differences between the use of journals and books were not as great, indicating that they have a closer value in both disciplines. On the average, over all types of documents, social scientists preferred to use journals in exploration and support of their own research.

Aside from the citations of journal articles and book materials, the number of theses and dissertations cited in the journal of education was higher than those in journals of other disciplines. Online resources such as websites or electronic files were cited more frequently in the political science journal, suggesting that political scientists use more digital literature as references in their research. Research reports were cited more in the journal of economics than in other disciplines, which indicates that economists tended to prove or support their own research by data or results provided by research reports. Furthermore, the fact that economists and scholars of management cited a few unpublished manuscripts and working papers showed the significance of informal and unpublished materials to these two disciplines.

Citation Location

The number and location of citations from the 116 articles complying with the IMRAD format were calculated to analyze the distribution of citations in structured research articles. There were 11,149 citations collected in the section of introduction (literature review included), methods and materials, results, and discussion.

The distribution of citations in different sections of an article may help to determine the status, research patterns and characteristics of a discipline. As Table 3 shows, citations appeared the most in the introduction section of articles in every discipline of social science. The Introduction may include literature reviews, and both sections need a few references for proving points or serving as motivations. In the six disciplines of social science, the highest number of citations in the introduction sections occurred in the journals of sociology and political science, while the lowest was in the journal of economics. For the method section, scholars of economics and management cite more frequently in the section of methods and materials. In contrast, the sociologists cite the least frequently.

Table 3. Distribution of citation location in social science journals of Taiwan.

Discipline	*Introduction*		*Methodology & Materials*		*Results*		*Discussion*		*Total*	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Management[1]	1,799	65.4	480	17.3	256	9.3	216	7.9	2,751	24.7
Economics[2]	499	54.6	164	17.9	189	20.7	62	6.8	914	8.2
Political Sci.[3]	931	70.5	171	13.0	165	12.5	53	4.0	1,320	11.8
Psychology[4]	1,048	58.6	198	11.1	222	12.4	320	17.9	1,788	16.0
Education[5]	1,888	64.1	282	9.6	295	10.0	481	16.3	2,946	26.4
Sociology[6]	993	69.4	63	4.4	254	17.8	120	8.4	1,430	12.8
Total	7,158	64.2	1,358	12.2	1,381	12.4	1,252	11.2	11,149	100

1. Journal of Management; 2. Academia Economic Papers; 3. Taiwan Political Science Review; 4. Chinese Journal of Psychology; 5. Bulletin of Educational Psychology; 6. Taiwanese Journal of Sociology

In the results section, economists tended to cite more articles for comparison and contrast. Aside from economics, the number of citations in the results section of the sociology journal

also high. In the management journal, descriptive statistics and quantitative analysis may be the major causes of its lower number of citations in the results section.

In the discussion section, the number of citations may reflect scholars' degree of concern about deliberations and evaluation of research outcomes. The top two numbers of citations in discussion section occurred in the journals of psychology and education.

Citation Type

In addition to the distribution of citation location, Peritz's classification scheme of citation type is used to classify articles cited in the sample journals. Mapping was made to inspect the relations between citation type and citation location and to analyze the differences among the six disciplines. The eight categories of citation classification scheme proposed by Peritz (1983, pp.304-305) are: 1. Setting the stage for the present study; 2. Background information; 3. Methodological; 4. Comparative; 5. Argumentative speculative, hypothetical; 6. Documentary; 7. Historical and 8. Casual.

Table 4. Distribution of citation type.

Citation type	*Sociology*		*Education*		*Psychology*		*Political Science*		*Economics*		*Management*		*Total*	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Setting the stage for the present study	271	56.7	153	53.5	235	58.6	311	56.3	133	48.9	411	63.8	1,514	57.5
Background information	44	9.2	15	5.2	21	5.2	23	4.2	9	3.3	13	2.0	125	4.7
Methodological	33	6.9	34	11.9	54	13.5	85	15.4	85	31.3	109	16.9	400	15.2
Comparative	70	14.6	35	12.2	68	17.0	46	8.3	38	14.0	72	11.2	329	12.5
Argumentative, speculative, hypothetical	45	9.4	48	16.8	23	5.7	17	3.1	5	1.8	38	5.9	176	6.7
Documentary	15	3.1	0	0.0	0	0.0	66	12.0	2	0.7	1	0.2	84	3.2
Historical	0	0.0	1	0.3	0	0.0	0	0.0	0	0.0	0	0.0	1	0.0
Casual	0	0.0	0	0.0	0	0.0	4	0.7	0	0.0	0	0.0	4	0.2
Total	478	-	286	-	401	-	552	-	272	-	644	-	2,633	-

Based on Table 4, the highest percentage of citations classified as "setting the stage for the present research" appeared in the journal of management (64%), and the lowest in the journal of economics (49%). Compared to other types of citation, citations that set the stage for the present study were significantly high in all six disciplines. The citation type of "background information" was most frequently found in the journal of sociology, while it was least frequent in the journal of management. The journal of economics contained the most methodological citations, which accounted for 31% of total citations, while the journal of sociology the least, which accounted for 7%; the interval between was rather large. Comparative citations were most found in the journal of psychology (17%) and the least in the journal of political science (8%). The journal of education included the most citations (17%), which were used in the presentation of argument, speculation, and hypothesis while the journal of political science the least (merely 3%). Documentary citations accounted for 12% of total citations in the journal of political science, which was the top among the six disciplines; whereas there was no such type of citations found in the journals of education and psychology. The citation types of "historical" and "casual" were hardly found in the journals of six disciplines, with only one historical citation in the journal of education and four casual citations in the journal of political science.

The distribution of citation type may reveal the research characteristics of a certain disciplines. For example, scholars of management tend to cite a large amount of literature to support or motivate their own research, whereas economists cite more methodological materials in their works, which indicates that research methods are valued more in economics. Political scientists tended to cite more raw data to support their studies; whereas scholars of education cited more articles for argumentation, speculation, and hypothesis. Comparative citations appeared the most in the journal of psychology, suggesting that psychological researchers tend to introduce other research in their own studies for comparison, correction, or corroboration.

Citation Type and Citation Location

According to Peritz's study, citation type was highly relevant to citation location. In this study, therefore, the relation between citation type and citation location in the six discipline sample journals was analyzed as follows.

Sociology

As Table 5 shows, in the journal of sociology, the number of citations that set the stage for the present study was 271, accounting for 56.7% of the total citations. Comparative citations accounted for 14.6% of the total citations, suggesting that the materials being cited in the journal articles were used to describe or support the present research. The citation type of "setting the stage for the present study" appeared primarily in the introduction section, while methodological citations that introduced the process of other research were mostly in the methods and materials section. In the results section, comparative, argumentative, speculative, and hypothetical citations accounted for the greatest number of citations. In the discussion session, comparative citations comprised the major part of total citations.

Table 5. Citations in Taiwanese Journal of Sociology by category and location.

Location / Category	Introduction		Methodology & Materials		Results		Discussion		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%
Setting the stage	271	86.0	0	0.0	0	0.0	0	0.0	271	56.7
Background information	30	9.5	1	4.0	13	18.8	0	0.0	44	9.2
Methodology	7	2.2	19	76.0	7	10.1	0	0.0	33	6.9
Comparative	0	0.0	0	0.0	23	33.3	47	68.1	70	14.6
Argumentative, speculative, hypothetical	3	1.0	0	0.0	23	33.3	19	27.5	45	9.4
Documentary	4	1.3	5	20.0	3	4.3	3	4.3	15	3.1
Historical	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Casual	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Total	315	100.0	25	100.0	69	100.0	69	100.0	478	100.0

Education

In the journal of education, the citation type of "setting the stage for the present study" accounted for the largest percentage of the total citations, 53.5%, as shown in Table 6. The distribution of methodological citations, comparative citations, and argumentative, speculative and hypothetical citations was rather even. Similar to the distribution in the sociology journal, all of the citations that set the stage for the present study appeared in the introduction section, and the citations in methods and materials section were mostly methodological citations, while there were few citations in the results section. As for the

discussion part, the numbers of comparative, argumentative, speculative and hypothetical citations, especially the last three types, greatly exceeded other types of citation, indicating that scholars of education often introduced other research for detailed exploration, or made further inference based on previous studies.

Table 6. Citations in Bulletin of Educational Psychology by category and location.

Category \ Location	Introduction		Methodology & Materials		Results		Discussion		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%
Setting the stage	153	90.5	0	0.0	0	0.0	0	0.0	153	53.5
Background information	12	7.1	0	0.0	0	0.0	3	3.8	15	5.2
Methodology	3	1.8	30	100.0	1	12.5	0	0.0	34	11.9
Comparative	0	0.0	0	0.0	1	12.5	34	43.0	35	12.2
Argumentative, speculative, hypothetical	0	0.0	0	0.0	6	75.0	42	53.2	48	16.8
Historical	1	0.6	0	0.0	0	0.0	0	0.0	1	0.3
Total	169	100.0	30	100.0	8	100.0	79	100.0	286	100.0

Psychology

In the journal of psychology, as Table 7 presented, over half of its citations were classified as the type of "setting the stage for the present studies" (58.6%). In the discussion section, comparative citations accounting for 71% of total citations appeared in the discussion section, which suggests that psychologists tend to cite other materials as comparisons to examine whether their research results were consistent with previous studies, or to correct previous research and hereafter propose their own unique results.

Table 7. Citations in Chinese Journal of Psychology by category and location.

Category \ Location	Introduction		Methodology & Materials		Results		Discussion		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%
Setting the stage	235	93.6	0	0.0	0	0.0	0	0.0	235	58.6
Background information	16	6.4	4	7.3	0	0.0	1	1.3	21	5.2
Methodology	0	0.0	48	87.3	6	30.0	0	0.0	54	13.5
Comparative	0	0.0	3	5.5	12	60.0	53	70.7	68	17.0
Argumentative, speculative, hypothetical	0	0.0	0	0.0	2	10.0	21	28.0	23	5.7
Total	251	100.0	55	100.0	20	100.0	75	100.0	401	100.0

Political Science

From Table 8, it is clear that "setting the stage for the present study" citations were the most numerous of the eight types of citation, accounting for 56% of the total citations in the journal of political science. The second most numerous were the methodological citations, though they comprised only 15% of total citations, while the percentage of other types of citations was even lower. Interestingly, political scientists cited much more statistical data in the introduction section, which indicates that they tended to use quantitative data or factual information to support their studies when writing introduction and literature review. As for the other locations, comparison was often made in the results section, while citations in the discussion section mostly served as bases for inference.

Table 8. Citations in Taiwan Political Science Review by category and location.

Category \ Location	Introduction		Methodology & Materials		Results		Discussion		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%
Setting the stage	305	79.0	6	6.3	0	0.0	0	0.0	311	56.3
Background information	16	4.1	6	6.3	1	1.9	0	0.0	23	4.2
Methodology	7	1.8	73	76.0	5	9.3	0	0.0	85	15.4
Comparative	0	0.0	3	3.1	40	74.1	3	18.8	46	8.3
Argumentative, speculative, hypothetical	0	0.0	0	0.0	4	7.4	13	81.3	17	3.1
Statistical data	58	15.0	4	4.2	4	7.4	0	0.0	66	12.0
Casual	0	0.0	4	4.2	0	0.0	0	0.0	4	0.7
Total	386	100.0	96	100.0	54	100.0	16	100.0	552	100.0

Economics

Though the "setting the stage for the present study" citations were more numerous than other types of citations in the journal of economics, its percentage was a bit lower than in other disciplines, accounting for only 49% of all the citations in the journal. Table 9 also shows that economists cited more methodological materials, accounting for 31% of all citations, indicating a preference for empirical study in the field of economics. Models or methods proposed by other research were frequently found in the studies of economics, and comparative citations were mostly made in the section of results, which is consistent with the inference that economists were used to comparing their research results with previous studies. However, few citations in the discussion section revealed little of the characteristics of citation types in the journal of economics.

Table 9. Citations in Academia Economic Papers by category and location.

Category \ Location	Introduction		Methodology & Materials		Results		Discussion		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%
Setting the stage	129	92.8	2	3.8	2	2.7	0	0.0	133	48.9
Background information	5	3.6	0	0.0	4	5.3	0	0.0	9	3.3
Methodology	5	3.6	49	94.2	29	38.7	2	33.3	85	31.3
Comparative	0	0.0	0	0.0	37	49.3	1	16.7	38	14.0
Argumentative, speculative, hypothetical	0	0.0	0	0.0	2	2.7	3	50.0	5	1.8
Statistical data	0	0.0	1	1.9	1	1.3	0	0.0	2	0.7
Total	139	100.0	52	100.0	75	100.0	6	100.0	272	100.0

Management

The relations between citation type and location in the journal of management can be seen in Table 10. The percentage of citations that set the stage for the present study was comparatively high (64%) in the journal of management, which was the only discipline whose percentage exceeded 60% among all six disciplines discussed in this study. Unlike economists, who were found to care more about methods and materials, scholars of management focused more on literature reviews, tending to project the importance of their research questions by contrasting them with previous studies. Yet they still valued the implementation of research

methods from other studies, according to the second top percentage (17%) of methodological citations. Comparative, argumentative, speculative and hypothetical citations also appeared in the section of discussion, while comparative citations accounted for more percentage (11%) of total citations in the journal of management.

Table 10. Citations in Journal of Management by category and location.

Location / Category	Introduction		Methodology & Materials		Results		Discussion		Total	
	No.	%	No.	%	No.	%	No.	%	No.	%
Setting the stage	400	97.3	10	8.3	0	0.0	1	1.6	411	63.8
Background information	6	1.5	7	5.8	0	0.0	0	0.0	13	2.0
Methodology	5	1.2	90	75.0	11	22.4	3	4.7	109	16.9
Comparative	0	0.0	12	10.0	22	44.9	38	59.4	72	11.2
Argumentative, speculative, hypothetical	0	0.0	0	0.0	16	32.7	22	34.4	38	5.9
Statistical data	0	0.0	1	0.8	0	0.0	0	0.0	1	0.2
Total	411	100.0	120	100.0	49	100.0	64	100.0	644	100.0

Table 11. Citations in social science journals in Taiwan by category and location.

Location / Category	Introduction (%)		Methodology & Materials (%)		Results (%)		Discussion (%)		Total (%)	
Setting the stage	89.87	3.07	0.45	0.27	56.3	89.87	3.07	0.45	0.27	56.3
Background information	5.37	3.9	4.33	0.85	4.85	5.37	3.9	4.33	0.85	4.85
Methodology	1.77	84.75	20.5	6.33	15.98	1.77	84.75	20.5	6.33	15.98
Comparative	0	3.1	45.68	46.12	12.88	0	3.1	45.68	46.12	12.88
Argumentative, speculative, hypothetical	0.17	0	26.85	45.73	7.12	0.17	0	26.85	45.73	7.12
Documentary	0.22	3.33	0.72	0.72	0.52	0.22	3.33	0.72	0.72	0.52
Historical	0.1	0	0	0	0.05	0.1	0	0	0	0.05
Statistical data	2.5	1.15	1.45	0	2.15	2.5	1.15	1.45	0	2.15
Casual	0	0.7	0	0	0.12	0	0.7	0	0	0.12
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

In sum, the percentages of "setting the stage for the present study" citations ranked first in the journals of all six disciplines, with management accounting for 63.8%, psychology 58.6%, sociology 56.7%, political science 56.3%, education 53.5%, and economics 48.9%. The percentage of methodological citations to total citations was 15.2%, which made the second high among the six journals, with economics accounting for 13.3%, management 16.9%, and political science 15.4%. As for the comparative citations, psychology (17%) and sociology (14.6%) covered more than other disciplines, while education exceeded other disciplines in the argumentative, speculative and hypothetical citations, with a percentage of 16.8%.

In Peritz's study, the citation type was highly relevant to the citation location, as confirmed by the results of this research shown in Table 11. In the introduction section, most citations belonged to the category of "setting the stage for the present study"; in the section of methods and materials, methodological citations appeared the most; as for the section of results and discussion, although the distribution of citation types varied among the six disciplines,

comparative, argumentative, speculative, and hypothetical citations were the most on the average. Overall, the research outcomes indicated that most social scientists of Taiwan complied with international writing format, and confirmed the hypothesis proposed by Peritz that the citation location was highly relevant to the citation type.

Summary and Discussions

This study explores and compares the distribution of article types of journals, languages of citation, citation years, document types of citations, citation types and locations of citations among citations in the top social science journals of six disciplines published in Taiwan and indexed in the Taiwan Social Sciences Citation Index (TSSCI). The following conclusions may be drawn from the results.

1. Journals and books were the most cited materials; English language articles were the most cited in social science studies in Taiwan.
2. Social scientists in Taiwan tended to cite materials published within the past 10 year, most citations in the sample journals were for articles with four to seven years of the journal publication, indicating that social scientists in Taiwan tend to cite the most recent articles.
3. The ratio of articles following IMRAD format was high in social science journal in Taiwan, suggesting that the top social science journals comply strictly with the IMRAD format of structured articles in Taiwan.
4. In Taiwan, citations in social science journals occurred the most in the introduction section, while the conclusions section had the least: The distribution of citations in different sections of an article may indicate the status and characteristics of a research domain. In this study, citations occurred most frequently in the introduction section for each of the social science disciplines. The introduction may include research background and literature review, and both sections need quite a few references for proving points or indicating motivation. For the methods section, economics and management had high percentage of citations, indicating that scholars in these two disciplines were used to adopting models, designations or operations from previously published research. In the results section, economists and psychologists tended to cite more articles for comparison and contrast. In general, citations appeared least frequently in the conclusions section, though the percentage rates were still a little higher in psychology and education, revealing their concern for further discussion and evaluation of research results.
5. Social scientists mostly cite to set the stage for their present studies: The “setting the stage for the present study” citations were the most frequently used in the sampled social science journals, accounting for 57.5% of all citations. From the distribution of citation type, it is clear that social scientists tended to cite in order to provide support or motivation for their own studies, which as shown by the large number of “setting the stage for the present study” type of citations. Scholars of economics, management and political science used to introduce methods and materials to compare or verify their findings. Psychologists and sociologists tended to compare their research results with previous studies, whereas scholars of education emphasized discussion greater than other sections.
6. Citation type is highly relevant to the citation location, which is consistent with the findings of Peritz’s study.

In this study, citation characteristics of social scientists in Taiwan were analyzed via bibliographic data such as types of cited materials and languages of citations. The results revealed the citation characteristics and information need of Taiwan’s social scientists, which could be valuable in collection development of libraries or refinement of information services. Under the assumption that citations indicate the actual use of materials, the distribution of publication years and citing half-life may serve as evidence for libraries to order or suspend

information resources (electronic journals, for instance), which could help to achieve similar goals on better budget allocation. Providing further exploration and examination of citations, this study is also expected to provide a better understanding of citation nature, and is anticipated to serve as a basis for future empirical studies.

There are limitations for the citation type determination by the textual analyst on the basis of the surrounding text. This is because, first, citation types may not be apparent simply by studying the text and, second, effective analysis sometimes requires specialist knowledge in the discipline of the texts being studied. Therefore, conducting an interview study with authors of the text to obtain their own views of citation types is suggested for further study. The small number of samples involved in this study preclude from making confident generalizations regarding the frequency of the citation types across these social science disciplines as a whole. Thus, the collection and analysis of a larger sample size is also suggested for further study.

Conclusion and Suggestion

The study is still to be improved owing to its restrictions and limitations. For better interpretation of the research trend, paradigm shifts and citation distribution of social sciences, it is suggested that the time frame, scope and quantity of sample collection be extended, including citations from both domestic and foreign articles. Co-research with experts and scholars in concerning disciplines are recommended as well. Even more, to reach a fuller apprehension of research features in academia by means of citation characteristics, samples in humanities and sciences may be examined in the future studies. Though Periz's classification scheme is known for its simplicity and directness, it is not quite suitable for those non-empirical studies. However, the Citation Content Analysis (CCA) framework proposed by Zhang, Ding and Milojevic (2013) may serve as solution to the problem, since it adopts both syntactic and semantic measurement of citation, which thus makes cross-field comparison possible. As for the essence of citation, the purposes and motives of citation are also valuable topics for further studying.

Acknowledgments

This work was supported by grant NSC1007-2410-H-004-153-MY2 from the National Science Council, Taiwan, R.O.C. Bibliometric data collected by Min-yee Lee, Graduate Institute of Library, Information and Archival Studies, National Chengchi University, Taiwan is very much appreciated.

References

Case, D.O. & Miller, J.B. (2011). Do bibliometricians cite differently from other scholars? *Journal of the American Society for Information Science and Technology*, *62*(3), 421-432.

Frost, C.O. (1979). The use of citations in library research: a preliminary classification of citation functions. *Library Quarterly*, *49*(4), 399.

Harwood, N. (2008). An interview-based study of the functions of citations in academic writing across two disciplines. *Journal of Pragmatics*, *41*(3), 497-518.

Moravcsik, M.J. & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science*, *5*, 86-92.

Peritz, B C. (1983). A classification of citation roles for the social sciences and related fields. *Scientometrics*, 5, 303-312.

Voos, H. and Dagaev, K.S. (1976). Are all citation equal? Or, did we op cit your idem? *The Journal of Academic Librarianship*, *1*(6), 19-21.

Zhang, G, Ding, Y. & Milojevic, S. Citation Content Analysis (CCA): A Framework for Syntactic and Semantic Analysis of Citation Content. *Journal of the American Society for Information Science and Technology*, *64*(7), 1490-1503.

University Citation Distributions

Antonio Perianes-Rodriguez[1] and Javier Ruiz-Castillo[2]

[1] *antonio.perianes@uc3m.es*
Universidad Carlos III, Department of Library and Information Science, SCImago Research Group, C/ Madrid, 128, 28903 Getafe, Madrid (Spain)

[2] *jrc@eco.uc3m.es*
Universidad Carlos III, Departamento de Economía, C/ Madrid, 126, 28903 Getafe, Madrid (Spain)

Abstract
In this paper we investigate the characteristics of the citation distributions of the 500 universities in the 2013 edition of the CWTS Leiden Ranking. We use a WoS dataset consisting of 3.6 million articles published in 2003-2008 with a five-year citation window, and classified into 5,119 clusters. The main findings are the following four. Firstly, The universality claim, according to which all university citation distributions, appropriately normalized, follow a single functional form, is not supported by the data. Secondly, nevertheless, the 500 university citation distributions are all highly skewed and very similar. Broadly speaking, university citation distributions appear to behave as if they differ by a relatively constant scale factor over a large, intermediate part of their support. Thirdly, citation impact differences between universities account for 3.85% of overall citation inequality. However, these differences are greatly reduced when university citation distributions are normalized using their MNCS values as normalization factors. Finally, the above results have important practical consequences. On one hand, we only need a single explanatory model for the single type of high skewness characterizing all university citation distributions. On the other hand, the similarity of university citation distributions goes a long way in explaining the similarity of the university rankings obtained with the MNCS and the top 10% indicator.

Conference Topic
Citation and co-citation analysis

Introduction
Universities constitute a key vehicle in the production of knowledge in contemporary societies. However, the evaluation of the quality, or the relevance of the research done by universities in a myriad of scientific fields is a very difficult problem. For the assessment of the performance of research units of all types during the last decades, academic bodies, public officials in charge of science policy, and specialists in the field of Scientometrics have been paying increasing attention to one observable aspect of research in all fields: the citation impact of publications in the periodical literature.

In this paper, we focus on this aspect of research for the 500 universities included in the 2013 edition of the CWTS Leiden Ranking (LR universities) (Waltman et al., 2012a). We use a Web of Science (WoS) dataset consisting of 3.6 million publications in the 2005-2008 period, the citations they receive during a five-year citation window for each year in that period, and a classification system consisting of 5,119 clusters (Ruiz-Castillo & Waltman, 2015).

The construction of university citation distributions in the all-sciences case requires the prior solution of two methodological problems: the assignment of responsibility for publications with two or more co-authors belonging to different institutions, and the aggregation of the citation impact achieved by research units working in different scientific clusters. We solve these problems using a fractional counting approach in the presence of co-authorship, and the standard field-normalization procedure where cluster mean citations are used as normalization factors.

Once these two problems have been solved, specialists typically debate the properties of alternative citation impact indicators. In this paper, we study a basic aspect of the research

evaluation problem that comes *before* the comparison of the advantages and shortcomings of specific indicators, namely, the characteristics of the university citation distributions themselves. These distributions arise from the interplay of a complex set of economic, sociological, and intellectual factors that influence in a way hard to summarize the research performance of each university in every field. In this scenario, it is well known that some universities are more productive or successful than others in terms of the number of publications and/or the mean citation that these publications receive. However, little is known concerning the shape of university citation distributions abstracting from size and mean citation differences. In order to contribute to this knowledge, in this paper we investigate the following four issues.

Firstly, we inquire whether university citation distributions are universally distributed. The universality condition, borrowed from statistical physics, means that, appropriately normalized, citation distributions follow a unique functional form within the bounds set by random variation. Radichhi *et al.* (2008) suggest a statistical test of this condition in their study of 14 WoS journal subject categories. According to this test, the universality condition is not satisfied for our 500 university citation distributions. This is consistent with previous results for large classification systems in WoS datasets consisting of complete field citation distributions that include publications with zero citations (Albarrán & Ruiz-Castillo, 2011, Albarrán *et al.*, 2011a, Waltman *et al.*, 2012a, Perianes-Rodriguez & Ruiz-Castillo, 2014).

Secondly, in view of the above finding, we ask: are at least university citation distributions as highly skewed and as similar among each other as previous results indicate for field citation distributions? Using the same size- and scale-independent techniques that have been used in previous research, we confirm that this is the case in our dataset. This result has been established at different aggregation levels, publication years, and citation window lengths, and independently of whether the problem of the multiple assignment of publications to sub-fields in WoS datasets is solved by following a multiplicative or a fractional approach (Glänzel, 2007, Radicchi *et al.*, 2008, Albarrán & Ruiz-Castillo, 2011, Albarrán *et al.*, 2012, Herranz & Ruiz-Castillo, 2012, Waltman *et al.*, 2012a, Radicci & Castellano, 2012, Li *et al.*, 2013, Ruiz-Castillo & Waltman, 2015, Perianes-Rodriguez & Ruiz-Castillo, 2014). Similar conclusions concerning the skewness and similarity of individual productivity distributions are found when authors are classified into 30 broad scientific fields (Ruiz-Castillo & Costas, 2014).

Thirdly, using the measuring framework introduced in Crespo *et al.* (2013), we investigate how important is the effect of differences in citation impact between LR universities in the overall citation inequality in the union of the 500 LR university citation distributions. Furthermore, we inquire up to what point this effect can be accounted for by scale factors captured by the universities' Mean Normalized Citation Score (*MNCS* hereafter). The answer is that citation impact differences between universities account for 3.85% of overall citation inequality –a much smaller percentage than what is found in the context of production and citation practice differences between scientific fields (Crespo *et al.*, 2013, 2014, Ruiz-Castillo & Waltman, 2015, Perianes-Rodriguez & Ruiz-Castillo, 2014). These differences are greatly reduced when university citation distributions are normalized using their *MNCS* values as normalization factors.

Finally, we discuss the implications of these results for the understanding of the high correlation between the university rankings according to two citation impact indicators: the *MNCS*, and the Top 10% indicator of scientific excellence (the $PP_{top\ 10\%}$ indicator hereafter), defined as the percentage of an institution's output included into the set formed by 10% of the world most cited papers in the different scientific fields. The latter indicator has been recently adopted by well-established institutions, such as the CWTS in the Netherlands, and SCImago in Spain.

The rest of the paper is organized into two Sections. The first section presents the empirical results, while the next section discusses further research.

Empirical results

The universality of university citation distributions

Let $\boldsymbol{c_i}$ be the LR university i field-normalized citation distribution. Note that, for each university, the mean citation of $\boldsymbol{c_i}$ is precisely the Mean Normalized Citation Score (MNCS hereafter). Let $\boldsymbol{c^*_i}$ be the normalized citation distribution of university i using the university MNCS as the normalization factor. Let $\boldsymbol{C^*}$ be the union of the universities' normalized citation distributions, $\boldsymbol{C^*} = \cup_i \{\boldsymbol{c^*_i}\}$, where publications are ranked in increasing order of the number of normalized citations. Let $\boldsymbol{X_z}$ be the set of publications in the top z% of distribution $\boldsymbol{C^*}$, and let $\boldsymbol{x_{zi}}$ be the publications in X_z that belongs to the i-th university, so that $\boldsymbol{X_z} = \cup_i \{\boldsymbol{x_{zi}}\}$. In the terminology of Radicchi *et al.* (2008), if the ranking is fair, or unbiased, the percentage of publications that the set $\boldsymbol{x_{zi}}$ represents within each university should be near z% with small fluctuations. Let N_c and N_i be, respectively, the number of universities and the number of publications in the i-th university. Assuming that publications of the various universities are scattered uniformly along the rank axis, for any value z% one would expect the average relative frequency of the number of articles in any university to be z% with a standard deviation $\sigma_z = \{[z(100 - z)\Sigma_i\ (1/N_i)]/N_c\}^{1/2}$, which is equation (2) in Radicchi *et al.* (2008).

Table 1. Percentage of publications in each sub-field that appear in the top z% of the global rank, together with the standard deviation, σ_z, and the coefficient of variation, σ_z/z.

Theoretical values			Normalised distribution		
z%	σ_z	σ_z/z	z%	σ_z	σ_z/z
(1)	(2)	(3)	(4)	(5)	(6)
1	0.20	0.20	0.96	0.29	0.30
5	0.43	0.09	4.95	0.90	0.18
10	0.59	0.06	10.00	1.46	0.15
20	0.79	0.04	20.03	2.41	0.12
30	0.91	0.03	30.04	3.11	0.10
40	0.97	0.02	40.00	3.49	0.09
50	0.99	0.02	49.88	3.76	0.08
75	0.86	0.01	74.73	4.08	0.05
90	0.59	0.01	88.94	4.08	0.05

For each z value in a certain sequence, column 2 in Table 1 presents the standard deviations σ_z, while column 3 is the theoretical coefficient of variation, namely, σ_z/z. Columns 4 to 6 contain the values for the average z, the standard deviation σ_z, and the coefficient of variation σ_z/z obtained empirically in distribution $\boldsymbol{C^*}$.

Although σ_z varies non-linearly with z, the theoretical coefficient of variation in column 3 raises from 0.01 to 0.20 when we proceed from z = 90% towards z = 1%. In the normalized case, the considerable differences with the theoretical values in column 6, above all for lower values of z, indicate the lack of universality for this set of 500 university citation distributions.

This conclusion contrasts with the universality claim in Chatterjee *et al.* (2014), who study 42 academic institutions across the world, their publications in four years, 1980, 1990, 2000, and 2010, and the citations they receive according to the WoS until July 2014. We should emphasize that this paper has a number of technical problems. The criterion for selecting their

42 academic institutions is not given, and there is no information on how the following three problems have been solved: the assignment of publications in WoS datasets to multiple journal subject categories, the assignment of responsibility for co-authored publications, and the all-sciences aggregation problem. Nevertheless, we will proceed discussing their results. Chatterjee *et al.* (2014) explain that, for each publication year, the university normalized citation distributions fit well to a lognormal for most of the range, although the poorly cited publications seem to follow another distribution, while the upper tail is better described by a power law. This is quite different from the claim that there is a single functional form for the entire domain of definition of the 42 institutions in their sample. Our statistical approach tests whether the universality claim is supported by the data over the entire domain of the 500 LR universities. In this sense, our results do not contradict each other. We both agree that the universality claim over the entire domain is not the case in our respective samples.

On the other hand, the main problem with the still unpublished version of Chatterjee *et al.* (2014) is that, in our opinion, their statistical methods are not clearly explained. Unfortunately, the authors do not explain the following three aspects: (i) how the partition of the domain into three segments is estimated for each university, and whether this partition is universal; (ii) which tests have been used to determine the functional form chosen in each segment versus possible alternatives; (iii) how the confidence interval for the power law parameter has been estimated, and which is the confidence interval for the lognormal parameters. As a matter of fact, the only clear evidence for the distributions collapse into a universal curve is the graphical illustration provided for a sample –whose selection is unexplained– of 24 of the original 42 academic institutions.

The skewness and similarity of university citation distributions

The skewness of citation distributions is assessed by simply partitioning citation distributions into three classes of articles with low, fair, and very high number of citations. For this purpose, we follow the Characteristic Scores and Scale (CSS hereafter) approach, first introduced in Scientometrics by Schubert *et al.* (1987). In our application of the CSS technique, the following two *characteristic scores* are determined for every university: μ_1 = mean citation, which in our context is equal to the *MNCS*, and μ_2 = mean citation for articles with citations greater than μ_1. We consider the partition of the distribution into three broad categories: (i) articles with a low number of citations, smaller than or equal to μ_1; (iii) fairly cited articles, with a number of citations greater than μ_1 and smaller than or equal to μ_2, and (iii) articles with a remarkable or outstanding number of citations greater than μ_2. For each citation distribution, we measure the percentages of publications in the three categories, as well as the percentages of the total citations accounted for by the three categories. The average, standard deviation, and coefficient of variation for the 500 university values of the percentages of publications, the percentages of the total citations in the three categories are included in Table 2.

The results are remarkable. In principle, differences in resources, intellectual traditions, organization, the structure of incentives, and other factors lead us to expect large differences between the 500 LR university citation distributions in different parts of the world. However, judging from the size of the standard deviations and the coefficient of variations for the 500 universities, we find that university citation distributions are extremely similar. At the same time, the distributions are highly skewed: on average, the MNCS values of the 500 universities is 12.9 percentage points above the median, while the 12.5 of outstanding articles account for 44.4% of all normalized citations.

Table 2. The skewness of citation distributions according to the CSS approach. Percentages of articles, and percentages of citations by category. Average, standard deviation, and coefficient of variation over the 500 LR universities, and results for the overall citation distribution.

	Percentage of articles in category:			Percentage of citations in category:		
	1	2	3	1	2	3
Average (Std. deviation)	62.9 (1.9)	24.6 (1.2)	12.5 (1.2)	22.9 (1.7)	32.7 (0.8)	44.4 (1.5)
Coefficient of variation	0.03	0.05	0.10	0.08	0.02	0.03

For the sake of robustness, we have conducted two more sets of computations. In the first place, in the presence of co-authorship we have assigned publications to universities in a multiplicative way. In the second place, we have studied the raw citation distributions without the benefit of any field-normalization procedure. Interestingly enough, the results are very similar to those obtained for field-normalized university citation distributions in the fractional case. Thus, we conclude that the characteristics of university citation distributions are robust to the way the assignment of publications to universities in the presence of co-authorship and the all-sciences aggregation problem are solved.

Finally, we should mention the results of two contributions closer to our own in which research publications are aggregated into the type of organization unit to which the authors belong. Firstly, Albarrán *et al.* (2015) study the partition of world citation distributions into 36 countries and two residual geographical areas using a dataset, comparable to ours, consisting of 4.4 million articles published in 1998-2003 with a five-year citation window for each year. They find that, at least in some broad fields and in the all-sciences case, the country citation distributions are not only highly skewed, but also very similar across countries –a result parallel to our own for the 500 LR universities. Secondly, Perianes-Rodriguez & Ruiz-Castillo (2015) study a set of 2,530 highly productive economists who work in 2007 in a selection of the top 81 economics departments in the world. Contrary to previous results for field or country citation distributions, we find that productivity distributions are very different across the 81 economics departments. However, the data in Perianes-Rodriguez & Ruiz-Castillo (2015) does not consist of department citation distributions of articles published in a certain period of time with a citation window of common length, but of the individual productivity of faculty members in each department, where individual productivity is measured as a quality index that weights differently the articles published up to 2007 by each researcher in four journal equivalent classes. Nevertheless, we cannot rule out that the similarity of citation distributions is a phenomenon present at certain aggregate levels. To settle this issue, we need more work at the department level with citation distributions articles published in a certain period of time with a common citation window.

The importance of citation impact differences between universities

Together with the assessment of the between-group variability concerning the shape of university citation distributions, we are interested in measuring how important are the citation impact differences between universities. Formally, this problem is analogous to the measurement of the importance of differences in production and citation practices between scientific fields. For the latter, Crespo *et al.* (2013) suggested to measure the impact of such differences on the overall citation inequality for the entire set of field citation distributions applying an additively decomposable citation inequality index to a double partition into scientific fields and quantiles. Similarly, in our case we measure how much of the overall citation inequality exhibited by the union of the 500 LR university citation distributions can be attributed to the citation impact differences between universities (this is also the approach adopted in Perianes-Rodriguez & Ruiz-Castillo, 2014a, to assess the effect of citation impact between countries).

For that purpose, we begin with the partition of, say, each university citation distribution into Π quantiles, indexed by π = 1,.., Π. In practice, in this paper we use the partition into percentiles, that is, we choose Π= 100. Assume for a moment that, in any university u, we disregard the citation inequality within every percentile by assigning to every article in that percentile the mean citation of the percentile itself, μ_u^{π}. The interpretation of the fact that, for example, $\mu_u^{\pi} = 2\,\mu_v^{\pi}$ is that, on average, the citation impact of university u is twice as large as the citation impact of university v in spite of the fact that both quantities represent a common underlying phenomenon, namely, the same *degree of citation impact* in both universities. In other words, for any π, the distance between μ_u^{π} and μ_v^{π} is entirely attributable to the difference in the citation impact that prevails in the two universities for publications with the same degree of excellence in each of them. Thus, the citation inequality between universities at each percentile, denoted by $I(\pi)$, is entirely attributable to the citation impact differences between the 500 LR universities holding constant the degree of excellence in all universities at quantile π. Hence, any weighted average of these quantities, denoted by *IDCU* (*I*nequality due to *D*ifferences in *C*itation impact between *U*niversities), provides a good measure of the total impact on overall citation inequality that can be attributed to such differences. Let $\boldsymbol{c_i}$ be university i citation distribution, and let $\boldsymbol{C}$ be the union of the universities citation distributions, $\boldsymbol{C} = \cup\,\{\boldsymbol{c_i}\}$. We use the ratio

$$IDCU/I(\boldsymbol{C}) \qquad (1)$$

to assess the relative effect on overall citation inequality, $I(\boldsymbol{C})$, attributed to citation impact differences between universities (for details, see Crespo *et al.*, 2013).

Finally, we are interested in estimating how important scale differences between university citation distributions are in accounting for the effect measured by expression (1). Following the experience in other contexts, we choose the university mean citations as normalization factors. To assess the importance of such scale factors, we use the relative change in the *IDPD* term, that is, the ratio

$$[IDCU - IDCU^*]/IDCU, \qquad (2)$$

where *IDCU** is the term that measures the effect on overall citation inequality attributed to the differences in university distributions after the normalization of university citation distributions using university mean citations as normalization factors (for details, see again Crespo *et al.*, 2013). The estimates for expressions (1) and (2) in our dataset are included in table 3:

Table 3. The effect on overall citation inequality, I(C), of the differences in citation impact between universities before and after MNCS normalization, and the impact of normalization on this effect.

	Normalization impact = 100 [*IDPD – IDCP*/IDCP*]	
Before MNCS normalization, 100 [*IDPU*/*I*(*C*)]	3.85 %	-
After MNCS normalization, 100 [*IDPU/*I*(*C*)]**	0.72 %	81.9 %

It is interesting to compare these figures with what was obtained in two instances in the previous literature. The first case concerns the partition into 36 countries and two residual geographical areas in the all-sciences case (Albarrán *et al.*, 2014), while the second case

refers to 219 WoS sub-fields (Crespo *et al.*, 2014). Two comments are in order. Firstly, the effect on overall citation inequality due to citation impact differences between the 500 LR universities (3.85%) is comparable to the effect due to citation impact differences between countries (5.4%). However, both of them are considerably smaller than the corresponding effect on overall citation inequality attributable to differences in production and citation practices across the 219 sub-fields (approximately 18%). Secondly, the reduction of the total effect generated by *MNCS* normalization in our dataset (81.9% of the total effect) is of a comparable order of magnitude to the same phenomenon in the context of country (85.2%) or sub-field citation distributions (83.2%).

It should be noted that these results summarize in a pair of scalars a complex phenomenon that takes place along the entire support of our university citation distributions. As a matter of fact, the term *IDCU* is simply a weighted average of the $I(\pi)$ terms, $\pi = 1,\ldots, 100$, that capture the effect on overall inequality of the citation impact differences between the 500 LR universities holding constant the degree of excellence in all universities at percentile π. Therefore, it is instructive to study how $I(\pi)$ changes with π both before and after the MNCS normalization. The results appear in Figure 1 (since $I(\pi)$ is very high for $\pi < 27$, for clarity these percentiles are omitted from Figure 1), which deserves the following two comments. Firstly, the strong impact of *MNCS* normalization is readily apparent. Secondly, it is useful to informally partition the support of our citation distributions into the following three intervals: [0, 57], [58, 96], and [98, 100]. In the first and the third one, $I(\pi)$ values are very high. This means that, since in these two intervals university citation distributions differ by more than a scale factor, the universality condition can hardly be satisfied in them. However, $I(\pi)$ is approximately constant for a wide range of intermediate values in the second interval. Thus, this is the range of values where the search for a single functional form in Chatterjee *et al.* (2014) may give good results in our dataset.

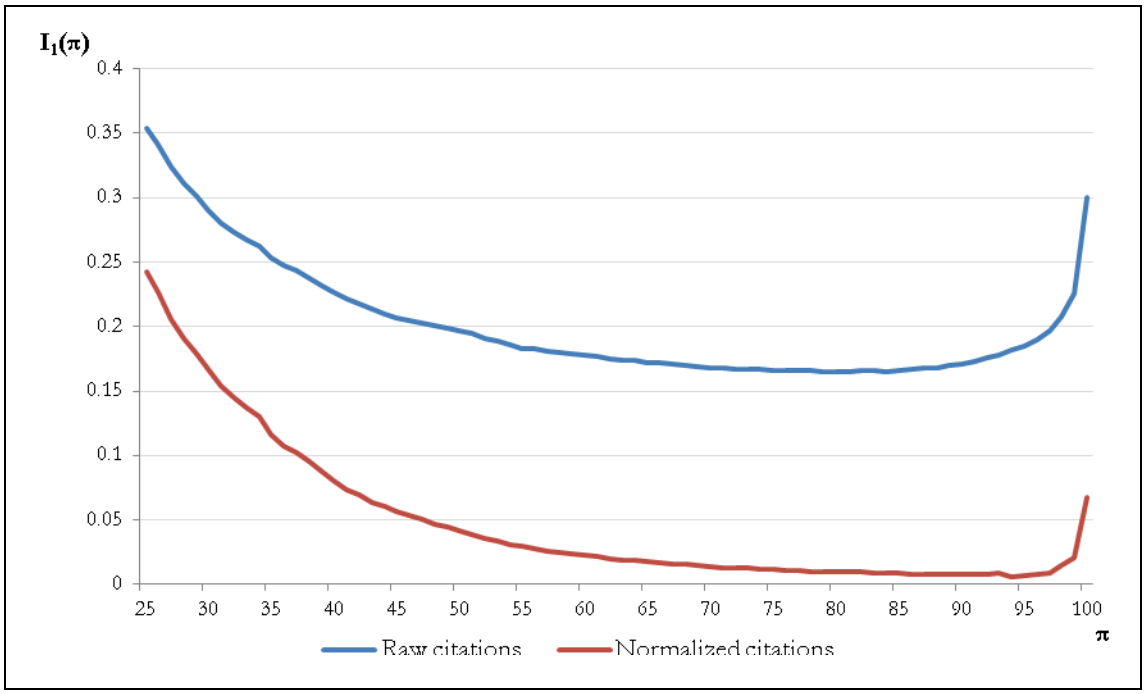


Figure 1. Citation Inequality Due to Differences in Citation Practices, I(π), as a function of π. Results for the [27, 100] quantile interval.

Implications of the results

Our results have two types of practical implications. In the first place, assume that the top,

intermediate, and worse universities have different types of citation distributions. In this case, we would need to build different models to explain the citation impact variability within the universities of the three types. On the contrary, since we have found that, although not universal, university citation distributions are rather similar, we need a single model to explain the high within-universities variability.

In the second place, recall that the move in the CWTS and SCImago rankings from an average-based citation impact indicator –such as the *MNCS*– towards a rank percentile approach that throws all the weight on the top *x*% of most cited papers –such as the $PP_{top\ 10\%}$ indicator– is surely due to the idea that, for highly skewed citation distributions, average-based indicators might not represent well the excellence in citation impact. However, the two rankings are rather similar: the Pearson correlation coefficient between university values is 0.981, while the Spearman correlation coefficient between ranks is 0.986. The situation is illustrated in Figure 2, where the positive slope indicates that to low (high) *MNCS* values there correspond lower (higher) $PP_{top\ 10\%}$ values.

We conclude that ordinal differences between the university rankings according to the MNCS and the $PP_{top\ 10\%}$ indicators are of a small order of magnitude. As a matter of fact, we find a strong, more or less linear relationship between the $PP_{top\ 10\%}$ and the MNCS in two other instances: for the 500 universities in the 2011/2012 edition of the Leiden Ranking (see Figure 2 in Waltman *et al.*, 2012b), and for the partition of the world into 39 countries and eight geographical areas studied in Albarrán and Ruiz-Castillo (2012). How can we explain these results? We have seen already that, university citation distributions behave as if they differ by a relatively constant scale factor over the [58, 96] percentile interval in their support. In this empirical scenario, it is not surprising that the MNCS values, which are reached at approximately the 63th percentile of citation distributions, and the $PP_{top\ 10\%}$ indicator that focus on the last 10 percentiles, provide very similar rankings. A convenient practical consequence is that the citation impact university ranking provided by the MNCS indicator is an adequate one. The $PP_{top\ 10\%}$ indicator would only add greater cardinal differences between the best and worse universities with relatively few re-rankings.

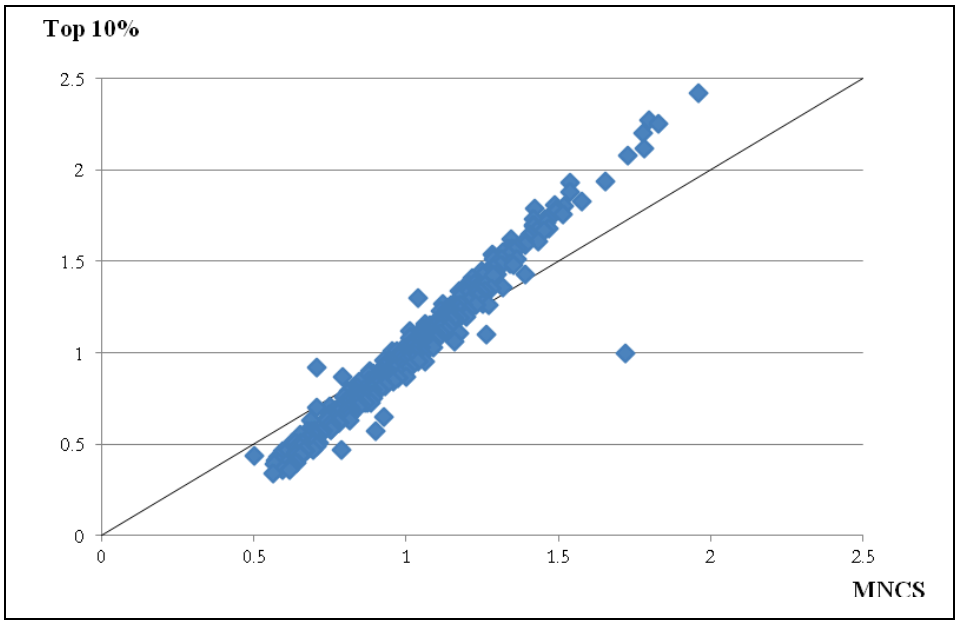


Figure 2. Scatterplot of the relation between the MNCS indicator and the PPtop 10% indicator for the 500 Leiden Ranking universities

It should be noted that further details concerning the following topics can be found in the Working Paper version of this paper, Perianes-Rodriguez & Ruiz-Castillo, 2014b): (i) the distribution of the total number of publications by universities; (2) the means μ_1 and μ_2, as well as the results of the CSSS approach for individual universities; (3) the graphical illustration of these results; (4) the measurement of the skewness of university citation distributions by means of a skewness index robust to extreme observations; (5) the robustness of all skewness results for the assignment of publications to universities in a multiplicative way, as well as the treatment of raw citation distributions without the benefit of any field-normalization procedure; (6) the re-rankings involved in the move from the MNCS towards the $PP_{top\ 10\%}$ indicator, as well as the cardinal differences between their values. In any case, the robustness of all of our results must be investigated with other datasets characterized by other publication years, and other citation windows, as well as other data sources different from the WoS.

Further research

Here are the possibilities for further research:

1. The effect on overall citation inequality attributable to the differences in citation impact between universities shows a characteristic pattern: broadly speaking, university citation distributions appear to behave as if they differ by a relatively constant scale factor over a large, intermediate part of their support. Consequently, it might be interesting to compute the exchange rates introduced in Crespo *et al.* (2013, 2014) to exploit this feature, and to use them as normalization factors. More generally, one could experiment with other normalization approaches that have been found useful in other contexts, notably the two parameter scheme introduced by Radicci & Castellano (2012).
2. Chatterjee *et al.*'s (2014) idea of fitting specific functional forms to university citation distributions in different intervals of their support is worth pursuing. The threshold determining the upper tail where a power law might be the best alternative could be estimated following the methods advocated in Clauset *et al.* (2009). Similar grid techniques could be applied to determine the lower bound of the interval where a lognormal might be the best alternative. In any case, standard methods should be used to test which specific functional form is best in each interval, as well as to estimate the parameters' confidence intervals (Thelwall & Wilson, 2014, and Brzezinski, 2015).
3. As we have seen in Section III.4, differences in citation impact between universities after MNCS normalization tend to rise when we reach the last few percentiles including the most highly cited articles. The question left for further research is how to complement average-based or $PP_{top\ 10\%}$ indicators with other measurement instruments that highlight the behavior of citation distributions over the last few percentiles. Given the important role of extreme observations in citation distributions, robustness of alternative high-impact indicators to these extreme situations will be an important element in the discussion.
4. Consider an array of citation distributions with a smaller number of scientific fields than in this paper in the columns, and the 500 LR universities in the rows. We already know much concerning field citation distributions and university citation distributions in the all-sciences case. A possible next step is to study the characteristics of university citation distributions column by column, that is, restricted to each field. The results will determine to what extent the similarities between citation distributions is a question depending on the aggregation level at which the study is conducted.

References

Albarrán, P., & Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology*, *62*, 40–49.

Albarrán, P. and Ruiz-Castillo, J. (2012). *The Measurement of Scientific Excellence Around the World*. Working Paper, Economic Series 12-08, Universidad Carlos III (http://hdl.handle.net/10016/13896).

Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011a). The measurement of low- and high-impact in citation distributions: technical results. *Journal of Informetrics*, 5, 48–63.

Albarrán, P., Crespo, J., Ortuño, I., & Ruiz-Castillo, J. (2011b). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88,* 385–397.

Albarrán, P., Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014). Differences in citation impact across countries. *Journal of the American Society for Information Science and Technology*. *66*, 512-525.

Brzezinski, M. (2015). Power laws in citation distributions: Evidence from Scopus. *Scientometrics*, *103*, 213-228.

Chatterjee, A., Ghosh, A., and Chakrabarty, B. K. (2014). *Universality of citation distributions for academic institutions and journals*, 29 September, arViv:1409.8029 [physics.soc-ph].

Clauset, A., C. R. Shalizi, and M. E. J. Newman (2009). Power-law Distributions In Empirical Data. *SIAM Review*, *51*, 661-703.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLoS ONE*, *8,* e58727.

Crespo, J. A., Herranz, N., Li, Y., & Ruiz-Castillo, J. (2014). The effect on citation inequality of differences in citation practices at the Web of Science subject category level. *Journal of the Association for Information Science and Technology*, *65*, 1244–1256.

Glänzel, W. (2007). Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, *1*, 92–102.

Herranz, N., & Ruiz-Castillo, J. (2012). Multiplicative and fractional strategies when journals are assigned to several sub-fields. *Journal of the American Society for Information Science and Technology*, *63*, 2195–2205.

Leydesdorff, L., Radicchi, F., Bornmann, L., Castellano, C., and de Nooye, W. (2012). Field-normalized Impact Factors: A Comparison of Rescaling versus Fractionally Counted Ifs. *Journal of the American Society for Information Science and Technology*, *27*, 292-306.

Li, Y., Castellano, C., Radicchi, F., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, *7*, 746–755.

Perianes-Rodriguez, A. & Ruiz-Castillo, J. (2015). Within- and between-department variability in individual productivity. The case of economics. *Scientometrics*, *102*: 1497-1520.

Perianes-Rodriguez, A. & Ruiz-Castillo, J. (2014). *University citation distributions*. Working Paper, Economic Series 14-26, Universidad Carlos III (http://hdl.handle.net/10016/19811).

Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*, e33833.

Radicchi, F., Fortunato, S., and Castellano, C. (2008), "Universality of Citation Distributions: Toward An Objective Measure of Scientific Impact", *PNAS*, 105: 17268-17272.

Ruiz-Castillo, J. (2014). The comparison of classification-system-based normalization procedures with source normalization alternatives in Waltman and Van Eck. *Journal of Informetrics*, *8*, 25–28.

Ruiz-Castillo, J. & Costas, R. (2014). The Skewness of Scientific Productivity. *Journal of Informetrics*, *8*, 917-934.

Ruiz-Castillo, J. & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*, 102-117.

Schubert, A., Glänzel, W., & Braun, T. (1987). A New Methodology for Ranking Scientific Institutions". *Scientometrics*, *12*, 267-292.

Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, *8*, 824-839.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*, 2378–2392.

Waltman, L., Van Eck, N. J., & Van Raan, A. F. J. (2012a). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Van Eck, N. J., Van Leeuwen, T. N., Van Raan, A. F. J., Visser, M. S., & Wouters, P. (2012b). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*, 2419–2432.

Exploration of the Bibliometric Coordinates for the Field of 'Geography'

Juan Gorraiz and Christian Gumpenberger

christian.gumpenberger, juan.gorraiz@univie.ac.at
University of Vienna, Library and Archive Services, Bibliometrics and Publication Strategies, Boltzmanngasse 5, A-1090 Vienna (Austria)

Abstract
This study is a bibliometric analysis of a highly complex research discipline, namely geography, in order to identify the most used and cited publication channels, to reveal publication strategies, and to analyse the discipline's coverage in the three main data sources for citation analyses: Web of Science, Scopus and Google Scholar. The results show very heterogeneous and individual publication strategies when considering the selection of adequate publication channels even in the same research fields. Monographs, journal articles (including proceedings papers) and book chapters are the most cited document types. Differences between research fields more related to the natural sciences than to the social sciences are clearly visible but not so considerable when taking into account the higher number of co-authors. General publication strategies are more established in the fields related to the natural sciences. Although an "iceberg citation model" is suggested, citation analyses for monographs, book chapters and reports (working papers) should be conducted separately and include complementary data sources, such as Google Scholar, in order to enhance the coverage and improve the quality of the citation analysis.

Conference Topics
Citation and co-citation analysis – Social Sciences

Introduction and background
From a bibliometric point of view, geography is a very challenging discipline, because it belongs to the natural sciences (geography, physical) as well as to the social sciences (geography), as it is clearly depicted in each edition of Journal Citation Reports (see Table 1).

Table 1. Category data of geography in both Editions of JCR (2013)

JCR EDITION 2013	*Category*	*Total Cites*	*Median IF*	*Aggregate IF*	*Aggregate Immediacy Index*	*Aggregate Cited Half-Life*	*# Journals*	*# Articles*
Sciences	GEOGRAPHY, PHYSICAL	159297	2.152	2.574	0.72	7.5	46	4972
Social Sciences	GEOGRAPHY	79207	1.059	1.612	0.343	7.4	76	3762

Table 1 shows very different citation characteristics according to the corresponding JCR edition. Furthermore, geography is a highly interdisciplinary field, very strongly related to geosciences, environmental sciences, ecology and remote sensing (natural sciences), or to economics, urban studies and political sciences (social science), as a quick search and refine analysis in WoS (Web of Sciences - core collection) illustrates.

Although there are many studies illustrating the differences between natural and social sciences and the different publication cultures depending on the discipline (e.g. Nederhof, 2006; Australian Research Council, 2012; Ossenblok et al., 2012; van Leeuwen, 2013; Moksony, 2014), no literature focusing on this specific could be retrieved by the authors.

The main research questions of this study are:

- What are the publication characteristics depending on the different research field?
- Can differences be observed concerning research fields? What is their time evolution?

- Which are the most used publication channels? Which document types are the most cited ones? Is it possible to identify publication strategies?
- What is the coverage in the three main citation data sources, Web of Science, Scopus and Google Scholar? Could Google Scholar be used as a complementary data source?

Data sources and methodology

This study is primarily based on publication data collected for three professorial appointments at the University of Vienna (Department for Geography): the first one, related to Geosciences and comprising of twelve candidates, and the second one, related to Social and Economic Geography and comprising of ten candidates, were performed during 2013. The third one, related to Demography and comprising of nine candidates, was performed in August 2014.

All the publication data were delivered directly by the applicants, whose identity has to remain anonymous. All bibliometric indicators added to the list of publications by the authors themselves, such as citation counts, impact factor or the h-index, were controlled or recalculated in order to enable a correct and comparable analysis (Gorraiz, J. & Gumpenberger, C., 2015). Document types used by the authors in their list of publications were manually reassigned to the following standard groups: Monographs (Books), Book chapters, Journal articles, Proceedings Papers, Conferences (including meeting abstracts and talks), Reports (Working Papers), Book Reviews, Edited Books and Journals Issues, and other publications (or Miscellaneous). A clear distinction between "Proceedings Papers" and "Conferences" was not always possible when relying on the lists of publications.

The main data source for coverage and citation analyses was Web of Science - Core Collection (WoS) including the Conference Proceedings and Book Citation Index. Since coverage in the usual multidisciplinary bibliographic and citation databases (Web of Science, Scopus) is very low and unsatisfactory for citation analyses, we have included Google Scholar (GS) as an additional data source in a first explorative attempt (Jacso, 2005; Kousha & Thelwall, 2007; Meho, & Yang, 2007; Gorraiz et al., 2013).

The analysis in GS was performed by using the Google Scholar Citation Profiles (applicants for the third appointment were invited to create their individual profiles and make them publicly available for a couple of weeks) as well as by applying the tool 'Publish or Perish' particularly designed for this purpose.

In spite of the fact that citations were checked and the percentage of self-citations was determined, citation analyses in GS should be taken with a pinch of salt. Google Scholar is not a database but a search engine, and therefore indexing remains non-transparent and documentation is lacking. That is why the analyses were also performed in Web of Science, including the Cited Reference Search (which means considering citations originating from Web of Science (WoS) 'core journals' to all document types without any restrictions), and in Scopus.

Publication windows were the last ten years (general for all authors, appointments no.1 and 2) and the career length of each applicant (for all appointments). In order to distinguish individual scientific career lengths, the year of the first publication activity is always included.

The observed citations window was identical for all applicants per professorial appointment procedure. It covers the date from publication until April - May 2013 for appointments no. 1 and 2, and until July - August 2014 for appointment procedure no.3.

Visibility analyses were performed according to the data in the Journal Citation Reports (JCR), Science Edition 2012 (appointments no. 1&2).

The quartiles (Q1= top 25%; Q2= top 25-50%; Q3= top 50-75% and Q4= top 75-100%) were calculated according to the 2-years impact factor (IF) in the corresponding WoS category.

Results

Comparison between appointments no.1 and no.2

Table 2 and 3 show the most important publication document types used by the candidates for both appointments. The spectrum is much more heterogeneous in the social sciences, where journal articles are not always the most common publication channel.

Table 2. Publication spectrum and WoS coverage according to provided publication list for appointment no.1 – Geosciences - 12 candidates. (In parenthesis, the number of document types indexed in WoS; PY=all years; *no distinction).

Candidate no.	1st Pub Year	Books	Edited Books/ Issues	Book Chapters	Proceedings & Conference Papers*	Book Reviews	Miscellaneous	Journal Articles (JA)
1	2004	1	0	5 (1)	14 (1)	0	3	28 (24)
2	2002	0	0	6 (1)	35 (3)	0	2	33 (30)
3	1996	13	7	12 (4)	26 (1)	0	0	38 (28)
4	1990	2	4 (2)	25 (6)	17	0	29	17 (11)
5	1998	4	2	1	6 (2)	0	65	75 (61)
6	1998	2	0	8 (2)	55 (2)	0	3	31 (21)
7	2007	4	0	1	41	0	1	35 (33)
8	1994	9	0	16	192	0	0	66 (53)
9	1999	0	0	7	13 (3)	0	5	28 (28)
10	2005	3	0	12	12(2)	10 (5)	10	18 (11)
11	2002	0	0	5 (1)	70	0	0	28 (18)
12	1994	1	0	2 (1)	8	0	1	51 (51)

Table 3. Publication spectrum and WoS coverage according to provided publication list for appointment no. 2 - Social & Economic Geography - 10 candidates. (In parenthesis, the number of document types indexed in WoS; PY=all years; *no distinction).

Candidate no.	1st Pub Year	Books	Edited Books/ Issues	Book Chapters	Proceedings & Conference Papers*	Book Reviews	Miscellaneous	Journal Articles (JA)
1	1999	3	2	8 (1)	2	8	50	72 (35)
2	2002	3	11	21	5 +*56	0	0	16 (8)
3	1991	7	0	19(1)	*87	0	13	37 (18)
4	1993	3	0	17 (2)	*67	19(9)	44	46 (24)
5	1994	7	2	16	2 + *34	0	9	31 (17)
6	2005	3	5	15	*42	0	5	15 (4)
7	1990	3	11	58	*4*	10	14	35 (22)
8	2005	1	1	5	**40*	0	9	20 (7)
9	2004	3 (1)	0	21 (7)	**10*	2	10	16 (11)
10	2000	3	1	17	*72	0	49	22 (11)

Miscellaneous were principally Reports and Working Papers in both appointments. Therefore this document type was considered separately in the second part of the study.

In appointment no. 2, other document types such as Films, Policy Briefs, Newspapers and Special Issues were mentioned but only individually. For two candidates (one in appointment no.1 and one in no.2), articles in other (non-scientific or non-peer-reviewed) journals were also assigned to the group Miscellaneous.

Concerning the coverage in WoS both tables corroborate the low coverage of books and book chapters in both editions of the Book Citation Index. For articles in peer-reviewed journals, the WoS coverage in appointment no.1 varies between 60 and 100% and the trend in the last 10 years was constantly increasing until it reached a quota of almost 90% for all candidates. In appointment no. 2, the coverage was lower, varying between about 30 and 60%, but a similar trend was also observed even if not as steep.
Tables 4 and 5 show the results of the visibility (publication strategies) and citation analyses performed for both appointments. Only publications indexed in WoS in the last ten complete years (2003-2012) were considered.

Table 4. Visibility (Q1 and %Q1) and citation analysis in WoS for appointment no. 1 – Geosciences - 12 candidates. (PY=2003 -2012, ARPP= Articles, Reviews & Proceedings Papers).

Candidate no.	*1st Pub Year*	*Publications*			*# Authors per Paper*	*Citations ARPP*			*h-Index*	*% Self-citations*	*Q1*	*% Q1*
		Total	*ARPP*	*per Y*		*Sum*	*per P*	*Max*				
1	2004	25	25	2.78	6.36	147	5.88	28	7	16.22%	16	69.57%
2	2002	28	28	2.80	4.93	181	6.46	36	7	24.31%	14	87.50%
3	1996	29	26	2.60	4.83	249	9.58	31	10	19.05%	14	53.85%
4	1990	11	7	0.70	2.73	29	4.14	21	3	12.50%	5	100.00%
5	1998	49	48	4.80	5.57	458	9.54	42	12	30.07%	34	72.34%
6	1998	18	18	1.80	3.72	180	10.00	44	7	7.78%	8	53.33%
7	2007	32	32	5.33	5.53	428	13.38	155	12	21.26%	20	62.50%
8	1994	31	29	2.90	5.06	598	21.36	110	15	7.18%	29	93.55%
9	1999	17	17	1.70	4.94	317	18.65	102	7	4.73%	6	42.86%
10	2005	16	11	1.38	2.94	40	3.64	24	3	10.00%	2	14.29%
11	2002	16	16	1.60	4.38	129	8.06	21	8	15.50%	9	60.00%
12	1994	36	26	2.60	4.69	294	11.31	44	12	17.06%	32	91.43%
	Mean	**25.67**	**23.583**	**2.582**	**4.64**	**254.2**	**10.166**	**54.8**	**8.583**	**15.47%**	**16**	**66.77%**

Table 5. Visibility (Q1 and %Q1) and citation analysis in WoS for appointment no. 2 - Social & Economic Geography - 10 candidates. (PY=2003-2012; ARPP= Articles, Reviews & Proceedings Papers).

Candidate no.	*1st Pub Year*	*Publications*			*# Authors per Paper*	*Citations ARPP*			*h-Index*	*% Self-citations*	*Q1*	*% Q1*
		Total	*ARPP*	*per Y*		*Sum*	*per P*	*Max*				
1	1999	22	15	1.50	1.14	122	8.13	53	6	11.02%	12	60.00%
2	2002	7	4	0.40	2.00	22	5.50	10	3	9.09%	0	0.00%
3	1991	12	9	0.90	1.75	352	39.11	94	7	3.13%	9	81.82%
4	1993	**23**	12	1.20	2.61	134	11.167	76	6	13.41%	7	31.82%
5	1994	13	9	0.90	2.23	76	8.44	34	4	3.13%	3	23.08%
6	**2005**	4	3	0.38	**1.00**	3	1.00	2	1	0.00%	0	0.00%
7	**1990**	18	13	1.30	2	36	2.77	11	3	24.32%	3	18.75%
8	**2005**	7	6	0.75	2.57	48	8.00	17	4	8.33%	1	14.29%
9	**2004**	17	14	1.56	1.82	259	18.50	149	5	8.33%	7	70.00%
10	2000	8	7	0.70	1.13	53	7.57	40	3	9.26%	1	12.50%
	Mean	**13.1**	**9.2**	**0.958**	**1.82**	**110.5**	**11.02**	**48.6**	**4.2**	**9.00%**	**4.3**	**31.22%**

These results corroborate the higher number of publications and citations in the discipline related to the natural sciences (about twice as many). But taking into account the number of co-authors and the percentage of self-citations, which is almost twice as high in the natural sciences, there is not really a considerable difference.

The visibility analysis (number of Q1- journal articles) shows that publishing in top journals with impact factor, result in a much higher visibility in the appointment related to natural sciences than in the one related to the social sciences.

Finally, tables 6 and 7 show that the citation differences, according to the aggregate impact factor of the main WoS category, are higher in appointment no.1 than in no.2.

Table 6. First and second research field according to WoS categories for appointment no. 1 - Geosciences – 12 candidates.

Candidate no.	First Research Field (2003-2012)		Second Research Field (2003-2012)
	WoS Category	IF aggregate 2012	WoS Category
1	Ecology	3.095	Environmental Sciences
2	Remote Sensing	1.845	Geosciences, Multidisciplinary
3	Water Resources	1.803	Geosciences, Multidisciplinary
4	Water Resources	1.803	Geosciences, Multidisciplinary
5	Soil Science	1.780	Geosciences, Multidisciplinary
6	Ecology	3.095	Forestry / Soil Science/ Environm. Sci.
7	Ecology	3.095	Forestry / Plant Sciences
8	Geosciences, Multidisciplinary	2.176	Geography, Physical
9	Geosciences, Multidisciplinary	2.176	Geography/ Water Resources
10	Geography, Physical	2.206	Geography / Remote Sensing
11	Water Resources	1.803	Soil Sciences /Environmental Sci.
12	Geochemistry & Geophysics	1.474	Oceanography/Geosciences, Multi.

Table 7. First and second research field according to WoS categories for appointment no. 2 - Social & Political Geography – 10 candidates.

Candidate no.	First Research Field (2003-2012)		Second Research Field (2003-2012)
	WoS Category	IF aggregate 2012	WoS Category
1	Geography	1.469	Industrial Relations & Labor
2	Geography	1.469	Environmental Sciences
3	Geography	1.469	Economics; Management
4	Geography	1.469	Environmental Studies; Economics
5	Geography	1.469	Economics
6	Geography	1.469	Geography, Physical
7	Geography	1.469	Urban Studies
8	Geography	1.469	Environmental Studies & Sciences
9	Economics	1.148	Geography; Planning & Development
10	Geography	1.469	Economics

Results obtained in appointment no. 3 (Demography &Population Geography)

Applicants were invited to create their individual Google Scholar Citations profiles and make them publicly available for a couple of weeks.

From the nine applicants:

- six created a GS Citation Profile

- two refused to create one
- one followed the invitation, but the profile was incomplete

The tool 'Publish or Perish', particularly designed for this purpose, was then used for collecting and checking the data.

First of all, two key aspects (Focus 1 and 2) of each candidate's publications were determined in GS (free keywords) and in Web of Science according to the assigned Subject Categories (WoS categories) in the database. The results are shown in Table 8.

Table 8. First and second research field in WoS categories and GS for appointment no. 3– 9 candidates.

Candidate no.	Google Scholar		Web of Science	
	Focus 1	**Focus 2**	**WoS Category 1**	**WoS Category 2**
1	Human Geography -	Area Studies - East Asia - Japan	Urban Studies	Area Studies
2	Human Geography	Population Geography	Geography	Geography
3	Migration Studies	Demographic Change	Geography	Geography, Physical
4	Migration	Urban Studies	Geography; Planning & Development	Urban Studies
5	Urbanization	Cross-border Mobility	Geography	Geography, Physical
6	Demography	Fertility	Demography	Geography
7	Demography	Population	Demography	Public, Environmental & Occupational Health
8	Population Geography	Migration and Labour Markets	Geography	Political Science
9	Resilience	Livelihood	Public, Environmental & Occupational Health	Geography, Physical

Table 9 represents the publication activity for each scientist according to the most relevant publication types. The data are based on the list of publications submitted by the candidates. In order to distinguish individual scientific career lengths, the year of the first publication activity has been included.

The results hint at very heterogeneous and individual publication strategies taking into account publication types. The three next sections contain coverage and citation analyses performed in the three considered data sources. Table 10 shows the percentage of coverage in Google Scholar for each publication type. Monographs (Books) and Edited Books or Issues are very well covered, probably due to the inclusion of Google Books (Kousha & Thelwall, 2009).

The coverage of Journal Articles is also much higher than in WoS or Scopus (see Table 11). Also of interest is the high coverage of Reports (Working Papers). Chapters in Books are not so well covered, but this is probably due to incidental incorrect citations.

Table 9. Publication spectrum (publication types) for appointment no. 3. (*no distinction).

Candidate no.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Total (excl. Conferences)	**58**	**73**	**36**	**121**	**73**	**80**	**75**	**60**	**42**
Monographs	5	1	4	4	3	3	3	2	1
Book Chapters	13	32	15	48	17	11	11	21	7
Journal Articles	20	20	5	21	17	44	28	27	20
Proceedings Papers*	2	0	2	1	8	0	8	0	0
Reports (Working Papers)	3	0	7	11	7	13	10	3	11
Book Reviews	8	0	2	8	2	0	0	0	0
Edited Books/Journals	5	20	1	11	5	6	3	3	2
Other Publications	*1*	*0*	*0*	*17*	*14*	*3*	*12*	*4*	*1*
Conferences*	64	94	33	94	4	38	109	90	34
1st Year Publication	**1998**	**1994**	**2000**	**1993**	**1999**	**1992**	**1999**	**1989**	**2000**

Table 10. Coverage (%) in Google Scholar for each publication type (Appointment no. 3) (*no distinction).

Candidate no.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Total (excl. Conferences)	**58**	**73**	**36**	**121**	**73**	**80**	**75**	**60**	**42**
GS Profile	Yes	Incom-plete	Yes	No	No	Yes	Yes	Yes	Yes
Total Pub (excl. Conf)	44.83%	52.05%	44.44%	57.02%	35.62%	72.50%	77.33%	68.33%	97.62%
Monographs	60.00%	100.00%	50.00%	75.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Book Chapters	16.67%	12.50%	40.00%	56.25%	35.29%	45.45%	90.91%	42.86%	100.00%
Journal Articles	85.00%	50.00%	60.00%	71.43%	41.18%	81.82%	82.14%	100.00%	100.00%
Proceedings Papers*			50.00%		25.00%		100.00%		
Reports	66.67%		28.57%	54.55%	28.57%	46.15%	60.00%	33.33%	90.91%
Book Reviews			50.00%	25.00%	100.00%				
Edited Books/Journals	20.00%	70.00%	100.00%	81.82%	80.00%	83.33%	100.00%	66.67%	100.00%
Other Publications				41.18%		33.33%	41.67%		100.00%
1st Year	**1998**	**1994**	**1998**	**1995**	**1999**	**1992**	**1999**	**1995**	**2002**

Table 11 shows the results of the coverage and citation analyses performed in WoS, including the Cited Reference Search, in Scopus and in Google Scholar. The higher coverage scores in WoS over those in Scopus are due to the inclusion of the Cited Reference Search. This enabled citations not only of journal articles and book indexed in WoS to be retrieved, but also of other books, reports and other document types cited by the core journals in WoS.

All sections include the same indicators for each data source: 1) number of indexed publications; 2) percentage of publications covered according to the provided publication list; 3) number of cited documents; 4) total number of citations; 5) number of citations per cited publication; 6) maximum number of citations attracted by a publication; 7) total h-index and 8) i-index (number of publications with more than 10 citations).

The percentage of self-citations was only calculated for GS, where the number of citations was of sufficient significance.

Table 11 confirms that the values of the main citation indicators (number of citations, citations per cited publication and h-index) are different in absolute values in GS, WoS and Scopus, but are comparable in terms of relative values. Spearman correlations performed for these indicators (number of citations, citations per cited publication and h-index) in the three data sources (WoS, Scopus and Google Scholar) were very strong (varying from 0.8 to 0.95).

A detailed coverage and citation analysis for the three most cited document types in Google Scholar, Monographs, Book Chapters and Journal Articles (see Table 12) is shown in Table 13.

Table 11. Coverage and citation analysis in the three data sources for each candidate (Appointment no. 3)

Candidate no		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Google Scholar	GS Profile available	Yes	Incom-plete	Yes	No	No	Yes	Yes	Yes	Yes
	Total Pub (excl. Conf)	26	38	22	74	26	60	60	55	44
	% covered in GS	44.83%	52.05%	44.44%	57.02%	35.62%	72.50%	77.33%	68.33%	97.62%
	# cited documents	20	15	16	60	14	53	43	33	23
	Total Citations	123	36	106	667	80	1026	699	320	142
	% Self-citations	5.69%	13.89%	15.09%	7.65%	7.50%	14.52%	16.45%	20.94%	21.13%
	Citations/Cited Pub	6.15	2.40	6.63	11.12	5.71	19.36	16.26	9.70	6.17
	Maximum Citations	20	6	49	86	16	144	165	128	14
	h-index	7	3	5	14	5	19	13	9	8
	i-index (more than 10 cit)	5	0	2	21	3	25	18	8	5
WoS + Cited Ref Search	Total Pub (excl. Conf)	13	11	7	31	10	47	35	15	26
	% covered in WoS + CRS	17.24%	8.22%	16.67%	22.31%	9.59%	53.75%	38.67%	13.33%	52.38%
	# cited documents	11	6	6	29	9	44	31	12	24
	Total Citations	30	6	16	86	17	435	102	39	60
	Citations/Cited Pub	2.73	1.00	2.67	2.97	1.89	9.89	3.29	3.25	2.50
	Maximum Citations	9	1	10	16	4	55	21	24	7
	h-index	4	1	2	6	3	12	5	2	4
	i-index (more than 10 cit)	0	0	1	2	0	14	2	1	0
Scopus	Total Pub (excl. Conf)	9	10	2	11	6	30	16	11	10
	% covered in Scopus	15.52%	13.70%	5.56%	9.09%	8.22%	36.25%	21.33%	18.33%	23.81%
	# cited documents	5	5	1	7	2	24	10	8	9
	Total Citations	22	6	2	35	3	384	58	50	27
	Citations/Cited Pub	4.40	1.20	2.00	5.00	1.50	16.00	5.80	6.25	3.00
	Maximum Citations	11	2	2	22	2	57	23	31	8
	h-index	2	1	1	2	1	11	4	4	3
	i-index (more than 10)	1	0	0	1	0	13	2	1	0
1st Year Publication		**1998**	**1994**	**2000**	**1993**	**1999**	**1992**	**1999**	**1989**	**2000**

Table 12. Summary of the three most cited publication types in Google Scholar (Appointment no. 3).

Document Type	% Coverage	% Cited	Citations/Cited P	Maximum Citations	% Self-citations
Book Chapters	48.74%	68.77%	6.21	86	23.04%
Journal Articles	74.62%	74.20%	10.06	144	11.22%
Monographs	87.22%	92.59%	21.17	165	9.76%

The results show that not always the same publication types are the most cited for each candidate. There are individual differences. A separate citation analysis of these publication types is then recommended for evaluation purposes.

Table 13. Detailed Citation analysis in Google Scholar for each candidate and the three most cited publication types (Appointment no. 3). (the three highest values for each document type are highlighted in different colours).

Candidate no.	Publication Types	Liste	Google Scholar										
			Publications							Citations			
		# P	1st year	# Total	# Not list	# Cited	% cited	% Coverage	# Total	Mean	# Max	# Self	% Self
1	**Monographs**	**5**	*1998*	**3**	**0**	**3**	**100.00%**	60.00%	19	6.33	7	2	10.53%
	Book chapters	**13**	*2001*	**3**	**1**	**3**	**100.00%**	15.38%	44	14.67	19	4	9.09%
	Journal articles	**20**	*1998*	**17**	**0**	**12**	**70.59%**	85.00%	58	4.83	20	1	1.72%
2	**Monographs**	**1**	*1994*	**2**	**1**	**2**	**100.00%**	100.00%	7	3.50	6		0.00%
	Book chapters	**32**	*1996*	**4**	**0**	**1**	**25.00%**	12.50%	2	2.00	2		0.00%
	Journal articles	**20**	*1998*	**10**	**0**	**9**	**90.00%**	50.00%	21	2.33	4	3	14.29%
3	**Monographs**	**4**	*2002*	**2**	**0**	**2**	**100.00%**	50.00%	55	27.50	49	2	3.64%
	Book chapters	**15**	*2003*	**6**	**0**	**4**	**66.67%**	40.00%	8	2.00	4	2	25.00%
	Journal articles	**5**	*2009*	**3**	**0**	**2**	**66.67%**	60.00%	10	5.00	6	0	0.00%
4	**Monographs**	**4**	*1996*	**3**	**0**	**2**	**66.67%**	75.00%	20	10.00	18	0	0.00%
	Book chapters	**48**	*1996*	**27**	**0**	**25**	**92.59%**	56.25%	313	12.52	86	20	6.39%
	Journal articles	**21**	*1996*	**15**	**0**	**14**	**93.33%**	71.43%	151	10.79	48	7	4.64%
5	**Monographs**	**3**	*1999*	**3**	**0**	**2**	**66.67%**	100.00%	25	12.50	16	4	16.00%
	Book chapters	**17**	*2001*	**6**	**0**	**4**	**66.67%**	35.29%	12	3.00	5	0	0.00%
	Journal articles	**17**	*2000*	**7**	**0**	**4**	**57.14%**	41.18%	25	6.25	12	1	4.00%
6	**Monographs**	**3**	*1992*	**3**	**0**	**3**	**100.00%**	100.00%	74	24.67	27	8	10.81%
	Book chapters	**11**	*1997*	**5**	**0**	**5**	**100.00%**	45.45%	11	2.20	4	5	45.45%
	Journal articles	**44**	*1996*	**36**	**0**	**34**	**94.44%**	81.82%	892	26.24	144	126	14.13%
7	**Monographs**	**3**	*2002*	**3**	**0**	**3**	**100.00%**	100.00%	249	83.00	165	10	4.02%
	Book chapters	**11**	*2005*	**11**	**1**	**8**	**72.73%**	90.91%	64	8.00	16	25	39.06%
	Journal articles	**28**	*1999*	**23**	**0**	**17**	**73.91%**	82.14%	278	16.35	66	68	24.46%
8	**Monographs**	**2**	*2003*	**2**	**0**	**2**	**100.00%**	100.00%	18	9.00	17	0	0.00%
	Book chapters	**21**	*1995*	**9**	**0**	**6**	**66.67%**	42.86%	36	6.00	15	10	27.78%
	Journal articles	**27**	*1999*	**27**	**0**	**18**	**66.67%**	100.00%	227	12.61	83	39	17.18%
9	**Monographs**	**1**	*2010*	**1**	**0**	**1**	**100.00%**	100.00%	14	14.00	14	6	42.86%
	Book chapters	**7**	*2005*	**7**	**0**	**2**	**28.57%**	100.00%	11	5.50	8	6	54.55%
	Journal articles	**20**	*2005*	**20**	**0**	**11**	**55.00%**	100.00%	68	6.18	13	14	20.59%

Conclusions and discussion

The main conclusions of this case study for the field geography can be summarized in the following points:

Differences between research fields more related to the natural sciences than to the social sciences are clearly visible. However, the higher productivity (number of publications per year) and citation counts, are relativized when also considering the higher number of co-authors and percentage of self-citations

- General publication strategies, especially these based on the impact factor, are still more evident in the fields related to the natural sciences
- The results hint at very heterogeneous and individual publication strategies considering the selection of adequate publication channels even in the same research fields
- Journal Articles and Book Chapters are the most used publication channels
- Monographs, Journal Articles (including Proceedings Papers) and Book Chapters are the most cited document types
- The coverage, especially books, is much higher in Google Scholar and suggests the recommendation of this data source as complementary one, although this data source is still a black box (no transparency, missing content information, etc.). In this study the accuracy of the citations in GS was very high (~95%). Nevertheless further

measures are needed to reduce the noise of Google Scholar data in order to increase the significance of this alternative data source for evaluative purposes.

- The values of the main citation indicators might differ in absolute values in GS, WoS and Scopus, but are comparable in terms of relative values.
- This fact suggests a "citation iceberg model" (see Figure 1). The citation analysis in WoS or Scopus shows only the 'visible part' but this is generally still related to and indicates the 'invisible part'.
- Therefore, citation analyses for monographs, book chapters and reports (working papers) should be conducted separately and require the inclusion of complementary data sources. Otherwise relevant publications can be easily missed, resulting in wrong interpretations.
- Peers still have to be aware of blind spots in 'citation analyses' (e.g. 'non cited' document types and publications) with potentially harmful consequences in evaluation exercises

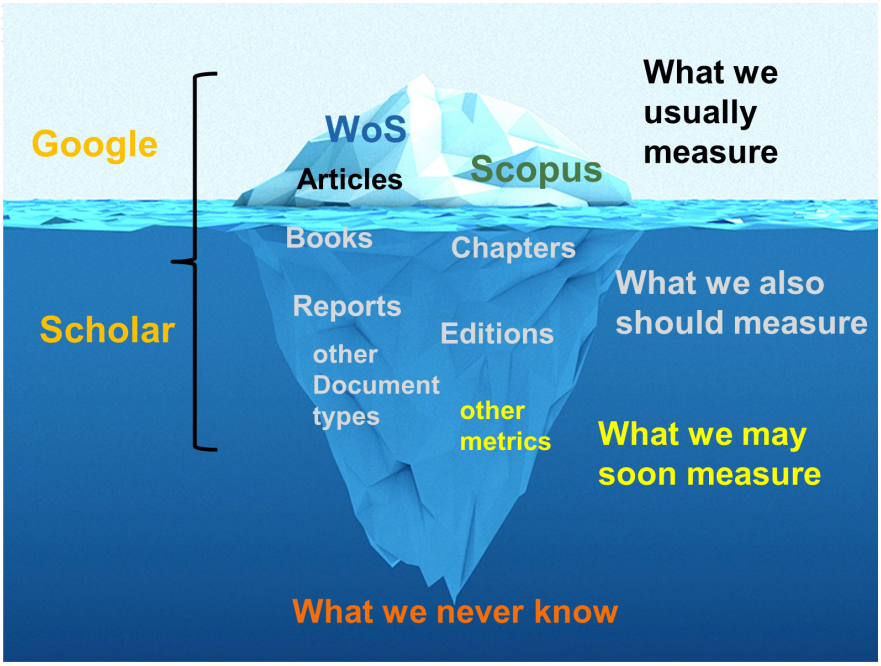


Figure 1. Citation "iceberg" model.

Finally, it should be stressed that citations can only used as a proxy for impact (and not for the quality) of publications produced in the 'publish or perish' community (i.e. the scientists who are committed to publishing their results). However, the scientific community is much broader and also comprises teaching academics as well as representatives from government or industry, who rather use than cite scientific output. Furthermore, our society has become progressively informed ('societal impact'). Unfortunately alternative metrics (like usage metrics and altmetrics) are still in their infancy (Kurtz M.J. & Bollen. J., 2010; Priem, J. et al., 2012; Gorraiz, J. et al., 2014; Hammarfelt, B., 2014) to measure the impact beyond citations and could not yet be applied to the described appointment procedures due to the current lack of available and reliable data.

Acknowledgments

We wish to thank Christian Buchmayer for his help collecting and disambiguating the data, and Prof. Thomas Glade for his support and suggestions.

References

Australian Research Council (2012). The Excellence in Research for Australia (ERA) Initiative, Retrieved January 5, 2015 from http://www.arc.gov.au/era/.

Gorraiz, J., Purnell, P.J. & Glänzel, W. (2013). Opportunities for and limitations of the book citation index. *Journal of the American Society for Information Science and Technology*, *64*(7), 1388-1398.

Gorraiz, J., Gumpenberger, C. & Schlögl, C. (2014). Usage Versus Citation Behaviours in Four Subject Areas. Scientometrics, *101*(2), 1077-1095.

Gorraiz, J. & Gumpenberger, C. (2015). A bibliometric model for the enhancement of professorial appointments, in press.

Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. *Scientometrics*, *101*(2), 1419-1430.

Jacso, P. (2005). Google Scholar: the pros and the cons. *Online Information Review*, *29*(2), 208-214.

Kousha, K., & Thelwall, M. (2009). Google book search: Citation analysis for social science and the humanities. *Journal of the American Society for Information Science and Technology*, *60*(8), 1537-1549.

Kousha, K. & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, *58*(7), 1055-1065.

Kurtz, M. J. & Bollen, J. (2010). Usage bibliometrics. *Annual Review of Information Science and Technology*, *44*(1), 1-64.

Meho, L. I. & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105-2125.

Moksony, F., Hegedus, R. & Csaszar, M. (2014). Rankings, research styles, and publication cultures: a study of American sociology departments. *Scientometrics*, *101*(3), 1715-1729.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*(1), 81-100.

Ossenblok, T.L.B., Engels, T.C.E. & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science - a comparison of publication patterns and incentive structures in Flanders and Norway (2005-9). *Research Evaluation*, *21*(4), 280-290.

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. Retrieved January 5, 2015 from arXiv preprint:1203.4745, http://arxiv.org/abs/1203.4745.

van Leeuwen, T.N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship? *Bibliometrie-Praxis & Forschung*, *2*, Retrieved January 5, 2015 from http://www.bibliometrie-pf.de/article/viewFile/173/218.

The Most-Cited Articles of the 21st Century

Elias Sanz-Casado[1,2], Carlos García-Zorita[1,2] and Ronald Rousseau[3,4]

[1,2] *elias.sanz@uc3m.es; czorita@bib.uc3m.es*
[1] Department of Library and Information Science. Laboratory of Metric Studies on Information (LEMI). Carlos III University of Madrid. C/Madrid 126, Getafe, 28903 Madrid, Spain

[2] *Research Institute for Higher Education and Science (INAECU).*
Carlos III University of Madrid-Autonomous University of Madrid. C/Madrid 126, Getafe, 28903 Madrid, Spain

[3,4] *ronald.rousseau@kuleuven.be*
[3] KU Leuven, Department of Mathematics, Celestijnenlaan 200B, B-3000 Leuven (Heverlee), Belgium
[4] Universiteit Antwerpen, IBW, Venusstraat 35, B-2000 Antwerpen, Belgium

Abstract
The aim of this paper is to collect the most-cited articles of the 21st century and to study how this group changed over time. Here the term "most-cited" is operationalized by considering yearly h-cores in the Web of Science. These h-cores are analysed in terms of authors, research areas, countries, institutions, journals and average number of authors per paper. We only consider publications of article or proceedings type. The research of some of the more prolific authors is on genetics and genomes publishing in multidisciplinary journals, such as *Nature* and *Science*, while the results show that writing a software tool for crystallography or molecular biology may help collecting large numbers of citations. English is the language of all articles in any h-core. The core institutions are largely those best placed in most rankings of world universities. Some attention is given on the relation between h-core articles and the information sciences. We conclude by stating that the notion of an h-core provides a new perspective on leading countries, articles and scientists.

Conference Topic
Citation and co-citation analysis

Introduction
The objective of this paper is to collect the most-cited articles of the 21st century and to study how this group changed over time. The term "most-cited" is operationalized by considering the h-core (Hirsch, 2005; Rousseau, 2006) in the Web of Science (WoS) for each period of time, starting with the period 2001-2005, continuing with 2001-2006 and ending with 2001-2013. These periods refer to the publication and the citation window. We recall that the h-core at a given moment in time, for instance on January 1, 2009, consists of the set of articles which at that time received a number of citations at least equal to their rank among all articles published during the period 2001-2008. This approach is different from the one taken in (Van Noorden et al., 2014) where a fixed number, concretely 100, of articles is considered. Furthermore, we study the papers making up the corresponding h-cores in terms of authors, research areas, countries, institutions, journals and average number of authors per paper.

Methodology
We have to point out that the 21st century starts on January 1, 2001. This implies that we only consider publications from 2001 on. Moreover, we only consider publications in Thomson Reuters' Web of Science (WoS) and we restrict ourselves to publications of article or proceedings type.
Although finding today's h-core for a set of articles in the Web of Science is easy, finding an h-core in the past needs some specific knowledge of the tools available in the WoS. First one retrieves the set for which one wants to determine the h-core (ending in the year Y). Its

articles are ranked from most cited to least cited. These are collected as a marked list. This is possible for at most 5,000 items. Clicking on Marked List shows this list and now, on this page, the system can provide a Citation Report, which is downloaded as an Excel file showing yearly citations for each of these records. Now we add the same data for the next 5,000 items (more was not necessary for our investigation). In this Excel file, we remove the columns corresponding to the year Y+1 and all later ones. In a next step we sum all remaining citations of each article. Sorting these sums from highest to lowest and comparing with a column of natural numbers leads to the h-index and the h-core. More details of this procedure are provided in (Rousseau & Zhang, 2014).

Results

The most-cited papers

The most-cited articles over the period 2001-2013 (the latest h-core) are shown in Table 1. It is clear that writing a software tool for crystallography or molecular biology may give one's paper a huge boost. The article by the National Cholesterol Education Program Expert Panel (2001) was the most-cited one from 2005 till 2008. From the year 2009 on Sheldrick's became the most-cited one.

Table 1. Most-cited articles over the period 2001-2013.

Rank	Article cited	Times cited
1	Sheldrick, G.M. (2008). A short history of SHELX. *Acta Crystallographica Section A*, 64, 112-122.	34,533
2	Livak, K.J. & Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. *Methods*, 25(4), 402-408.	24,796
3	Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24(8), 1596-1599.	17,049
4	Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Zhang, Y., Dubonos, S.V., Grigorieva, I.V. & Firsov, A.A. (2004). Electric field effect in atomically thin carbon films. *Science*, 306(5696), 666-669.	12,512
5	Ronquist, F. & Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19(12), 1572-1574.	11,185
6	National Cholesterol Education Program Expert Panel (Group author; includes 28 members). (2001). Executive summary of the Third Report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III). *JAMA-Journal of the American Medical Association*, 285(19), 2486-2497.	11,160
7	Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D – Biological Crystallography*, 60(special issue 1), 2126-2132.	10,392
8	Huelsenbeck, J.P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8), 754-755.	10,317
9	Spek, A.L. (2003). Single-crystal structure validation with the program PLATON. *Journal of Applied Crystallography*, 36, 7-13.	9,920
10	Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, 5(2), 150-163.	9,175

Time evolution of h-index and h-cores

The difference between the h-index and the number of items in the h-core is due to the possible existence of more than one document with the same number of citations as the h-index, as illustrated in Table 2. For the year 2005, for example, there were five articles with 359 citations.

Table 2. H-indices and h-cores for the periods 2001-2005 till 2001-2013.

End year	h-index	# articles in the h-core
2005	359	363
2006	441	442
2007	526	527
2008	614	616
2009	704	704
2010	800	800
2011	902	902
2012	1014	1014
2013	1122	1122

It is obvious that only a small percentage of articles included in the WoS belongs to the h-core of a specific period. In order to show the evolution of the ratio of the h-core with respect to all articles we put their values for the period 2001-2004 equal to 100. Figure 1 shows the total number of papers in each period and the number of papers in each h-core when this rescaling has been performed. Linear regression is almost perfect for the two lines: all publications (R^2= 0,9982) and h-core (R^2= 0,9967). For this reason we can forecast the 21st century h-index for, at least, the next years to come. This would lead to an h-core of 1195 documents in 2014 and 1290 in the year 2015.

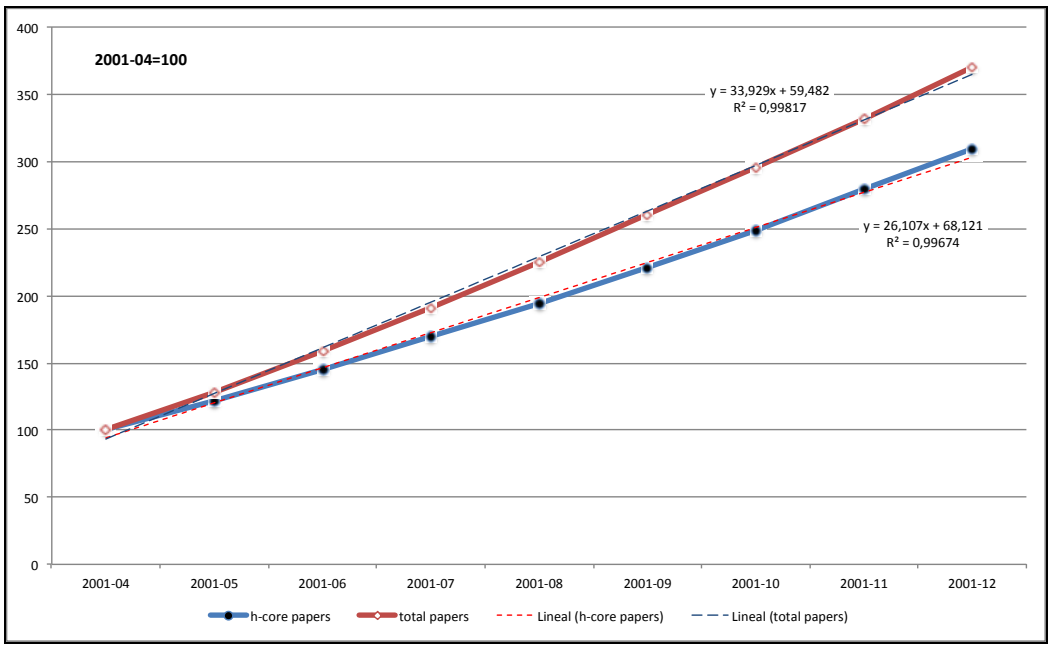


Figure 1. Evolution of the h-core.

In Table 3, we show the number of articles published in the years 2001 to 2011 included in each of the h-cores. For each h-core these numbers follow the order of publication, i.e. most

articles are published in the year 2001 and least in the latest year included in the core. Core13 has exactly the same number of articles published in 2001 as in 2002 (209 articles), while it does not contain articles published in 2013.

Table 3. Evolution of h-cores.

Year of Publication	Core-05	Core-06	Core-07	Core-08	Core-09	Core-10	Core-11	Core-12	Core-13
2001	196	210	218	217	217	213	213	209	209
2002	116	137	158	173	187	197	201	205	209
2003	43	72	96	120	138	151	159	163	169
2004	7	21	41	62	82	99	117	138	146
2005	1	2	11	31	49	74	93	110	121
2006			3	9	17	35	56	70	95
2007				3	10	23	36	47	58
2008				1	3	6	19	39	54
2009					1	2	6	21	32
2010							2	9	19
2011								3	8
2012									2
Total	363	442	527	616	704	800	902	1014	1122

Table 4 shows the number of articles in the h-core (on the diagonal) and on the last line the number of unique articles in the union of all h-cores until the year indicated on top of the column. The other numbers refer to the number of articles originally belonging to the core referred to on the left, but which do not anymore belong to the h-core. We note that there is one article that left the core (in 2007) but re-entered (in 2008) and from then on stayed in the core. This paper is:

> Minokoshi, Y., Kim, Y., Peroni, O., Fryer, L., Muller, C., Carling, D., & Kahn, B. (2002). Leptin stimulates fatty-acid oxidation by activating AMP-activated protein kinase. *NATURE*, *415* (6869), 339–343. doi:10.1038/415339a

Table 4. H-cores and h-core losses

	Core-05	Core-06	Core-07	Core-08	Core-09	Core-10	Core-11	Core-12	Core-13
Core-05	363	9	9	9	9	9	9	9	9
Core-06		442	13	12	12	12	12	12	12
Core-07			527	17	17	17	17	17	17
Core-08				616	15	15	15	15	15
Core-09					704	26	26	26	26
Core-10						800	27	27	27
Core-11							902	24	24
Core-12								1014	22
Core-13									1122
Total	363	451	549	654	757	879	1008	1144	1274

H-cores characteristics

All articles in any h-core are written in English. We note that the 2001-2005 h-core contains one article that was later retracted (Chang and Roth, published in *Science*, which has now 533 citations and had 359 citations by the end of 2005, being the last one in the 2005 core). Some of the more prolific authors (E.S. Lander, M.J. Daly, R.A. Gibbs, J. Wang) perform research on genetics and genomes publishing in multidisciplinary journals, such as *Nature* and *Science*, often in hyper co-authored papers (with dozens and even hundreds of authors). A. Jemal and E. Ward publish yearly statistics on cancer, which all enter the h-core. R. Collins and R. Peto work on internal medicine and publish almost exclusively in *Lancet*. The fields of

nanotechnology and grapheme research are represented by C.M. Lieber and Nobel Prize winners A.K. Geim and K.S. Novoselov (Table 5).

Table 5. Authors with highest number of papers in the h-core (Authors with more than 7 papers in the latest core).

Author	Core-05	Core-06	Core-07	Core-08	Core-09	Core-10	Core-11	Core-12	Core-13
Lander, ES	11	13	14	15	16	17	17	19	18
Wang, J	7	8	8	9	9	10	10	14	14
Jemal, A	4	4	5	6	7	8	9	10	12
Collins, R	5	6	7	8	9	11	11	11	10
Daly, MJ	4	5	6	6	7	10	10	12	10
Peto, R	4	5	7	8	8	9	9	9	10
Lieber, CM	5	6	7	7	7	8	8	9	10
Ward, E	3	3	4	5	6	7	8	9	10
Gibbs, RA	2	3	3	3	3	4	4	11	10
Geim, AK				3	3	5	6	8	10
Novoselov, KS				3	3	5	6	8	10
Thun, MJ	5	5	6	7	8	9	9	9	9
Altshuler, D	4	4	5	5	6	8	8	10	9
Abecasis, GR	2	2	2	2	2	4	4	9	9
Golub, TR	4	5	6	8	8	9	9	8	8
Murray, T	5	5	6	7	8	8	8	8	8
Gabriel, SB	3	3	3	3	4	5	5	9	8
Li, Y	1	2	3	3	4	7	7	8	8
Bartel, DP	1	1	1	3	3	4	4	7	8

The multidisciplinary areas (which include journals such as *Nature*, *Science* and *PNAS*), and the ones related to general and internal Medicine (such as *Lancet* or the *New England Journal of Medicine*) occur the most in each of the cores, as illustrated in Table 6.

Table 6. H-cores in different research areas (Areas with more than 10 papers in the last core).

Research area	Core-05	Core-06	Core-07	Core-08	Core-09	Core-10	Core-11	Core-12	Core-13
Science & Technology - Other Topics	39,1%	38,0%	35,3%	34,9%	32,8%	33,4%	32,7%	32,0%	31,9%
General & Internal Medicine	27,8%	26,2%	26,4%	25,0%	24,6%	23,1%	21,6%	20,4%	20,0%
Biochemistry & Molecular Biology	8,3%	9,0%	8,3%	9,7%	10,1%	10,6%	11,4%	12,8%	13,3%
Physics	5,5%	5,0%	4,9%	4,5%	5,0%	5,5%	6,4%	6,9%	7,0%
Chemistry	0,8%	1,4%	2,1%	1,9%	3,0%	3,9%	5,3%	6,0%	6,1%
Computer Science	2,5%	3,6%	4,7%	4,2%	4,5%	4,5%	5,1%	5,3%	5,5%
Cell Biology	4,1%	4,3%	4,0%	4,5%	4,5%	5,0%	5,2%	5,3%	5,1%
Engineering	1,4%	1,6%	3,0%	3,4%	3,6%	3,5%	3,8%	3,6%	3,9%
Biotechnology & Applied Microbiology	2,2%	3,4%	2,8%	3,1%	3,4%	3,1%	3,3%	3,8%	3,8%
Materials Science	0,6%	0,7%	0,8%	0,8%	1,7%	2,1%	3,0%	3,4%	3,8%
Oncology	2,8%	2,3%	2,3%	2,4%	2,7%	2,9%	2,5%	2,6%	2,9%
Genetics & Heredity	3,6%	3,4%	3,4%	3,2%	3,7%	3,4%	3,2%	3,3%	2,8%
Mathematics	0,8%	1,8%	1,7%	1,8%	1,7%	1,8%	2,0%	2,5%	2,7%
Mathematical & Computational Biology	0,8%	2,0%	1,7%	1,9%	1,8%	1,8%	1,8%	2,4%	2,4%
Research & Experimental Medicine	3,0%	3,2%	3,4%	3,2%	3,1%	2,9%	2,5%	2,5%	2,2%
Crystallography	0,8%	0,7%	0,9%	1,1%	1,3%	1,5%	1,6%	1,8%	2,0%
Neurosciences & Neurology	0,3%	0,7%	0,8%	0,8%	0,9%	1,4%	1,4%	1,9%	2,0%
Astronomy & Astrophysics	2,5%	2,9%	2,5%	2,3%	2,1%	2,1%	1,9%	1,9%	1,6%
Cardiovascular System & Cardiology	1,4%	1,8%	1,9%	1,8%	1,6%	1,4%	1,4%	1,5%	1,5%
Evolutionary Biology	0,0%	0,2%	0,2%	0,5%	0,7%	0,9%	1,2%	1,4%	1,5%
Immunology	2,8%	3,2%	3,2%	2,4%	2,7%	2,1%	1,8%	1,6%	1,3%
Biophysics	0,0%	0,0%	0,4%	0,5%	0,9%	1,0%	1,0%	1,1%	1,3%
Environmental Sciences & Ecology	0,3%	0,5%	0,4%	0,2%	0,4%	0,9%	0,9%	1,1%	1,3%
Radiology, Nuclear Medicine & Medical Imaging	0,0%	0,2%	0,4%	0,5%	0,6%	0,8%	1,0%	1,2%	1,2%
Endocrinology & Metabolism	1,4%	1,1%	1,1%	1,5%	1,4%	1,3%	1,1%	1,0%	1,1%

Table 7 shows a list of most used sources, where we observe, together with the mentioned multidisciplinary journals, the presence of medicine-related journals, including the specialized journal, *CA-A Cancer Journal for Clinicians*, whose presence is due to the systematic publication of the highly-cited annual statistics on cancer (all of them are in core 13). Other journal in the top positions, such as *Physical Review Letters* or *Nature Materials* occur less frequently.

Table 7. Journals of h-core publications (sources with 10 or more papers).

Source Titles	Core-05	Core-06	Core-07	Core-08	Core-09	Core-10	Core-11	Core-12	Core-13
NATURE	19,6%	17,4%	15,6%	15,9%	14,6%	14,9%	14,6%	14,4%	13,9%
SCIENCE	15,2%	16,1%	15,6%	15,1%	14,1%	14,0%	13,4%	12,9%	12,7%
NEW ENGLAND JOURNAL OF MEDICINE	16,5%	15,4%	15,2%	14,9%	14,8%	14,0%	13,1%	12,1%	11,9%
LANCET	5,2%	5,0%	5,1%	4,5%	4,4%	4,4%	4,1%	3,7%	3,6%
JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	5,5%	5,2%	5,1%	4,7%	4,4%	3,9%	3,5%	3,3%	3,1%
PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	3,6%	3,8%	3,6%	3,4%	3,3%	3,5%	3,4%	3,2%	3,1%
CELL	0,8%	0,7%	0,9%	1,5%	1,8%	2,4%	2,7%	2,9%	2,9%
NUCLEIC ACIDS RESEARCH	3,3%	2,9%	2,5%	2,8%	2,4%	2,1%	2,2%	2,5%	2,6%
BIOINFORMATICS	0,8%	1,6%	1,3%	1,1%	1,1%	1,1%	1,1%	1,5%	1,6%
PHYSICAL REVIEW LETTERS	3,6%	2,5%	2,3%	1,8%	1,6%	1,1%	1,4%	1,5%	1,4%
CA-A CANCER JOURNAL FOR CLINICIANS	1,4%	1,1%	1,3%	1,3%	1,3%	1,3%	1,2%	1,2%	1,2%
NATURE MATERIALS	0,0%	0,0%	0,0%	0,2%	0,6%	0,9%	1,2%	1,4%	1,2%
ACTA CRYSTALLOGRAPHICA SECTION D-BIOLOGICAL CRYSTALLOGRAPHY	0,0%	0,0%	0,4%	0,5%	0,7%	0,9%	0,9%	1,0%	1,2%
NATURE MEDICINE	1,7%	1,8%	1,5%	1,6%	1,6%	1,5%	1,4%	1,3%	1,2%
CIRCULATION	1,1%	1,4%	1,5%	1,3%	1,1%	1,0%	0,9%	1,0%	1,1%
IEEE TRANSACTIONS ON INFORMATION THEORY	0,8%	0,7%	1,3%	1,1%	1,1%	1,1%	1,0%	0,9%	1,0%
JOURNAL OF CLINICAL ONCOLOGY	0,8%	0,7%	0,6%	0,8%	0,9%	1,0%	0,8%	0,7%	0,9%
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY	0,0%	0,2%	0,6%	0,5%	0,6%	0,6%	0,9%	1,0%	0,9%
NATURE GENETICS	2,5%	2,0%	1,9%	1,6%	1,8%	1,6%	1,4%	1,3%	0,9%

We observe that the shares of the top journals such as *Nature*, *Science* and the *NEJM* are slowly declining over the years, while the share of *Cell* is increasing. This corresponds with recent findings (Lozano et al., 2012; Larivière et al., 2014; Acharya et al., 2014) that more and more highly-cited publications are published in journals that do not have the highest impact factors, say "non-elite journals". Of course, this is as such not surprising as the number of publications world-wide increases faster than the publication opportunities provided by so-called elite journals.

In Table 8 we show the distribution of countries in the h-cores, where an article is classified as belonging to a country if at least one author has an address in this country. The first place goes to the USA. If, however, we consider the European Union (EU-28) as one entity then it leads the rankings in all except one year. Our results correspond to those obtained by King (2004) for the percentage of documents published by USA in the 1% most cited papers. Our results are also similar to those found by Leydesdorff et al. (2014). In their work the EU-28 gains gradually in the top-10% segment at the expense of the USA, and one can expect a cross-over between the EU28 and the USA in the near future within the top-10% segment. However, the distance between the U.S. and the EU is much larger in the top-1% segment.

Also here we see that the top performers (USA, EU-28 and Germany) lose in the share of h-core articles. This observation also holds for the Netherlands and most Scandinavian countries. England and Scotland consolidate their share, while Brazil and New Zealand show an increase. Although China's share in publications shows an exponential growth (Jin & Rousseau, 2005; Zhou & Leydesdorff, 2006, 2008) its share in h-core papers is much lower and shows at best a small increase in the latest years, after a decrease in the period 2008-2009.

Core institutions are shown in Table 9. Leading institutions are those that one can find in most rankings of world universities, although The University of Texas (Austin) is only 39th in the latest ARWU ranking.

Table 8. Countries of publication (with 10 or more papers in the latest core).

Countries	Core-05	Core-06	Core-07	Core-08	Core-09	Core-10	Core-11	Core-12	Core-13
European Union	78,8%	76,9%	76,5%	76,8%	73,6%	75,3%	73,8%	75,8%	76,0%
USA	75,2%	75,1%	75,5%	74,8%	75,1%	74,5%	73,1%	72,0%	71,7%
England	18,2%	19,0%	17,5%	17,9%	17,3%	17,6%	17,1%	17,9%	17,8%
Germany	14,0%	13,6%	13,5%	12,5%	11,9%	12,0%	12,2%	12,2%	11,7%
France	8,5%	8,8%	9,1%	9,3%	8,9%	8,8%	8,2%	8,3%	8,5%
Canada	9,9%	9,0%	8,7%	8,6%	8,1%	8,6%	8,0%	8,4%	8,3%
Japan	7,4%	8,8%	8,3%	8,3%	7,7%	7,9%	7,3%	7,8%	7,7%
Italy	5,8%	5,7%	6,1%	5,8%	5,5%	6,4%	6,3%	6,4%	6,1%
Switzerland	5,5%	4,8%	5,1%	5,2%	4,8%	5,1%	5,2%	5,6%	6,0%
Netherlands	6,9%	6,3%	5,7%	5,7%	5,5%	5,5%	5,1%	5,7%	5,8%
Australia	5,0%	5,2%	5,1%	5,4%	5,3%	5,4%	5,5%	5,3%	5,7%
Sweden	5,2%	5,4%	5,1%	5,4%	5,3%	5,3%	5,4%	5,5%	5,3%
Spain	3,6%	3,4%	3,6%	3,4%	3,0%	3,1%	3,2%	3,5%	3,8%
Belgium	4,1%	3,8%	3,6%	4,1%	4,0%	4,0%	3,7%	3,7%	3,7%
Scotland	2,8%	2,7%	3,2%	3,4%	3,3%	3,5%	3,3%	3,3%	3,1%
Denmark	3,6%	3,2%	3,0%	3,1%	2,8%	3,1%	3,0%	2,7%	2,8%
Finland	3,3%	2,7%	2,8%	2,3%	2,1%	2,3%	2,3%	2,6%	2,6%
Peoples R China	2,2%	1,8%	1,9%	1,5%	1,4%	1,8%	1,8%	2,5%	2,4%
Austria	2,2%	1,8%	1,9%	1,9%	1,8%	2,0%	2,2%	2,1%	2,1%
Israel	1,4%	1,6%	1,9%	1,6%	1,6%	1,6%	1,6%	1,7%	1,6%
Norway	1,7%	1,6%	2,3%	2,1%	1,8%	1,9%	1,8%	1,6%	1,5%
Russia	1,4%	0,7%	0,9%	1,1%	1,1%	1,3%	1,4%	1,5%	1,5%
South Korea	1,1%	0,9%	0,8%	1,0%	0,9%	0,9%	1,1%	1,4%	1,5%
Poland	1,1%	0,9%	0,8%	1,1%	1,4%	1,3%	1,3%	1,3%	1,4%
Ireland	1,4%	1,4%	1,5%	1,3%	1,3%	1,6%	1,3%	1,3%	1,3%
Brazil	0,8%	0,7%	0,8%	1,0%	1,0%	1,0%	1,1%	1,1%	1,2%
New Zealand	0,3%	0,5%	0,6%	0,6%	0,9%	1,0%	1,0%	1,0%	1,2%
Taiwan	1,1%	0,9%	0,9%	1,0%	1,0%	0,9%	0,7%	0,7%	0,9%

Table 9. Core institutions restricted to those with 25 or more papers in the latest core.

Institution	Core-05	Core-06	Core-07	Core-08	Core-09	Core-10	Core-11	Core-12	Core-13
Harvard Univ	37	47	52	63	69	80	86	97	106
MIT	16	18	23	29	33	41	43	53	56
Univ Calif Berkeley	17	22	28	34	39	39	49	54	54
Univ Texas	11	16	20	25	30	35	39	41	45
Johns Hopkins Univ	12	17	19	26	29	34	33	40	43
Univ Washington	21	25	30	36	38	38	38	39	42
Univ Michigan	10	12	18	20	20	27	27	35	41
Univ Cambridge	11	13	16	20	22	26	29	34	39
Univ Oxford	15	14	16	18	19	24	27	34	39
Stanford Univ	15	21	24	24	26	26	33	37	38
Brigham & Womens Hosp	13	18	24	29	32	32	31	34	35
Univ Calif Los Angeles	13	19	19	20	21	24	26	28	35
Univ Calif San Diego	9	12	13	15	18	23	25	29	32
Columbia Univ	3	4	8	13	15	19	22	28	31
Massachusetts Gen Hosp	9	11	13	15	18	24	25	27	31
Univ Calif San Francisco	13	14	18	21	22	23	25	28	29
Univ Penn	13	13	14	15	17	19	19	25	26
Duke Univ	8	9	11	12	17	18	18	23	25
NCI	12	14	16	20	21	24	25	27	25
Univ Pittsburgh	7	9	11	16	16	18	19	22	25

In table 10 we have calculated average co-authorship values of articles in h-cores by research areas. For several research areas these values are higher than the co-authorship values of all publications: for example, in Clinical Medicine the co-authorship value for all publications was 4.5 authors per document and 5 in Bioscience and Biomedical Research (Bordons & Gómez 2000; Glänzel & Schubert, 2005). For several research areas these values are higher than the co-authorship values expected from previous research. For example, in Clinical Medicine the co-authorship value for all publications was 4.5 authors per document and 5 in Bioscience and Biomedical Research (Bordons & Gómez 2000; Glänzel & Schubert, 2005).

Table 10. Average numbers of authors for papers in the h-cores by research areas (areas with more than 10 papers in 2013).

Research Area	Core-05	Core-06	Core-07	Core-08	Core-09	Core-10	Core-11	Core-12	Core-13	Average
Science & Technology - Other Topics	15,5	16,1	14,6	13,9	14,7	14,5	14,5	17,0	15,9	15,3
General & Internal Medicine	19,8	20,4	23,4	25,6	24,2	25,9	22,7	22,1	22,1	23,1
Biochemistry & Molecular Biology	8,2	8,6	8,3	8,5	8,4	7,9	7,3	7,5	7,4	7,8
Physics	52,2	45,0	40,4	37,9	31,3	19,4	15,3	13,6	49,6	31,0
Chemistry	4,0	3,8	4,5	4,4	4,8	5,4	5,2	5,2	5,2	5,1
Computer Science	3,6	3,3	3,0	3,0	3,2	3,1	3,0	3,1	3,0	3,1
Cell Biology	11,4	11,8	11,7	10,9	10,8	10,7	10,2	11,1	11,1	10,9
Engineering	3,8	3,6	3,1	2,9	2,8	2,9	2,8	2,8	2,8	2,9
Biotechnology & Applied Microbiolog	6,8	5,9	7,0	7,4	7,4	6,5	6,0	5,6	5,4	6,2
Materials Science	4,5	3,3	6,5	5,6	5,0	5,2	5,6	5,8	6,3	5,7
Oncology	10,6	10,6	9,8	10,1	10,8	11,2	11,1	11,2	11,1	10,8
Genetics & Heredity	7,1	6,7	8,4	8,0	7,5	7,0	6,5	6,2	5,9	6,9
Mathematics	3,3	3,9	3,9	3,5	4,3	3,9	3,8	3,7	3,6	3,8
Mathematical & Computational Biolo	3,3	3,8	3,8	4,7	5,3	5,0	4,7	4,3	4,3	4,5
Research & Experimental Medicine	11,5	12,1	11,6	11,6	11,0	11,5	11,8	11,4	11,4	11,5
Crystallography	3,3	3,3	3,0	2,6	2,6	3,4	3,1	4,2	5,1	3,7
Neurosciences & Neurology	16,0	10,7	8,8	8,6	8,7	8,5	8,3	7,6	7,8	8,3
Astronomy & Astrophysics	41,8	30,7	30,7	37,3	35,9	37,5	38,8	46,5	45,8	38,8
Cardiovascular System & Cardiology	12,6	10,5	8,8	10,1	10,1	9,7	9,8	11,9	13,5	10,9
Evolutionary Biology		2,0	2,0	3,0	3,4	3,1	2,6	2,7	2,9	2,8
Immunology	8,1	7,3	7,2	7,4	7,5	7,7	7,6	7,6	7,8	7,6
Biophysics			2,5	2,3	3,3	4,0	3,8	5,1	5,9	4,5
Environmental Sciences & Ecology	7,0	4,0	4,0	7,0	5,0	3,1	2,9	2,6	2,8	3,2
Radiology, Nuclear Medicine & Medical Imagin		6,0	4,5	5,7	6,3	5,0	4,3	5,0	5,5	5,1
Endocrinology & Metabolism	7,2	7,2	6,8	8,4	6,9	6,9	6,9	6,6	5,9	6,9

Areas with an average of less than 5 authors (in 2013) are: computer science, engineering, mathematics, mathematical and computational biology, crystallography, evolutionary biology, biophysics and environmental sciences & ecology. Areas with an average larger than 15 are: science & technology – other topics, general & internal medicine, physics and astronomy & astrophysics.

The 21st century h-core (2001-2013) and the information sciences

Only one article classified by Thomson Reuters as *Information science and library science* belongs to this h-core, namely Venkatesh, V., Morris, M.G., Davis, G.B. et al. (2003). User acceptance of information technology: toward a unified view. *MIS Quarterly*, 27(3), 425-478 (cited 2261 times in total).

Yet, other ones were used and cited in *Information science and library science* articles. We list those that were cited at least 30 times by ILS researchers (on December 25, 2014).

1. Hirsch, J.E. (2005). An index to quantify an individual's research output. *Proceedings of the National Academy of Sciences of the USA*, 102(46), 16569-16572. Cited 682 times by ILS researchers.

2. Venkatesh, V., Morris, M.G., Davis, G.B. et al. (2003). User acceptance of information technology: toward a unified view. *MIS Quarterly*, 27(3), 425-478. Cited 595 times.
3. Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the USA*, 98(2), 404-409. Cited 118 times.
4. Blei, D.M., Ng, A.Y. & Jordan, M/I. (2003). Latent Dirichlet allocation. *Journal of Machine-Learning Research*, 3(4-5), 993-1022. Cited 93 times.
5. Zhara, S.A. & George, G. (2002). Absorptive capacity: a review, reconceptualization, and extension. *Academy of Management Review*, 27(2), 185-203. Cited 91 times.
6. Berners-Lee, T., Hendler, L. & Lassila, O. (2001). The semantic web. *Scientific American*, 284(5), 28-37. Cited 64 times.
7. Newman, M.E.J., Strogatz, S.H. & Watts, D.J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 62(2), article number 026118. Cited 60 times
8. Girvan, M. & Newman, M.E.J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the USA*, 99(12), 7821-7826. Cited 50 times.
9. Newmann. M.E.J. & Girvan, M. (2004). Finding and evaluating community structure in networls. *Physical Review E*, 69(2), article number 026113. Cited 36 times
10. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. Cited 30 times.

Besides Hirsch's famous article on the h-index (Hirsch, 2005), we see also Berners-Lee's article on the semantic web (Berners-Lee et al., 2001) and note the fact that Mark Newman occurs four times in this ILS h-core.

Conclusions

-Using the notion of an h-core provides a new perspective on leading countries, articles and scientists.

-The scientific contribution to the h-cores by the EU-28 is slightly higher than the USA's.

-The trend of annual h-cores since 2001 can predict future values of this indicator.

Of course, the view provided in this contribution is highly biased in favor of certain research areas such as General & Internal Medicine, or Biochemistry & Molecular Biology, and certain methodologies (writing heavily used software programs). Yet, it is a fact of life that these areas provide today's leading research. One should clearly realize that publishing highly cited research is different from realizing outstanding intellectual achievements.

References

Acharya, A., Verstak, A., Suzuki, H., Henderson, S., Iakhiaev, M., Chiung, C., Lin, Y. & Shetty, N. (2014). Rise of the rest: the growing impact of non-elite journals. Retrieved October 8, 2014 from arXiv:1410:2217v1.

Bordons, M., & Gómez, I. (2000). Collaboration networks in science. In: B. Cronin & H.B. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. Information Today Inc, Medford, NJ, 197–213.

Glänzel, W., & Schubert, A. (2005). Analysing scientific networks through co-Authorship. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 257–276). Dordrecht: Kluwer Academic Publishers.

Hirsch, J.E. (2005). An index to quantify an individual's research output. *Proceedings of the National Academy of Sciences of the USA*, *102*(46), 16569-16572.

Jin, BH. & Rousseau, R. (2005). China's quantitative expansion phase: exponential growth but low impact. In: P. Ingwersen & B. Larsen (Eds.) *Proceedings of ISSI 2005* (pp. 362-370). Stockholm: Karolinska University Press,.

King, D. A. (2004). The scientific impact of nations. *Nature*, *430*(6997), 311–316.

Larivière, V., Lozano, G.A., & Gingras, Y. (2014). Are elite journals declining? J*ournal of the Association for Information Science and Technology, 65*(4), 649-655.

Leydesdorff, L., Wagner, C.S. & Bornmann, L. (2014). The European Union, China, and the United States in the top-1% and top-10% layers of most-frequently cited publications: Competition and collaborations. *Journal of Informetrics*, *8*(3), 606-617.

Lozano, G.A., Larivière, V. & Gingras, Y. (2012). The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology*, *63*(11), 2140-2145.

Rousseau, R. & Zhang, L. (2014). How to determine the h-index of a set of publications in the WoS? *ISSI Newsletter*, *10*(3), 63-65.

Van Noorden, R., Maher, B. & Nuzzo, R. (2014). The Top-100 papers. *Nature*, *514*(7524), 550-553.

Rousseau, R. (2006). New developments related to the Hirsch index. *Science Focus 1*(4), 23-25 (in Chinese). An English version is available at: http://eprints.rclis.org/7616/

Zhou, P. & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. *Research Policy*, *35*(1), 83-104.

Zhou, P. & Leydesdorff, L. (2008). China ranks second in scientific publications since 2006. *ISSI Newsletter*, *4*(1), 7–9.

An International Comparison of the Citation Impact of Chinese Journals with Priority Funding

Ping Zhou[1*] and Loet Leydesdorff[2]

[1] *pingzhou@zju.edu.cn*
Department of Information Resources Management, School of Public Affairs, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058 (China)

[2] *loet@leydesdorff.net*
Amsterdam School of Communication Research (ASCoR), University of Amsterdam, PO Box 15793, 1001 NG Amsterdam (The Netherlands); http://www.leydesdorff.net

Abstract
We have investigated the citation impact of four pairs of journals in four subject categories including the category of multidisciplinary journals, journals in environmental sciences, applied mathematics, as well as metallurgy and metallurgical engineering. Each pair is composed of one Chinese journal and one leading international journal in the same subject category. Comparison is done between the selected Chinese and international journals in each pair. The four Chinese journals are selected because of priority funding by the Chinese CIU Plan in categories A and B. Compared with leading international journals in the same subject category, citation impacts of the four Chinese journals in their relevant environments are low, although they have been improving from 2004 to 2013. Leading international journals are more intensively and systematically cited than Chinese ones in the same subject category of the JCR. Regarding the CIU Plan, the level of funding seems not to follow exactly the citation impacts: Journals receiving larger amounts of funding do not necessarily perform better in citation impact, and journals receiving the same amount of subsidy may have different citation performances.

Keywords:
Citation and co-citation analysis

Introduction
Right after the United States, China has been the second largest producer of scientific publications since 2006 (Zhou & Leydesdorff, 2008; ISTIC, 2013). With citation impact rising continuously China jumped to the fifth position in 2013 in terms of national total citation impact from the eighth in 2010 (ISTIC, 2013), two years earlier in reaching the target set by the Ministry of Science and Technology (MOST) of China in the 12th National Plan for the Development of Science and Technology (NPDST). In terms of total citations received by disciplines, however, China's performance was not evenly distributed: chemistry, materials science, engineering technology, mathematics, computer science, and physics performed best by taking the second position in the world total (ISTIC, 2013).

In addition to being a second largest producer of academic papers, China is also the second largest publishing nation of academic journals. Of the 9,884 journals, approximately 5,300 are in science and technology (Liu, 2012; Yao et al., 2014). Nevertheless, international visibility of Chinese journals is still low (Jin & Rousseau, 2004; Leydesdorff & Jin, 2005; Zhou & Leydesdorff, 2007a, 2007b; ISTIC, 2014). In 2013, only 162 Chinese journals (i.e., about 3% of China's total S&T journals) were indexed in the Science Citation Index (SCI) of Thomson Reuters. Journals to be indexed in the SCI are required to satisfy basic criteria, and thus one can expect these 162 Chinese journals to be of relatively higher quality among the 5,300 Chinese S&T journals. Nevertheless, most of the SCI indexed Chinese journals do not perform well in terms of citation impact as measured by the Impact Factor. Take the data of 2011 for example, of the 114 Chinese journals indexed in the SCI, only four were in the first

quartile and 23 in the second of the corresponding subject categories of JCR 2011 (Liu, 2012).

The administrative structure of Chinese journals is special, and sometimes, confusing because of the involvement of both government agencies and the practical management by editorial boards. Administration at the national level is carried out by the General Administration of Press and Publication (GAPP) that is directly led by the State Council of China. At the provincial/regional level, the Administration of Press and Publication (APP) is responsible in each province or municipality. In addition to making regulations and policies relevant to journal publication and development, the GAPP is responsible for the approval of new journals and regular censorship; provincial APPs are responsible for administration and controls (including censorship) of local journals.

Practical management of Chinese academic journals is carried out by the editorial boards affiliated to research institutes, universities, and academic associations/societies. These institutions are affiliated to respective government agencies. Different governmental agencies are responsible for different sets of journals with different policies aiming at quality improvement with a special focus on international visibility. For example, at the national level are projects such as 'Journal Phalanx of China' of the GAPP, the 'Development Strategy Research for Competitive S&T Journals' of the Ministry of Science and Technology (MOST), and the 'Key Academic Specific Foundation' of the National Natural Science Foundation of China (NSFC). Years have passed since these projects were adopted, but the original targets of raising journal quality and international visibility have remained too far to reach.

In November 2013, in order to fasten the process towards international visibility of Chinese journals, six government agencies including the China Association for Science and Technology (CASST), the Ministry of Finance, the Ministry of Education (MOE), The State Press and Publication Administration (SPPA), the Chinese Academy of Sciences (CAS), and Chinese Academy of Engineering (CAE) jointly issued a unified standard of journal selection and funding: the International Impact Upgrading Plan for Chinese S&T Journals (abbreviated as CIU Plan). The CIU Plan is carried out in two steps. The objective of the first step is to raise the Journal Impact Factor (JIF) of a selected set of Chinese journals published in English to Quartile 1 and 2 of the Impact Factor in the Journal Citation Reports (JCR), by the end of the 12th Five-Year Plan (2011-2015), and to establish a journal set in the English language that can represent research frontiers or dominant fields of China, or in fields in which China does not yet have its own journals. The second step is to form a world top-journal set to which China has independent intellectual property rights by the year 2020.

Candidate journals must be in English and under the management of the above listed six government agencies. To ensure high-quality journals to be funded, the selection scheme combines bibliometric indicators, expert reviews, and a response by editorial boards. Journals being funded are classified into four categories, namely A, B, C and D. Those in categories A, B and C already have English version and are funded for three years. The funding amount in categories A, B, and C are respectively 2 million RMB or 322,092 US$, 1 million RMB (US$ 161,046), and 0.5 million RMB (US$ 85,230), respectively. Journals in category D are those that do not but will have an English edition; they receive 0.5 million RMB each. Of the nearly 5,300 scholarly journals in science and technology, only 76 are covered by the CIU Plan, among which 66 are in the categories of A, B, and C (Yao et al., 2014).

Journals receiving the largest funding are distributed among different Subject Categories and with different performances as measured by Journal Impact Factor (JIF) in the *Journal Citation Reports*. The rank of *Nano Research* is the highest whereas that of the *Journal of Zhejiang University-Science A* is the lowest. Questions arise such as: Are these journals selected because they outperform the rest of Chinese journals in the same subject category

based on the selection scheme mentioned above? How do they perform in comparison with their past, and their international counterparts?

Comparative studies between Chinese and international journals have been done before (Li, 2006; Zhou, et al., 2010; Jin & Leydesdorff, 2005; Zhou & Leydesdorff, 2007a, 2007b). Based on data of the *Journal Citation Reports* (*JCR*) of Thomson Reuters and the China Scientific and Technical Papers and Citations Database (CSTPCD) of the Institute of Scientific and Technological Information of China (ISTIC), Zhou and Leydesdorff (2007a, 2007b), for example, compared journal-journal citation relations from different perspectives, and found that international visibility of high-quality Chinese journals was low. These studies were based on data of ten or more years ago (i.e., *JCR* 2003 and 2004). The situation has changed given China's rapid development in science and technology and its increasing R&D investment during the last ten years (MOST, 2012; NBS, 2013). The CIU Plan further stimulated our interests in mapping an updated picture of the citation performance of Chinese journals in the international scholarly community. To highlight scholarly impact the current study mainly focuses on the citation impact environments of Chinese journals supported by the CIU Plan.

Methods and materials

We use routines developed by Leydesdorff & Cozzens (1992): aggregated journal-journal citation matrices of the environment of a seed journal can be harvested from *JCR* data. A seed journal is the one under investigation and acts as a starter to run the routines. Any journal indexed in the Science Citation Index (*SCI*) or Social Science Citation (*SSCI*) can be used as a seed. The relevant citation networks of the seed journal is determined by including all journals which cite or are cited by the seed journal to the extent of a contribution of (e.g.) 1% of its citation rate (He & Pao, 1986; Leydesdorff, 1986). By default the threshold is 1%, but this can be changed so as to include an appropriate number of journals in a local citation environment. For a network with too many journals, one may raise the threshold to reduce the size of the network, and vice versa.

Each journal in a network is represented by a node, which can be a circle or an ellipse in a Pajek map. The size of an ellipse is determined by the corresponding journal's contribution to the citing or citation impact environment in the year under investigation. The distinction of the vertical and horizontal size of the ellipse, informs the reader about the extent to which within-journal (self-) citations participate in the citation impact (Leydesdorff, 2007; Zhou & Leydesdorff, 2007). Note that within-journal citations can be author self-citations or citations among authors publishing in the same journal. Citation excluding journal self-citations can be considered as a measure of inter-journal communication.

In a citation impact environment, a journal's node size in the representation is determined by the logarithm of its contribution to the total number of citations in a local environment during the year under investigation. Citation counts are total of a journal during the current year; citation counts are combined for both the *SCI* and *SSCI*.

Many programs such as VOSviewer, Pajek, or Gephi can be used to visualize journal citation networks. In this study, we use Pajek because it serves the purpose of illustrating relative cited size of individual journals in local environments. Data of a citation impact environment can be imported into Pajek after being generated by the routines. The cosine between two vectors (Salton & McGill, 1983) is used to measure the similarity between the distributions for the various journals included in a citation environment (Leydesdorff, 2007). A visualized citation network showing strength of citation relations between journals in a local environment can thus be obtained.

Table 1. Journals to be investigated.

Journal Pair	*Journal Title*	*Country*	*Items in 2012*	*CIU Plan Category*	*JIF 2013*	*Rank in JIF*	*Quartile in Category*	*Category Name*
1	Chinese Science Bulletin	China	631	A	1.365	14/55	Q2	Multidisciplinary Sciences
	Science	USA	832		31.47	2/55	Q1	
2	Journal of Environmental Sciences-China	China	281	A	1.922	95/216	Q2	Environmental Sciences
	Environment International	USA	199		5.664	7/216	Q1	
3	Journal of Computational Mathematics	China	42	B	1.049	73/251	Q2	Mathematics, applied
	Foundations of Computational Mathematics	USA	23		2.152	13/251	Q1	
4	Acta Metallurgica Sinica	China	215	B	0.548	42/75	Q3	Metallurgy & Metallurgical Engineering
	Acta Materialia	USA	681		3.940	1/75	Q1	

In 2004, 71 Chinese journals were indexed in the *JCR*. Only a few journals satisfied the above three conditions; four journals were selected for the current study. For horizontal comparison, both Chinese and foreign journals must be in the same subject category of the *JCR*. Furthermore, the foreign journals do not have to be ranked first in the corresponding subject categories, but they should be in the first Quartile of Impact Factors and in the same subject category of the JCR as the selected Chinese journals. Table 1 lists journals satisfying the above conditions and will be used to study.

Results

Cited patterns of the selected journals will be investigated. The threshold is set at 1%, which means in a seed journal's citation environment, only journals contributing to 1% or more of the seed journal's total citations will be included. Due to the page limit of the ISSI 2015, only the results of the first two pairs of journals listed in Table 1 will be presented in detail. Conclusions and discussion, however, are based on the results of the four pairs of journals.

Chinese Science Bulletin versus Science

Chinese Science Bulletin. Only 10 journals contributed at least 1% of the total citation counts of *Chinese Science Bulletin* (*CSB*) in 2004, and these journals were all from China. In other words, visibility of *CSB* among foreign journals that were indexed in the *SCI*/*SSCI* was very low. As a multidisciplinary journal, citation impact of *CSB* was multidisciplinary with specific impacts in the geosciences, geology, and chemistry (Fig. 1a). In the citation impact environment of *CSB*, citation to *CSB* was highest even if within-journal citations were excluded. Within-journal citations of some Chinese journals took high proportions in their total citations, among which journals like *Acta Physica Sinca* and *Advances in Atmospheric*

Sciences were most obvious. In terms of Impact Factor, however, *Acta Geologica Sinica-English Edition* (2.150), *Science in China Series D – Earth Sciences* (0.909), *Acta Chimica Sinica* (0.895), and *Acta Petrologica Sinica* (0.805) performed relatively better than *CSB* (0.683) (Fig. 1a).

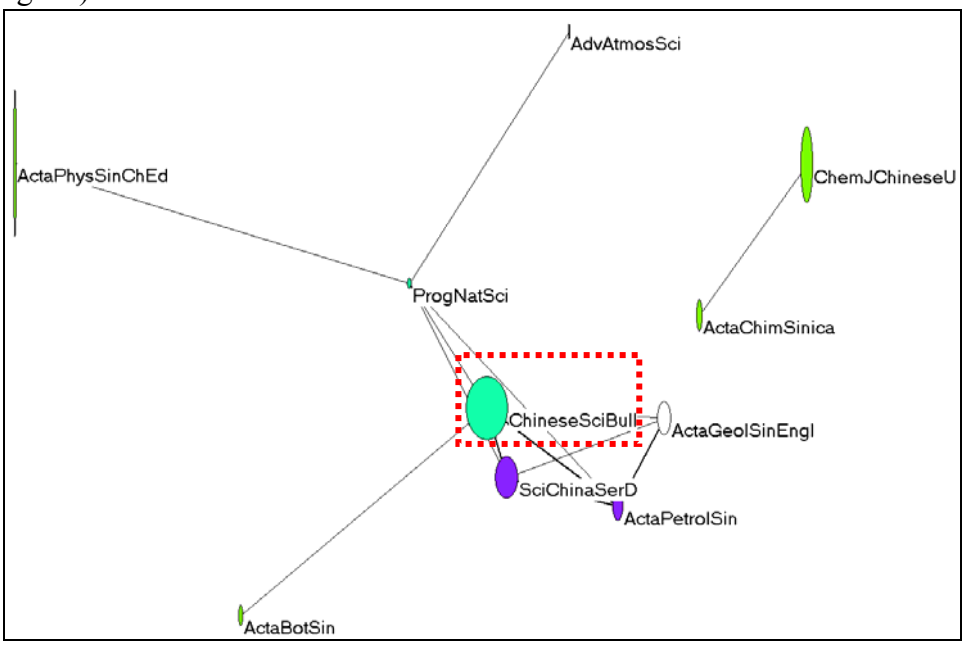


Figure 1a. Citation impact environment of *Chinese Science Bulletin* in 2004 (threshold = 1%, cosine ≥ 0.2).

Citation impact of *CSB* was enlarged to 13 journals in 2013 in terms of number of journals contributing at least 1% to the total citations of *CSB*. Most importantly, of these 13 journals eight were from other countries, which is a significant progress for Chinese journals in terms of citation impact on foreign journals compared to the year 2004. Within-journal citations contributed the most to the total citations of *CSB*. Citation impact of *CSB* on disciplines was similar to that in 2004 – involving multidisciplinary areas, geosciences, geology, and chemistry (Fig. 1b).

Impact Factor value of *CSB* were increased from 0.683 in 2004 to 1.365 in 2013.With the addition of foreign journals in the citation impact environment of *CSB*, journals with the highest citation impact is no longer *CSB* itself as in the year 2004; but instead, foreign journals such as the *Journal of Geophysical Research*, *Lithos*, and *Precambrian Research*, take the lead. In other words, in the citation impact environment of the Chinese journal *CSB*, citation impact of foreign journals was higher than that of Chinese journals. In terms of within-journal citations, *Journal of Geophysical Research* and *PLoS ONE* are most pronouncedly present. The heavy within-journal citations made the node of *PLoS ONE* a vertical line - citations from other journals in this environment were almost negligible (Fig. 1b).

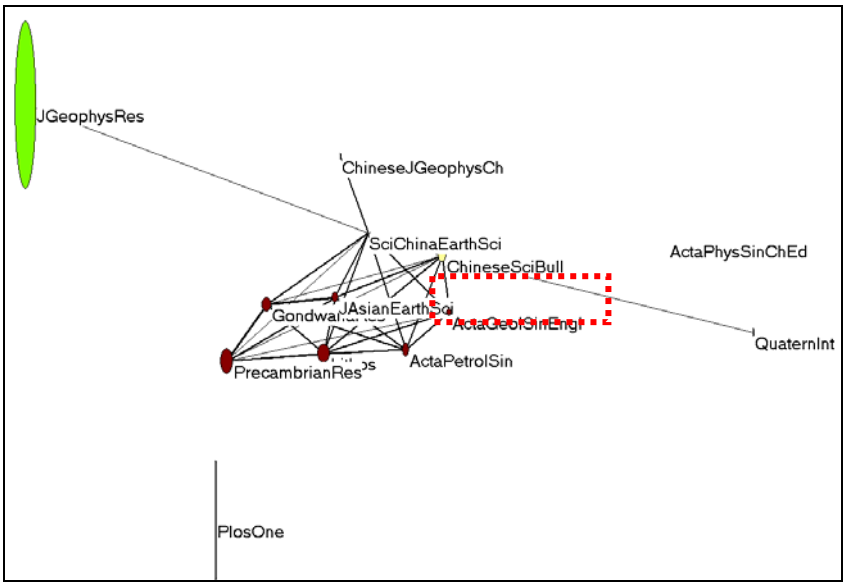


Figure 1b. Citation impact environment of *Chinese Science Bulletin* in 2013 (threshold = 1%, cosine ≥ 0.2).

Science. The citation impact network of *Science* was very much focused in 2004: Three journals contributed mostly to the citations of *Science*, and none of these three was from China. Except within journal citations of *Science*, the other two top contributors were *Journal of Biological Chemistry* (*JBC*) and *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS*) (Fig. 2a). Unlike the multidisciplinary journal *Chinese Science Bulletin* with distinct impact on geosciences and geology, citation impact of *Science* was more in biochemistry, in addition to impact in multiple disciplines. In terms of citation impact in the citation environment of *Science*, all the three journals are high with *JBC* having the highest impact. When within-journal citations are excluded, however, *PNAS* performed the best, and *Science* came next. In other words, compared with *JBC, PNAS* and *Science* had higher visibility in other journals. The distinct performance of citation impact of *JBC* and *PNAS* might largely be attributed to their high volumes of publications. In 2003, publications of *JBC*, *PNAS*, and *Science* were 6,585, 3084, and 845, respectively. In terms of average citation impact measured by the Impact Factor, however, *Science* performed the best (IF = 31.85), and followed by *PNAS* (IF = 10.452) and *JBC* (IF = 6.355).

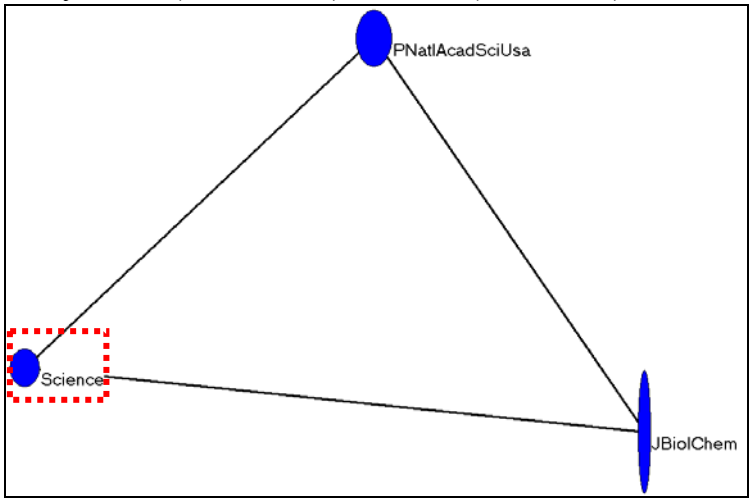


Figure 2a. Citation impact environment of *Science* in 2004 (threshold = 1%, cosine ≥ 0.2).

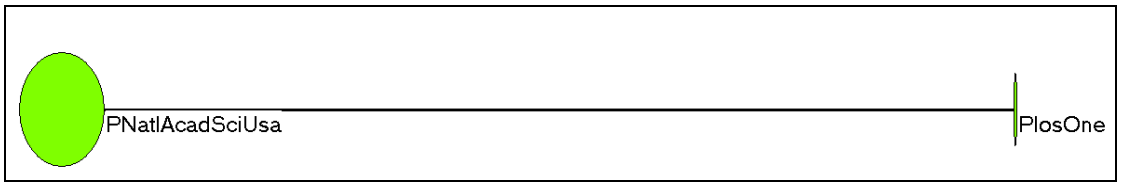


Figure 2b. Citation impact environment of *Science* in 2013 (threshold = 1%, cosine ≥ 0.2).

In the citation environment of *Science* in 2013, the percentage of within-journal citations of *Science* declined to less than 1% of its total citations. As a result, *Science* did not appear in its citation impact environment. In other words, the citation impact of *Science* was even more concentrated than in 2004. Impact Factor value of *Science* had increased from 31.853 in 2004 to 34.463 in 2013. Again, no Chinese journals appeared in this environment. *Science* was mostly cited by two multidisciplinary journals – *PNAS* and *PLoS ONE*, implying the multidisciplinary citation impact of *Science* with no distinct field emphasis like the situation in 2004. The high total citation impact of *PNAS* and *PLoS ONE* can be partially attributed to their high volume of publications: In 2013 *PLoS ONE* published 31,496 papers, which was eight times of that of the *PNAS* (3,901) and 37 times of that of *Science* (841). In terms of average citation impact (i.e., JIF), however, *Science* performed the best (31.477), and *PNAS* (9.809) came next. Average citation impact of *PLoS ONE* was the lowest (3.534), and furthermore, with heavy within-journal citations (Fig. 2b).

In summary, *Science* is widely cited in many journals in a range of different disciplines. When the threshold is set at 1%, however, only two or three journals are left in the citation impact environment of *Science*. In other words, these journals cited *Science* more intensively than other journals.

Journal of Environmental Sciences-China versus Environment International

Journal of Environmental Sciences-China. By 2004, the *Journal of Environmental Sciences-China* (*JES*) only received in total 193 citations of which 27 within-journal citations contributed the most; the other citations were scattered among journals in the environmental sciences, geosciences, chemistry, and biosciences. Although journals contributing 1% or more to *JES*'s total citation were mostly foreign and were as many as 26, these journals cited *JES* for only two or three times. In other words, except within-journal citations, there were no other journals citing *JES* systematically. Impact Factors of journals citing the *JES* were also low, between the highest of *Applied Catalysis B- Environmental* (4.042) citing *JES* six times in total and the lowest (0.172) of *Journal of the Chemical Society of Pakistan* citing *JES* four times (Fig. 3a). In other words, the *JES* had very low impact on other journals, citation impact in terms of Impact Factors of those citing *JES* occasionally was also very low.

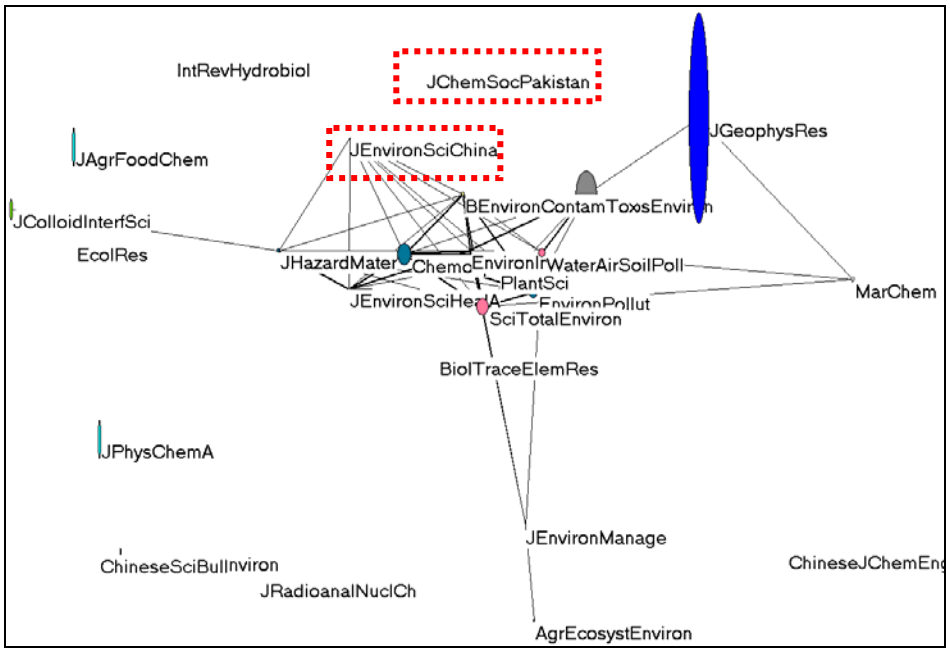


Figure 3a. Citation impact environment of *Journal of Environmental Sciences-China* in 2004 (threshold = 1%, cosine ≥ 0.2).

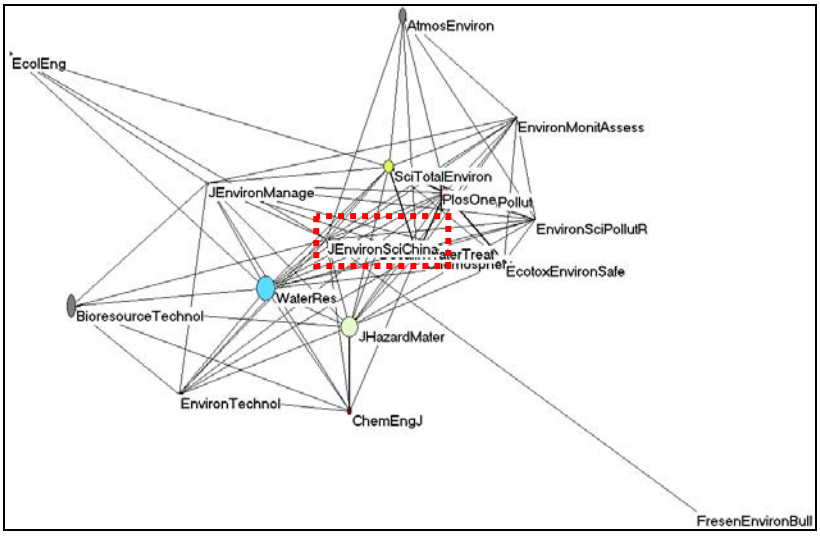


Figure 3b. Citation impact environment of *Journal of Environmental Sciences-China* in 2013 (threshold = 1%, cosine ≥ 0.2).

Performance of *JES* had been improved significantly in 2013, in addition to a large increase of the Impact Factor value from 0.254 in 2004 to 1.922 in 2013. Compared with the citation impact environment in 2004, the number of journals citing *JES* was less (i.e., 18 journals) but each contributed more citations. Journals citing *JES* were mostly foreign, although within-journal citations were still the first contributor. Instead of being cited occasionally like it was ten years ago, *JES* received more focused citation from other journals, and citation impact was more focused instead of scattering among different disciplines. For example, the foreign journal *Environmental Science and Pollution Research* contributed 28% of *JES*'s total citation by 2013, but did not appear in the citation environment of *JES* in 2004. Furthermore, journals citing *JES* had higher citation impact than those in 2004 ranging from 0.527 to 5.323. Citation relations among journals in the citation impact environment of *JES* formed closer relationship and thus interlinked with one another (Fig. 3b).

Environment International. In 2004, the citation impact of the *Environment International* was concentrated on environmental science. The journal contributing most to the total citations of *Environment International* was *Environmental Science & Technology*. Within-journal citations played much less a role than that of the *Journal of Environmental Sciences-China*. Impact Factors of journals citing the *Environment International* were much higher than those of the *Environment International*. For example, the Impact Factor of *Environmental Science & Technology*, the largest citation contributor to the citation impact of *Environment International*, was 3.557, which was even higher than that of *Environment International* (2.335). In other words, the *Environment International* had significant citation impact on high-quality journals. In the citation environment of *Environment International*, its citation impact was negligible whereas that of *Environmental Science & Technology* was highest (Fig. 4a).

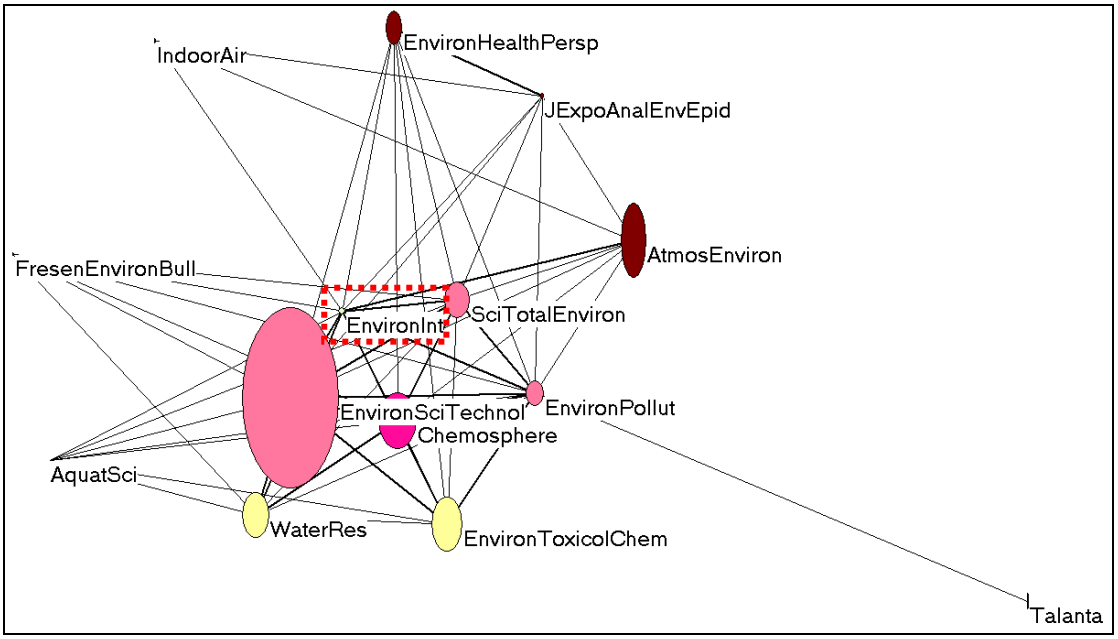


Figure 4a. Citation impact environment of *Environment International* in 2004 (threshold = 1%, cosine ≥ 0.2).

From 2004 to 2013, the Impact Factor value of *Environment International* increased from 2.335 to 5.664. Citation impact on number of journals extended from 14 to 18. Journals citing *Environment International* most frequently were *Chemosphere* (IF = 3.499) and *Science of the Total Environment* (IF = 3.163). Impact Factors of journals contributing at least 1% to the citation of *Environment International* were ranging from 1.679 to 5.664. In the citation impact environment of *Environment International*, the citation impact of *Environment International* itself became visible whereas that of *Environmental Science & Technology* was still the highest (Fig. 4b).

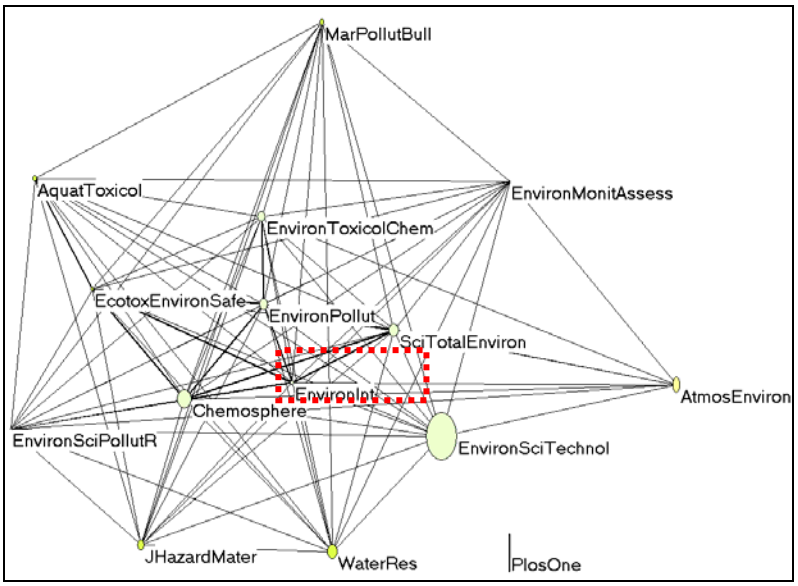


Figure 4b. Citation impact environment of *Environment International* in 2013 (threshold = 1%, cosine ≥ 0.2).

Conclusions and discussion

We have carried out a comparative study on journal citation impact between four pairs of journals in multiple disciplines, environmental sciences, applied mathematics, as well as metallurgy and metallurgical engineering. The four Chinese journals are selected because of additional funding by the Chinese CIU Plan in categories A and B. In Category A are *Chinese Science Bulletin* (*CSB*) and *Journal of Environmental Sciences-China* (*JES*), and in Category B are *Journal of Computational Mathematics* (*JCM*) and *Acta Metallurgica Sinica* (*AMS*). Leading foreign journals were used as matched pairs with the four Chinese journals. These are *Science*, *Environment International*, *Foundations of Computational Mathematics*, and *Acta Materialia* respectively.

International visibility of *CSB* was very low in 2004 although being indexed in the SCI and with a citation impact only on Chinese journals. The situation has been improved ten years later in 2013. More foreign journals cited *CSB*, but this may be by Chinese authors. Citation impact measured by Impact Factor of *CSB* has also been increased, but is still a long distance away from the best. Compared with *CSB*, *Science* has citation impact on higher quality journals measured by Impact Factor, and was cited more intensively with just two or three multidisciplinary journals contributing most to the citation counts of *Science*. By the year 2013, most citations to *Science* were from two multidisciplinary journals - *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS*) and *PLoS ONE*.

Within-journal citations were the first contributor of *CSB*, whereas this is not the case for *Science*. As a multidisciplinary journal, *CSB* did not appear in the citation impact environment of *Science*, implying a weak contribution of references in *CSB* to *Science*. On the other hand, the absence of *Science* in the citation environment of *CSB* implies that *CSB* has a long way to go before coming into the sight of authors publishing in *Science*.

Although being cited by foreign journals in 2004, citations received by the *Journal of Environmental Sciences-China* (*JES*) remained occasional. The situation has improved ten years later in 2013. Citation impact of *JES* has been increased significantly, but is still far behind that of the leading foreign journals in the same subject category. Compared with the *JES*, *Environment International* has citation impact on journals with higher quality measured by Impact Factor. The citation impact of the *Environment International* was more focused: Fewer journals contributing to 1% of the total citations of *Environment International* but each

journal contributed more; within-journal citations of *Environment International* were less significant to total citation counts than that of the *JES*.

Similar to the *Journal of Environmental Sciences-China*, the citation impact of the *Journal of Computational Mathematics* (*JCM*) was very low and was distributed among many journals in 2004. The situation was improved in 2013 with citation impact of the *JCM* being increased significantly, but still far behind that of leading foreign journals in the same subject category. The starting point of *Foundations of Computational Mathematics* was not high in 2004 because of a short history of being indexed in the *SCI*. Compared with the *JCM*, *Foundations of Computational Mathematics* (*FCM*) has citation impact on journals with higher quality measured by Impact Factor. Citation impact of *FCM* is also more focused: Fewer journals contributing to 1% of the total citations. Within-journal citations of *Foundations of Computational Mathematics* contributed less to its total citation than that of the *JCM*.

In 2004 the citation impact of *Acta Metallurgica Sinica* (*AMS*) was low and scattered among many journals, most of which were from China. Within-journal citation was rather heavy and became even heavier in 2013. Citation impact had been improved slightly in 2013 but was still very low. Furthermore, journal quality measured by Impact Factors of journals citing *AMS* had not been improved during 2004-2013. In contrast to *AMS*, *Acta Materialia* was able to generate citation impact in journals with higher quality measured by Impact Factors. Similar to *Acta Metallurgica Sinica*, within-journal citations of *Acta Materialia* also contributed first to its own total citation.

In general, the citation impact of leading Chinese journals has improved during the period 2004-2013, but there is still a long distance to catch up with leading foreign journals. Although being funded under Category B in the CIU Plan, *Journal of Computational Mathematics* performed as well as the other two in a higher rank of category – Category A of the CIU Plan. Being funded at the same level under Category B, the *Journal of Computational Mathematics* performed better than *Acta Metallurgica Sinica*. Foreign journals of higher Impact Factor are more intensively cited than Chinese journals at a given threshold (e.g., 1%) in the same subject category of the JCR, which may imply a positive correlation between journal quality and citation intensity in a specialist citation environment. In other words, journals with higher Impact Factor in the same subject category may be cited more intensively, or by a relatively stable number of journals in their citation impact environment across different years.

Acknowledgement

The study was supported by National Natural Science Foundation of China (NSFC) with Grant Number 71473219 and the Planning Office of Philosophy and Social Sciences of Hangzhou City, Zhejiang Province, with Grant Number B14TD02. The authors thank Thomson Reuters for access to the JCR data.

References

He, C., & Pao, M.L. (1986). A discipline-specific journal selection algorithm. *Information Processing & Management*, *22*(5), 405–416.

ISTIC (2013). 中国科技论文统计结果 2013 (Statistical Data of Chinese S&T Papers). Beijing: Institute of Scientific and Technical Information of China.

ISTIC (2014). 中国科技论文统计结果 2014 (Statistical Data of Chinese S&T Papers). Beijing: Institute of Scientific and Technical Information of China.

Jin, B.H.& Leydesdorff, L (2005), 中国科技期刊引文网络：国际影响和国内影响分析(Citation networks of Chinese S&T journals: analysis on international and domestic impact), 中国科技期刊研究*(Chinese Journal of Scientific and Technical Periodicals)*, *16*(2): 141–146.

Leydesdorff, L. (1986). The development of frames of references. *Scientometrics, 9*, 103–125.

Leydesdorff, L., & Cozzens, S.E. (1993). The delineation of specialties in terms of journals using the dynamic journal set of the Science Citation Index. *Scientometrics, 26*, 133–154.

Leydesdorff, L., & Jin, B.H. (2005). Mapping the Chinese Science Citation Database in terms of aggregated journal–journal citation relations. J*ournal of the American Society of Information Science & Technology, 56*(14), 1469–1479.

Leydesdorff, L. (2007). Visualization of the Citation Impact Environments of Scientific Journals: An Online Mapping Exercise, *Journal of the American Society for Information Science and Technology, 58*(1):25-38.

Li, L. (2012). 解读"科技期刊国际影响力提升计划" (Understanding the International Impact Upgrading Plan for Journals in Science and Technology). 科技日报 (Science and Technology Daily), Page 3, December 27, 2012.

Li, ZX (2006). How to establish a first-class international scientific journal in China? *World Journal of Gastroenterology, 12* (43): 6905-6908.

MOST (2012). China Science & Technology Statistics Database. http://www.sts.org.cn/sjkl/kjtjdt/index.htm

NBS (2013). Expenditure on R&D (100 million yuan) in: China Statistical Year Book. Available at: http://data.stats.gov.cn/english/easyquery.htm?cn=C01Ren, S.L. (2005). Editing scientific journals in Mainland China. European Science Editing, 31(1), 8–9.

Yao, Z.C., Luo, Z.F., Jin, X.Y. & Duan, R.Y. (2014). 《中国科技期刊国际影响力提升计划》资助期刊的分析与展望(New start, new task, new development: Analysis and expectation on journals funded by the International Impact Upgrading Plan for China's S&T journals) (in Chinese). 编辑学报 (*Acta Editologica*), *26*(4): 342-346.

Zhou, P., Leydesdorff, L. & Wu, Y.S. (2005).中国科技期刊引文环境的可视化(Visualization of the Citation Impact Environments in the CSTPC Journal Set), 中国科技期刊研究 (Chinese Journal of Scientific and Technical Periodicals), *16*(6), 773-780.

Zhou, Ping & Loet Leydesdorff (2007a). A Comparison Between the China Scientific and Technical Papers and Citations Database and the Science Citation Index in Terms of Journal Hierarchies and Inter-journal Citation Relations, *Journal of the American society for Information science and Technology*, *58*(2):223–236.

Zhou, Ping & Loet Leydesdorff (2007b). The citation impacts and citation environments of Chinese journals in mathematics, *Scientometrics, 72*(2).

Zhou, P. & Leydesdorff, L. (2008). China ranks second in scientific publications since 2006. *ISSI Newsletter*, *4*(1): 7-9.

Zhou, P. (2009). Mapping knowledge production and scholarly communication in China. PhD Thesis. University of Amsterdam. http://dare.uva.nl/document/2/64587

Zhou, Ping, Xinning Su & Loet Leydesdorff (2010), A comparative study on communication structures of Chinese journals in the social sciences. *Journal of the American society for Information science and Technology, 61*(7):1360–1376.

Research Data Explored: Citations versus Altmetrics

Isabella Peters[1], Peter Kraker[2], Elisabeth Lex[3], Christian Gumpenberger[4], and Juan Gorraiz[4]

[1] *i.peters@zbw.eu*
ZBW Leibniz Information Centre for Economics, Düsternbrooker Weg 120, D-24105 Kiel (Germany) & Kiel University, Christian-Albrechts-Platz 4, D-24118 Kiel (Germany)

[2] *pkraker@know-center.at*
Know-Center, Inffeldgasse 13, A-8010 Graz (Austria)

[3] *elex@know-center.at*
Graz University of Technology, Knowledge Technologies Institute, Inffeldgasse 13, A-8010 Graz (Austria)

[4] *christian.gumpenberger, juan.gorraiz@univie.ac.at*
University of Vienna, Vienna University Library, Dept of Bibliometrics, Boltzmanngasse 5, A-1090 Vienna (Austria)

Abstract
The study explores the citedness of research data, its distribution over time and how it is related to the availability of a DOI (Digital Object Identifier) in Thomson Reuters' DCI (Data Citation Index). We investigate if cited research data "impact" the (social) web, reflected by altmetrics scores, and if there is any relationship between the number of citations and the sum of altmetrics scores from various social media-platforms. Three tools are used to collect and compare altmetrics scores, i.e. PlumX, ImpactStory, and Altmetric.com. In terms of coverage, PlumX is the most helpful altmetrics tool. While research data remain mostly uncited (about 85%), there has been a growing trend in citing data sets published since 2007. Surprisingly, the percentage of the number of cited research data with a DOI in DCI has decreased in the last years. Only nine repositories account for research data with DOIs and two or more citations. The number of cited research data with altmetrics scores is even lower (4 to 9%) but shows a higher coverage of research data from the last decade. However, no correlation between the number of citations and the total number of altmetrics scores is observable. Certain data types (i.e. survey, aggregate data, and sequence data) are more often cited and receive higher altmetrics scores.

Conference Topic
Altmetrics, Citation and co-citation analysis

Introduction
Recently, data citations have gained momentum (Piwowar & Chapman, 2010; Borgman, 2012; Torres-Salinas, Martín-Martín, & Fuente-Gutiérrez, 2013). This is reflected, among others, in the development of data-level metrics (DLM), an initiative driven by PLOS, UC3 and DataONE[1], to track and measure activity on research data, and the recent announcement of CERN to provide DOIs for each dataset they share through their novel Open Data portal[2]. Data citations are citations included in the reference list of a publication that formally cite either the data that led to a research result or a data paper[3]. Thereby, data citations indicate the influence and reuse of data in scientific publications.
First studies on data citations showed that certain well-curated data sets receive far more citations or mentions in other articles than many traditional articles (Belter, 2014). Citations, however, are used as a proxy for the assessment of impact primarily in the "publish or perish" community; to consider other disciplines and stakeholders of research, such as industry, government and academia, and in a much broader sense, the society as a whole, altmetrics

[1] http://escholarship.org/uc/item/9kf081vf
[2] https://www.datacite.org/news/cern-launches-data-sharing-portal.html
[3] http://www.asis.org/Bulletin/Jun-12/JunJul12_MayernikDataCitation.html

(i.e. social media-based indicators) are emerging as a useful instrument to assess the "societal" impact of research data or at least to provide a more complete picture of research uptake, besides more traditional usage and citation metrics (Bornman, 2014; Konkiel, 2013). Previous work on altmetrics for research data has mainly focused on motivations for data sharing, creating reliable data metrics and effective reward systems (Costas et al., 2012).

This study contributes to the research on data citations in describing their characteristics as well as their impact in terms of citations and altmetrics scores. Specifically, we tackle the following research questions:

- How often are research data cited? Which and how many of these have a DOI? From which repositories do research data originate?
- What are the characteristics of the most cited research data? Which data types and disciplines are the most cited? How does citedness evolve over time?
- To what extent are cited research data visible on various altmetrics channels? Are there any differences between the tools used for altmetrics scores aggregation?

Data sources

On the Web, a large number of data repositories are available to store and disseminate research data. The Thomson Reuters Data Citation Index (DCI), launched in 2012, provides an index of high-quality research data from various data repositories across disciplines and around the world. It enables search, exploration and bibliometric analysis of research data through a single point of access, i.e. the Web of Science (Torres-Salinas, Martín-Martín & Fuente- Gutiérrez, 2013). The selection criteria are mainly based on the reputation and characteristics of the repositories[4]. Three document types are available in the DCI: data set, data study, and repository. The document type "repository" can distort bibliometric analyses, because repositories are mainly considered as a source, but not as a document type.

First coverage and citation analyses of the DCI have been performed April-June 2013 by the EC3 bibliometrics group of Granada (Torres-Salinas, Jimenez-Contreras & Robinson-Garcia, 2014; Torres-Salinas, Robinson-Garcia & Cabezas-Clavijo, 2013). They found that data is highly skewed: Science areas accounted for almost 80% of records in the database and four repositories contained 75% of all the records in the database; 88% of all records remained uncited. In Science, Engineering and Technology citations are concentrated among datasets, whereas in the Social Sciences and Arts & Humanities, citations often refer to data studies.

Since these first analyses, DCI has been constantly growing, now indexing nearly two million records from high-quality repositories around the world. One of the most important enhancements of the DCI has undoubtedly been the inclusion of "figshare[5]" as new data source which led to an increase of almost a half million of data sets and 40.000 data studies (i.e. about one fourth of the total coverage in the database).

Gathering altmetrics data is quite laborious since they are spread over a variety of social media platforms which each offer different applications programming interfaces (APIs). Tools, which collect and aggregate these altmetrics data come in handy and are now fighting for market shares since also large publishers increasingly display altmetrics for articles (e.g., Wiley[6]). There are currently three big altmetrics data providers: ImpactStory[7], Altmetric.com, and PlumX[8]. Whereas Altmetrics.com and PlumX focus more on gathering and providing

[4] http://thomsonreuters.com/data-citation-index, http://thomsonreuters.com/products/ip-science/04_037/dci-selection-essay.pdf
[5] http://figshare.com
[6] http://eu.wiley.com/WileyCDA/PressRelease/pressReleaseId-108763.html?campaign=wlytk-41414.4780439815
[7] https://impactstory.org
[8] https://plu.mx

data for institutions (e.g., publishers, libraries, or universities), ImpactStory's target group is the individual researcher who wants to include altmetrics information in her CV.

ImpactStory is a web-based tool, which works with individually assigned permanent identifiers (such as DOIs, URLs, PubMed IDs) or links to ORCID, Figshare, Publons, Slideshare, or Github to auto-import new research outputs like e.g. papers, data sets, slides. Altmetric scores from a large range of social media-platforms, including Twitter, Facebook, Mendeley, Figshare, Google+, and Wikipedia[9], can be downloaded as .json or .csv (as far as original data providers allow data sharing). With Altmetric.com, users can search within a variety of social media-platforms (e.g., Twitter, Facebook, Google+, or 8,000 blogs[10]) for keywords as well as for permanent identifiers (e.g., DOIs, arXiv IDs, RePEc identifiers, handles, or PubMed IDs). Queries can be restricted to certain dates, journals, publishers, social media-platforms, and Medline Subject Headings. The search results can be downloaded as .csv from the Altmetric Explorer (web-based application) or via the API. Plum Analytics or Plum X (the fee-based altmetrics dashboard) offers article-level metrics for so-called artifacts, which include articles, audios, videos, book chapters, or clinical trials[11]. Plum Analytics works with ORCID and other user IDs (e.g., from YouTube, Slideshare) as well as with DOIs, ISBNs, PubMed-IDs, patent numbers, and URLs. Because of its collaboration with EBSCO, Plum Analytics can provide statistics on the usage of articles and other artifacts (e.g., views to or downloads of html pages or pdfs), but also on, amongst others, Mendeley readers, GitHub forks, Facebook comments, and YouTube subscribers.

Methodology

We used DCI to retrieve the records of cited research data. All items published in the last 5.5 decades (1960-9, 1970-9, 1980-9, 1990-9, 2000-9, and 2010-4) with two or more citations (Sample 1, n=10,934 records) were downloaded and analysed. The criterion of having at least two citations is based on an operational reason (reduction of the number of items) as well as on a conceptual reason (to avoid self-citations). The following metadata fields were used in the analysis: available DOI or URL, document type, source, research area, publication year, data type, number of citations and ORCID availability[12]. The citedness in the database was computed for each decade considered in this study and investigated in detail for each year since 2000. We then analysed the distribution of document types, data types, sources and research areas with respect to the availability or non-availability of DOIs reported by DCI.

All research data with two or more citations and with an available DOI (n=2,907 items) were analysed with PlumX, ImpactStory, and Altmetric.com and their coverage on social media platforms and the altmetric scores was compared. All other items with 2 or more citations and an available URL (n=8,027) were also analysed in PlumX, the only tool enabling analyses based on URLs, and the results were compared with the ones obtained for items with a DOI. We also analysed the distribution of document types, data types, sources and research areas for all research data with 2 or more citations and at least one altmetric score (sample 2; n=301 items) with respect to the availability or non-availability of the permanent identifier DOI reported by DCI (items with DOI and URL or items with URL only).

[9] http://feedback.impactstory.org/knowledgebase/articles/367139-what-data-do-you-include-on-profiles
[10] http://support.altmetric.com/knowledgebase/articles/83335-which-data-sources-does-altmetric-track
[11] http://www.plumanalytics.com/metrics.html
[12] The DCI field "data type" was manually merged to more general categories; e.g. "survey data in social sciences" was merged with the category "survey data".

Table 1. Results of DCI-based citation and altmetrics analyses for the last 5.5 decades.

DCI	*1960-69*	*1970-79*	*1980-89*	*1990-99*	*2000-09*	*2010-14*
total # items	6 040	23 712	43 620	186 965	2 096 023	1 627 668
# items with > 2 citations	5	110	360	956	4 727	4 777
# items with at least 1 citation	5	4207	7519	43749	239867	218440
uncited %	99.9%	82.3%	82.8%	76.6%	88.6%	86.6%
items with DOI and >= 2 cit	4	107	343	846	1381	226
% with DOI and >=2 cit	0.8	97.27%	95.28%	88.49%	29.22%	4.73%
with Altmetrics Data (PlumX)	1	5	14	40	114	20
%	25.0%	4.7%	4.1%	4.7%	8.3%	8.8%
items with URL only and >= 2 cit	1	3	17	110	3 346	4551
% with URL only and >=2 cit	0.2	2.73%	4.72%	11.51%	70.78%	95.27%
with Altmetrics Data (PlumX)	1	1	8	11	54	33
%	100.0%	33.3%	47.1%	10.0%	1.6%	0.7%

Results and discussion

General Results

Table 1 gives an overview of the general results obtained in this study. The analysis revealed a high uncitedness of research data, which corresponds to the findings of Torres-Salinas, Martin-Martin and Fuente-Gutiérrez (2013). A more detailed analysis for each year (see Table 2) shows, however, that the citedness is comparatively higher for research data published in recent years (published after 2007) although the citation window is shorter.

Table 2. Evolution of uncitedness in DCI in the last 14 years.

PY	*Items*	*uncited*	*% uncited*
2000	28282	18152	64.18%
2001	36397	25367	69.70%
2002	64781	51464	79.44%
2003	115997	93538	80.64%
2004	141065	122802	87.05%
2005	212781	178146	83.72%
2006	299443	275216	91.91%
2007	362405	333136	91.92%
2008	398931	364236	91.30%
2009	435941	394099	90.40%
2010	390957	349623	89.43%
2011	270932	224790	82.97%
2012	492534	428752	87.05%
2013	448489	386507	86.18%
2014	24756	19556	78.99%

The results also show a very low percentage of altmetrics scores available for research data with two or more citations (see Table 1). But, two different trends can be observed: the percentage of data with DOI referred to on social media-platforms is steadily increasing while the percentage of data with URL only is steadily decreasing in the same time frame.

The percentage of research data with altmetrics scores in PlumX, the tool with the highest average in this study, is lower than expected (ranging between 4 and 9%) but actually has doubled for data published in the last decades, which confirms the interest in younger research data and an increase in social media activity of the scientific community in recent years.

Table 3. Overview on citation distribution of Sample 1 (n=10,934 items).

items with at least 2 citations	*Document Type*	*# items*	*Total Citations*	*Mean Citations*	*Maximum Citations*	*Standard Deviation*	*Variance*
all	Data set	5641	17984	3.19	121	3.38	11.46
	Data study	5242	91623	17.48	1236	50.22	2521.67
	Repository	51	10076	197.57	3193	618.73	382824.45
	Total	10934	119683	10.95	3193	56.39	3179.49
with DOI	Data set	342	977	2.86	52	3.86	14.93
	Data study	2565	53293	20.78	1236	63.44	4024.45
	Total	2907	54270	18.67	1236	59.88	3585.92
with URL only	Data set	5299	17007	3.21	121	3.35	11.23
	Data study	2677	38330	14.32	272	32.59	1062.31
	Repository	51	10076	197.57	3193	618.73	382824.45
	Total	8027	65413	8.15	3193	54.80	3003.30

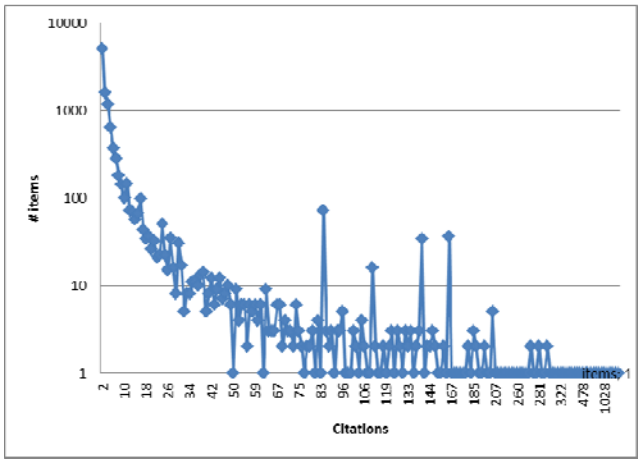


Figure 1. Citation distribution of Sample 1 (logarithmic scale).

Results for Sample 1

Table 3 shows the citation distribution of Sample 1 (10,934 items with at least two citations in DCI) for items with DOI or URL only separated according to the three main DCI document types (data set, data study, and repository[13]). The results reveal that almost half of the data studies have a DOI (48.9%) but only few data sets do so. Data studies are on average more often cited than data sets (17.5 vs. 3.2 citations per item), and data studies with a DOI attract more citations (mean values) than those with a URL (20 vs. 14 citations per item).

There were only few repositories (51) in the data set; it is the document type, which attracts the most citations per item. This finding is in line with the results of Belter (2014) who also found aggregated data sets – Belter calls them "global-level data sets" – to be more cited. However, such citing behaviour has a negative side effect on repository content (i.e., the single data sets) since it is not properly attributed in favour of citing the repository as a whole.

The high values of the standard deviation and variance illustrate the skewness of the citation distribution (see Figure 1). Almost half of the research data (4,974 items; 45.5%) have only two citations. Six items, two repositories and four data studies, from different decades (PY=1981, 1984, 1995, 2002, 2011, and 1998, sorted by descending number of citations) had more than 1,000 citations and account for almost 30% of the total number of citations.

[13] Table 3 includes repositories as document type to illustrate the citation volume in DCI.

Table 4 shows the top 10 repositories by the number of items. Considering the number of citations, there are three other repositories which account for more than 1,000 citations each: Manitoba Centre for Health Policy Population Health Research Data Repository (29 items; 1,631 citations), CHILDES - Child Language Data Exchange System (1 item; 3,082 citations), and World Values Survey (1 item; 3,193 citations). Interestingly, although "figshare" accounts for almost 25% of the DCI, no item from "figshare" was cited at least twice in DCI. We also noted that the categorization of "figshare" items is missing. All items are assigned to the Web of Science category (WC) "Multidisciplinary Sciences" or the Research Area (SU) "Science & Technology/Other Topics" preventing detailed topic-based citation analyses. Furthermore, only nine items from Sample 1 were related to an ORCID, three data sets with a DOI, and three data sets and data studies with a URL.

Table 4. Analysis of Sample 1 by sources (repositories) (n=10,934 items).

Sources (with DOI)	*# items*	*# citations*	*Sources (with URL)*	*# items*	*# citations*
Inter-university Consortium for Political and Social Research	2530	53041	miRBase	3456	10209
Worldwide Protein Data Bank	229	458	Cancer Models Database	864	2698
Oak Ridge National Laboratory Distributed Active Archive Center for Biogeochemical Dynamics	108	508	UK Data Archive	836	25479
Archaeology Data Service	21	75	European Nucleotide Archive	361	1346
3TU.Datacentrum	8	22	Gene Expression Omnibus	353	754
SHARE - Survey of Health, Ageing and Retirement in Europe	4	151	National Snow & Ice Data Center	298	2796
World Agroforestry Centre	3	6	Australian Data Archive	264	2469
Dryad	2	4	Australian Antarctic Data Centre	249	1621
GigaDB	2	5	nmrshiftdb2	219	445
			Finnish Social Science Data Archive	183	913

Considering their origin, considerable differences were reported in Sample 1 for items with or without a DOI (see Table 4). All twice or more frequently cited research data with a DOI are archived in nine repositories, while 92 repositories account for research data without a DOI.

Table 5 shows that there are big differences between the most cited data types when considering research data with a DOI or with a URL. Survey data, aggregate data, and clinical data are the most cited ones of the first group (with a DOI), while sequence data and numerical and individual level data are the most cited data types of the second group (with a URL). Apart from survey data, there is no overlap in the top 10 data types indexed in DCI. Similar results were obtained when considering data sets and data studies separately.

Disciplinary differences become apparent in the citations of DOIs and URLs as well as in the use of certain document types. As shown in Table 6 it is more common to refer to data studies via DOIs in the Social Sciences than in the Natural and Life Sciences, where the use of URLs for both data studies and data sets is more popular. Torres-Salinas, Jimenez-Contreras and Robinson-Garcia (2014) also report that citations in Science, Engineering and Technology citations are concentrated on data sets, whereas the majority of citations in the Social Sciences and Arts & Humanities refer to data studies. Table 6 suggests that these differences could be related to the availability of a DOI.

Table 5. Analysis of Sample 1 by data types (manually merged), top 10 types (n=10,934 items).

Data Types (with DOI)	*# items*	*# citations*	*Data Types (with URL only)*	*# items*	*# citations*
survey data	1734	43686	sequence data	3408	10458
administrative records data	302	3326	profiling by array, gen, etc	352	752
aggregate data	274	9440	Individual (micro) level	240	9024
event/transaction data	210	2400	Numeric data	216	4317
clinical data	118	3469	Structured questionnaire	155	673
census/enumeration data	109	1019	survey data	127	1315
protein structure	95	190	Seismic:Reflection:MCS	47	185
observational data	30	575	statistical data	41	1352
program source code	10	116	Digital media	40	290
roll call voting data	8	236	EXCEL	25	101

Table 6. Sample 1 by research areas and document types, top 10 areas (n=10,934 items).

with DOI					with URL only				
	# Items		# citations			# Items		# citations	
Research Area	Data set	Data study	Data set	Data study	Research Area	Data set	Data study	Data set	Data study
Criminology & Penology		471		4403	Genetics & Heredity	4658	159	14024	571
Sociology		432		7930	Meteorology & Atmospheric Sciences	91	298	493	2796
Government & Law		352		10399	Biochemistry & Molecular Biology; Genetics & Heredity		353		754
Demography		317		9178	Sociology		286		1994
Health Care Sciences & Services		290		8170	Physics	5	214	10	435
Biochemistry & Molecular Biology	229		458		Business & Economics; Sociology		143		12665
Business & Economics		204		3083	Biochemistry & Molecular Biology; Spectroscopy	129		383	
Environmental Sciences & Ecology; Geology	108		508		Oceanography; Geology	114		353	
Education & Educational Research		69		1881	Demography; Sociology		103		5673
Family Studies		68		2268	Sociology; Demography; Communication		84		393

Results for Sample 2

Sample 2 comprises all items from DCI satisfying the following criteria: two or more citations in DCI, a DOI or a URL and at least one altmetrics score in PlumX (n=301 items). Table 7 shows the general results for this sample. The total number of altmetrics scores is lower than the number of citations for all document types with or without a DOI. Furthermore, the mean altmetrics score is higher for data studies than for data sets.

Tables 8 and 9 show the distributions of data types and subject areas in this sample. Most data with DOI are survey data, aggregate data, event over transaction data, whereas sequence data and images are most often referred to via URL only (see Table 8). Microdata with DOI and spectra with URL only are the data types with the highest altmetrics scores per item. Concerning subject areas the results of Table 9 are very similar to the results of Table 6. Given the small sample size it is, however, notable that in some subject areas, e.g. Archaeology, research data receive more interest in social media (i.e. altmetrics scores), than via citations in traditional publications. This is confirmed by the missing correlation between citations and altmetrics scores for this sample (see Figure 2). Both cases clearly demonstrate that altmetrics can complement traditional impact evaluation. Nevertheless, coverage of

research data in social media is still low, e.g. from the nine repositories whose data studies and data sets were cited twice in DCI and had a DOI (see Table 4), only five items had altmetrics scores in PlumX, and only one DOI item of Sample 2 included an ORCID.

Table 7. Citation and altmetrics results of Sample 2 (n=301 items) according to document type. *8 items with URL found in PlumX could not properly be identified (broken URL, wrong item, etc.)

	Document Type	*# items*	*Total Citations*	*Mean Citations*	*Maximum Citations*	*Standard Deviation*	*Variance*
with DOI	Data set	15	173	11.53	52	13.75	189.12
	Data study	179	6716	37.52	1135	107.36	11525.43
	Total	194	6889	35.51	1135	103.40	10691.82
	Document Type	*# items*	*Total Scores*	*Mean Scores*	*Maximum Scores*	*Standard Deviation*	*Variance*
	Data set	15	34	2.27	6	1.75	3.07
	Data study	179	710	3.97	64	7.42	55.09
	Total	194	752	376.00	748	526.09	276768.00
	Document Type	*# items*	*Total Citations*	*Mean Citations*	*Maximum Citations*	*Standard Deviation*	*Variance*
with URL only	Data set	24	172	7.17	46	10.12	102.41
	Data study	31	779	25.13	272	51.67	2669.65
	Repository	44	9677	219.93	3193	662.92	439464.20
	Total*	99	10628	107.35	3193	451.61	203954.50
	Document Type	*# items*	*Total Scores*	*Mean Scores*	*Maximum Scores*	*Standard Deviation*	*Variance*
	Data set	24	428	17.83	378	76.75	5890.23
	Data study	31	664	21.42	213	53.25	2835.65
	Repository	44	3961	90.02	1150	198.53	39415.70
	Total*	99	5319	49.71	1150	139.82	19549.38

Table 8. Citation and altmetrics overview of Sample 2 (n=301 items) according to their data type (Field DY; no aggregated counts, "document type" "repository" (34 items) not included.

Data Type (with DOI)	*# items*	*total citations*	*mean citations*	*total scores*	*mean scores*	*Data Type (with URL only) **	*# items*	*total citations*	*mean citations*	*total scores*	*mean scores*
survey data	110	5276	47.96	353	3.21	miRNA sequence data	15	71	4.73	21	1.40
aggregate data	26	793	30.50	80	3.08	FITS images; spectra; calibrations; redshifts	4	248	62	16	4.00
event/transaction data	19	414	21.79	43	2.26	statistical data	3	333	111	22	7.33
administrative records data	13	125	9.62	58	4.46	Expression profiling by array	3	6	2	4	1.33
clinical data	11	314	28.55	26	2.36	Sensor data; survey data	2	51	25.5	10	5.00
census/enumeration data	8	90	11.25	14	1.75	Quantitative	2	35	17.5	10	5.00
observational data	4	99	24.75	7	1.75	images	1	20	20	3	3.00
Longitudinal data; Panel Data; Micro data	2	79	39.50	46	23.00	images; spectra	1	4	4	102	102.00
roll call voting data	2	178	89.00	3	1.50	table	1	9	9	1	1.00
machine-readable text	1	5	5.00	1	1.00	redshifts; spectra	1	5	5	213	213.00
program source code	1	2	2.00	1	1.00	images; spectra; astrometry	1	2	2	90	90.00

Table 9. Citation and altmetrics overview of Sample 2 according to their subject area.

with DOI				*with URL only*			
Subject Areas	*# items*	*# citations*	*# scores*	*Subject Areas*	*# items*	*# citations*	*# scores*
Sociology	35	1226	213	Genetics & Heredity	26	492	654
Government & Law	28	793	53	Meteorology & Atmospheric Sciences	15	166	28
Criminology & Penology	22	317	42	Astronomy & Astrophysics	9	933	427
Health Care Sciences & Services	14	1498	70	Biochemistry & Molecular Biology; Genetics & Heredity	5	22	557
Environmental Sciences & Ecology; Geology	14	171	33	Cell Biology	4	13	383
Demography	12	433	28	Health Care Sciences & Services; Business & Economics	3	335	68
Family Studies	10	166	26	Genetics & Heredity; Biochemistry & Molecular Biology	2	27	36
Archaeology	10	47	139	Business & Economics	2	35	10
Education & Educational Research	9	661	40	Health Care Sciences & Services	2	423	2
International Relations	9	384	46	Communication; Sociology; Telecommunications	2	51	10

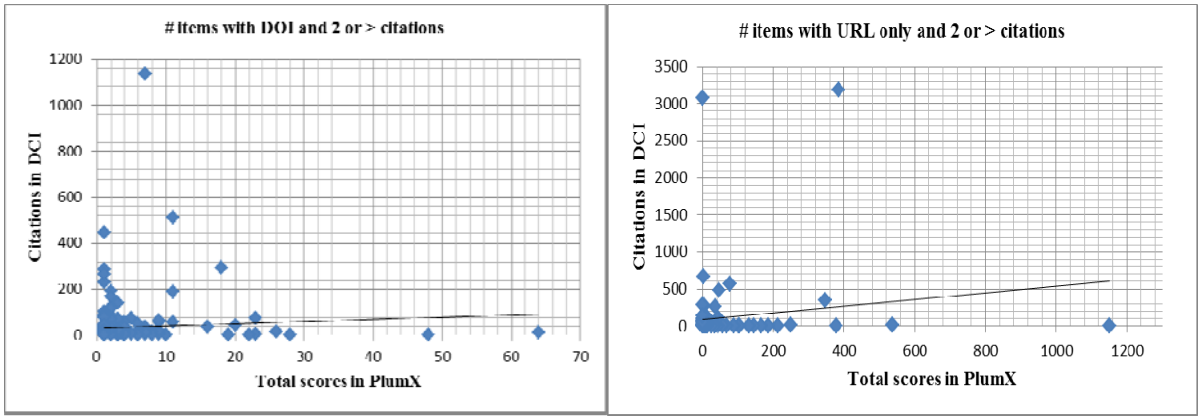


Figure 2. Citations DCI versus scores in PlumX for items with (left) and without (right).

Selected altmetrics scores and comparison of the results of three altmetrics tools

Table 10 shows the general results obtained in PlumX according to PlumX's aggregation groups (i.e., captures, social media, mentions, and usage) for all document types and with or without DOI. While DOIs for data sets seem to be important in order to get captures (mainly in Mendeley), a URL is sufficient for an inclusion in social media tools like Facebook, Twitter, etc.

The top 10 research data-DOIs attracting two or more citations and with at least one entry in PlumX are shown in Table 11. We can observe that cited research data attracts more citations than altmetrics scores, and that there is no correlation between highly cited and highly scored research data.

The comparison of altmetrics aggregation tools also revealed that ImpactStory only found Mendeley reader statistics for the research data: 78 DOIs had 257 readers. Additionally, ImpactStory found one other DOI in Wikipedia. ImpactStory found five items, which have

not been found by PlumX, although they all solely relied on Mendeley Data. The Mendeley data scores were exactly the same in PlumX and in ImpactStory. On the other hand, PlumX found 18 items that were not available via ImpactStory. These research data were distributed on social media platforms (mostly shares in Facebook) and one entry has been used via click on a Bitly-URL (Usage:Clicks:Bitly).The tool Altmetric.com found only one from 194 items.
As already reported in Jobmann et al. (2014), PlumX is the tool with the highest coverage of research products found on social media-platforms. Whereas Mendeley is well covered in ImpactStory, no other altmetrics score were found for the data set used in this study.

General Conclusions

Most of the research data still remain uncited (approx. 86%) and total altmetrics scores found via aggregation tools are even lower than the number of citations. However, research data published from 2007 onwards have gradually attracted more citations reflecting a bias towards more recent research data. No correlation between citation and altmetrics scores could be observed in a preliminary analysis: neither the most cited research data nor the most cited sources (repositories) received the highest scores in PlumX.
In the DCI, the availability of cited research data with a DOI is rather low. A reason for this may be the increase of available research data in recent years. Furthermore, the percentage of cited research data with a DOI has not increased as expected, which indicates that citations do not depend on this standard identifier in order to be processed by the DCI.

Table 10. PlumX altmetrics scores for all document types with or without DOI.

		with DOI			*with URL only*			
	Document Type	*Data set*	*Data study*	*Total*	*Data set*	*Data study*	*Repository*	*Total*
	# items	15	179	**194**	24	31	44	**99**
Captures	***Sum***	32	471	**503**	0	0	30	**30**
	Mean	2.13	2.63	**2.59**	0.00	0.00	0.68	**0.28**
	Max	6	48	**48**	0	0	23	**23**
Social Media	***Sum***	1	220	**221**	407	281	3060	**3890**
	Mean	0.07	1.23	**1.14**	16.96	9.06	69.55	**36.36**
	Max	1	58	**58**	366	119	1008	**1008**
Mentions	***Sum***	1	13	**14**	13	62	433	**629**
	Mean	0.07	0.07	**0.07**	0.54	2.00	9.84	**5.88**
	Max	1	4	**4**	12	31	119	**120**
Usage	***Sum***	0	6	**6**	8	321	438	**770**
	Mean	0.00	0.03	**0.03**	0.33	10.35	9.95	**7.20**
	Max	0	6	**6**	4	187	92	**187**
Total entries		**34**	**710**	**744**	**428**	**664**	**3961**	**5319**
% Captures		**94.1%**	**66.3%**	**67.6%**	**0.0%**	**0.0%**	**0.8%**	**0.6%**
% Social Media		**2.9%**	**31.0%**	**29.7%**	**95.1%**	**42.3%**	**77.3%**	**73.1%**
% Mentions		**2.9%**	**1.8%**	**1.9%**	**3.0%**	**9.3%**	**10.9%**	**11.8%**
% Usage		**0.0%**	**0.8%**	**0.8%**	**1.9%**	**48.3%**	**11.1%**	**14.5%**

Nevertheless, data studies with a DOI attract more citations than those with a URL. Despite the low number of research data with a DOI in general, surprisingly, the DOI in cited research data has so far been more embraced in the Social Sciences than in the Natural Sciences.
Furthermore, our study shows an extremely low number of research data with two or more citations (only nine out of around 10,000) related to an ORCID. Only three of them had a DOI

likewise. This illustrates that we are still a far cry from the establishment of permanent identifiers and their optimal interconnectedness in a data source.

The low percentage of altmetrics scores for research data with two or more citations corroborates a threefold hypothesis: First, research data are either rarely published or not findable on social media-platforms, because DOIs or URLs are not used in references thus resulting in a low coverage of items. Second, research data are not widely shared on social media by the scientific community so far, which would result in higher altmetrics scores[14]. Third, the reliability of altmetrics aggregation tools is questionable as the results on the coverage of research data on social media-platforms differ widely between tools. However, the steadily increasing percentage of cited research data with DOI suggests that the adoption of this permanent identifier increases the online visibility of research data and inclusion in altmetrics tools (since they heavily rely on DOIs or other permanent identifiers for search).

A limitation of our study is that the results rely on the indexing quality of the DCI. Our analysis shows that the categorisation in DCI is problematic at times. This is illustrated by the fact that all items from figshare, which is one of the top providers of records, are categorised

Table 11. Top 10 Research Data with DOI according to the total scores in PlumX.

DOI	*SO*	*PY*	*Captures :Readers: Mendeley*	*Social Media:+ 1s:Google+*	*Social Media :Shares:Facebook*	*Social Media :Likes :Facebook*	*Social Media: Tweets :Twitter*	*Mentions: Comments: Facebook*	*# total Scores*	*# Citations*
10.5284/1000415	ADS	2012	2		13	45		4	64	13
10.3886/icpsr13580	IUC	2005	48						48	3
10.5284/1000397	ADS	2011			14	12		2	28	2
10.3886/icpsr06389	IUC	2007	25	1					26	14
10.6103/share.w4.111	SHARE	2004			8	15			23	74
10.6103/share.w4.111	SHARE	2010			8	15			23	5
10.3886/icpsr13611	IUC	2006	22						22	3
10.3886/icpsr02766	IUC	2007	20						20	44
10.5284/1000381	ADS	2009		2	3	10	3	1	19	2
10.3886/icpsr09905	IUC	1994	18						18	295
10.3886/icpsr08624	IUC	2010	16						16	36
10.3886/icpsr04697	IUC	2009	11						11	510
10.3886/icpsr06716	IUC	2007	11						11	59
10.3886/icpsr20240	IUC	2008	11						11	190
10.3886/icpsr20440	IUC	2007	3				7		10	3

into “Miscellaneous”. Also, the category “repository” is rather a source than a document type. Such incorrect assignments of data types and disciplines can easily lead to wrong interpretations in citation analyses. Furthermore, it should be taken into account that citation counts are not always traceable.

Finally, citations of research data should be studied in more detail. They certainly differ from citations of papers relying on these data with regard to dimension and purpose. For example, we found that entire repositories are proportionally more often cited than single data sets, which was confirmed by a former study (Belter, 2014). Therefore, it will be important to study single repositories (such as figshare) in more detail. It is crucial to further explore the real meaning and rationale of research data citations and how they depend on the nature and structure of the underlying research data, e.g., in terms of data curation and awarding of DOIs.

[14] figshare lately announced a deal with Altmetric.com which might increase the visibility of altmetrics with respect to data sharing: http://figshare.com/blog/The_figshare_top_10_of_2014_according_to_altmetric/142

Also, little is known about how data citations complement and differ from data sharing and data usage activities as well as altmetrics.

Acknowledgments

This analysis was done within the scope of e-Infrastructures Austria (http://e-infrastructures.at/). The authors thank Dr. Uwe Wendland (Thomson Reuters) and Stephan Buettgen (EBSCO) for granted trial access to Data Citation Index resp. PlumX. The Know-Center is funded within the Austrian COMET program – Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth, and the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

Belter, C.W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLoS ONE, 9*(3): e92590. doi:10.1371/journal.pone.0092590

Bornman, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics, Retrieved January 1, 2015 from http://arxiv.org/abs/1406.7091

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*, 1059-1078.

Costas, R., Meijer, I., Zahedi, Z. & Wouters, P. (2012). The value of research data – Metrics for data sets from a cultural and technical point of view. A Knowledge Exchange Report. Retrieved January 1, 2015 from http://www.knowledge-exchange.info/datametrics.

Jobmann, A., Hoffmann, C.P., Künne, S., Peters, I., Schmitz, J. & Wollnik-Korn, G. (2014). Altmetrics for large, multidisciplinary research groups: Comparison of current tools. *Bibliometrie - Praxis und Forschung*, *3*, Retrieved January 1, 2015 from http://www.bibliometrie-pf.de/article/viewFile/205/258.

Konkiel, S. (2013). Altmetrics . A 21st-century solution to determining research quality. *Information Toda*y, *37*(4), Retrieved January 1, 2015 from http://www.knowledge-exchange.info/datametrics http://www.infotoday.com/OnlineSearcher/Articles/Features/Altmetrics-A-stCentury-Solution-to-Determining-Research-Quality-90551.shtml.

Piwowar, H.A. & Chapman, W.W. (2010). Public sharing of research datasets: A pilot study of associations. *Journal of Informetrics*, *4*, 148-156.

Torres-Salinas, D., Robinson-Garcia, N. & Cabezas-Clavijo, Á. (2013). Compartir los datos de investigación: Una introducción al 'Data Sharing'. *El profesional de la información*, *21*, 173-184.

Torres-Salinas, D., Martín-Martín, A. & Fuente-Gutiérrez, E. (2013). An introduction to the coverage of the Data Citation Index (Thomson-Reuters): Disciplines, document types and repositories. EC3 Working Papers, 11, June 2013. Retrieved January 1, 2015 from http://arxiv.org/ftp/arxiv/papers/1306/1306.6584.pdf.

Torres-Salinas, D., Jimenez-Contreras, E. & Robinson-Garcia, N. (2014). How many citations are there in the Data Citation Index? *Proceedings of the STI Conference*, Leiden, The Netherlands. Retrieved January 1, 2015 from http://arxiv.org/abs/1409.0753.

Stopped Sum Models for Citation Data

Wan Jing Low, Paul Wilson and Mike Thelwall

W.J.Low@wlv.ac.uk, PaulJWilson@wlv.ac.uk, M.Thelwall@wlv.ac.uk
Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of
Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY (UK)

Abstract
It is important to identify the most appropriate statistical model for citation data in order to maximise the power of future analyses as well as to shed light on the processes that drive citations. This article assesses stopped sum models and compares them with two previously used models, the discretised lognormal and negative binomial distributions using the Akaike Information Criterion (AIC). Based upon data from 20 Scopus categories, some of the stopped sum models had lower AIC values than the discretised lognormal models, which were otherwise the best. However, very large standard errors were produced for some of these stopped sum models, indicating the imprecision of the estimates and the impracticality of the approach. Hence, although stopped sum models show some promise for citation analysis, they are only recommended when they fit better than the alternatives and have manageable standard errors. Nevertheless, their good fit to citation data gives evidence that two different, but related, processes drive citations.

Conference Topic
Citation and co-citation analysis

Introduction

Fitting statistical models to citation data is useful both to understand the citation process itself (de Solla Price, 1976) and to identify the factors that affect the citedness of academic papers (Bornmann, Schier, Marx, & Daniel, 2012; Didegah & Thelwall, 2013). For example, negative binomial regression previously has been used to analyse factors underlying patent citations (Maurseth & Verspagen, 2002). The choice of statistical model is not straightforward (Bookstein, 2001), however, because citation data is typically highly skewed (de Solla Price, 1976) with a heavy tail (i.e., with particularly many articles having high citation counts) which makes it difficult to identify and fit the best distribution (Clauset, Shalizi, & Newman, 2009). Nevertheless, it has recently been shown that the distribution of citations to articles from an individual Scopus category and year follows a hooked power law or a discretised lognormal distribution substantially better than a power law (Thelwall & Wilson, 2014a) and that, on this basis, (discretised) ordinary least squares regression on the log of the citation data, after adding 1 to cope with the problem of uncited articles, is applicable and is probably the best available regression method (Thelwall & Wilson, 2014b). It should be noted that although the data is well fitted by the discretised lognormal distribution, it should not be assumed that it was derived from that distribution, as models should not be regarded as literal descriptions of nature (Hesse, 1953). Moreover, it is useful to assess additional statistical models in case a more powerful model can be found as well as to shed light on the processes underlying citation, which are still far from fully understood. This paper investigates stopped sum models for citation data for the first time. These have very different underlying assumptions to the lognormal distribution but can result in similar shaped distributions.

Hence, should citation data fit them well, the results would have both practical and theoretical implications for citation analysis.

Stopped sum distributions

Stopped sum distributions were initially developed by Neyman to model the number of larvae in a field (Neyman, 1939). Neyman viewed the distribution of larvae as resulting from two population waves. The first 'parent' (or primary wave) distribution was followed by a distribution of 'offspring' (or secondary wave), whereby the numbers in the secondary wave would be dependent on the numbers in the primary wave; the overall population being the sum of the populations from the two waves (Johnson, Kemp, & Kotz, 2005, pp. 381–382). The two waves can have completely different statistical distributions. If, for example, the primary wave distribution is Poisson and the secondary wave distribution is negative binomial, the overall distribution is known as a *Poisson stopped sum negative binomial (NB) distribution*. Here stopped sum models are explored due to their potential to model citation data as two waves, the primary wave and secondary wave. Given that the overall number of citations that an article receives might come from a similar two waves process, the primary wave representing citations received shortly after a journal article has been published, and the secondary wave, perhaps overlapping with the first to some extent, representing the citations received as a result of scientists discovering an article because of its previous citations, either directly by following citations or indirectly because more cited articles are ranked more highly in some citation databases.

The stopped sum models for citation counts could also be appropriate if the two waves occurred simultaneously instead of sequentially. For example, for the Poisson stopped sum negative binomial model, one of the wave distributions follows the Poisson distribution and the other wave follows the negative binomial distribution at the same time.

The original model proposed by Neyman (1939) assumed that zero counts in the primary wave will automatically be followed by zero counts in the second wave. Hence, if X follows the Poisson stopped sum NB distribution, P(X=0) is just P(X=0) under the Poisson distribution.

For citation counts of one or more, the stopped sum assumes that this can only be a result of a non-zero citation in the primary wave. For example, a citation count of 3 can only arise as a result of one of the three combinations:

- 3 citations in the primary wave, 0 citation in the secondary wave; or
- 2 citations in the primary wave, 1 citation in the secondary wave; or
- 1 citation in the primary wave, 2 citations in the secondary wave.

The Poisson stopped sum NB distribution will therefore have the following probability mass function (p.m.f.):

$$P(X=y)=\begin{cases} e^{-\lambda} & if\ y=0 \\ \sum_{j=1}^{y}\frac{e^{-\lambda}\lambda^{j}}{j!} * \binom{y-j+\alpha-1}{\alpha-1} p^{\alpha}(1-p)^{y-j} & if\ y\geq 1, \text{and } p=\frac{\alpha}{\mu+\alpha} \end{cases}$$

The other stopped sum distributions that are considered include the NB stopped sum Poisson distribution:

$$P(X=y)=\begin{cases} p^{\alpha} & if\ y=0 \\ \sum_{j=1}^{y}\binom{y+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y}*\frac{e^{-\lambda}\lambda^{y-j}}{(y-j)!} & if\ y\geq 1 \end{cases}$$

and the NB stopped sum NB distribution:

$$P(X=y)=\begin{cases} p^{\alpha} & if\ y=0 \\ \sum_{j=1}^{y}\binom{y+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y}*\binom{y-j+\theta-1}{\theta-1}q^{\theta}(1-q)^{y-j} & if\ y\geq 1 \end{cases}$$

where $p=\frac{\alpha}{\mu+\alpha}$ in all cases.

The Poisson stopped sum Poisson distribution was considered but because very large AICs were obtained indicating a poor fit for citation data we do not discuss it further here.

Modified stopped sum distributions

In the study made by Neyman in 1939, the restriction of having zero counts in the primary wave resulting in zero counts in the secondary wave was necessary, but in the case of citation analysis, it is feasible that a zero citation count in the first population wave could be followed by a non-zero count in the second. This can occur due to the limitations of the citation database used to analyse the citations. For example, an article may be uncited in Scopus, but cited in Google Scholar, and its Google Scholar citations could attract new second wave citations. Hence a modified stopped sum is also considered, where, for example, 3 citations could arise from 0 citations in the primary wave and 3 citations in the secondary wave. The modified Poisson stopped sum NB distribution for this case has p.m.f.:

$$P(X=y)=\sum_{j=0}^{y}\frac{e^{-\lambda}\lambda^{j}}{j!}*\binom{y-j+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y-j} \quad where\ y\geq 0 \text{ and } p=\frac{\alpha}{\mu+\alpha}$$

Using similar adjustments, the modified NB stopped sum Poisson distribution has p.m.f.:

$$P(X=y)=\sum_{j=0}^{y}\binom{y+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y}*\frac{e^{-\lambda}\lambda^{y-j}}{(y-j)!} \quad where\ y\geq 0 \text{ and } p=\frac{\alpha}{\mu+\alpha}$$

Whilst the modified NB stopped sum NB distribution has p.m.f.:

$$P(X=y)=\sum_{j=1}^{y}\binom{y+\alpha-1}{\alpha-1}p^{\alpha}(1-p)^{y}*\binom{y-j+\theta-1}{\theta-1}q^{\theta}(1-q)^{y-j}$$

$$where\ y\geq 0 \text{ and } p=\frac{\alpha}{\mu+\alpha}$$

Note that the modified Poisson stopped sum Poisson distribution is equivalent to a Poisson distribution, and hence is not considered here.

Research Questions

1. Do stopped sum models fit citation count data better than discretised lognormal and negative binomial models?
2. If so, which stopped sum model produces the most consistent results?

Methods

Data from 20 different subject areas were selected from Scopus in order to assess the models for a wide range of different disciplines. This is important because citation patterns are known to vary considerably between disciplines. This data has previously been analysed in Thelwall and Wilson (2014). Each subject area is a single Scopus category and consists of all documents of type article that were published in 2004, giving ten years for the articles to attract citations.

Fitting statistical models

The models were fitted using the R software (R Core Team, 2014). The MASS package (Venables & Ripley, 2002) was used to fit the negative binomial distribution. As there are no known statistical packages readily available to model the proposed stopped sum distributions, the parameters of the distributions were estimated by maximum likelihood estimations methods. AIC is a commonly used statistic for model selection, the model with the lowest AIC usually being regarded as the model that best fits the data (Bozdogan, 2000).

$$AIC = -2 \times \log(L) + (2 \times p)$$

Hence the AIC may be regarded as a penalised version of the loglikelihood, where L is the likelihood of the model and p is the number of parameters estimated. For example, both the Poisson stopped sum NB and NB stopped sum Poisson will have p=3, as there is one parameter (λ) in the Poisson wave and two parameters (NB mean, μ and size, α) in the NB wave. The NB stopped sum NB model will have p=4 as two parameters (μ and α) are estimated in each of the NB waves. Whilst opinions differ, when selecting the 'best' model, it has been suggested that a difference of 6 between the AICs will be large enough to imply a significant difference between the models (Burnham & Anderson, 2003).

Standard errors

Standard errors were computed to reflect the precision with which the proposed statistical models estimate the relevant parameters (Dodge, 2003, p. 386). For the negative binomial models, standard errors were obtained directly from the model fitting software. For the discretised lognormal, the standard errors were obtained by bootstrapping.

For other models the standard errors were calculated using the Hessian matrix, which is the matrix of the second derivatives of the log-likelihood function. The Hessian matrix can also be obtained whilst estimating the parameters for the corresponding distributions using the optim function in R (R Core Team, 2014). Suppose that L represents the log-likelihood function of a stopped sum distribution with two parameters, say λ and μ, then the Hessian matrix is given by $\begin{pmatrix} \frac{\partial^2 L}{\partial \lambda^2} & \frac{\partial^2 L}{\partial \mu \partial \lambda} \\ \frac{\partial^2 L}{\partial \mu \partial \lambda} & \frac{\partial^2 L}{\partial \mu^2} \end{pmatrix}$, and the standard errors for λ and μ are calculated as the square root of the main diagonal of the inverse of the negative Hessian matrix (Ruppert, 2011, pp. 166–167). At 95% confidence interval can be computed by parameter estimate ± 1.96*standard error.

Results

The modified negative binomial stopped sum negative binomial distribution (NBNB) produced the lowest AIC for 13 out of 20 subjects. The next most successful models are the NB stopped sum NB and the discretised lognormal. The Poisson stopped sum NB and the modified NB stopped sum Poisson each fitted 'best' for only one subject (see Table 3 in Appendix).

Parameter estimates for stopped sum distributions

The estimated parameters for Tourism and Soil will be discussed for the proposed stopped sum distributions. These subjects were selected as they are examples of subjects, which return parameter estimates and errors for all the fitted distributions. From Table 1, when Tourism is fitted with the Poisson stopped sum NB model, one wave follows the Poisson distribution with mean, λ=3.22, whilst the other wave follows a negative binomial distribution with mean, μ=18.77 and size, α=0.57; thus the negative binomial wave has a variance of 640.19, since the negative binomial variance equals $\frac{\mu^2}{\alpha} + \mu$. However, when fitted with the NB stopped sum Poisson model, one wave follows a negative binomial distribution with mean, μ=21.53, size, α=0.98, and variance=495.77, whilst the other wave follows a Poisson distribution with mean, λ=0.01. The estimated means (μ) in both negative binomial waves are relatively larger than the estimated means (λ) in the Poisson waves, suggesting that the majority of citation counts for Tourism derive from the negative binomial wave. This supports the interpretation that the two waves occur simultaneously, instead of sequentially, as mentioned above. It is also interesting to note that the sum of the estimated means from the Poisson waves and negative binomial waves of these stopped sum models are approximately equal to the estimated mean when Tourism is fitted solely with the negative binomial model.

When fitted with the NB stopped sum NB model, the estimated mean for Tourism in the primary NB wave (13.48) is larger than that of the secondary NB wave (8.25), suggesting that the majority of citation counts for Tourism derive from the primary wave. Furthermore, the sum of the estimated means from the NB stopped sum NB model for Tourism is also approximately equal to the estimated mean when Tourism is fitted with the negative binomial model only.

Similar results were obtained for Soil. When citation counts for Soil are fitted with the Poisson stopped sum NB model and NB stopped sum Poisson model, the mean estimates in the NB waves are much larger than those of the Poisson waves, suggesting that the majority of citation counts from Soil derive from the NB wave. Moreover, the sum of the estimated means for the stopped sum models is approximately equal to the estimated mean for the negative binomial model only (which is 16.93).

Table 1. Estimated parameters for the NB, Poisson stopped sum NB, NB stopped sum Poisson and NB stopped sum NB models.

	Negative binomial		*Poisson stopped sum NB*			*NB stopped sum Poisson*			*NB stopped sum NB*			
Sub.	*mu*	*size*	λ_1	*mu2*	*size2*	*mu1*	*size1*	λ_2	*mu1*	*size1*	*mu2*	*size2*
Tour.	21.53	0.98	3.22	18.77	0.57	21.53	0.98	0.01	13.48	1.30	8.25	0.10
Soil	16.93	0.74	2.27	16.09	0.56	16.87	0.74	0.06	13.78	0.82	3.46	0.04

Table 2 compares estimated parameters for the NB distribution against those of the modified stopped sum distributions. For the modified versions, the estimates of the Poisson stopped sum NB are similar to those of the NB stopped sum Poisson distributions. Similarly to the stopped sum distributions, Tourism and Soil depends largely on the wave that derives from

the NB distribution, as the λ estimates are relatively lower than the mu estimates. Furthermore, the sum of the two mu estimates for the modified NB stopped sum NB distributions (21.533 and 16.931) are also similar to the estimates from the NB distribution.

Table 2. Estimated parameters for the NB, modified Poisson stopped sum NB, modified NB stopped sum Poisson and modified NB stopped sum NB models.

	Negative binomial		*Modified Poisson stopped sum NB*			*Modified NB stopped sum Poisson*			*Modified NB stopped sum NB*			
Subj.	*mu*	*size*	λ_1	*mu2*	*size2*	*mu1*	*size1*	λ_2	*mu1*	*size1*	*mu2*	*size2*
Tour.	21.53	0.98	1.41	20.12	0.75	20.12	0.75	1.41	14.75	0.35	6.79	1.17
Soil	16.93	0.74	0.11	16.82	0.72	16.81	0.72	0.11	4.92	0.08	12.01	0.75

Standard errors for stopped sum distributions

Figures 1 and 2 show the mean and size estimates for the primary and secondary waves of the modified NB stopped sum NB distributions. Visual, Literature and Rehab were excluded as standard errors could not be obtained as a result of a singular hessian matrix.

Although the modified NB stopped sum NB distribution gave the lowest AIC, the model produced very large standard errors, resulting in large confidence intervals, as shown in Figures 1 and 2, indicating that this modified NB stopped sum NB model is impractical. This result could possibly be due to the nature of citations, which differs from that of the larvae studied by Neyman. With larvae and their offspring it is clear which wave of population a larvae originates from, this is not the case with citations – usually it will be far from clear cut which wave a given citation might belong to, which in turn leads to difficulty estimating the mean number of citations for that wave, and hence the large associated standard errors.

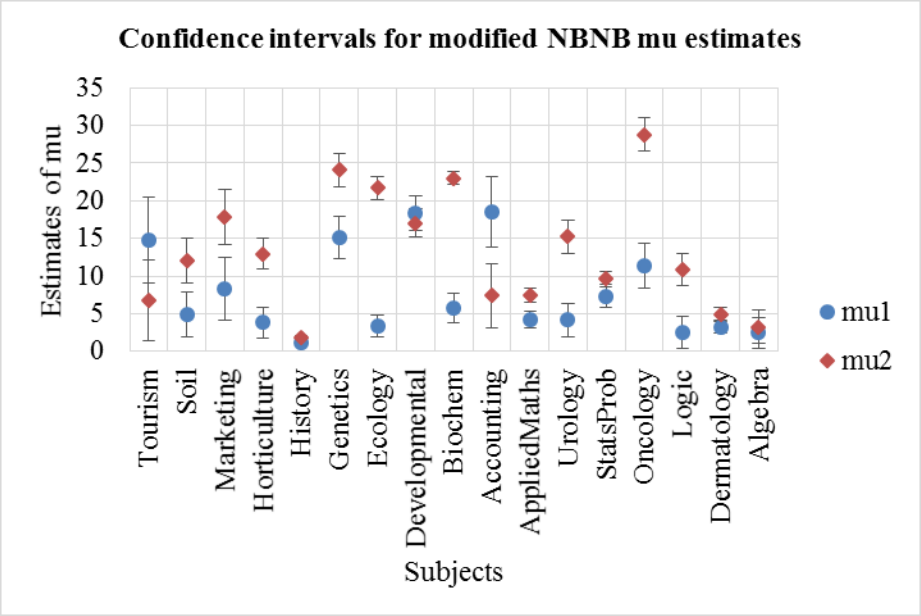


Figure 1. Mean (mu) estimates for the modified NB stopped sum NB distribution for both primary and secondary waves with 95% confidence intervals.

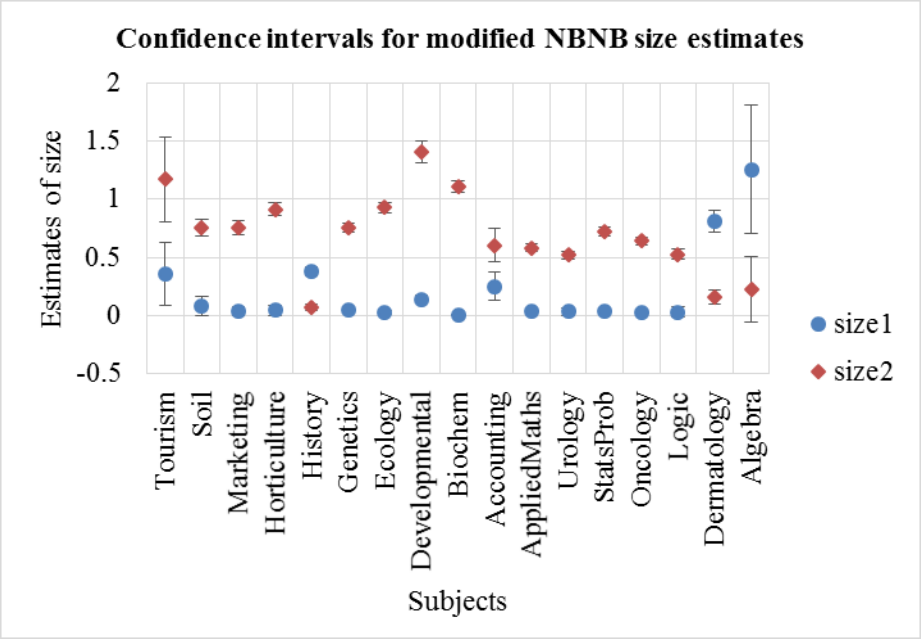


Figure 2. Size estimates for the modified NB stopped sum NB distribution for both primary and secondary waves with 95% confidence intervals.

A further examination of the modified NBNB stopped sum model was carried out with simulations using some known fixed parameters, and similar results were obtained. Moreover, simulations were carried out on all the other stopped sum models and similar results were also obtained for the NBNB stopped sum distribution. Hence it can be concluded that both the stopped sum and modified NBNB stopped sum models are impractical when modelling data with no covariates. Further studies should be conducted to see if adding covariates would change the reliability of the model.

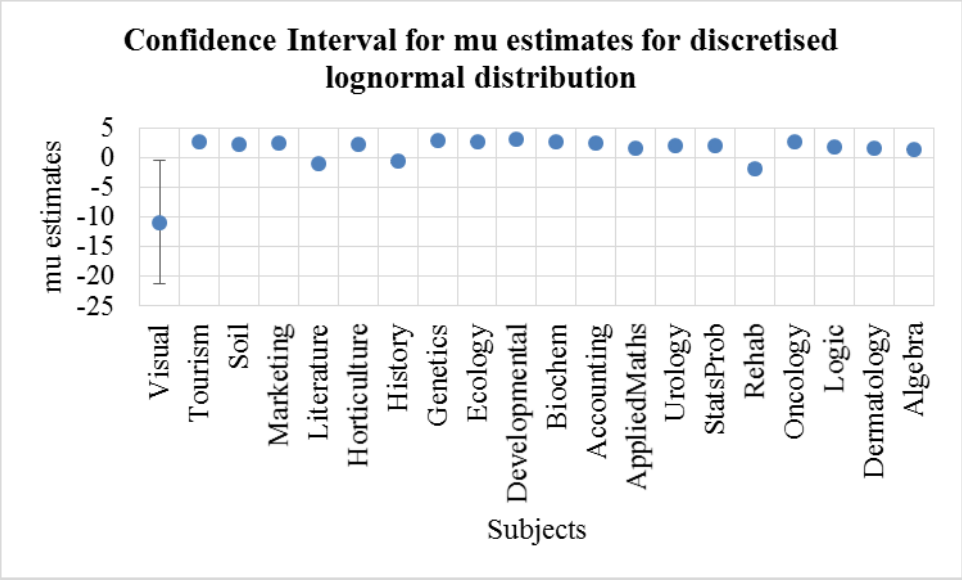


Figure 3. Mu estimates for the discretised lognormal distribution with 95% confidence intervals.

On the other hand, the 95% confidence interval for all subjects except Visual for the discretised lognormal distribution (Fig. 3) are much narrower compared to that of the

modified NB stopped sum NB distribution. This indicates that the discretised lognormal distribution is more suitable in practice.

Conclusions

This paper tested stopped sum distributions for modelling citation data for the first time and also introduces a modification to allow the 'waves' to occur simultaneously rather than sequentially. However, given that the standard errors for the stopped sum distribution tend to be very large it is doubtful whether these distributions are useful for citation data even though they produce the lowest AIC. For example, out of all the tested distributions, the modified NB stopped sum NB distribution produced the lowest AIC, but the large standard errors suggests that it is an unsuitable model as its parameter estimates are too unreliable for predictions or conclusions based upon the model to be meaningful.

Overall, the results suggest that for covariate free data, the discretised lognormal distribution is much more suitable for regressing citation data from a single subject and year. Nevertheless, on a theoretical level, the good fits found for some of the stopped sum models give evidence that there are (at least) two important and separate processes that govern the citing practices of authors. For one of these processes, existing citations are irrelevant for new citations, and for the other, they are relevant.

References

Bookstein, A. (2001). Implications of ambiguity for scientometric measurement. *Journal of the American Society for Information Science and Technology*, *52*(1), 74–79. doi:10.1002/1532-2890(2000)52:1<74::AID-ASI1052>3.0.CO;2-C

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? *Journal of Informetrics*, *6*(1), 11–18.

Bozdogan, H. (2000). Akaike's Information Criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, *44*(1), 62–91. doi:10.1006/jmps.1999.1277

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multi-model inference: A practical information-theoretic approach* (2nd ed., p. 520). Springer.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. doi:10.1137/070710111

De Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. doi:10.1002/asi.4630270505

Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. *Journal of Informetrics*, *7*(4), 861–873. doi:10.1016/j.joi.2013.08.006

Dodge, Y. (2003). *The Oxford dictionary of statistical terms*. (S. D. Cox, D. Commenges, A. Davison, P. Solomon, & S. Wilson, Eds.) (1st ed., p. 498). Oxford: Oxford University Press.

Hesse, M. B. (1953). Models in Physics. *The British J. for the Philosophy of Science*, *4*(15), 198–214.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distribution* (3rd ed., p. 672). Wiley-Interscience.

Maurseth, P. B., & Verspagen, B. (2002). Knowledge spillovers in Europe: A patent citations analysis. *Scandinavian Journal of Economics*, *104*(4), 531–545. doi:10.1111/1467-9442.00300

Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics*, *10*(1), 35–57. doi:10.1214/aoms/1177732245

R Core Team. (2014). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Ruppert, D. (2011). *Statistics and data analysis for financial engineering* (p. 638). New York: Springer.

Thelwall, M., & Wilson, P. (2014a). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, *8*(4), 824–839. doi:10.1016/j.joi.2014.08.001

Thelwall, M., & Wilson, P. (2014b). Regression for citation data: An evaluation of different methods. *Journal of Informetrics*, *8*(4), 963–971. doi:10.1016/j.joi.2014.09.011

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Appendix

Table 3. AIC for all subjects for each distribution

Subjects	*Discretised lognormal*	*Negative binomial*	*Poisson stopped sum NB*	*NB stopped sum Poisson*	*NB stopped sum NB*	*Modified Poisson stopped sum NB*	*Modified NB stopped sum Poisson*	*Modified NB stopped sum NB*	*Number of articles*
Visual	7902	7928	7916	7930	7865	7920	7920	**7865**	4096
Tourism	4956	4980	4980	4982	4969	4964	4964	**4955**	608
Soil	33470	33344	33458	33345	33287	33344	33344	**33282**	4347
Marketing	**12917**	13073	13025	13073	12941	13015	13015	12932	1550
Literature	11624	11635	**11618**	11637	11622	104485	11624	25449	5000
Horticulture	23058	23093	23165	23095	23001	23067	23067	**22992**	3009
History	19797	19994	19849	19996	19824	19880	19880	**19795**	5000
Genetics	45622	46014	45997	46002	45474	45982	45982	**45471**	5000
Ecology	42787	42343	42441	42335	42253	42366	42793	**42240**	5000
Developmental	40985	41604	41340	41558	40979	41385	41385	**40956**	4541
Biochem	42901	43690	43540	43638	**42675**	43659	43659	42680	5000
Accounting	9927	9933	9924	9931	9914	9929	9929	**9896**	1178
AppliedMaths	33504	33739	33704	33741	33460	33685	33685	**33441**	5000
Urology	38932	38621	38793	38623	**38560**	38623	38623	38563	5000
StatsProb	**36696**	37416	37177	37418	36742	37186	37186	36706	5000
Rehab	28086	27531	27622	27533	27628	27483	**27483**	28322	5000
Oncology	42577	42620	42679	42607	**42196**	42660	42684	42225	4646
Logic	32258	32044	32164	32046	32012	32045	32045	**32010**	4547
Dermatology	19608	19774	19671	19776	19675	19692	19692	**19606**	3184
Algebra	**2968**	2991	2973	2993	2977	2978	2978	2972	528

Table 4. Estimated parameters of negative binomial distribution with the stopped sum distributions

	Negative binomial		*Poisson stopped sum NB*			*NB stopped sum Poisson*			*NB stopped sum NB*			
Subjects	*mu*	*size*	*lambda1*	*mu2*	*size2*	*mu1*	*size1*	*lambda2*	*mu1*	*size1*	*mu2*	*size2*
Visual	0.66	0.17	0.28	1.61	0.34	0.66	0.17	0.00	0.60	0.19	0.26	0.00
Tourism	21.53	0.98	3.22	18.77	0.57	21.53	0.98	0.01	13.48	1.30	8.25	0.10
Soil	16.93	0.74	2.27	16.09	0.56	16.87	0.74	0.06	13.78	0.82	3.46	0.04
Marketing	26.13	0.63	2.63	24.97	0.43	26.02	0.62	0.12	20.34	0.76	6.16	0.01
Literature	0.79	0.32	0.40	1.18	0.33	0.79	0.32	0.00	0.41	9.22	1.16	0.31
Horticulture	16.72	0.83	2.52	15.15	0.54	16.71	0.83	0.01	14.27	0.94	2.62	0.02
History	2.90	0.30	0.75	4.08	0.27	2.90	0.30	0.00	1.26	0.75	3.12	0.12
Genetics	39.23	0.61	2.71	38.78	0.50	38.96	0.60	0.28	24.30	0.80	15.85	0.04
Ecology	25.02	0.86	2.52	24.17	0.79	24.73	0.84	0.31	22.61	0.76	2.60	0.32
Developmental	35.45	0.93	4.03	31.86	0.60	34.56	0.86	0.90	17.95	1.52	17.73	0.12
Biochem	28.81	0.84	3.21	26.60	0.61	28.08	0.79	0.75	22.86	1.12	6.09	0.01
Accounting	25.89	0.64	2.46	25.36	0.50	25.66	0.63	0.26	12.93	0.87	14.03	0.12
AppliedMaths	11.71	0.50	1.68	12.20	0.39	11.71	0.50	0.00	8.20	0.63	4.28	0.03
Urology	19.39	0.51	1.80	20.69	0.50	19.47	0.51	0.00	15.49	0.56	4.60	0.03
StatsProb	16.93	0.54	2.12	16.62	0.36	16.93	0.54	0.00	10.50	0.77	7.21	0.03
Rehab	9.29	0.23	0.83	14.56	0.37	9.28	0.23	0.00	0.83	89.55	14.56	0.37
Oncology	40.23	0.55	2.34	41.68	0.53	39.94	0.54	0.33	25.50	0.68	16.33	0.05
Logic	13.40	0.53	1.67	14.21	0.49	13.37	0.53	0.00	11.59	0.56	2.19	0.02
Dermatology	8.07	0.65	1.79	7.44	0.37	8.06	0.65	0.01	1.83	41.25	7.39	0.36
Algebra	5.75	0.90	1.90	4.46	0.37	5.74	0.90	0.01	1.94	42.31	4.41	0.36

Table 5. Estimated parameters of negative binomial distribution with the modified stopped sum distributions

	Negative binomial		*Modified Poisson stopped sum NB*			*Modified NB stopped sum Poisson*			*Modified NB stopped sum NB*			
Subjects	*mu*	*size*	*lambda1*	*mu2*	*size2*	*mu1*	*size1*	*lambda2*	*mu1*	*size1*	*mu2*	*size2*
Visual	0.66	0.17	0.04	0.62	0.14	0.62	0.14	0.04	0.60	0.19	0.06	0.00
Tourism	21.53	0.98	1.41	20.12	0.75	20.12	0.75	1.41	14.75	0.35	6.79	1.17
Soil	16.93	0.74	0.11	16.82	0.72	16.81	0.72	0.11	4.92	0.08	12.01	0.75
Marketing	26.13	0.63	1.02	25.11	0.50	25.11	0.50	1.02	8.35	0.03	17.78	0.76
Literature	0.79	0.32	11.82	11.99	0.00	0.72	0.24	0.07	4.65	2.71	3.85	0.00
Horticulture	16.72	0.83	0.50	16.24	0.73	16.18	0.72	0.53	3.82	0.05	12.90	0.91
History	2.90	0.30	0.20	2.70	0.21	2.70	0.21	0.20	1.08	0.38	1.82	0.07
Genetics	39.23	0.61	0.43	38.81	0.57	38.81	0.57	0.43	15.12	0.04	24.12	0.75
Ecology	25.02	0.86	0.00	23.60	0.91	18.21	0.80	0.00	3.36	0.02	21.67	0.93
Developmental	35.45	0.93	2.56	32.89	0.69	32.89	0.69	2.56	18.40	0.14	17.04	1.41
Biochem	28.81	0.84	0.69	28.12	0.76	28.12	0.76	0.69	5.79	0.01	23.02	1.11
Accounting	25.89	0.64	0.34	25.55	0.60	25.55	0.60	0.34	18.48	0.25	7.40	0.60
AppliedMaths	11.71	0.50	0.28	11.44	0.44	11.44	0.44	0.28	4.26	0.04	7.45	0.58
Urology	19.39	0.51	0.02	19.37	0.51	19.37	0.51	0.02	4.17	0.03	15.21	0.52
StatsProb	16.93	0.54	0.78	16.16	0.41	16.15	0.41	0.78	7.19	0.04	9.74	0.72
Rehab	9.29	0.23	0.09	9.19	0.21	9.19	0.21	0.09	5.71	0.00	25.74	0.20
Oncology	40.23	0.55	0.00	45.66	0.54	34.70	0.57	0.00	11.43	0.02	28.81	0.64
Logic	13.40	0.53	0.04	13.37	0.52	13.37	0.52	0.04	2.52	0.03	10.88	0.53
Dermatology	8.07	0.65	0.60	7.48	0.47	7.48	0.47	0.60	3.22	0.81	4.85	0.16
Algebra	5.75	0.90	0.84	4.91	0.55	4.91	0.55	0.84	2.48	1.25	3.27	0.23

Differences in Received Citations over Time and Across Fields in China

Siluo Yang[1], Junping Qiu[1], Jinda Ding[2] and Houqiang Yu[1]
[1] *58605025@qq.com*
School of Information Management, Wuhan University, Wuhan 430072, China
[2] *djdhyn@126.com*
Department of library, information and archives, Shanghai University, Shanghai 200444, China

Abstract
We analyse and compare the difference in discipline level of the received citations over a period of time and across fields in China by implementing the diachronous methods of bibliometrics. The citations of 896,645 papers from the Chinese Citation Database (1994 to 2013) that comprised four disciplines, namely, Philosophy, Library and Information Science (LIS), Physics, and Mechanical Engineering, are collected. Results indicate the following conclusions. First, the received citations strongly differ across various fields and over time. Second, the average of the received citations after a given year has an identical change. The number initially increases rapidly, and then declines slightly in the recent years. Uncitedness rate decreases in the early stage of the study period, whereas the rate stabilises or increases slightly in the recent years. Third, the average of the received citations peak after seven and nine years in mechanical engineering and philosophy, respectively, whereas both physics and LIS peak after three years. The span from the year of publication to the cited peak is relatively stable in LIS for 20 years. However, the span decreases in the early stage of the study period, and then stabilizes in the recent years for the other three disciplines. Recently, all four disciplines indicate relatively consistent citation trends. These results highlight the recent evolution of Chinese research systems towards relatively steady states.

Conference Topics
Citation and Co-citation Analysis; Country-level Studies

Introduction
Citing is a fundamental academic behavior among scholars. Citing shows the use of previous research, presents the processes of scientific inheritance and communication, and manifests respect for other scientific researchers (Yang et al., 2010). In the 20th century, citing other works became common in writing scholarly or scientific papers (Kaplan, 1965). Analysis of citing behavior is an important field and method in information science. At present, citation analysis is widely used to evaluate scientific works, initiate scholarly communication, analyse academic behavior, and process information retrieval (Hirsch, 2005; Hammarfelt, 2011; Ketzler & Zimmermann, 2013; Ding et al., 2014).

Information scientists have extensively investigated the distributions and changes of citing behavior (Finardi, 2014). According to the general theory of human behavior, we design the framework of citation behavior analysis. Figure 1 shows a four-dimensional model of citing behavior analysis. This model integrates analytical dimensions in terms of level (who), method (how), perspective (when), and content/topic (what and why). The combination of different dimensions can display the citing behavior in multiple functions and aspects. According to the analysis perspective, citing behaviors mainly include synchronic and diachronic distributions that fundamentally designate and refer to completely different characteristics of scientific literature (Nakamoto, 1988). Synchronic analysis is generally more common than other analytical approaches to citing behaviors (Heistermann et al., 2014). Line and Sandison (1974) proposed the diasynchronous analysis, a kind of synchronous analysis, which studies the synchronous distribution of cited documents at different time periods. Larivière et al. (2008) studied the evolution of yearly synchronous scores computed from 1900 to 2004. Their study showed the increase in average and median ages of cited literature, whereas the price index decreases over time. However, Egghe (2010) argued that

"Larivière' results do not have a special informetric reason but that they are just a mathematical consequence of a widely accepted simple literature growth model."

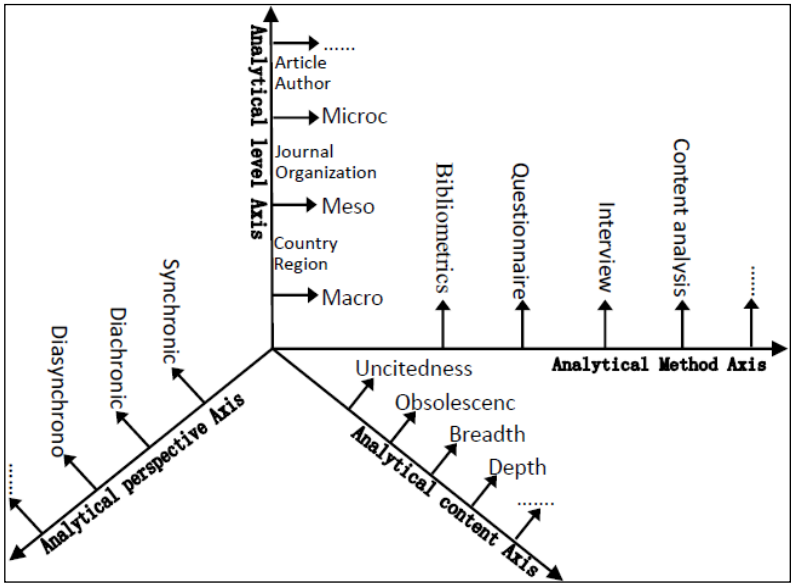


Figure 1. Four-dimensional model of citing behavior analysis.

Diachronic analysis consists of analyzing the distribution of citations gained over time by a publication within a given year by subsequent literature. However, this analysis is generally ignored because of the unavailability of data and the difficulty in implementation. Nevertheless, diachronic analysis has certain advantages, including its appropriateness for citation distribution (Bouabid & Larivière, 2013). Some papers focused on citation distribution and its evolution based on diachronic analysis. First, Finardi (2014) plotted the mean received citations against the time gap (in years) between the publication of the cited article and received citations. Afterwards, he established that citations follow different trends in various fields or disciplines. Some scholars studied the time gap between the publication of a scientific work, as well as the first citation it received (Bornmann & Daniel, 2010). Egghe et al. (2011) proposed a first-citation-speed index, which is utilised for a set of papers, based on the number of publication times and the initial citation. Bouabid and Larivière (2013) recently used a diachronous model to study life expectancy changes and to identify variations in life expectancy between countries and scientific fields based on the citations received by papers.

Second, studies focused on one intriguing aspect of citation analysis, which is the distribution of uncitedness. Schwartz (1997) defined uncitedness as the inability of papers to be cited in citation indexes within five years after their publication. Stern (1990) claimed that although most papers are eventually cited, a number of papers in various scientific disciplines are never cited. Pendlebury (1991) established that the lowest rates of uncitedness occurred among physics and chemistry papers. Garfield (1998) opined that knowing the number of uncited papers and clearly defining these prior to interpretation are important. Egghe et al. (2011) discovered that Nobel laureates and Fields medalists cover a large fraction (10% or more) of uncited publications. A positive correlation was found between the h-index and the number of uncited articles as well.

Lastly, some researchers investigated changes in citing behavior in the context of the overall situation. Larivière et al. (2008) studied the evolution of the aging phenomenon, particularly on how the age of cited literature has changed in over 100 years of scientific activity. They discovered that the average and median ages of cited literature underwent several changes during the period. Evans (2008) showed that as more journal issues are offered online, fewer

journals and articles are cited, and a large part of these citations refer to a small number of journals and articles. Larivière et al. (2009) challenged the conclusion of Evans (2008) and argued that the dispersion of citations is, in fact, increasing. Yang et al. (2010) studied citing behavior by employing three measures of citation concentrations using the Chinese Citation Database (CCD). The concentration of citations was claimed to be declining, and cited papers are broad and diverse. In our view, the diachronic analysis of citation behaviour has two main aspects: the citation change of papers **published** in different years and the citation change of papers **cited** in different years. However, scholars have yet to analyse received citations over a long period of time and across various fields in China.

Since 1978, when the reforms and opening up policies were implemented, China has experienced unprecedented changes. Chinese science exhibited remarkable progress as well. With the popularity of the Internet and development of computer networks in recent years, social environment and scientific research underwent significant changes (Zhou et al., 2009; Yang, 2010). In China, What is the exact general distribution of citation? What are the advancements in citation behaviour in Internet era? Are there differences in citation behaviour across various scientific fields in China?

Our research aims to discover the citation distribution trends over time in different scientific fields in China. Specifically, we focus on the following: (1) the general differences of citation distributions among disciplines, (2) the citation or uncitedness characteristic of papers **published** in different years (For example, papers published in 2000, 2001, 2002... are cited respectively after 5 years, that is, 2004, 2005, 2006...), and (3) the citation characteristic of papers **cited** in different years (For example, a paper published in 2000 is cited in 2000, 2001, 2002...).

Methods and data

Data sample

China has the following citation databases: Chinese Science Citation Database, Chinese Social Sciences Citation Index, Chinese Humanities and Social Science Citation Database, Chinese Science and Technology Paper Citation Database, CQVIP Citation Database, and CCD. In this study, we used CCD as our data resource. CCD collects all references for the China National Knowledge Infrastructure (CNKI) and performs deep data excavation on the citation relationship between studies. Furthermore, CCD provides a citation statistical analysis function based on authors, institutions, publishers, and journals. CCD is one of the products of CNKI (http://www.cnki.net/), and the database covers 6,642 journals while its web version has more than 8200 journals. CCD only contains Chinese journals. Tsinghua University and Tsinghua Tongfang Holding Group first launched CNKI in June 1999. CNKI is the key project of the national informatization construction in China, which established the most comprehensive system of academic knowledge resources (CNKI, 2014). CNKI comprises more than 90% of the knowledge resources in China, which is the broadest in titles and type coverages, as well as the most in-depth in years of coverage in the country. The oldest paper dates back to 1979. This database is updated daily.

We analysed publications and citations from 1994 to 2013, which spans 20 years, to identify publishing and citing patterns at the discipline level. This period was chosen because it is recent and 20 years is sufficiently long in performing the comparisons. All papers from 1994 to 2013 were collected in July 2014. The papers covered four disciplines based on the classification system of CNKI: philosophy, library and information science (LIS), physics, and mechanical engineering. These disciplines, respectively, represent the humanities, social sciences, science, and engineering. The LIS is somewhat peculiar given its evolution towards forms of publication and citation that are closer to the hard sciences. However, we are highly

familiar with this subject because many related research also use LIS as an example. We considered citation types including journals, books, dissertations, meetings, and newspapers. To verify the consistency of the data, we downloaded the data again after a week. We consulted the database provider several times regarding data access issues (i.e., the exact time of database upgrade per day and the range and scope of the citation database). The database is only appropriate for a country, and only reflects the situation in China. Thus, results may differ when international databases are used for comparison.

Methodology

Three aspects of related indicators of received citations across fields and over time are presented. The three aspects involve six equations.

Generally, the papers published in year *i* were cited in year *j*. Both *i* and *j* are from 1994 to 2013, and *j*>=*i*. P_j represents the number of papers published in year *i*. C_j represents the number of citations in year *j*, which were obtained from the papers published in year *i*. We analyse the general situation of the papers cited and published every year and analysed them using the following equations.

1) The average number of citations obtained by each paper from the published year to year *m* (*m* equals to 2013 in this study), and the average number of citations obtained by each paper in each year.

F1: $\frac{\sum_{j=i}^{m} C_j}{P_i}$ expresses the average number of citations obtained by each paper from the published year to year *m*.

F2: $\frac{F1}{n}$ expresses the average number of citations obtained by each paper in each year, where *n* represents the distance between published years *i* and *m*, that is, *n* = *m*-*i*+*1*.

2) Percentage of uncited papers within a given time period.

F3: $(1 - \frac{P^c}{P_i}) \times 100$, P^c is the number of papers cited at least once within a given time period after publication. The time span of one, two, or all years are set. In the case of three years, all papers published in 2003 are referred to as P_{2003}. We attempted to determine how many of the papers are uncited after three years (between 2003 and 2005). The time period ends in 2005 for the three-year perspective (including the publication year).

3) Time evolution of the average received citations.

We obtain Equation 4 by the methodology described in Finardi (2014).

F4: $MEAN_k = \frac{C_j}{P_i}$ expresses the average number of citations in year *j*, which were obtained by the papers published in year *i*. That is, the received citations of each paper in year *j* after being published for *x* (*x*=*j*-*i*+*1*) years (including the published year). At a constant value of *x*, which can be changed or assigned between 0 and 19 in the empirical analysis, we can obtain a series of $MENN_k$. For example, if we set *x* equals to 3, then we have $MEAN_1 = \frac{C_{1996}}{P_{1994}}$, $MEAN_2 = \frac{C_{1997}}{P_{1995}} \ldots MEAN_{18} = \frac{C_{2013}}{P_{2011}}$, where *k* is from 1 to *N* and *N* is dependent on *x* that equals to 2013-1993-*x*+1(*x* is the time distance between the published and cited years *i* and *j*, respectively).

F5: $AMEAN_x = \frac{\sum_{k=1}^{N} MEAN_k}{N}$ expresses the average of means among different occurrences from papers published in several years. By this equation, any possible bias because of the use of citations received in a single year may be avoided. The final result is the plot of $AMEAN_x$ vs. x.

F6: $CAC_x = \frac{\sum_{j=i}^{i+x} C_j}{P_i} \Big/ x$ expresses the cumulating average number of citations that each paper has received during x years, beginning its publication in year i (including the published year). For example, if i equals 2000 and x equals 3, the number of citations received at 2000, 2001, and 2002 from the papers published in 2000 will be summed, and then the cumulating average values of received citations of each paper per year will be calculated.

Result and discussion

Overview

A total of 896,645 papers in philosophy, LIS, physics, and mechanical engineering that were published from 1994 to 2013 were collected. The upper left curve in Figure 2 shows that 41,793,391 papers were published across all fields in CCD for the past 20 years (1994 to 2013). The number of papers steadily increased each year, from 927,684 in 1994 to 3,478,490 in 2013. The curve shows that the growth pattern is an S-shape and has three stages (i.e., slow, rapid, and slow growth). The growth of scientific papers slowed down after 2008. The progress of LIS and philosophy papers remains consistent with those of the other fields. However, a downward trend in physics and a highly irregular trend in mechanical engineering in the recent years are observed. Instead of using typical journals, we selected sample papers in the selected disciplines by an artificial category classification of the database. Numerous papers in China are being published in international journals, especially those in the science and technology field, resulting in changes in the growth rate in Chinese journals.

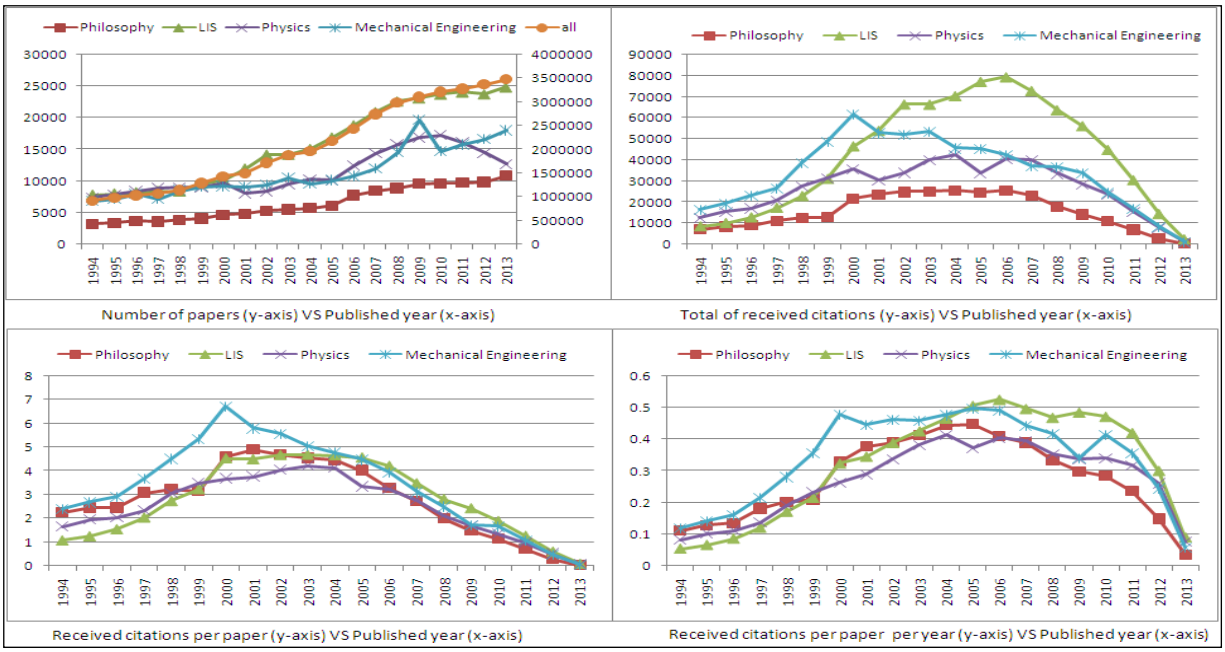


Figure 2. Overall situations of received citations across four disciplines.

Figure 2 shows the overall situations of received citations across four disciplines in CCD. The curves of the received citations exhibit an arch shape (i.e., the middle is high and the end is low). A paper published a long time ago generally has increased chances of receiving citations because of the cumulative phenomenon. However, Figure 2 exhibits the trend of received citations in all four subjects as from increasing in the early periods of the study period to decreasing in the recent years. This phenomenon is caused by two reasons. First, the number of published papers and references for each paper increases each year. The rapid updating of information and the increase in the received citations of each paper can lead to the increase in the number of citations (Price, 1965). Therefore, the cumulative effect of received citations is weakened. Second, people are generally interested in and use the latest research as reference. Researchers strive to make their papers novel. Thus, papers published in previous years have become irrelevant. Figure 2 also exhibits that the received citations of each paper each year (bottom right corner of the figure) eliminate the accumulation phenomenon and display the advantages of papers published in the recent years. The curves of the total received citations and the number of papers published in a specific year are generally consistent. LIS indicated the largest number of papers and received citations in the recent years, whereas philosophy recorded the lowest.

Citation and uncitedness characteristics of papers published in different years

Figures 3 and 4 show the average of the received citations after a paper is published in a given year. In the case of five years window, all papers published in 2000 were taken as the research sample; we determined the average number of times that these papers were cited in 2004. For clarity of presentation, Figure 3 displays only the received citations in four fields after 1, 2, 5, and 10 years. The curves exhibit an identical change (i.e., an initial rapid increase and then a slight decline in the recent years) and indicate that the average of the received citations (published in the recent years) failed to increase. The rapid growth of the average of the received citations in the early stages of the study period changes to a relatively stable development phase because of the slow growth in the number of published papers, the development of the Internet, and the widespread use of open-access and e-print materials. However, whether a special informetric reason or merely a mathematical consequence of a simple literature growth model exists, this phenomenon requires further validation and

investigation (Egghe, 2010). The average of the received citations exhibits significant differences among the four disciplines in various time spans. The maximum value was attained by LIS after one, two, and three years compared with the other three disciplines in each publication year. However, this value slowly decreased, and LIS attained the minimum value each year after 10 years. Physics and mechanical engineering show the exact opposite of LIS. That is, after 10 years, the maximum value of the average of the received citations was achieved.

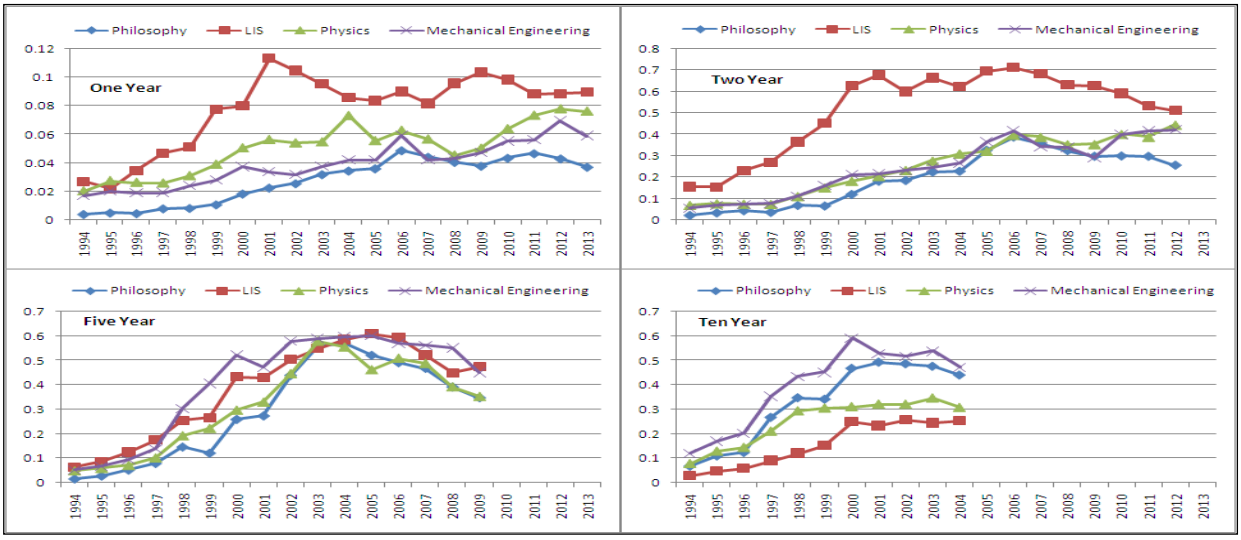


Figure 3. Received citations of each paper each year (y-axis) vs. published year (x-axis) (Part I).

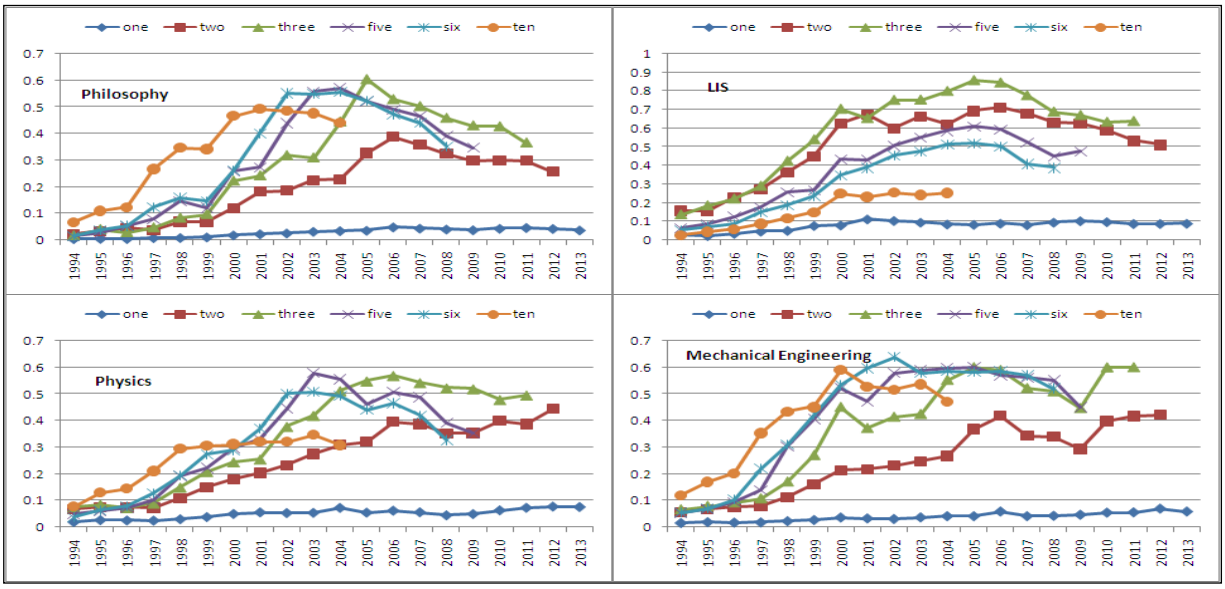


Figure 4. Received citations of each paper each year (y-axis) vs. published year (x-axis) (Part II).

Figure 4 illustrates the received citations by discipline and clarifies the situations of various time spans in each field. Philosophy, physics, and mechanical engineering papers published in the early stages of the study period received more citations in six and ten-year windows than in the recent years. Generally, recently published papers have more citations of papers published from the last three years, which implies that the life expectancy of scientific

literature is generally becoming shorter. Papers on LIS (published almost all year) received more citations in two and three-year windows.

The uncitedness results are presented in four citation windows representing one, two, five, and all years after the publication year. Figures 5 and 6 show that the uncitedness rate generally decreases in the early stages of the study period, and then stabilises or increases slightly in the recent years. This phenomenon is due to the following reason. First is the emergence of databases and networks that provided researchers with additional opportunities to find articles for citation and that allowed equal access to all documents. However, the development of databases has entered a period of relative stability in recent years and the uncitedness rate changes slowly as well. Second, the steady increase in the number of published articles and references for each paper decreases the uncitedness rates in the early stages of the study period. However, the rates of both published articles and references relatively stabilised in the recent years. Third, CCD, which is used and promoted in a wide range of areas, was established in 1999. As CCD became increasingly stable, its data updates became timely in recent years. After the reform, the opening up, and the development of science and technology, research conditions and environments significantly improved. The state of scientific research has become steady in recent years in China.

A number of studies showed that the uncitedness rate is lowest in the sciences, high in the social sciences, and highest in arts and humanities (Hamilton, 1991). However, Figures 5 and 6 display contrasting results. The uncitedness rates in LIS are significantly lower than the other three disciplines in the one-, two-, and five-year citation windows in almost all publication years. A possible reason for this phenomenon is the privileges and required expertise in accessing and using documents (especially online information retrieval) in LIS. Papers published in the recent year exhibit high uncitedness rates for Philosophy in the one-, two-, and five-year citation windows. However, the low uncitedness rates in the all-year citation window showed more documents being cited in this discipline.

Figure 6 shows the uncitedness situation by discipline. The curves exhibit the same trend for all four disciplines. The uncitedness rates in the one-year window are relatively stable, while in the two-year window, the uncitedness rates decrease rapidly and decline sharply in the five-year window. However, the all-year window is special because different results were obtained for papers in different publication years. For example, papers published in 1994, 2000, and 2008 are in the all-year citation window, particularly 20, 14, and 6, respectively. Consequently, the two curves of the five- and all-year windows move gradually closer.

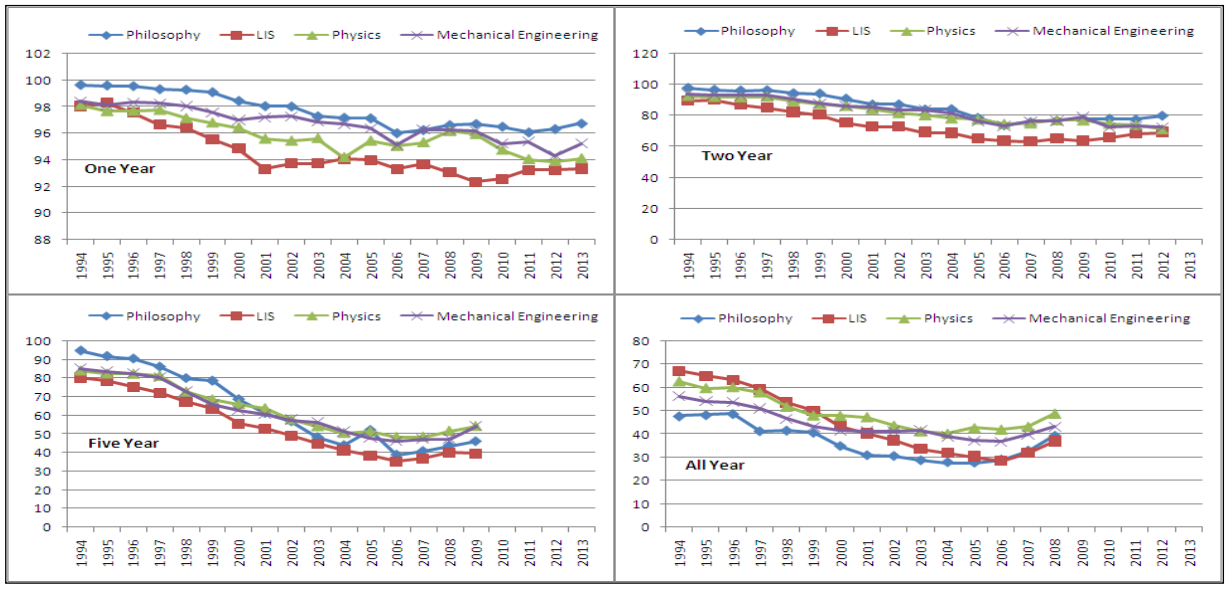


Figure 5. Number of uncited articles (y-axis) vs. published year (x-axis) (Part I).

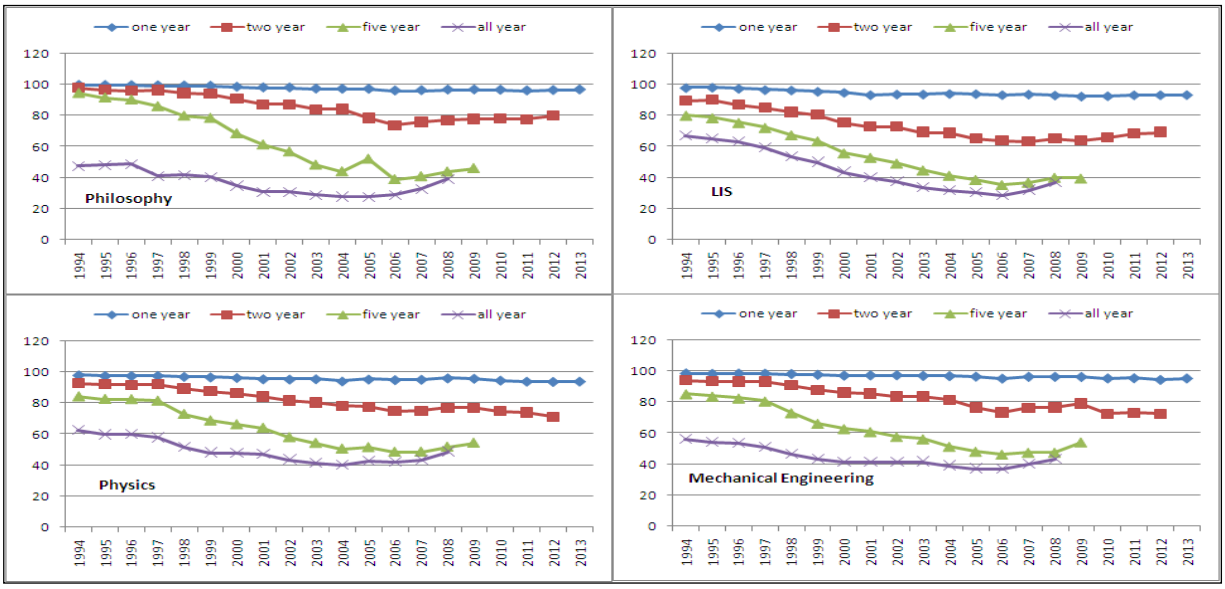


Figure 6. Number of uncited articles (y-axis) vs. published year (x-axis) (Part II).

Citation characteristics of papers cited in different years

Figure 7 shows the mean of the received citations of each paper after a given time period using Equations 5 and 6. The average value avoids possible biases that are caused by using the received citations in a single year. The curve shows the values from 1 to 20 years after publication.

Figure 7 presents the average of the received citations over time. The typical citation curve starts with a rapid increase during the initial years followed by a peak, and then a slow but steady decrease (Larivière et al., 2008). LIS and physics had a similar trend in terms of the average of the received citations. These disciplines peaked at three years after publication, as observed by Finardi (2014) and Bouabid & Larivière (2013). However, physics steadily decreases and LIS rapidly decreases, which created a steep curve. The times of cited peak values are distinct among different disciplines. The trend of mechanical engineering presents a peculiar behaviour because a peak is not exhibited. Instead, the received citations increase in the first three to five years and then stabilise at high values. Citations of mechanical engineering papers continue for a long time after their publication. Figure 7 also suggests that philosophy has a different citation path, with the continuous growth from one to eight years, peak at nine years, and a subsequent slight decrease. This trend is because philosophy information can be accessed and used for a long time, with slow obsolescence.

Figure 7 shows that notable differences exist between the trends of the mean of the received citations in different fields. Consequently, we can conclude that clear differences exist among other specific fields of natural and social sciences. However, further evidence must be obtained by using longer time periods and increasing the number of disciplines compared with that in this study. The maximum values of the average of the received citations peaked after seven years in mechanical engineering and nine years in philosophy. The journal impact factor (IF) only considers citations received in the first two or five years after publication (i.e., 2-years IF or 5-years IF). Thus, high citation values are not captured in the IF computation. The following reasons can explain the particular trends in mechanical engineering and philosophy. Papers published in both disciplines increased from 1994 to 2013, resulting in a parallel growth in the number of citations. Moreover, referring to old literature is preferred in both disciplines, resulting in stable citation curves.

The curves at the right of Figure 7 represent the cumulating value. The curves of the right and left categories in Figure 7 are relatively consistent. However, the curves on the right are smoother than the curves on the left, and the corresponding peaks lag for several years because of the average cumulative effect. For example, in the case of x=3 (x-axis) in Equation 6, we calculated the number of received citations published after one, two, and three years, and then calculated the average values of the received citations of each paper each year.

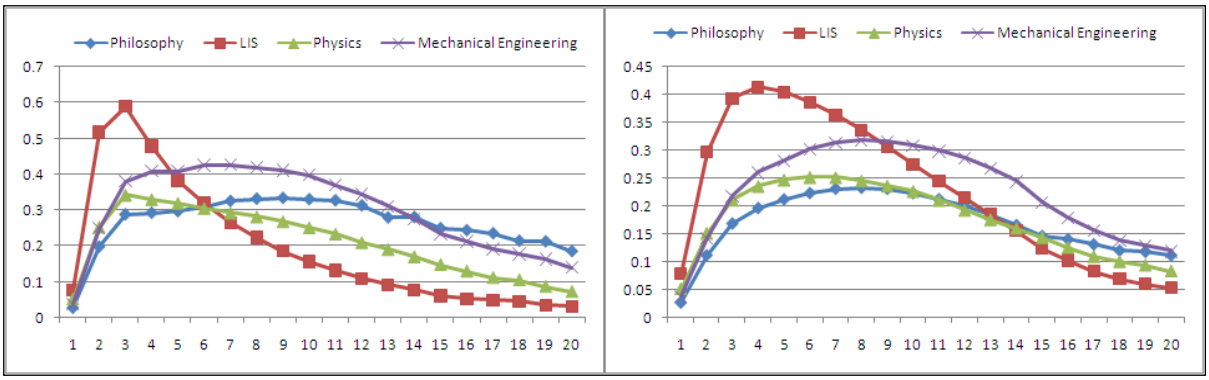


Figure 7. $AMEAN_x$ **(y-axis) vs.** x **(x-axis).**

Figure 8 shows the received citations of each paper each year within the identified time period. We selected the publication years of 1994, 1998, 2002, and 2008 as representatives. The data for the other years of publication showed the same trends. However, these were not included in this paper. The trend in LIS is completely different from those of the other three disciplines. LIS presents a peak at two or three years, which slightly decreases in all cited years. The curves of the other three disciplines are relatively consistent. The received citations of papers published in 1994, 1998, and 2002 increase tremendously and peak in 2006 before slightly decreasing. However, a big difference is observed in the received citations for papers published in different years (i.e., 1994, 1998, and 2002). We can conclude that the early publication years tend to have late citation peaks. For example, the received citations of philosophy papers published in 1994 exhibited their peak 14 years after publication (2007), whereas papers published in 2002 exhibited their peak six years after publication (2007). In general, all four disciplines possess a relatively consistent citation trend in recent years.

Figure 9 shows the situation of the received citations by discipline. Philosophy papers published in the early part of the study period still received many citations. These old papers are not excluded from the science system. Thus, they remain to have a relevant contribution. The citation curves in LIS are consistent in the different cited years. However, the curves of the other three disciplines exhibit a similar trend; papers in these three disciplines became more quickly obsolete in general in recently. Furthermore, many curves peak between 2006 and 2008.

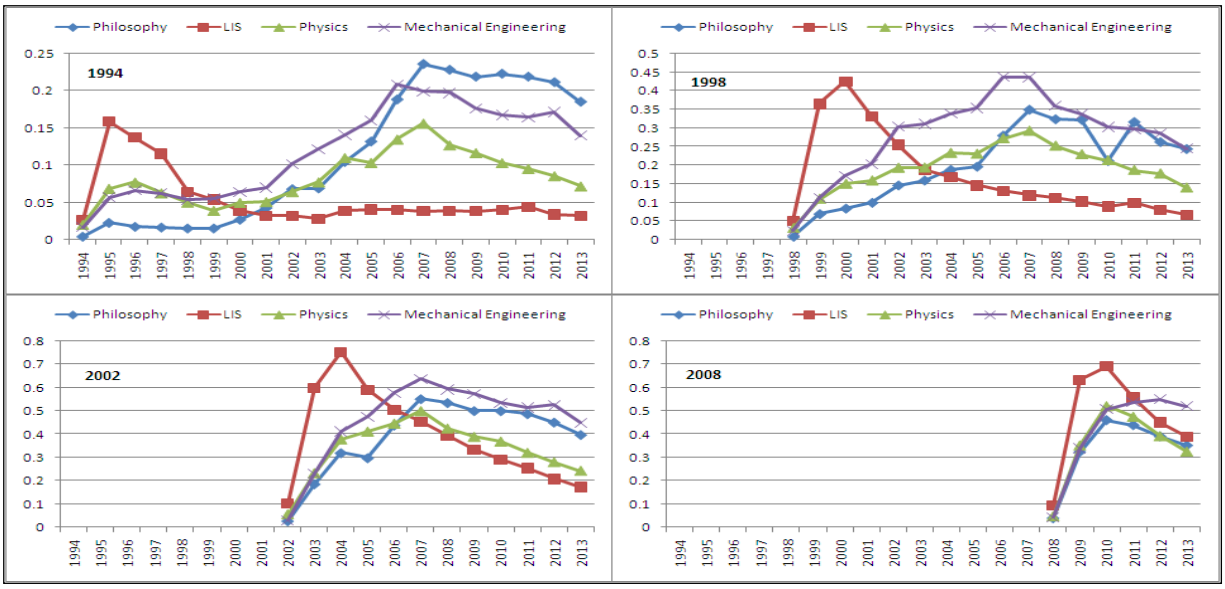


Figure 8. Received citations of each paper each year (y-axis) vs. cited year (x-axis) (Part I).

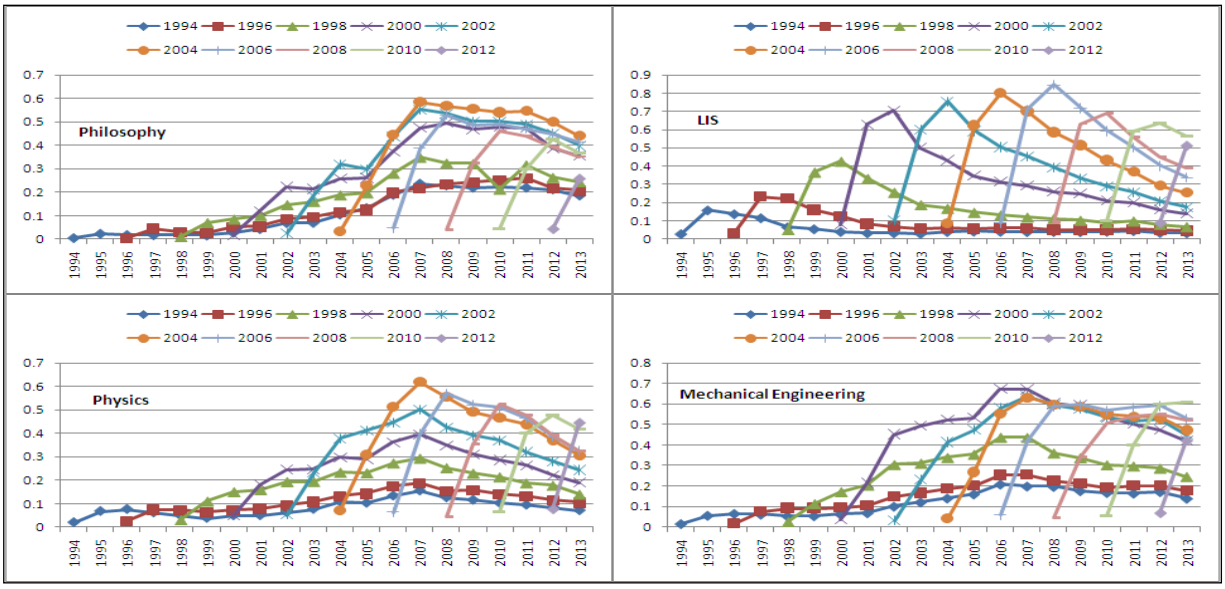


Figure 9. Received citations of each paper each year (y-axis) vs. cited year (x-axis) (Part II).

Conclusion and further research

A total of 896,645 papers on philosophy, LIS, physics, and mechanical engineering, which were published from 1994 to 2013, were collected. This study analysed the differences of these papers in terms of the received citations across fields and over time in China. The following conclusions were derived from the results. First, the growth of published papers is generally S-shaped and undergoes three stages (i.e., slow growth and rapid growth). The curves of the received citations of each paper exhibit an arch shape (i.e., the middle is high and the end is low). The cumulative phenomenon of received citations is not obvious. Second, the average of the received citations in a given year window changes identically, initially increases rapidly, and then slightly decreases in the recent years. The average of the received citations exhibits significant differences among the four disciplines in various time spans. In one-, two-, and three-year windows, a maximum value is observed in LIS in each published

year. The value slowly decreases until the LIS obtains a minimum value within the 10-year windows. However, physics and mechanical engineering exhibit an exactly opposite change. Third, the uncitedness rate generally decreases in the early stages of the study period, but stabilises or increases slightly in recent years. The uncitedness rates in the one-year window are relatively stable, but decreases rapidly in the two-year window and drops sharply in the five-year window. Fourth, notable differences exist among the trends of the mean of the received citations of the different fields. The maximum values of the average of the received citations peak after seven years for mechanical engineering, nine years for philosophy, and three years for both physics and LIS. These results are similar to those obtained by Finardi (*2014*) and Bouabid & Larivière (*2013*). Lastly, citation characteristics of papers cited in different years. LIS citations are completely different from those of the other three disciplines. LIS citations peak at two or three years and then slightly decrease in all cited years. The curves of the other three disciplines are similar. Papers published in the early stages of the study period have a later cited peak. In the recent years, all four disciplines possess a relatively constant citation trend. Generally, Chinese research systems evolve into a relatively steady state from a rapid growth and then change in the early period.

This study has analysed comprehensively the received citations across fields and over time in a systematic manner. As a result, consistent conclusions are drawn. For future research, we intend to perform the following. First is we will measure the received citations at the discipline level by implementing diachronous methods. We will consider synchronic methods and combine the two methods. Aside from the discipline level, other levels (e.g., journals, authors, countries, papers, agencies) will also be analysed. We intend to study citations based on literature units and analyse large-scale samples using probability statistics. Second is we will increase the number of disciplines. We will choose additional representative samples from other disciplines for a comprehensive statistical analysis. Furthermore, we will select other document databases such as international document databases, to verify the pattern and characteristic changes in the received citations. Third is we will increase the level of examination and improve the measured indicators of distribution and evolution of the received citations. The measurement methods of the received citations can be enhanced, and an in-depth analysis of the specific distribution of highly cited papers will be conducted. Lastly, a detailed and in-depth study will be implemented to check the factors that affect citation evolution and examine the cause and effect of these changes (e.g., the effect of the growth in number of papers on received citations). Furthermore, we will determine how to handle the trend and changes in the distribution of the received citations.

Acknowledgments

This research is funded by A Foundation for the Author of National Excellent Doctoral Dissertation of PR China.

References

Bornmann, L. & Daniel, H. (2010). The citation speed index: A useful bibliometric indicator to add to the h index. *Journal of Informetrics*, *4*(3), 444-446.

Bouabid, H., & Larivière, V. (2013). The lengthening of papers' life expectancy: a diachronous analysis. *Scientometrics*, *97*(3), 695-717.

CNKI, http://oversea.cnki.net/kns55/support/ en/about_cnki.aspx

Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X. & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. *Journal of the American Society for Information Science & Technology*.

Egghe, L. (2010). A model showing the increase in time of the average and median reference age and the decrease in time of the Price Index. *Scientometrics*, *82*(2), 243-248.

Egghe, L. Bornmann, & L. Guns R. (2011). A proposal for a first-citation-speed-index. *Journal of Informetrics*, *5*(1), 181-186.

Egghe, L. Guns, R. & Rousseau, R. (2011). Thoughts on uncitedness: Nobel laureates and Fields medalists as case studies. *Journal of the American Society for Information Science and Technology*, *62*(8), 1637-1644.

Evans, J.A. (2008). Electronic publication and the narrowing of science and scholarship. *Science*, *321*(5887), 395-399.

Finardi, U. (2014). On the time evolution of received citations, in different scientific fields: An empirical study. *Journal of Informetrics*, *8*(1), 13-24.

Hamilton, D. (1991). *Research Papers: Who's Uncited Now?* Science, 251:25. Available: http://www.garfield.library.upenn.edu/commentaries/tsv12(14)p10y19980706.pdf

Hammarfelt, B. (2011). Citation analysis on the micro level: The example of Walter Benjamin's Illuminations. *Journal of the American Society for Information Science and Technology*, *62*(5), 819-830.

Heistermann, M., Francke, T., Georgi, C., & Bronstert, A. (2014). Increasing life expectancy of water resources literature. *Water Resources Research*. *50*(6), 5019-5028.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *PNAS*, *102*(46), 16569-16572.

Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. *American Documentation*, *16*(3), 179-184.

Ketzler, R. & Zimmermann, K.F. (2013). A citation-analysis of economic research institutes. *Scientometrics*, *95*(3), 1095-1112.

Larivière, V. Archambault É. & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: from exponential growth to steady-state science (1900-2004). *Journal of the American Society for Information Science and Technology*, *59*, 288-296.

Larivière, V., Gingras, Y. & Archambault, É. (2009). The Decline in the Concentration of Citations, 1900–2007. *Journal of the American Society for Information Science and Technology*. *60*(4), 858-862.

Line, M.B., & Sandison, A. (1974). "Obsolescence" and changes in the use of literature with time. *Journal of Documentation*, *30*, 283-350.

Nakamoto, H. (1988). Synchronous and dyachronous citation distributions. In L. Egghe & R. Rousseau (Eds.), Informetrics 87/88 (pp. 157-163). Amsterdam: Elsevier.

Price, D.J. (1965). Networks of scientific papers. *Science*, *149*, 510-515.

Schwartz, C. A. (1997). The rise and fall of uncitedness. *College & Research Libraries*, *58*(1), 19-29.

Yang, S. Ma, F. Song Y. & Qiu, J. (2010). A longitudinal analysis of citation distribution breadth for Chinese scholars. *Scientometrics*, *85*(3), 755-765.

Zhou, P.,Thijs, B. & Glanzel, W.(2009).Is China also becoming a giant in social sciences? . *Scientometrics*, *73*(3), 593-621.

The Rise in Co-authorship in the Social Sciences (1980-2013)

Dorte Henriksen[1]

[1] *dh@ps.au.dk*
Danish Centre for Studies in Research & Research Policy, Department of Political Science, Business and Social Sciences, Aarhus University, Denmark.

Abstract
This paper examines the rise in co-authorship in the Social Sciences over a 33-year period. We investigate the development in co-authorship in different research areas and discuss how the methodological differences in these research areas and changes in academia affect the tendency to co-author articles. The study is based on bibliographic data about 4.5 million peer review articles published in the period 1980-2013 and indexed in the 56 subject categories of the Web of Science's (WoS) Social Science Citation Index (SSCI). Results show that in the majority of the subject categories we can document a rise in the mean number of authors and that there are disciplinary differences in how much the number of authors has increased. The most substantial rise in the mean and median number of authors has happen in subject categories, where the research often is based on the use of experiments, large data set, statistical methods and/or team-production models.

Conference Topic
Citation and Co-citation Analysis

Introduction
This paper explores the rise in co-authorship in the social sciences. The study is based on all the articles registered from 1980-2013 in the Web of Science's (WoS) Social Science Citation Index (SSCI). Several studies have examined the rise in the number of authors in different research fields. The studies vary in design, but the majority of the bibliometric studies can be categorized as studies either based on bibliographic data from a national database (Lariviere, Gingras, & Archambault, 2006; Ossenblok, Verleysen, & Engels, 2014) or a selection of journals (Cronin, Shaw, & La Barre, 2003; Fisher, Cobane, Vander Ven, & Cullen, 1998; Hudson, 1996; Norris, 1993; White, Dalgleish, & Arnold, 1982). The study by Wuchty, Jones, and Uzzi (2007) is one of the few studies, that examined the increase in research collaboration by using bibliographic data about research articles from multiple fields collected from the subject categories in WoS. However, their study is based on a sample of research articles and not an exhaustively data collection of the research articles indexed in WoS. Furthermore, Wuchty et al. (2007) do not clarify how many articles in their study that are indexed in either Science and Engineering, Social Sciences or Arts and Humanities. This paper is the first study of the rise of co-authorship in the social sciences to use a large sample of time series data based on all of the publications in SSCI, thus the study cover multiple fields of the social science. The study is therefore not bias by national publication tendencies or the selection of journals. The disadvantage of a data set restricted to articles from SSCI is that other publication types and a substantial share of journals are excluded (Hicks, 2005; Ossenblok et al., 2014; Piro, Aksnes, & Rørstad, 2013). However, we believe that the larger data sample compensate for these data limitations. Hence, the objective of this

paper is to document the rise in co-authorship in the social sciences and discuss the factors that could have influenced this evolution.

The increasing focus on authorship can partly be attributed to the growing importance of and attention paid to a researcher's publication record, which is influential in the considerations for employment, promotion, funding and increases in salary (Biagioli, 2012; Costa & Gatz, 1992; Weingart, 2005). Thus, there is a tendency to measure and assess researchers' based on their quantitative research output instead for the content of this output. This creates incentives to "game" the system to improve one's resume by co-producing publications. This is especially the case, when the performance-based research funding systems use whole counts instead of fractionalizing (Butler, 2003; Ossenblok et al., 2014), so the reward for producing a publication does not have to be shared. Hence, the instrumental uses of performance-based funding systems affect the researchers' publishing behavior, including their definitions, perceptions and practices of authorship (e.g. Ossenblok, Engels, & Sivertsen, 2012). However, the rises in co-authorship and research collaboration are also affected by other factors that influence the research community. The rise can be a result of the increasing tendency to perform large scale research projects executed as team-production models. These projects require greater human and financial resources, a larger data collection effort and often more advanced technical and statistical analyses, hence leading to more specialization and division of labor in the research process (Beaver, 2001; Birnholtz, 2006; Cronin et al., 2003; Fisher et al., 1998; Hudson, 1996; Moody, 2004; Rennie, Yank, & Emanuel, 1997; White et al., 1982). These types of projects are often associated with natural and medical sciences, where there is a strong tradition for working in the fore mention team-production model. However, the increasing tendency to work with large scale data set, the rise in using quantitative methods and in some cases experiments have generated a similar team-production model in the social sciences (Cronin et al., 2003; Hudson, 1996; Moody, 2004). Furthermore, studies have found that researchers in the more quantitative research areas of social science is more likely to collaborate (Fisher et al., 1998). Others have pointed at the increasing mobility of researchers that has made it possible and desirable to expand inter-institutional collaborations (Melin, 2000; White et al., 1982) while the development of communication and information technology have enabled geographically disperse researchers to collaborate, by making it easier to communicate, analyze and exchange data (Beaver, 2001; Fisher et al., 1998; Melin, 2000). Furthermore, the growing number of people working in academia has created more collaboration opportunities (Fisher et al., 1998; Lee, 2000; Melin, 2000), especially the increase in PhD students have created more opportunities for research advisors to collaborate and co-author with their students (Fisher et al., 1998; Price, Dake, & Oden, 2000). However, this tendency has given rise to issues regarding honorary or gift authorship in academia and some studies suggest that research advisors may be inappropriately demanding co-authorship with their students (Rennie et al., 1997). This is disputed by Costa and Gatz (1992), who found that students willingly are giving their advisors inappropriate authorship credit even though the advisors do not fulfill the journal guidelines and requirements for co-authorship. However, they do suggest that the willingness to offer co-authorship can be affected by a power imbalance between advisors and advisees, especially because of the increase in PhD students being subsidized by grants held by their advisors. In this paper we will document the evolution of co-authorship and research

collaboration by presenting evidence for the increase in the number of authors per publication.

Method

The bibliometric data used in this study were collected from the Centre for Science and Technology Studies (CWTS) enhanced version of Thomson Reuters' WoS database in December 2014. We collected bibliographic information for 4,466,134 articles from 99,752 journal issues published in 1980 to 2013 and registered in WoS' SSCI 56 subject categories. These 56 subject categories have in our analysis been grouped into 6 overall subject categories. The grouping of the categories is based on the topics of each subject category described in the SSCI scope notes (SSCI, 2012). Hence, there are differences in how many categories there has been group together, and the similarity of the research areas. The Social Sciences, Interdisciplinary group consist of a variety of subject categories and do not have the similar thematic relationship as the other groups.

- **Management, Planning & Geography** (Geography, Planning & Development, Urban Studies, Environmental Studies, Management, Transportation)
- **Political Sciences, Business and Law** (Criminology & Penology, Business, Business, Finance, Economics, Public administration, International Relations, Law, Political Science
- **Psychology** (Psychology, Mathematical, Psychology, Psychoanalysis, Psychology, Experimental, Psychology, Social, Psychology, Educational, Psychology, Applied, Psychology, Biological, Psychology, Clinical, Psychology, Developmental, Psychology, Multidisciplinary, Psychiatry
- **Social Health Sciences** (Public Environmental & Occupational Health, Substance Abuse, Gerontology, Health Policy & Services, Rehabilitation, Education, Special, Nursing, Ergonomics)
- **Social Sciences, Interdisciplinary** (Social Sciences, Biomedical, Family Studies, Information Science & Library Science, Social Sciences, Interdisciplinary, Hospitality, Leisure, Sport & Tourism, Industrial Relations & Labor, Social Sciences, Mathematical Methods, Communication, Linguistics, Ethics, History & Philosophy of Science, History of Social Sciences, History)
- **Sociology & Anthropology** (Anthropology, Area Studies, Social Work, Education & Educational Research, Women's Studies, Demography, Social Issues, Sociology, Ethnic Studies, Cultural Studies)

Our study limits the relevant types of publications to journal articles, though we know that the publication pattern in the social sciences is more varied (Lariviere et al., 2006; Ossenblok et al., 2014), thus letters, book chapters and books are an essential part of the scholarly communication in some fields of the social sciences. Unfortunately, the Thomson Reuters Book Citation Index (BCI), part of the WoS core collection, do not have as systematic and exhaustively bibliographic information about books compared to the SSCI's information about journal articles. The BCI do only cover the time period from 2006-present, while SSCI have bibliographic data from 1900 to present, so by choosing to only include journal articles we can set a larger time frame for this study.

Results

In the follow subsections we will present the data showing the increase in number of authors per publication. For each group we will present a figure demonstrating the development in the different subject categories[1]. Our data show that the fields of social sciences have experienced a mean 114 percent increase in the number of authors during the last 33 years, hence there have been added 1,2 authors more to each publication. However, there are large differences in how much the number of authors has risen, with the lowest increase being in the History subject category with a minimal change (0.1 authors) to the highest mean increase in Psychiatry (3 authors).

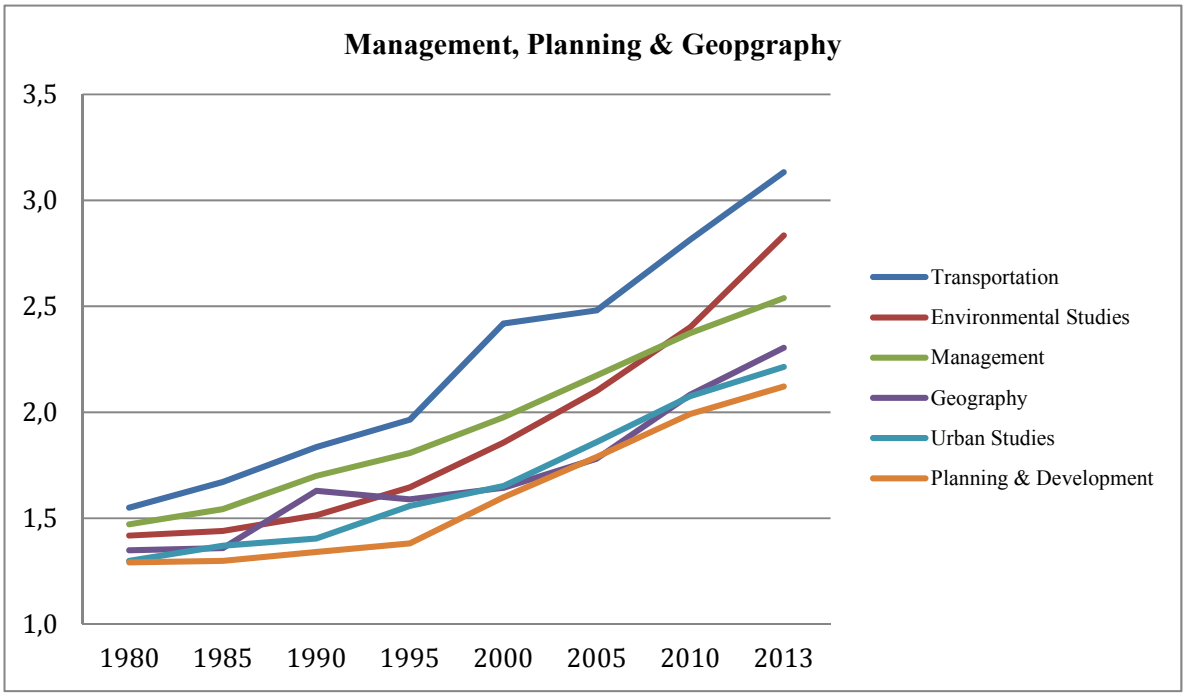


Figure 1. The mean number of authors per publication from 1980-2013 of the group Management, Planning & Geography.

The six categories group as Management, Planning & Geography consist of 373,372 publications. Figure 1 shows the evolution in numbers of authors. The mean numbers of author have increase 71% to 102% or 0.8 – 1.6 authors during the 33 year time period. The mean numbers of authors in 1980 are in the range of 1.3-1.6 authors and have increased in 2013 to 2.1-3.1 authors. The median number of authors is 1 in all categories in 1980. In 2013 the median number of authors has risen to 3 in the category Transportation, while the remaining categories have a median of 2. Even though the category Transportation does not cover civil engineering per se, the close relation with the above mentioned research field can explain some of the increase in co-authorship in this category. The subject categories in this group have all similarities to research fields

[1] We have in this article, because of the space limit, decided to present the development of co-authorship in six figures. The data behind the study will be presented in more details at the conference and are also available if requested.

in science and technology, and are probably influenced by collaboration and publication tendencies dominating these fields.

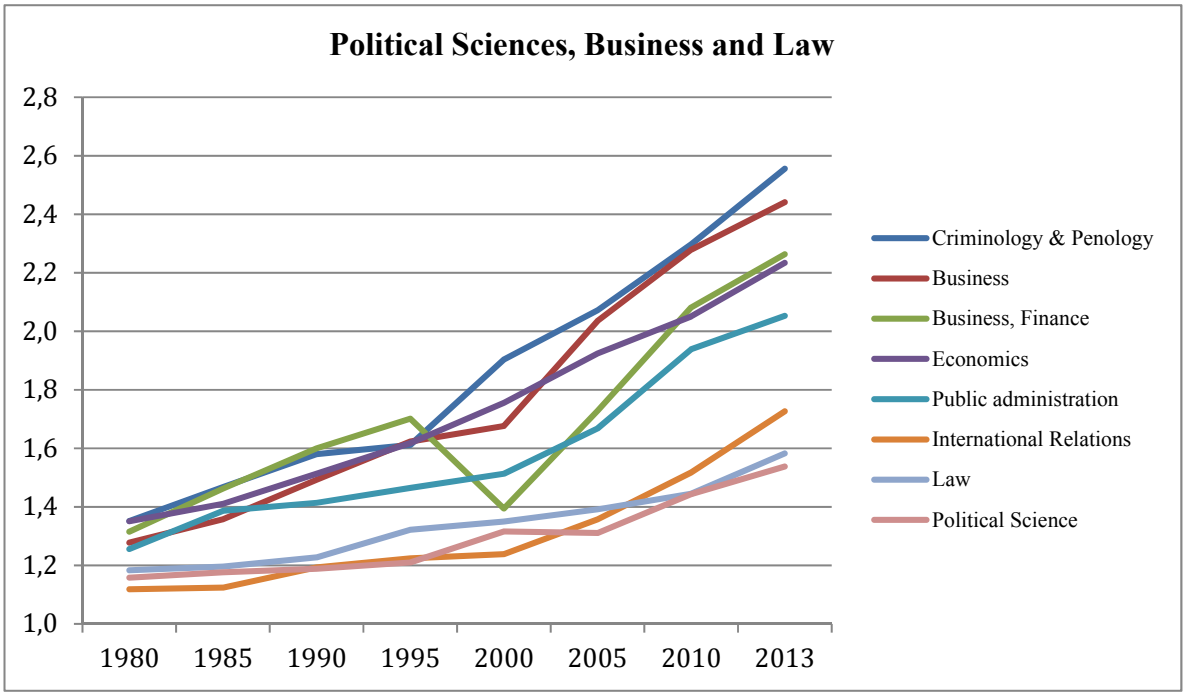


Figure 2. The mean number of authors per publication from 1980-2013 of the group Political Sciences, Business and Law.

The 1,011,725 publications belonging to the Political Sciences, Business and Law show a rise between 38% to 89% in the mean numbers of authors. The mean numbers of authors are between 1.1 - 1.4 in 1980 to 1.5 - 2.6 in 2013 (see figure 2). In Business, Business, Finance, Economics, Criminology & Penology and Public Administration have the median number of authors increase from 1 to 2 authors during the 33 years, while in the remaining categories the median number of authors is 1 during the time period. The greater rise in mean number of authors in the categories Criminology & Penology, Business, Business, Finance, Economics, and Public Administration could be because of the greater use of statistics and register/survey data (Fisher et al., 1998; Hudson, 1996). Political Science is the category in this group with the highest amount of publications (n = 172,625) and covers a broad range of research, thus the lower increase and mean number of authors is probably because areas of Political Science have similarities with research fields in the humanities. The same is the case for the category Law that draws on methods often associated with humanities, such as text analysis.

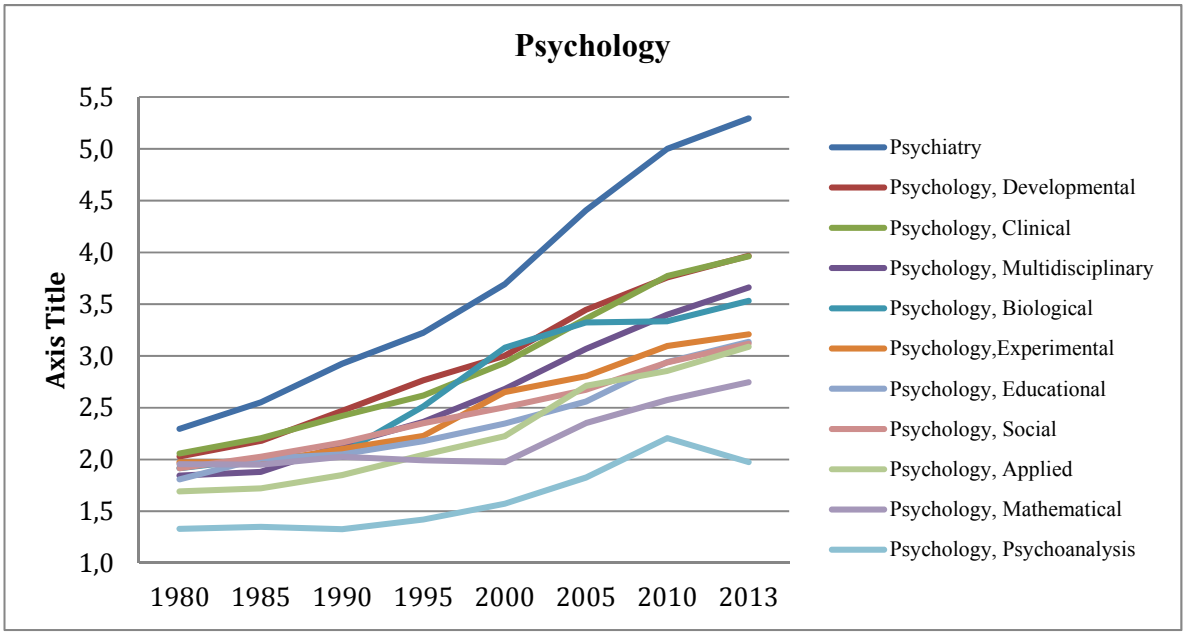


Figure 3. The mean number of authors per publication from 1980-2013 of the group Psychology.

We have collected 1,101,234 publications categorized as Psychology. During the 33 years the mean increase in number of authors is in the range from 0.6 to 3 authors or from 40 % to131%. The mean numbers of authors in 1980 are between 1.4-2.3 authors, this have in 2013 increased to between 2-5.3 authors. The categories of Psychology have all increased the number of authors in the byline during the 33 years, though it is not a constant increase as can been seen in figure 3. The category with the lowest increase is Psychoanalysis, the subject category with a publication and collaboration behavior closest to the humanities, and a mean number of authors in 2013 at 2 authors. Psychoanalysis is the only subject category in the Psychology group where the median have remain constant at 1. In the other end of the scale we have Psychiatry, a subject category with close relations to the medical research fields and therefore a similar collaboration and publication pattern. The mean number of authors in this category is 5.3 authors and the median is 5. Psychology, Mathematical have constantly had a median at 2, while Psychology, Applied have had an increase in the median number of authors from 1 to 3 and Psychology, Clinical have had an increase in median authors from 2 to 4. Psychology, Experimental, Psychology, Social, Psychology, Educational, Psychology, Development, Psychology, Biological and Psychology, Multidisciplinary have had an increase in the median number authors from 2 to 3 authors.

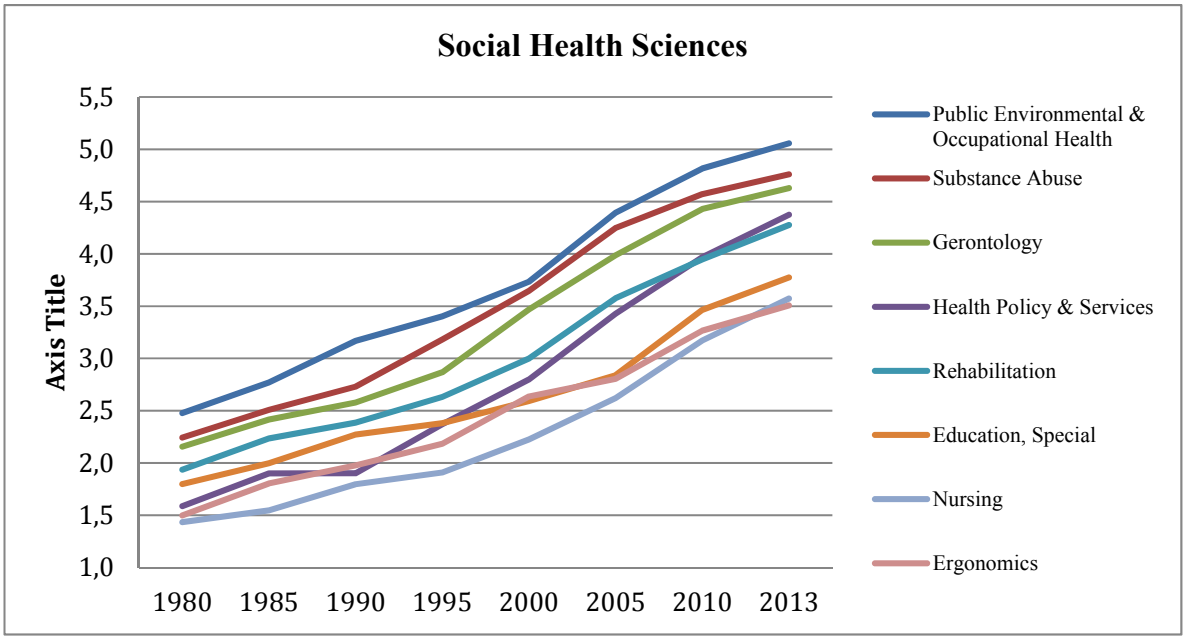


Figure 4. The mean number of authors per publication from 1980-2013 of the group Social Health Sciences.

The data of the categories Social Health Sciences is based on 824,125 publications. The mean number of authors per publication in the Social Health Sciences categories has risen between 104% to 176% or 2-2.6 authors. Figure 4 shows how there have been a substantial increase in all seven subject categories during the 33 years. The median number of authors in 1980 is 1 in the categories Ergonomics, Health Policy & Services and Nursing and 2 in the categories Rehabilitation, Public Environmental & Occupational Health, Substance Abuse, Gerontology and Educational, Special. In 2013 the median numbers of authors have risen to 3 authors in Ergonomics, Nursing and Education, Special and to 4 in the remaining categories. The mean numbers of authors in the Social Health Sciences are between 1.4-2.5 authors in 1980 and have risen to 3.5-5.1 authors in 2013. The average numbers of authors are general quite high in Social Health Sciences compared to other subject categories in the Social Sciences and the subject categories have a publication and collaboration pattern similar to the health and life sciences.

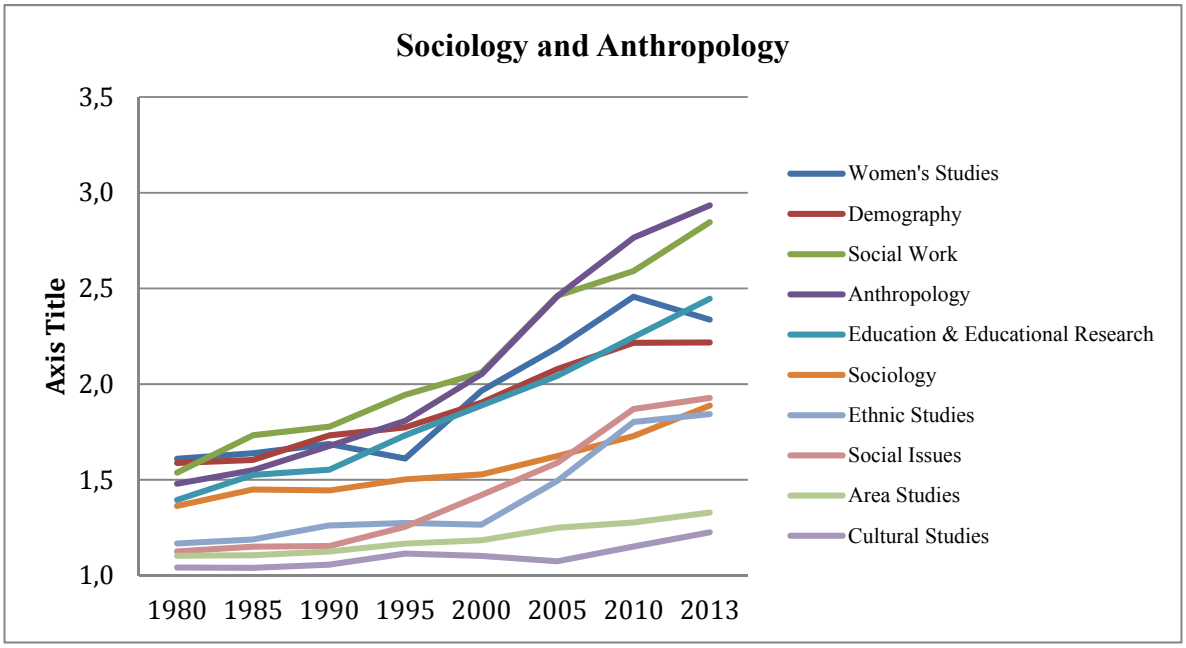


Figure 5. The mean number of authors per publication from 1980-2013 of the group Sociology and Anthropology.

In our data set we have 514,504 publications categorized in the 10 subject categories of Sociology and Anthropology, and the mean percentage increases in numbers of authors are between 17% to 98%. In Figure 5 is the increase in number of authors demonstrated. There have been minimal changes in the mean number of authors in the subject category Cultural Studies and Area Studies, while the categories Social Issues and Ethnic Studies have increased with 0.6-0.8 authors. All of these fore mention categories have a median at 1 in the whole time period. The median has risen to 2 authors for Education & Educational Research, Anthropology, Social Work, Sociology, Women's Studies and Demography. These categories, except Sociology, have a mean number of authors between 1.4-1.6 authors in 1980, which has increased to 2.2-2.9 in 2013. The mean number of authors has only increased with 0.5 for Sociology.

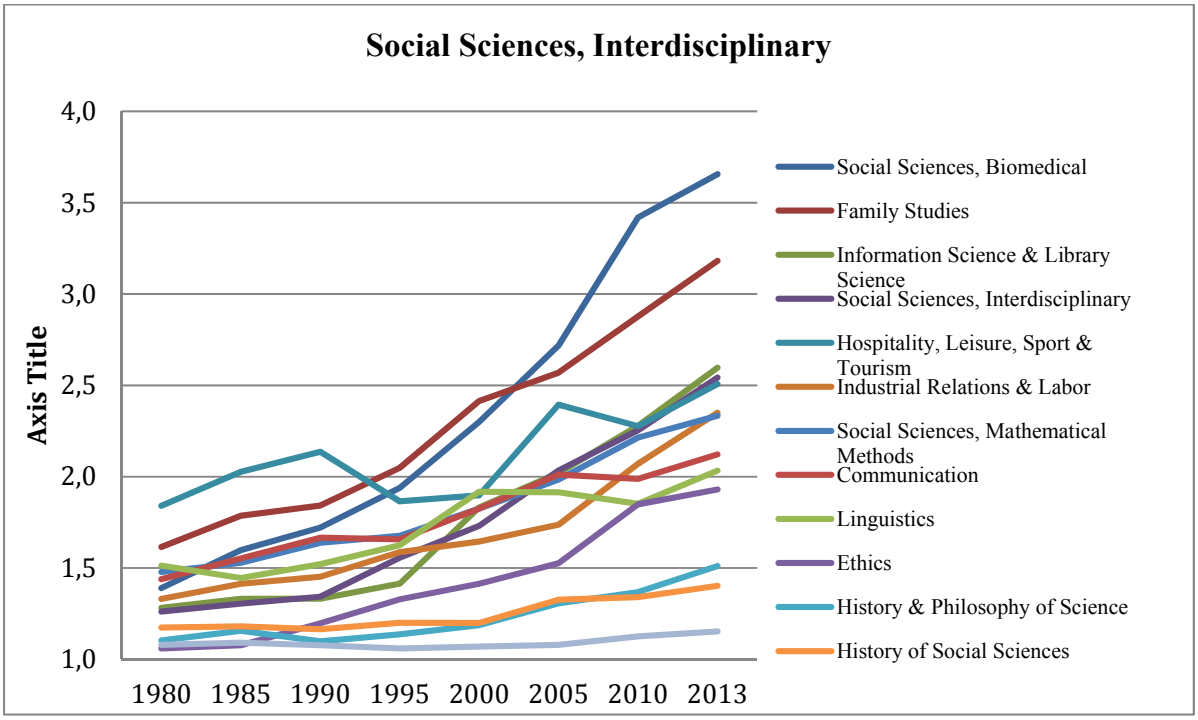


Figure 6. The mean number of authors per publication from 1980-2013 of the group Social Sciences, Interdisciplinary.

694,752 publications are indexed in the categories in the group Social Sciences, Interdisciplinary. The mean increases in numbers of authors are between 6.8%-163% or between 0.1-2.3 authors. Figure 6 demonstrates how much the increase in the numbers of authors varies from 1980 to 2013. The category History hardly had any changes in the mean number of authors and the median remain constantly at 1 during the time period. The median also remains at 1 author in the categories History of Social Science, History & Philosophy of Science and Ethics, while the mean rises from 1.1-1.2 authors in 1980 to 1.4-1.9 authors in 2013. The median increases from 1 to 2 authors in the categories Communication, Information Science & Library Science, Industrial Relations & Labor, Linguistics, Social Sciences, Interdisciplinary and Social Sciences, Mathematical Methods and the mean numbers of authors increases from 1.3-1.5 authors to 2-2.5 authors. The median is constant at 2 authors in Hospitality, Leisure, Sport & Tourism, where mean number of authors rise from 1.8 authors to 2.5 authors. The median increases from 1 author in 1980 to 3 authors in 2013 in the categories Family Studies and Social Sciences, Biomedical and the mean numbers of authors rises from 1.4-1.6 authors to 3.2-3.7 authors. In this very mixed group we can see how the categories with research closest to the humanities such as History, History of Social Science, History & Philosophy of Science and Ethics have a lower rise in the number of mean authors, while the categories Family Studies and Social Sciences, Biomedical, that both are methodological close to the life and medical sciences have had a substantial high rise in number of authors.

Discussion

In this study we document the evolution of co-authorship in the social sciences and find that the majority of research fields have had substantial increases in the numbers of authors per publication. During the 33 years the increase is equal to one author or more in

31 out of 56 subject categories, and in further five subject categories, the increase is nearly 1 author (0.9). We detect a similar increase when we include the median increase in the number authors, where the median number of authors has increased by one or more authors in 42 out the 56 subject categories. The increases in the number of authors have not happened in the same degree in all areas of the social sciences and illustrate how heterogeneous the research fields of social sciences are. The articles indexed in the four subject categories History, Cultural Studies, Area Studies and History of Social Sciences have only had a mean increase in the number of authorship between 0.1-0.2, and could be categorized as status quo during the 33 years. The percentage increases in the mean number of authors in the subject categories varies from 6.8% (History) to 175.6% (Health Policy & Services).

The results of this study confirm that there is an increasing tendency to co-author and collaborate and is in line with the tendency detected in previous studies of co-authorship and collaboration (e.g. Bebeau & Monson, 2011; Fisher et al., 1998; Ossenblok et al., 2014). Namely that the number of authors per publication has increased in the social sciences and that the largest increases have occurred in the fields with use of experiments, large data set, statistical methods and/or team-production models, such as the Social Health Sciences and parts of Psychology. A good example in our study of how the methodological differences affect the collaboration patterns is the subject categories group as Psychology. The subject categories Psychology, Psychoanalysis and Mathematical are both examples of research domains dominated by theory building and abstract concepts and with methodological relationships to research fields often defined as belonging to the humanities. The opposite are Psychiatry and Developmental Psychology, where the research are more experimental and empirical, and often sampled in collaboration with other researchers. Hence, the greatest rises in number of authors have occurred in subject categories containing research fields using quantitative methods and with a close relationship to the medical and life sciences or the natural sciences. An additional explanation for the rise in co-authorship in the majority of the subject categories is the increasing tendency for supervisors to co-author with students (Costa & Gatz, 1992; Fisher et al., 1998; Price et al., 2000).

Conclusion

As mentioned in the introduction, most of the bibliometric studies about co-authorship and research collaboration in the social sciences have been focusing on the trends and patterns in particular research fields or countries and have been based on data collected from a selection of journals in one or few research fields or national databases. In this study we use a larger sample of articles to confirm there is a rise in co-authorship in the majority of the research fields in the social sciences, and that in more than half of the subject categories the mean number of authors has increased by one or more authors.

Few of these studies undertake a deeper investigation of the rise of co-authorship and research collaboration (Costa & Gatz, 1992; Fisher et al., 1998), and the explanations offered for the rise is often speculative and anecdotal or borrowed from the “hard” sciences. We have discussed some of the factors that influence the researchers’ collaboration behavior and the rise in co-authorship. However, our explanations are based on the fore mention studies, and we therefore suggest that the next step is a thoroughly

investigation of the effects of these factors in the fields we have documented a rise in co-authorship.

Acknowledgement

I thank my supervisor Jesper Schneider for comments and subtracting data from the CWTS database. Furthermore, I thank Søren Midtgaard for his suggestions and comments.

References

Beaver, D. D. (2001). Reflections on scientific collaboration, (and its study): past, present, and future. *Scientometrics, 52*(3), 365-377. doi: 10.1023/a:1014254214337

Bebeau, M., & Monson, V. (2011). Authorship and Publication Practices in the Social Sciences: Historical Reflections on Current Practices. *Science and Engineering Ethics, 17*(2), 365-388. doi: 10.1007/s11948-011-9280-4

Biagioli, M. (2012). Recycling Texts or Stealing Time?: Plagiarism, Authorship, and Credit in Science. *International Journal of Cultural Property, 19*(3), 453-476. doi: 10.1017/s0940739112000276

Birnholtz, J. P. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *Journal of the American Society for Information Science and Technology, 57*(13), 1758-1770. doi: 10.1002/asi.20380

Butler, L. (2003). Modifying publication practices in response to funding formulas. *Research Evaluation, 12*(1), 39-46. doi: Doi 10.3152/147154403781776780

Costa, M. M., & Gatz, M. (1992). Determination of Authorship Credit in Published Dissertations. *Psychological Science, 3*(6), 354-357. doi: DOI 10.1111/j.1467-9280.1992.tb00046.x

Cronin, B., Shaw, D., & La Barre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology, 54*(9), 855-871. doi: 10.1002/asi.10278

Fisher, B. S., Cobane, C. T., Vander Ven, T. M., & Cullen, F. T. (1998). How many authors does it take to publish an article? Trends and patterns in political science. *Ps-Political Science & Politics, 31*(4), 847-856. doi: 10.2307/420730

Hicks, D. (2005). The Four Literatures of Social Science. In H. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 473-496): Springer Netherlands.

Hudson, J. (1996). Trends in multi-authored papers in economics. *Journal of Economic Perspectives, 10*(3), 153-158.

Lariviere, V., Gingras, Y., & Archambault, E. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics, 68*(3), 519-533. doi: DOI 10.1007/s11192-006-0127-8

Lee, W. M. (2000). Publication trends of doctoral students in three fields from 1965-1995. *Journal of the American Society for Information Science, 51*(2), 139-144. doi: 10.1002/(sici)1097-4571(2000)51:2<139::aid-asi5>3.0.co;2-1

Melin, G. (2000). Pragmatism and self-organization - Research collaboration on the individual level. *Research Policy, 29*(1), 31-40. doi: Doi 10.1016/S0048-7333(99)00031-1

Moody, J. (2004). The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213-238.

Norris, R. P. (1993). Authorship Patterns in Cjnr - 1970-1991. *Scientometrics, 28*(2), 151-158. doi: Doi 10.1007/Bf02016897

Ossenblok, T. L. B., Engels, T. C. E., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science-a comparison of publication patterns and incentive structures in Flanders and Norway (2005-9). *Research Evaluation, 21*(4), 280-290. doi: DOI 10.1093/reseval/rvs019

Ossenblok, T. L. B., Verleysen, F. T., & Engels, T. C. E. (2014). Coauthorship of journal articles and book chapters in the social sciences and humanities (2000–2010). *Journal of the Association for Information Science and Technology, 65*(5), 882-897. doi: 10.1002/asi.23015

Piro, F. N., Aksnes, D. W., & Rørstad, K. (2013). A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology, 64*(2), 307-320. doi: 10.1002/asi.22746

Price, J. H., Dake, J. A., & Oden, L. (2000). Authorship of health education articles: Guests, ghosts, and trends. *American Journal of Health Behavior, 24*(4), 290-299.

Rennie, D., Yank, V., & Emanuel, L. (1997). When authorship fails - A proposal to make contributors accountable. *Jama-Journal of the American Medical Association, 278*(7), 579-585. doi: 10.1001/jama.278.7.579

SSCI. (2012). Scope Notes 2012. Social Science Citation Index. Retrieved 10-01-2015, 2015, from http://ip-science.thomsonreuters.com/mjl/scope/scope_ssci/#BF

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics, 62*(1), 117-131. doi: 10.1007/s11192-005-0007-7

White, K. D., Dalgleish, L., & Arnold, G. (1982). Authorship Patterns in Psychology - National and International Trends. *Bulletin of the Psychonomic Society, 20*(4), 190-192.

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science, 316*(5827), 1036-1039. doi: 10.1126/science.1136099

The Recurrence of Citations within a Scientific Article

Zhigang Hu[1], Chaomei Chen[2] and Zeyuan Liu[1]

[1] *huzhigang@dlut.edu.cn, liuzy@dlut.edu.cn*
WISE Lab, Dalian University of Technology, 116024 Dalian (China)

[2] *cc345@drexel.edu*
College of Computing and Informatics, Drexel University, 19104 Philadelphia (USA)

Abstract
Although listed at the tail of a scientific article only once, a reference is usually cited repeatedly inside the full text of the article. In this research, we investigated the universality of recurring citations in Journal of Informetrics. About 1/4 references are repeatedly cited. For these repeatedly cited references, their citation location and citation context for the first and subsequent times are examined separately. Normally, recurring citations of a same reference tend to be located in the same section instead of different ones. It proves that, even if a reference is cited for multiple times in a single citing paper, it is still focus on the same topic in the same section most of the time. We also explored the reason why recurring citations are happening. By comparing the contexts of two kinds of citations, the first-time citations and the succeeding citations, we found that, for a specific reference, its first-time citation is usually not as intentional as the succeeding citations. Just because of the relative importance of the succeeding citations compared to the first-time citation, recurring citations are reasonable and necessary.

Conference Topic
Citation and co-citation analysis

Introduction
Citations are essential components for scientific articles. Traditional citation analysis is more like reference analysis, since only references listed at the tail of the article are researchers' concern. Citations, which indicate the locations and context where references are cited, are almost ignored in previous research. The reference analysis is much easier and effective most of the time, but in the meanwhile, some important information might be neglected. For example, where are these references are cited inside the citing papers? How are the citations distributed among different sections? By investigating the citation location and the citation context, however, we can understand not only the pattern how references are cited, but also the reason why authors cite it like that.

Nowadays, full-text citation analysis, which is about how references are cited in the body of citing papers, is just beginning (Ding, Liu, Guo, & Cronin, 2013; Hu, Chen, & Liu, 2013; Liu, Zhang, & Guo, 2013; Zhang, Ding, & Milojević, 2013). During to the increasingly availability of structured full texts such as XML-formatted articles, researchers began to turned their attention from references to citations in the body of articles. For example, Ding et al. have examined the distribution of references across text and find that most highly cited works appear in the Introduction and Literature Review sections of citing papers (Ding et al., 2013). Hu et al. visualize the location distribution of citation instances, especially those to highly-cited references. The results show that citations are usually distributed very uneven inside the full texts of scientific articles (Hu et al., 2013).

In full-text citation analysis, recurring-citation is an interesting issue. Recurring-citation refers to the phenomenon that a reference is cited more than once in a citing paper. Take this paper for example, we cite the reference of (Hu et al., 2013) in the first sentence of last paragraph for the first time, and then cite it again in the last sentence of the same paragraph for the second time. In this paper, we call this reference a repeatedly cited reference, or a reference with recurring citations. Recurring-citation is a common fact in citation behaviour. In our previous research, we find that, sometimes, a reference might be repeatedly cited as many as nine times in a single paper (Hu et al., 2013).

In this research, we will investigate the phenomenon of recurring citations. Our concern is the universality and the pattern of recurring-citations, including: (1) how common recurring citations are in scientific articles? (2) where the recurring citations of a single reference are usually located inside the paper? (3) what the difference is between its first-time citation and the succeeding ones? In the end, the reason why recurring-citation is necessary will also be discussed.

Data and Methods

To detect recurring citations of a reference inside a citing paper, the full text of the citing paper need to be observed. There are two types of full texts: one is in unstructured format such as PDFs, which is human-friendly; the other is in structured format such as XMLs/HTMLs, which is machine-friendly. Compared with PDFs, structured full texts, e.g., XMLs, are much easier to process by computer. For example, XML-formatted full texts can be parsed directly using an existing function: *xml_parse()* in PHP. Thus, it is very straightforward to identify citations inside a citing paper. Nowadays, structured XML-formatted full texts are available and downloadable in almost every bibliographic database, such as Elsevier, Springer, John & Wiley, and especially, the open access online journals like PLoS ONE. In this study, the data of full texts was sourced from Elsevier ConSyn (http://consyn.elsevier.com), a content syndication system developed by Elsevier. Since 2011, Elsevier ConSyn provided downloadable articles in XML format.

In Elsevier ConSyn, we retrieved and downloaded all the full texts of 350 articles published in Journal of Informetrics (JOI) from 2007 to 2013. Journal of Informetrics is chosen as the case in this study because it is published by Elsevier and belongs to the field of library and information science. By our own developed program, we parsed these XML-format full texts and extract all the citations inside them. Since each citation instances is clearly marked with a XML tag, i.e. <*ce:cross-ref*>, they can be recognized and extracted easily. All the attributions of each citation, including its location and its citees, were recorded and import into database tables.

By looking into citations' citees, we achieved the cited times of each reference inside each citing paper. If the cited times is equal to one, it means the reference is one-time cited inside this citing paper. While if cited more than once, the reference is considered as repeatedly cited or recurrently cited. In this research, we will count the frequency of each type of reference, e.g., once-cited, twice-cited, triple-cited, etc. In this way, the universality and intensity of recurring-citation can be estimated accordingly.

For repeatedly cited references, their citation locations will be studied. The location of citation can be measured by, from macro to micro scales: character, word, sentence, paragraph and section. In this study, we chose the measurement at the largest scale: section. We will calculate the count of citations in each section and see how citations are

distributed in different sections. Generally, a scientific article is made up of four sections, namely *Introduction*, *Data and Methods*, *Results,* and *Discussion and Conclusions*. It is called IMRaD structure usually (see e.g. Agarwal & Yu, 2009; Swales, 1990) To some extent, citation location can reveal the citation motivation. If we are aware of the section where a citation is located, the role of the citation can be figured out to some extent. For instance, if a reference is cited in the section of *Data and Methods*, usually section II, it is probably a helpful citation relevant in the aspect of methodology; while if it located in the section III or the section of *Results*, the citation is more likely about comparable results.

Besides the location distribution of recurring citations, we also examined the difference between a reference's first-time citation and the succeeding ones. We extracted the context when a reference is cited for the first time and when it is cited again in the following parts. The first-time and the succeeding citation contexts will be compared in terms of the count of their citees inside. The more citees/references a citation contains, the less important each citee/reference is. The citation with many citees/references, such as the one in the first sentence of the second paragraph, is called perfunctory citation (Cano, 1989; Oppenheim & Renn, 2004; Pham & Hoffmann, 2003; Voos & Dagaev, 1976), which means authors decide not to cite the citees/references seriously in an excluded way. In this research, we are interesting in which one, the first-time citation or the succeeding citation, is more likely to be perfunctory citation for a multiple-cited reference.

Results and Discussion

The universality of recurring citations

Firstly, we examined how common recurring citations are in the Journal of Informetrics. Among all the 11,327 references inside the 350 articles, 8,417 (74.3%) of them were cited once in a single citing article. The other 2,910 references (account for 25.7%) were cited twice or more, including 1,726 (15.2%) twice-cited references, 613 (5.4%) triple-cited references, and 571 (5.0%) references cited for four times or more. Although one-time citation is the main citation pattern undoubtedly, the phenomenon of recurring-citation cannot be ignored in both frequency and intensity.

Figure 1 shows the frequency distribution of references of each kind, companied with a distribution graph in double logarithmic coordinates. As it shown in the best fitting line, the frequency distribution of multiple citations follows a power law ($y= 21557\ x^{-3.479}$, $R^2=0.9679$), which is a very common law in the field of bibliometrics, such as the distribution of scientific productivity (Lotka, 1926) or keywords (Zipf, 1949). Obviously, it is not accidental that the frequency distribution of recurring citations is in this pattern.

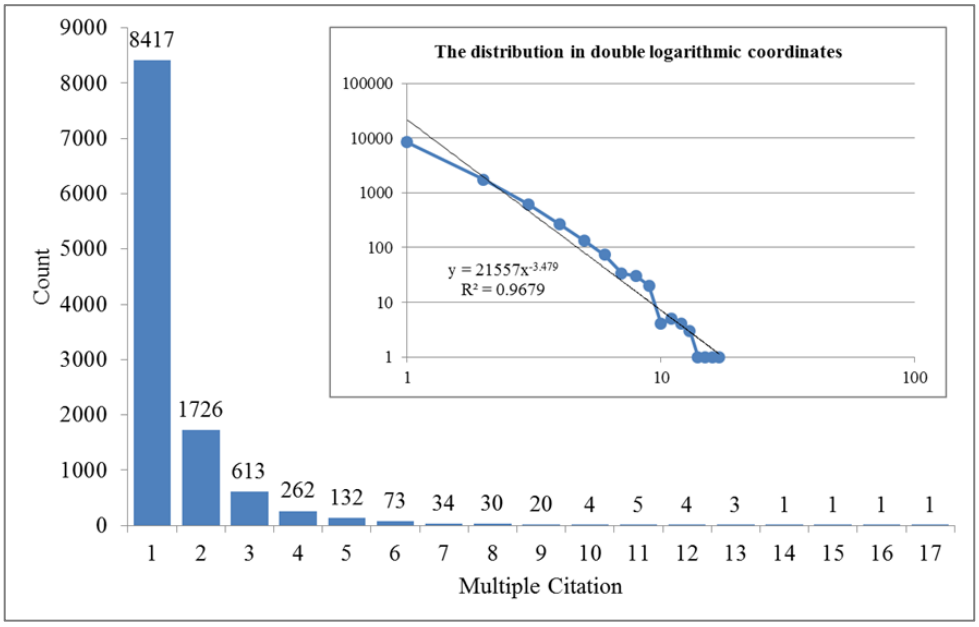


Figure 1. The count of references by their multi-citation times.

The locations of recurring citations

The location pattern of recurring citations is the focus of this research. In this part, we will investigate the location distribution of multi-citation by section. In Journal of Informetrics, 92 articles (26.3% of all) adopt IMRaD structure, which is most used form to organize articles in our research. Thus, we selected all these 92 four-section articles in IMRaD structure as cases, and explored how citations are distributed in the four different sections.

As shown in Figure 2, among all the 3035 citations in these 92 articles, 1238 (40.8%) citations are located in Section I, or the section of *Introduction*; 760 (25.0%) of them are located in Section II (or *Methods*); 769 (25.3%) citations is in the sections of *Results*; and 268 (8.8%) in *Discussion and Conclusions*. This mode of section distribution of citations meets our expectation on citation locations, since it is the widely accepted fact that authors are likely to cite most in the section of *Introduction*.

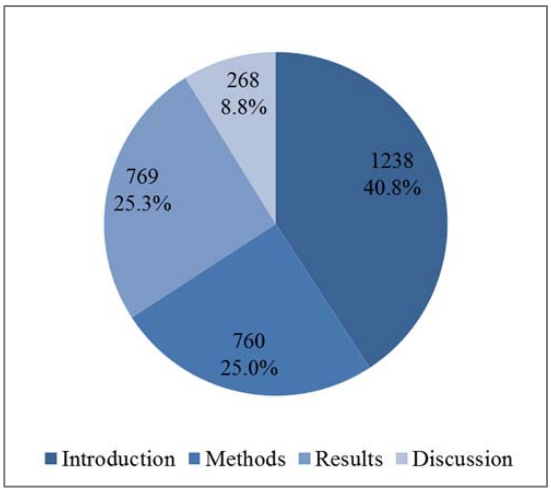


Figure 2. The count of citations in each section.

Based on the section distribution of citations, we are then able to investigate the section combination distribution of a reference's recurring citations. For each repeatedly cited reference, its recurring citations could be located in any sections, either the same section or the different sections. Since twice cited references are the simplest and most common (59.3%) types of repeatedly-cited references, they were chosen for calculating combined-section distribution.

For each twice-cited reference, we recorded the located sections of both citations. The counts of the 10 types of section combinations are shown in Table 1. Among all the 796 twice-cited references, the most common ones are those cited in Section I for both the first and the second time. 224 (28.2% of all) references belong to this type. 124 (15.6%) references are cited in Section I for the first time and Section II for the second time. References that are cited in section IV twice are least common (18 or 0.8%). Totally, 444 (55.8%) references are cited in the same section twice, while 350 (44.2%) ones are cited in the difference sections.

Table 1. The combined section distribution of the twice citations of references

Located Section of the second citation / Located section of the first citation	*Sec I*	*Sec II*	*Sec III*	*Sec IV*
Sec I	**224** 28.2%			
Sec II	**124** 15.6%	**90** 11.3%		
Sec III	**68** 8.6%	**42** 5.3%	**112** 14.1%	
Sec IV	**64** 8.1%	**24** 3.0%	**28** 3.5%	**18** 2.3%

Although more twice cited references are cited in the same section, we cannot say that a reference's multiple citations tend to be located in the same sections except that the expected proportion of the multiple citations located in the same section is calculated and compared. Thus, we assume that a reference's twice citations are located independently and randomly, just like two arbitrary citations in the article. Under this hypothesis, the expected distribution of section combinations of twice citations can be calculated as follow:

(Sec I, Sec I) : (Sec I, Sec II) : (Sec I, Sec III) : (Sec I, Sec IV)

: (Sec II, Sec II) : (Sec II, Sec III) : (Sec II, Sec IV)

: (Sec III, Sec III) : (Sec III, Sec IV)

: (Sec IV, Sec IV)

= 40.8%×40.8% : 40.8%×25.0%×2 : 40.8%×25.3%×2 : 40.8%×8.8%×2
: 25.0%×25.0% : 25.0%×25.3%×2 : 25.0%×8.8%×2
: 25.3%×25.3% : 25.3%×8.8%×2
: 8.8%×8.8%

=16.6% : 20.4%: 20.7% : 7.2% : 6.3% : 12.7% : 4.4% : 6.4% : 4.5% : 0.8%

Figure 4 shows the expected and observed proportions of the section combinations of each kind. If the expected values match the observed well, it means the twice citation are located independently and randomly indeed; otherwise, it means that there is a certain tendency in how to cite a reference twice. In Figure 4, we have not seen the match between the expected and observed values. For example, based on our initial hypothesis, the proportion of (Sec I, Sec I) should be 16.6%, not even closed to 28.2% as observed; the proportion of (Sec I, Sec III) should be 20.7%, while the observed value is 8.6%, which is much lower. Neither of them presents the match as assumed.

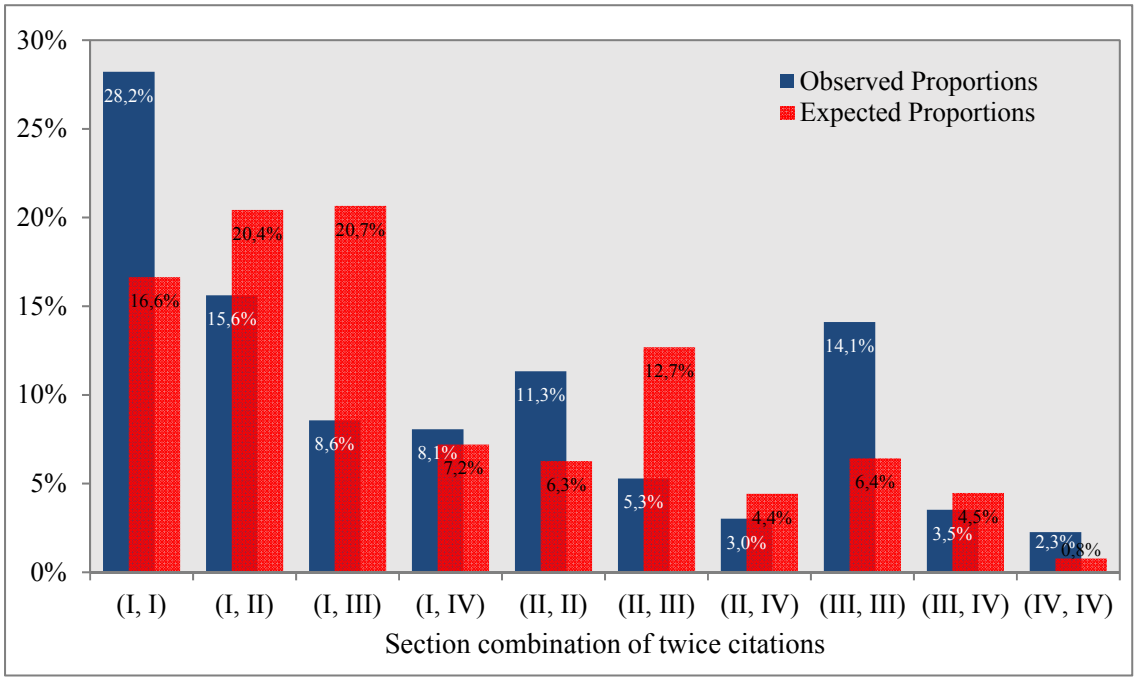


Figure 3. The distribution of the expected and observed proportions of located section combinations of twice citations.

According to the comparison between the expected and observed values, these 10 combinations can be divided into two classes: the above-expectation combinations and the below-expectation ones. The combinations of (Sec I, Sec I), (Sec II, Sec II), (Sec III, Sec III), (Sec IV, Sec IV) and (Sec I, Sec IV) belong to the former. Their observed proportions are higher than expected significantly. The other 5 combinations, i.e., (Sec I, Sec II), (Sec I, Sec III), (Sec II, Sec III), (Sec II, Sec IV) and (Sec III, Sec IV), belongs to the latter. From this division, we can see that, the references with twice citations located in the same section are preferable to those with twice citations located in different sections. The only exceptions are the references cited inside (Sec I, Sec IV), which have an above-expectation proportion (2.3% v.s. 0.8%), though its twice citations located in the different sections.

Why do authors tend to cite a reference multiple times inside the same section? The explanation could be simple. Normally, a reference is only helpful for a single topic, usually existing in a concentrated part of an article, such as a section. Few references are necessary for several different topics, or in different sections. That is why references are

preferred to be cited in a single section. This explanation also interprets why the combination of (Sec I, Sec IV) is an exception. The first and the fourth section, although farthest away with each other, are actually discussing about the same topic at the same level, i.e., the hindsight and foresight of research questions.

The context of recurring citations

We have revealed how common recurring citations are and where these recurring citations are usually located, and now we will examine their contexts. Firstly, the citation contexts of repeatedly cited references for the first and the succeeding times were extracted separately. There are totally 11,448 first-time citation contexts and 5,469 succeeding ones extracted. We will explore the difference between these two groups of citation contexts in terms of citation intensity, which can be estimated by how many citees they contained.

The count of citees contained in each citation context is calculated one by one. Averagely, a citation contains 1.94 citees, or put it another way, authors cite 1.94 references once at a time. As it shown in Figure 6, although most citations (64.2% of all) cite only one single citee/reference, there are still more than 1/3 of citations contain two or more citees/reference. 1457 (8.7%) citations cite even five or more references once.

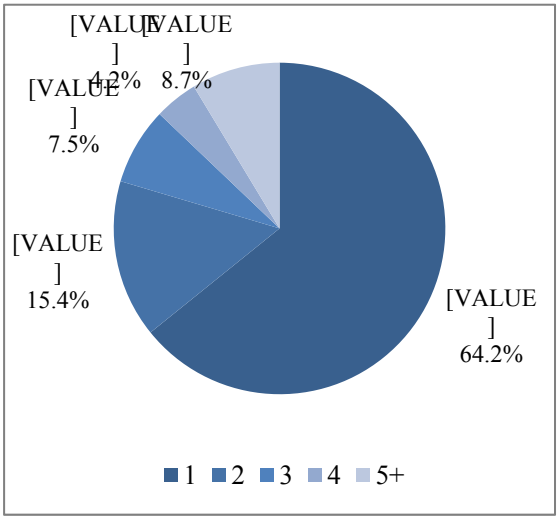


Figure 4. The distribution of citations by the count of contained citees.

Separately, the counts of citees contained by the first-time and succeeding citations are investigated. The first-time citations contain 2.13 citees on the average, while the succeeding citations contain 1.94 citees. Figure 5 shows the specific distribute of both of them by their count of contained citees. For the first-time citations, totally 38.5% of citations cited two citees or more; while for succeeding citations, only 30.1% did. It means the first-time citations are more likely to be perfunctory citations than the succeeding citations. In other words, authors normally cite a reference more casually and perfunctorily for the first time; and then cite it again in the following paragraphs more formally and solemnly. In other words, usually, authors just mention a reference in the beginning, and then seriously use it when citing it later again.

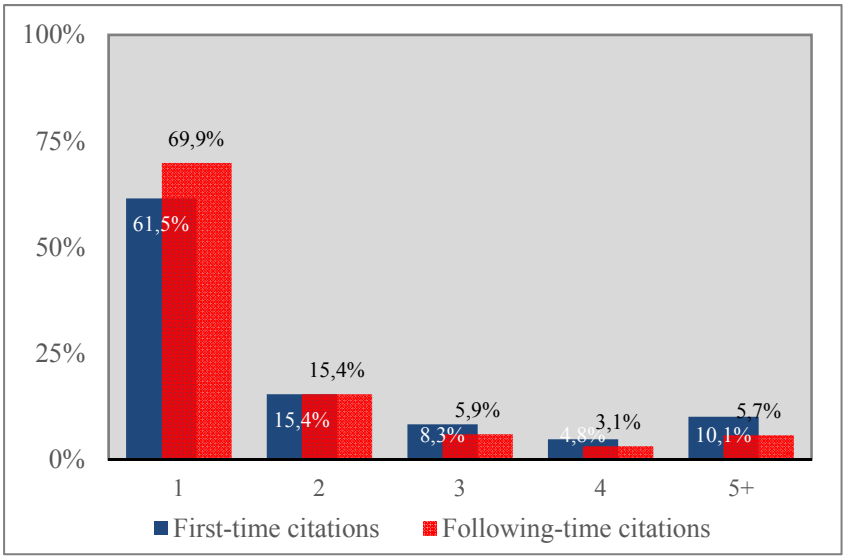


Figure 5. The distribution of the first-time and succeeding citations by their count of contained citees.

Conclusions

Recurring citations are common in scientific publications. In Journal of Informetrics, about 1/4 references are repeatedly cited in citing papers. Although not the mainstream of citation pattern, recurring-citation is undoubtedly a phenomenon that cannot be ignored in full-text citation analysis, an increasing hot research field in recent year.

In this study, we investigate the recurring-citation phenomenon in two perspectives: the citation location and the citation context. In citation location analysis, we find that a reference's recurring citations tend to be located in the same section or closely with each other. It shows that a reference is only cited in a single topic normally. When the topic switches, the reference has little chance to be cited again.

The context of recurring citations contexts are also examined in terms of their citation intensity. As it shown in the result, for a repeatedly cited reference, its first-time citation is usually kind of perfunctory. The reference is always cited accompanied with other references together. When it is cited another time in the following part of the citing paper, the citations are more exclusively and solemnly. Precisely because the succeeding citations are usually more importantly, recurring citations are reasonable and necessary inside scientific articles.

Acknowledgments

The research was supported by the Natural Science Foundation of China (NSFC) under Grant 71103022, 61301227 and 71473028.

References

Agarwal, S., & Yu, H. (2009). Automatically classifying sentences in full-text biomedical articles into Introduction, Methods, Results and Discussion. *Bioinformatics*, *25*(23), 3174–80. doi:10.1093/bioinformatics/btp548

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science*, *40*(4), 284–290.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics*, *7*(3), 583–592. doi:10.1016/j.joi.2013.03.003

Hu, Z., Chen, C., & Liu, Z. (2013). Where are citations located in the body of scientific articles? A study of the distributions of citation locations. *Journal of Informetrics*, *7*(4), 887–896.

Liu, X., Zhang, J., & Guo, C. (2013). Full‐text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, *64*(7), 1852–1863.

Lotka, A. J. (1926). The frequency distribution of scientific production. *Journal of the Washington Academy of Science*, *16*, 317–323.

Oppenheim, C., & Renn, S. P. (2004). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, *51*(5), 225–231.

Pham, S. B., & Hoffmann, A. (2003). A new approach for scientific citation classification using cue phrases. *AI 2003 Advances in Artificial Intelligence*, 759–771.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. *Journal of Advanced Composition* (Vol. 11, p. 272). Cambridge: Cambridge University Press.

Voos, H., & Dagaev, K. S. (1976). Are all citations equal? Or did we op. cit. your idem? *Journal of Academic Librarianship*, (1), 20–21.

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science*, *64*(7), 1490–1503.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley Press.

Do Authors With Stronger Bibliographic Coupling Ties Cite Each Other More Often?

Ali Gazni[1, 2] and Fereshteh Didegah[3]

[1]*ali.gazni@isc.gov.ir*
[1]Islamic World Science Citation Center, Shiraz (Iran)
[2]Regional Information Center for Science and Technology, Shiraz (Iran)

[3]*fdidgah@gmail.com*
Statistical Cybermetrics Research Group, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, (UK)

Abstract
Author bibliographic coupling is extended from bibliographic coupling concept and holds the view that two authors with more common references are more related and have more similar research interests. This study aims to examine the association between author bibliographic coupling strength and citation exchange in Information Science & Library Science and more specifically, in imetrics. The results show that there is a positive and significant association between these two factors in Information Science & Library Science and also in imetrics; however, the correlation is more significant among imetricians. This confirms the Merton's norm of universalism versus constructivists' particularism. A closer investigation of bibliographic coupling and citation networks among thirty highly cited imetricians shows that Thelwall, M. is in strong bibliographic coupling and citation relationships with the majority of authors in the network. He and Bar-Ilan have the strongest ABC and citation relationships in the network. Rousseau, R., Glänzel, W., Bornmann, L., Bar-Ilan, J., and Leydesdorff, L. are also in strong ABC relations with each other as well as other authors in the network.

Conference Topic
Citation and co-citation analysis

Introduction
Bibliographic coupling (BC), first introduced by Kessler in 1963, refers to the number of common references between two articles. The more the number of common references between two articles, the more intellectually related they are.

In contrast with co-citation analysis (CA) requiring strength signals (number of citations), BC could help in research fronts detection even with weak signals (Glänzel & Czerwon, 1996). Kuusi and Meyer (2007) claimed that BC has never been used for exploring technology foresight and rare studies used it for research evaluation purposes. However, they used BC for anticipating technological breakthroughs. Yan and Ding (2012) compared different types of networks, including citation based and non-citation based networks at institutional level, and found that BC and AC networks have high similarity and also found that AC has a high similarity with citation networks. Boyack, Börner and Klavans (2009) applied BC to mapping the structure and evolution of research publications in Chemistry. Soó (2014) proposed age-sensitive BC, so if two documents share recent references, they are more related than those sharing older references. Hence, not only the number of common references, but also their age, influences the extent of relatedness between two research works. Van Raan (2005) also reported that intellectual relatedness between two documents could be better obtained through using common

references that are more recent. BC is an effective way for science mapping, research fronts detection and information retrieval (See Jarneving, 2007, 2005; Morris, Yen, Wu, & Tesfaye, 2003; Qiu, 2007). Peters, Braam and van Raan (1995) investigated chemical engineering publications and found that publications with common citations to highly cited papers are more related. White et al. (2004) claim that intellectual ties based on shared references could serve as a better predictor for citations between authors than social ties.

Author bibliographic coupling (ABC), first proposed by Zhao and Strotmann (2008), is extended from BC concept and holds the view that two authors with more common references have more similar research interests. They mentioned that BC is fixed when two articles are published but ABC is constantly evolving over time as the two authors' oeuvre grows. Ma (2012) stated that ABC has an advantage in providing a more comprehensive and concrete map of intellectual structure of the fields and detecting their research fronts in comparison to author co-citation analysis (ACA). The very few studies on ABC did only an author coupling analysis of intellectual structure of few subject fields. For example, using a combination of ACA and author bibliographic coupling analysis (ABCA), Zhao and Strotmann (2014) sought to predict future research trends in information science (IS). They studied research fronts and knowledge bases of IS and also the structural evolution of IS between two 5-year periods (2001-2005 and 2006-2010). They found ABCA an appropriate method to investigate authors' specific research interests in IS and suggested using ACA and ABCA together to better investigate intellectual structure of a subject domain. The same combined method was used in Byun and Chung (2012) to study the research trends of authors in social welfare science; they also suggested using both ACA and ABCA together to investigate traditional and future research trends of a specific domain.

The extent to which two authors are coupled through common references is measured by ABC strength which has different methods to calculate it: Simple, minimum and combined methods (Ma, 2012). Rousseau (2010) also proposed a simple method for calculating the relative ABC by dividing the number of common references between two authors by the total number of their references. Frequency of common references was simply used to measure ABC strength in this study.

No research on the association between ABC strength of two authors and number of citations exchanging between them is found, so this study seeks to examine this relationship in Information Science & Library Science (IS&LS) and more specifically, in imetrics. Therefore it aims to examine the correlation between ABC strength measured by the number of common references between two authors and the number of citations exchanged between them.

Research questions

According to the normative theory of citation, citations are indicators of the cognitive or intellectual influence of a scientific work (Merton, 1973). In a scientific paper, citations can be concept markers (Small, 1978), however, and can transfer knowledge and help with its enlargement (Merton, 1988). As a result, methods like CA have been used for mapping intellectual structure in science (Small, 2004), where BC is used for the same purpose. Hence, common references between pairs of documents, authors, journals or institutions show the extent to which they are related. For instance, two authors who

share a larger number of common references are likely to do research on a narrow area and exchange a high number of citations. Counting citations between two authors with different BC strengths, not only could support Robert K. Merton's norm of universalism versus constructivists' particularism, but also shows any possible difference by the number of common references as a measure of relatedness and types of authors (i.e. highly cited vs. less cited authors).

The theories of citation, normative view vs. social constructivist view, will be examined through answering these questions. The normative theory of citation holds that citations reflect the scientific quality and merits of research outputs because citers use them to reward the works of their colleagues (Small, 2004; White, 2004; MacRoberts & MacRoberts, 1987; Merton, 1973) whereas the social constructivist theory holds that authors use the references to support their own claims and points made. This latter theory emphasises factors affecting citations other than the quality and content of the cited article (White, 2004; Baldi, 1998; Gilbert, 1977).

Given that BC shows relatedness, a positive association between the number of common references and number of citations between two authors will confirm that citations are made for the matter of 'relatedness' and are not perfunctory.

To reach the research goals, this study seeks to answer these questions:

1. Do two authors with a higher number of common references cite each other more often?
2. Is the above association stronger for highly cited authors than other authors?

Methodology

Data collection:

Documents published during 1990-2012 in the journals of Information Science & Library Science (IS&LS) were extracted from Thomson Reuters Web of Science (WoS). This time period is current and consists of a reasonable number of years for investigating the relationship between number of common references and citations exchanged between authors. WoS indexes the mainstream of research and the most prestigious journals in different fields of science; however, a large number of journals in WoS come from a small number of international publishers (Didegah & Gazni, 2011).

Author names disambiguation:

The author names were disambiguated by improving Gazni & Thelwall (2014) method, resulting in 98.2% precision and 92.7% recall. The co-authorship network of authors was used for the improvement. For example, A is a disambiguated author and B is his/her co-author. The papers written by both A and B as co-authors were appended to A's articles. Author names' disambiguation will improve the accuracy of research on author level analysis by distinguishing one name that belongs to several different people and conflating the name variants of a single person.

Calculations:

To make the processing manageable, a random sample of 385 authors with any properties out of all authors who have at least one paper in the journals of IS&LS during 1990-2012

was chosen. The number of common references between these 385 authors and all other authors in the field were counted, where the joint papers were eliminated either for counting the number of common references or for counting the number of citations made and received between each pair of authors. Only citations made and received from the journals in the field were processed for either counting the number of citations between authors or counting the number of common references among them. A list of authors who have at least one common reference with the authors in the sample, and also exchanged citations with them, was created for each author in the sample. For a closer investigation of the association between the number of common references and citations between pairs of authors and also of ABC networks, a sample of highly cited authors in imetrics was taken into account. For this purpose, thirty highly cited imetricians introduced in Abrizah and colleagues (2014) were selected for further analysis. The main reason for taking this sample into account is that these are prolific authors in a specific domain, publishing for a long time and have an excellent knowledge of the domain, its publications and researchers. This is while in the sample of authors from IS&LS, there may be less prolific authors, such as students who publish for a short period of time and then disappear from the research area, and their unfamiliarity with the area will affect their reference and citation behaviours. Therefore, a sample of thirty highly cited imetricians is a consistent sample for showing the association between ABC strength and citation exchange between pairs of authors.

Results

The association between number of common references (BC strength) and number of exchanged citations between pairs of authors in IS&LS

Spearman correlation was tested for the association between the number of common references and the number of citations exchanging between pairs of authors. The results show positive significant correlations between the number of times two authors cited each other and the number of common references between them. The correlation was tested for different groups of pairs of authors with one to 300 common references; it is stronger for the groups of authors with 300 common references than those with a single common reference (Table 1). Therefore, as the number of common references between two authors increases, the number of citations between them also increases. Table 1 shows the increase trend; however, the correlation fluctuated as the number of common references increases but tends to increase. To put it in another way, when the bibliographic coupling strength is stronger between two authors, they tend to cite each other more often. Author bibliographic coupling strength shows how strongly two authors are intellectually related. So, more intellectually related authors cite each other more often. This result confirms the normative theory of citation holding the view that authors cite relevant works, and citations reflect scientific merit and quality.

Table 1. Spearman correlation between ABC strength and number of citations in IS&LS.

No of common refs	Spearman correlation
1	0.31
10	0.36
20	0.35
30	0.38
40	0.37
50	0.37
60	0.39
70	0.38
80	0.36
90	0.39
100	0.4
150	0.46
200	0.47
250	0.58
300	0.61

ABC strength and citation relationship among thirty highly cited authors in imetrics

Thirty highly cited authors in imetrics identified in Abrizah and colleagues (2014) were chosen for a closer investigation of research goals. The main research question on the association between ABC strength and number of exchanged citations was also examined for this group of highly cited authors. Spearman correlation test shows a strong positive association between the number of common references and the number of citations between the authors (Spearman's rho= 0.771, p-value< 0.001), once more confirming the significance of content relevance in citation behavior and normative view of citations.

Moreover, all ABC relations are mapped between each pair of highly cited authors (See Fig. 1). Based on the results, all thirty authors are in BC relationships with all or some of other authors in the network except for Griffith, BC. During 1990-2012, he has published 4 papers in imetrics and has no common references with any of the highly cited authors.

Thelwall, M. is in strong BC relationships with all other authors except with Vanleeuwen, T.N. (only one common reference) and VanRaan, A.F.J. (three common references). He and Bar-Ilan, J have shared the highest common references in the network (4,527 common references) and they have exchanged a large number of citations in the network (118 citations). Thelwall, M. has more than 100 common references with 18 authors in the network. He is also in a strong BC relationship with Vaughan, L. (2,725 common references). Thelwall, M. has also exchanged the highest number of citations in the network with Vaughan, L. (195 citations). He has also strong BC ties with seven others, Leydesdorff, L., Ingwersen, P., Rousseau, R., Cronin, B., Glänzel, W., and Egghe, L., respectively.

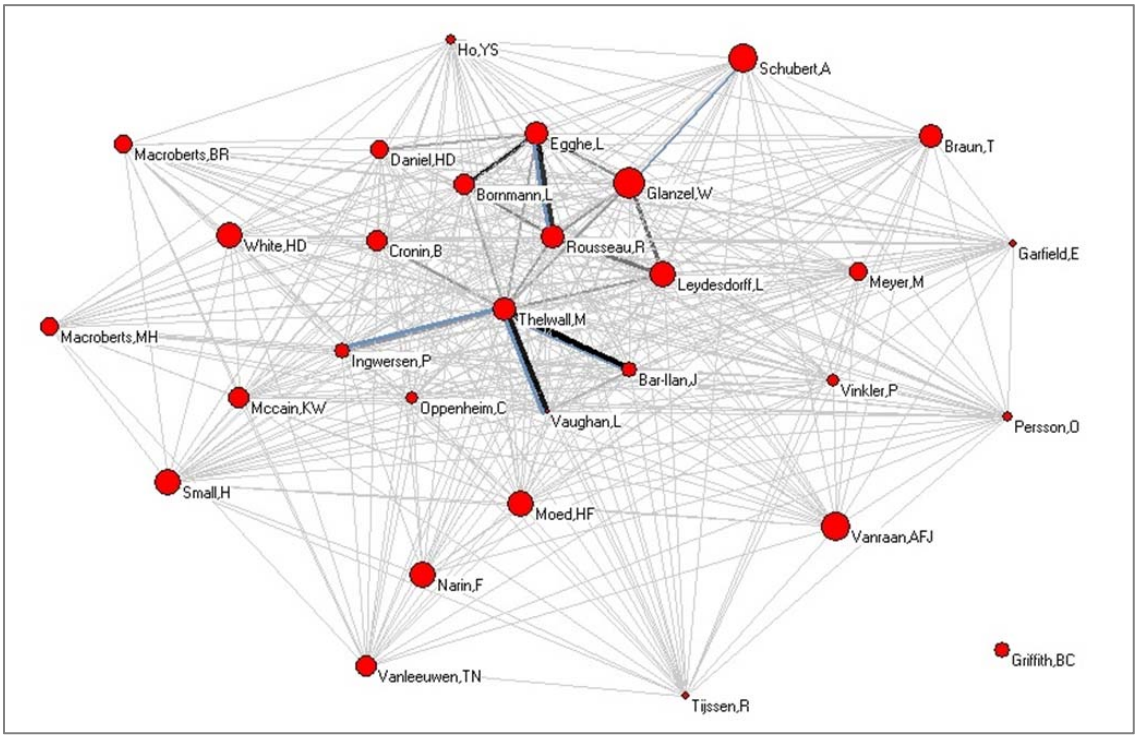


Figure 1. ABC among highly cited authors in imetrics; the black lines show ABC relations and the width of the lines shows ABC strength between pairs of authors; the blue lines show the strongest citation relations in the network and the width of the lines shows the number of citations exchanged between pairs of authors; the size of vertices shows the number of other highly cited authors in the network that each author is in an ABC relation with.

Another strong ABC relationship, and also citation relationship, is seen between Rousseau, R and Egghe, L. (2,270 common references and 175 exchanged citations). Rousseau, R is also in strong BC relationships with other authors in the network. He has strong BC ties with Leydesdorff, L., Bornmann, L., Glänzel, W., and Thelwall, M., respectively.

Glänzel, W., Bornmann, L., Bar-Ilan, J., and Leydesdorff, L. are also in strong BC relationships with other authors in the network. They also have strong citation relationships with each other as well as other highly cited authors.

The correlation between ABC strength and citation exchange in imetrics in comparison with IS&LS

The correlation between the number of common references and the number of citations for top thirty imetricians was examined first amongst themselves and then between them and all other authors in IS&LS with whom they are in BC or citation relationships. As shown in Figure 2, a stronger relationship exists between the authors in the first group than in the second one and regarding the top thirty imetricians, the correlation varies from one author to another one.

For each highly cited imetrician, the proportions of common references with each in-group authors was estimated. Fig. 3 shows that each highly cited author is in a BC relationship with 27 other in-group authors. For example, about 24% of references of

each author are common with one other author. The author distribution of the number of common references with other authors demonstrates a core-scatter shape.

Core references in imetrics

We tried to go further than author couples for common references and identified a number of common references between three and more authors. The thirty highly cited authors in imetrics were examined for this purpose.

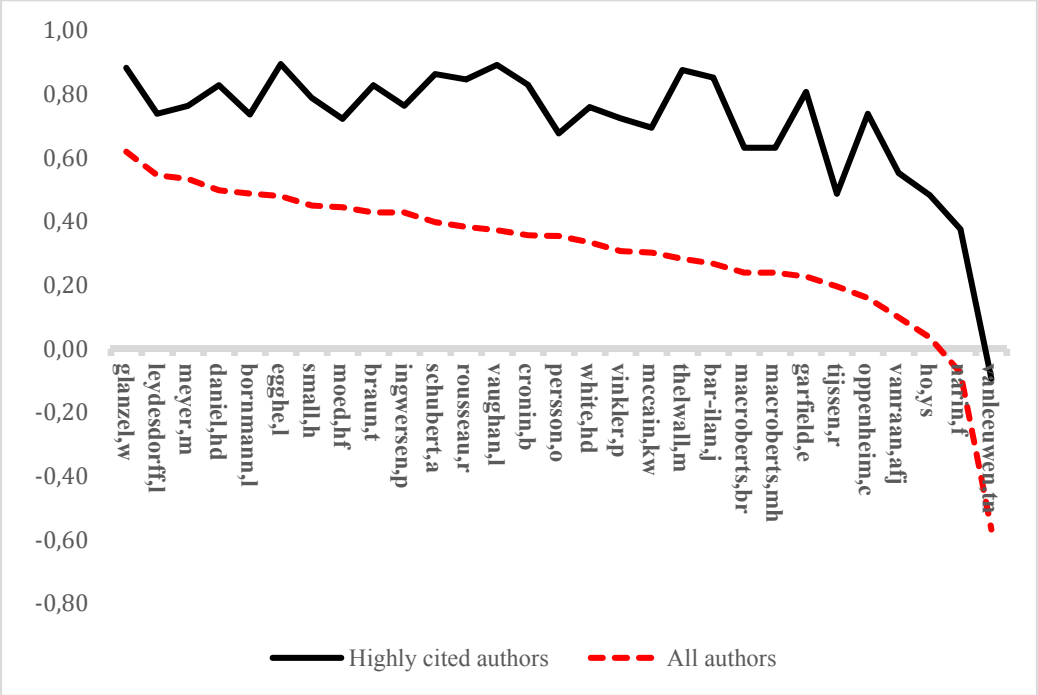


Figure 2. ABC strength and citation correlation between highly cited authors and all authors in IS&LS.

The interesting result is that seventeen highly cited imetricians have one reference in common. The common reference is Hirsch's paper on H-index (Hirsch, J.E. (2005): An index to quantify an individual's scientific research output. *Proceedings of the national academy of sciences of the United States of America*, 102 (46)). Egghe, L., Rousseau, R., and Bornmann, L. have cited this paper more than thirty times in their publications showing that the H-index is one of their common research interests. It is interesting to note that Egghe, L. and Rousseau, R also have the strongest citation relationship with each other in the network (seventeen5 citations have been exchanged between them) and these two imetricians are also in a strong citation relationship with Bornmann, L. with Bornmann, L. being the fourth top author in citation relationships with both Egghe, L. and Rousseau, R. The strong citation relationships between these authors are mainly due to their similar research interests, one of which is H-index. Twelve highly cited authors have simultaneously five references in common which are listed in Table 2. Eleven authors have nine references and ten authors have eleven references in common.

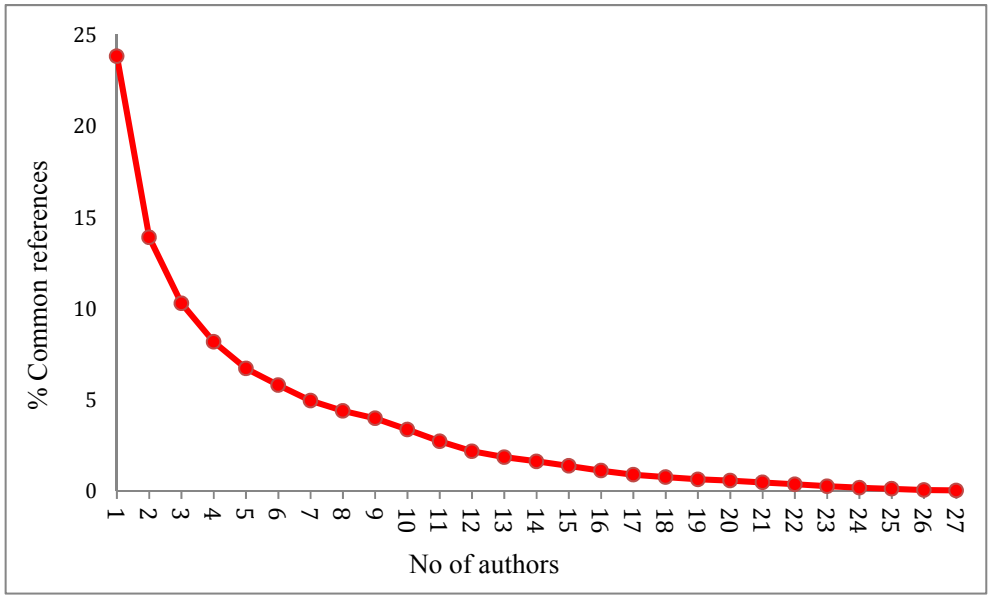


Figure 3. The proportion of common references between each of thirty highly cited imetricians and other in-group authors.

Table 2. Five common references between twelve highly cited imetricians.

VanRaan, A.F.J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3).
Meho, L. & Cronin, B. (2006). Using the h-index to rank influential information scientists. *JASIS&T*, 57(9).
Glänzel, W., Thijs, B., & Schlemmer, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58(3).
Macroberts, B.R. & Macroberts, M.H. (1996). Problems of citation analysis. *Scientometrics*, 36(3).
Moed, H.F., Vanleeuwen, T.N., & Debruin, R.E. (1995). New bibliometrics tools for the assessment of national research performance- database description, overview of indicators and first applications. *Scientometrics*, 33(3).

Discussion and conclusion

This study examined the association between author bibliographic coupling strength and the number of times authors cited each other. The results of the study on authors in IS&LS showed that there is a positive and significant correlation between ABC and exchanged citations between two linked authors confirming that authors are citing related authors and relevant research works in their field (Table 1). This finding opposes the social constructivist view holding that authors cite others for some other reasons than relevance or rewarding the cited author, but it confirms the normative theory of citations. A group of thirty highly cited authors in imetrics were also examined for this purpose. The result of the association between ABC and the number of citations shows a positive strong correlation between ABC and exchanged citations between imetricians. Therefore, highly cited authors in imetrics are in strong BC relationships with whom they also have strong citation relationships.

The number of common references between pairs of authors was accepted as a measure of relatedness between them. Therefore relatively, the higher number of common

references between two authors, especially in a long-term period, could show the extent to which they are working in similar research areas; however, authors may change their research interests over time due to changes in the research fields. The higher number of citations between two authors with higher number of common references, when they are not co-authors, could probably show that they cite each other since they may work on similar research areas and also for the matter of relevancy.

ABC relations between the thirty highly cited imetricians were examined and mapped and strong relationships were determined. Thelwall, M. and Bar-Ilan, have the strongest ABC relationship in the network; they are also in a strong citation relationship. Rousseau, R., Glänzel, W., Bornmann, L., Bar-Ilan, J., and Leydesdorff, L. are also in strong ABC relations with each other as well as other authors in the network. In an investigation of the number of common references in groups of two and more imetricians, smaller groups have a larger number of references in common while larger groups have fewer numbers of common references. For example, seventeen imetricians have only one reference in common while some two-author groups have more than a thousand common references. The latter groups presumably work on narrow research areas. Larger groups with fewer number of common references suggest membership in a wider research area. The results show that a maximum of seventeen authors have one reference on H-index in common. Authors citing this single paper are also in strong citation relationship with each other.

Comparing the correlation between number of common references and number of exchanged citations for highly cited imetricians and all authors in IS&LS related to Fig. 2 shows that number of common references between imetricians increases the probability of higher citations between them more than that of IS&LS. Moreover, ABC relationship or common references with each single author may result in different number of citations with him/her.

Intuitively, considering the core-scatter distribution of citations to papers in the science network, an author probably has common references with a large number of other authors, while he/she probably has more common references with a fewer number of other authors (Fig. 3). The author presumably has more related research interests with the latter group of authors where some of them may belong to the same research community.

The number of common references and citations between pairs of authors could be also influenced by the number of papers published by the authors. For example, two authors may have five common references whilst the first author only published a single paper during his/her entire research life and the second one published more than twenty papers. The first author will have fewer common references with any other authors in the field than the second author and he/she will have less opportunity to cite other authors due to his/her short research life. So authors' research lifetime in the science network (e.g. newcomers, students, faculty members and professional researchers) does matter. Authors with a longer research life have more chances to know other researchers in similar research fields and they also have extra opportunities to focus on more specific and narrow research topics, compared to authors with a shorter research lifetime. Hence, a stronger association between the number of common references and citations exchanged between authors is found for the former group.

Science network and its attributes are continuously changing over time and a research specialty may appear or disappears after a while; authors may also change their research interests during their research lifetime. In the current study, a longer time span is used to

show that clustering authors, based on more recent common references, may be replaced by a shorter one, which could result in a stronger relationship between the bibliographic coupling network and the citation network. According to the results of current studies, authors with a longer research lifetime and more citations demonstrate a stronger relationship between their number of common references and citations. However, even weak ties in bibliographic coupling networks could also be used for research front detection purposes. Bibliographic coupling is not enough for mapping intellectual structure of science and measuring relatedness by itself. Thus, as with previous studies, it is better to be combined with other methods, such as co-citations, to realise better results.

References

Abrizah, A., Erfanmanesh, M., Rohani, V. A., Thelwall, M., Levitt, J. M., & Didegah, F. (2014). Sixty-four years of informetrics research: productivity, impact and collaboration. *Scientometrics*, *101*(1), 569-585.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: a network-analytic model. *American Sociological Review*, *63*, 829-46.

Boyack, K. W., Börner, K., & Klavans, R. (2009). Mapping the structure and evolution of chemistry research. *Scientometrics*, *79*(1), 45-60.

Byun, J-H & Chung, E-K (2011). Domain Analysis on Electrical Engineering in Korea by Author Bibliographic Coupling Analysis. *Journal of Information Management*, *42*(4), 75-94.

Didegah, F., & Gazni, A. (2011). The extent of concentration in journal publishing. *Learned Publishing*, *24*(4), thirty3-310.

Gazni, A. & Thelwall, M. (2014). The long-term influence of collaboration on citation patterns. *Research Evaluation*, doi: 10.1093/reseval/rvu014.

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science*, *7*, 113-122.

Glänzel, W., & Czerwon, H.J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional, and institutional level. *Scientometrics*, *37*, 195–221.

Jarneving, B. (2005). Acomparison of two bibliometric methods for mapping of the research front. *Scientometrics*, *65*, 245–263.

Jarneving, B. (2007). Bibliographic coupling and its application to research front and other core documents. *Journal of Informetrics*, *1*, 287–307.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation*, *14*(1), 10-25.

Kuusi, O., & Meyer, M. (2007). Anticipating technological breakthroughs: Using bibliographic coupling to explore the nanotubes paradigm. *Scientometrics*, *70*(3), 759-777.

Ma, R. (2012). Author bibliographic coupling analysis: A test based on a Chinese academic database. *Journal of Informetrics*, *6*(4), 532-542.

MacRoberts, M. H. & MacRoberts, B. R. (1987). Another test of the normative theory of citing. *Journal of the American Society for Information Science*, *38,* 305-306.

Merton, R. K. (1973). The normative structure of science. The sociology of science: Theoretical and empirical investigations, 267.

Merton, R. K. (1988). The Matthew effect in science II: cumulative advantage and the symbolism of intellectual property. *Isis*, 79, 606-23.

Morris, S. A., Yen, G., Wu, Z., & Tesfaye, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, *54*(5), 413–422.

Peters, H. P., Braam, R. R., & van Raan, A. F. (1995). Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*, *46*(1), 9-21.

Qiu, J. P. (2007). *Methods of citation analysis*. In Informetrics. Wuhan: Wuhan University Press.

Rousseau, R. (2010). Bibliographic coupling and co-citation as dual notions. In L. Birger (Ed.), *A Festschrift in honour of Peter Ingwersen, special volume of the e-zine of the ISSI, June 2010*, 173–183.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science*, *8*(3), 327-340.

Small, H. (2004). On the shoulders of Robert Merton: towards a normative theory of citation. *Scientometrics*, *60*, 71-79.

Soós, S. (2014). Age-sensitive bibliographic coupling reflecting the history of science: The case of the Species Problem. *Scientometrics*, *98*(1), 23-51.

Van Raan, A. F. (2005). Reference-based publication networks with episodic memories. *Scientometrics*, *63*(3), 549-566.

Yan, E., & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, co-citation networks, topical networks, co-authorship networks, and co-word networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313-1326.

White, H. D. (2004). Reward, persuasion, and the Sokal Hoax: a study in citation identities. *Scientometrics*, *60*(1), 93-120.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070–2086.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006–2010: An author co-citation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology*, *65*(5), 995-1006.

The Research of Paper Influence Based on Citation Context -- A Case Study of the Nobel Prize Winner's Paper

Shengbo Liu[1], Kun Ding[1], Bo Wang[2], Delong Tang[1], Zhao Qu[1]

[1] *liushengbo1121@gmail.com, dingk@dlut.edu.cn, tangdl@mail.dlut.edu.cn, qz_031@mail.dlut.edu.cn*
WISElab, Dalian University of Technology, No. 2, Linggong Road, Ganjingzi district, Dalian, 116024, China

[2] *bowang1121@gmail.com*
School of Management, Dalian Ploytechnic University, #1 Qinggongyuan, Dalian, 116000,China

Abstract
Citation context was used to measure the influence of highly cited papers. The themes of citation context were analyzed with bibliometrics methods. The citation context was classified into three categories as positive, negative and neutral. And the neutral citations were also classified into three sub categories, related work in background or introduction, theoretical foundation, and experimental foundation. The citation contexts of a highly cited paper of O'Keefe were extracted as the experiment data set. The results showed that the co-occurrence method was very useful for describing the themes of the citation contexts. The citation contexts of the selected paper were divided into five themes. The classification of citation contexts could provide more information about how and why a paper was highly cited. There was no negative citation in this experiment, and more than 10% citation contexts were positive citation. About 50% of the neutral citations were belonging to related work in background or introduction. The detailed influence of the target paper was also illustrated in our research.

Conference Topic
Citation and co-citation analysis

Introduction
Citation frequency is a commonly used indicator to measure the importance of a paper. Recently, *Nature* asked Thomson Reuters, which now owns the SCI, to list the 100 most highly cited papers published from 1900 to 2014. The results revealed some surprises, many of the world's most famous papers do not rank in the top 100 (Van Noorden, Maher, & Nuzzo, 2014). John P. A. Ioannidis and colleagues surveyed the most-cited authors of biomedical research for their views on their own influential published work. The results showed that the most important paper was indeed one of author's most-cited ones. But they described most of their chart-topping work as evolutionary, not revolutionary (Ioannidis, Boyack, Small, Sorensen, & Klavans, 2014). Although the citation frequency is an important indicator to measure the influence of a paper, it is hard to reveal why others always cited this paper and what influence it makes. Citation context refers to the text surrounding the references (Henry Small, 1982). It could provide more detailed information about citation.

In this paper, we take John O'Keefe's (Nobel Prize winner in Physiology or Medicine 2014) most influence paper as instance. The influence of this paper will be analyzed based on citation context. Our analysis will provide a richer understanding of which knowledge claims made by O'Keefe have had the greatest impact on later work.

Related work

Citation context analysis

Citation context can be defined as the sentences that contain the citation of a particular reference. For example, the sentence "This comparison is made using BLASTX [18]" is the citation context of reference [18].

Citation content can be used to identify the nature of a citation. The attributions and functions of a cited paper can be identified from the semantics of the contextual sentences (A. Siddharthan, Teufel, S., 2007). Nanba and Okumura (Nanba, 1999, 2005) collected citation context information from multiple papers cited by the same paper and generated a summary of the paper based on this citation context information. They also extracted citing sentences from citation contexts and generated a review. Elkiss et al. (Elkiss, 2008) generated the citation summarization based on citation context to describe the topic of cited paper. Mei (Mei, 2008) and Mohammad (Mohammad, 2009) found that the summarization of citation contexts is very different from the abstract of the cited reference. Liu and Chen(Liu & Chen, 2013) studied the differences between latent topics in abstracts and citation contexts. The results showed that topics from citing sentences tend to include more specific terms than topics from abstracts of citing papers. Nakov (Nakov, 2004) referred to citation contexts as citances – a set of sentences that surrounding a particular citation. Citances can be used in abstract summarization and other Natural Language Processing (NLP) tasks such as corpora comparison, entity recognition, and relation extraction. Small (H. Small, 1979) studied the context of co-citation and analyzed the context in which the co-citation paper mentioned. In addition, he analyzed the sentiment of the co-citation context (H. Small, 2011).

Anderson (Anderson, 2010) analyzed the citation contexts of a classic paper in organizational learning which was published by Walsh and Ungson in the Academy of Management Review. The results provided a richer understanding of which knowledge claims made by Walsh and Ungson have been retrieved and have had the greatest impact on later work in the area of organizational memory, and also what criticisms have been leveled against their claims. Chang(Chang, 2013) compared the citing topics of *Little Science*, *Big Science* in natural sciences and humanities and social sciences through citation context. He found that the citing topics in natural sciences and humanities and social sciences were very similar, but the cited motivation had some differences.

The classification and function of citation context

Citation context contains the direct related information between cited paper and citing paper. It could be used to reveal the nature of a citation. The cited motivation of each citation is different, so the value of each citation will be different. For example, some of the citation contexts support the claims in the cited paper, and some of them may take the opposite opinion about the views or methods in the cited paper. Spiegel-Rösing (Spiegel-Rösing, 1977) studied the citation context of Science Study in 1977 and classify the citation context into 13 categories, including use the data of cited paper, use the method of cited paper, compare the work of cited paper and citing paper and so on. In order to provide more information for literature management, Teufel reclassified the above 13 categories into four categories, (1) Explicit statement of weakness, (2) Contrast or comparison with other work, (3) Agreement /usage /compatibility with other work, (4) A

neutral category(Teufel, Siddharthan, & Tidhar, 2006) . Cue phrases were used to identify the category of each citation context. The similar method was also employed in Liu's (Liu et al.) work in which the citation context was classified as positive citation, negative citation, and neutral citation. Other people like Small (Henry Small, 1982), McCain (McCain & Turner, 1989), Siddharthan (A. Siddharthan & Teufel, 2007), Swales (Swales, 1990) also did some work about citation context classification.

Data and Method

Our procedure consists of three major components, 1. Data collection and preprocessing, 2. Theme analysis of citation context, and 3. The classification of citation context. Details are explained in corresponding sections.

Data collection and preprocessing

The 2014 Nobel Prize in Physiology or Medicine is awarded to Dr. John M. O'Keefe, Dr. May-Britt Moser and Dr. Edvard I. Moser for their discoveries of nerve cells in the brain that enable a sense of place and navigation. The scientific background was introduced in the document "The Brain's Navigational Place and Grid Cell System" . The keywords this document were selected manually and used to retrieve the award field in Web of Science. The search query was shown as follows：

TI=(hippocamp* AND (place OR Position* OR spatial)) OR (("grid cell*" OR Position* OR Navigation* OR spatial OR place) And ("entorhinal cortex" OR brain OR cerebral))

The time period was from 1945 to 2014, and 4441 papers were collected.

The citation context collection was built through three steps. First, the paper with the first author O'Keefe and the highest citation frequency was selected. Second, the papers which cited the chosen paper were downloaded with full text. Actually, we could just find less than 20% full text papers. Last, the citation contexts of the chosen papers were extracted from the full text for further analysis. The extraction method has been introduced in our previous work (Liu & Chen, 2013).

The theme analysis of citation context

The theme analysis includes two tasks. One is counting the frequency of noun phrases appeared in citation contexts. Another is mapping the co- occurrence network of noun phrases.

Part-of-speech is needed before extract noun phrases. There are many tools to label part-of-speech, such as PosTagger, CLAWS POS tagger. Stanford Log-linear Part-Of-Speech Tagger (Toutanova & Manning, 2000) was employed in this work, which was developed by NLP group of Stanford University. The noun phrase formation rules was designed with the same method described in Wang's paper (Wang, Liu, Ding, Liu, & Xu, 2014). When counting the frequency of noun phrases. If one citation context contains two same noun phrases, it will count once.

In bibliometrics analysis, co-occurrence method was often used to detect subjects/themes (Hofer, Smejkal, Bilgin, & Wuehrer, 2010; Lee, 2008; Zhang et al., 2012). But few of the related works use this method to detect the theme of citation context. Pajek software was employed to mapping the noun phrases co-occurrence network of citation context. We expect to identify the citing themes through drawing the co-occurrence map.

The classification of citation context

Following the work of Spiegel-Rösing (Spiegel-Rösing, 1977) and Teufel (Teufel et al., 2006), citation contexts will be classified into three categories as positive, negative and neutral. Table 1 shows the description of each category. We divided the positive category into three sub categories and the negative category into two sub categories.

Table 1. The description of each category

Category		Description
Positive	(1)	Affirm or praise the cited work
	(2)	Apply the methods, tools or databases of the cited paper
	(3)	Comparison of methods and results
Negative	(1)	Point out the weakness of the citation
	(2)	Contain negative cue words
Neutral	(1)	Contain no cue words

To our knowledge, the proportion of neutral citations occupy more than others. So we will classify the neutral citation into three sub categories based on the citation motivation.

(1) Related work in background or introduction. Introduce the related work with no comments.
(2) Theoretical foundation. Concepts, principles, methods, or results which will be used in citing paper.
(3) Experimental foundation. Including experimental conditions, processes, environment, and results.

Results and discussion

Target paper detecting

Table 2 shows top ten highly cited papers in Nobel Prize award field. The highest cited paper was "PLACE NAVIGATION IMPAIRED IN RATS WITH HIPPOCAMPAL - LESIONS" which published in *Nature* in 1982. It has been cited 3589 times. Although this paper got highly cited in Nobel Prize award field, it did not appear in "Scientific background" document, which was the instruction of why the winner got this prize. The author Morris R.G.M did not get Nobel Prize. The Nobel Prize was given to the author of the second highest cited paper "HIPPOCAMPUS AS A SPATIAL MAP - PRELIMINARY EVIDENCE FROM UNIT ACTIVITY IN FREELY-MOVING RAT". The result is similar to the work of Van Noorden (Van Noorden et al., 2014) that the Nobel Prize winner's paper did not get the highest citation frequency.

O'Keefe who is the Nobel Prize winner had three papers ranked in top ten high cited papers in Nobel Prize award field. The highest cited paper had been cited 1812 times. This paper was selected as the target paper. The seminal work of this paper was the discovery of "place cell".

It is hard to download all the 1812 citing papers. So 200 citing papers with full text were selected in our experiment. There were 228 citing sentences. The target paper was average cited 1.14 times in each citing paper.

Table 2. Top ten high-cited papers in Nobel Prize award field.

Author	Title	Journal	Year	Cited frequency
Morris, R. G. M., P. Garrud, et al	PLACE NAVIGATION IMPAIRED IN RATS WITH HIPPOCAMPAL-LESIONS	Nature	1982	3589
Okeefe, J. and Dostrovs.J	HIPPOCAMPUS AS A SPATIAL MAP - PRELIMINARY EVIDENCE FROM UNIT ACTIVITY IN FREELY-MOVING RAT	Brain Research	1971	1812
Okeefe, J. and M. L. Recce	PHASE RELATIONSHIP BETWEEN HIPPOCAMPAL PLACE UNITS AND THE EEG THETA-RHYTHM	Hippocampus	1993	1033
Tsien, J. Z., P. T. Huerta, et al	The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory	Cell	1996	919
Grant, S. G. N., T. J. Odell, et al	IMPAIRED LONG-TERM POTENTIATION, SPATIAL-LEARNING, AND HIPPOCAMPAL DEVELOPMENT IN FYN MUTANT MICE	Science	1992	827
Hafting, T., M. Fyhn, et al	Microstructure of a spatial map in the entorhinal cortex	Nature	2005	773
Cohen, L., S. Dehaene, et al	The visual word form area - Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients	Brain	2000	755
Burgess, N., E. A. Maguire, et al	The human hippocampus and spatial and episodic memory	Neuron	2002	669
Packard, M. G. and J. L. McGaugh	Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning	Neurobiology of Learning and Memory	1996	666
Okeefe, J	PLACE UNITS IN HIPPOCAMPUS OF FREELY MOVING RAT	Experimental Neurology	1976	657

The themes of citation context

299 noun phrases were extracted from the citation contexts. Table 3 listed twenty high frequency noun phrases. The term “place cell” got the highest frequency of 76, because the most contributing work of the target paper was the discovery of place cell. Hippocampus, environment, rat, fire, neuron were all the important terms in target paper. Some of the terms were not mentioned in the target paper, such as cognitive map and ca3.

Table 3. Top twenty high cited papers in Nobel Prize award field.

No.	Noun phrase	Frequency	No.	Noun phrase	Frequency
1	place cell	76	11	discovery	17
2	hippocampus	74	12	place field	15
3	environment	55	13	rodent	13
4	rat	44	14	ca3	13
5	animal	40	15	space	12
6	cell	31	16	ca1	12
7	location	29	17	position	11
8	fire	25	18	pyramidal cell	9
9	cognitive map	19	19	region	9
10	neuron	18	20	navigation	9

Figure 1 showed the co-occurrence map of the noun phrases. Each node represents a noun phrase. The size of the node was proportional to the number of terms co-occurred with it. We set the co-occurrence threshold as more than once and got 71 nodes in the map.

The map could divide into five parts manually based on the relationship of terms. Part A was mainly involving navigation, which was not mention too much in cited paper. It was the following research of place cell. Part B was related to neuron region, including CA1 and CA3. CA1 was discussed in the cited paper, but CA3 was found in the later work. Part C was related to experimental process about firing pattern of rat. Part D was the experimental environment. The definition of place field was widely cited. Part E was about the concept of place cell.

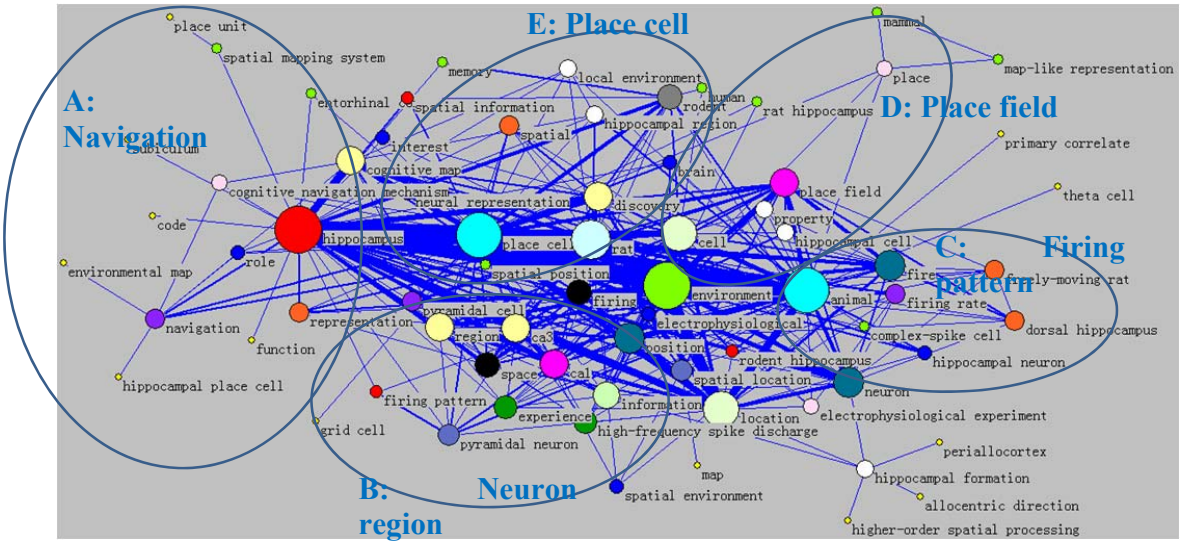


Figure 1. Co-occurrence map of the noun phrases.

Table 4. Example of positive citations

No.	Positive citation
1	The discovery of place cells [1]-[5] in the hippocampal regions of rats **consolidated** the idea that hippocampus probably represents a cognitive map of the local environment of an animal……
2	The concept of cognitive map for navigation, carried out mainly by Tolman [10], was **fuelled** by the discovery of the so-called place cells in the hippocampus of the rat and has widely increased our understanding of cognitive navigation mechanisms [11]
3	The **breakthrough** came in 1971 with the discovery of the rat s cognitive map in the cells of the hippocampus [16]……
4	The idea of the formation of a cognitive map was first proposed by Tolman [45] in the late 40s and was later **supported by the discovery** of place cells by o keefe and dostrovsky [35]
5	The **striking discovery** of place cells in the rat hippocampus [51] has triggered a wave of interest on spatial learning that holds until today

Table 5. Sub categories distribution of neutral citations

Category	Related work	Theoretical foundation	Experimental foundation
Counts	114	49	41

The classification results

The classification results showed that most of the citations were neutral citation. There was no negative citation in our datasets. 24 of 228 citation contexts were positive citations and 204 citations were neutral citations. Table 4 listed some examples of positive citations.

The sub categories distribution of neutral citations was shown in table 5. Nearly half of the citations were cited as related work. Theoretical foundation had 49 citations, and most of them were related to place cell or place field. 41 of 204 neutral citations were classified into experimental foundation, including ca1 neuron fire experiment, rodent studies and so on.

Conclusion and discussion

Citation context was used to measure the influence of paper in this research. The influence was identified from two aspects, the theme of the citation context and the classification of the citation context. The results showed that the traditional bibliometrics methods could be utilized in identify the themes of citation context. The citation contexts were divided into five themes in our experiment. The classification results showed that there were no negative citations of O'Keefe's most influential paper. More than 10% citation contexts were positive citations.

Through the citation context analysis of the influence paper, the detailed influence of the high influence paper could be revealed. The influence themes are more wide than the abstract of the target paper and the proportion of the positive citations takes more account than it appears in some journals (Liu et al., 2014).

There is only one case study in this paper. Although we could get some insightful results from this case study, comparative experiments are still needed in our future work.

Acknowledgments

This research is supported by National Natural Science Foundation of China (grant number 61272370), the specialized research fund for doctoral tutor (20110041110034), and the ISTIC-THOMSON REUTERS Joint Laboratory Open Foundation (IT201002). Part of the research was conducted during Shengbo Liu's visiting doctoral studentship at the iSchool at Drexel University. Thanks to the reviewers for the kindly suggestions.

References

Anderson, M. H. & Sun, P.Y.T. (2010). What have scholars retrieved from Walsh and Ungson (1991)? A citation context study. *Management Learning, 41*(2), 131-145.

Chang, Y.-W. (2013). A comparison of citation contexts between natural sciences and social sciences and humanities. *Scientometrics, 96*(2), 535-553.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D. & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American society for information science and technology, 59*(1), 51-62.

Hofer, K. M., Smejkal, A. E., Bilgin, F. Z., & Wuehrer, G. A. (2010). Conference proceedings as a matter of bibliometric studies: the Academy of International Business 2006–2008. *Scientometrics, 84*(3), 845-862.

Ioannidis, J. P. A., Boyack, K. W., Small, H., Sorensen, A. A., & Klavans, R. (2014). Bibliometrics: Is your most cited work your best? *Nature, 514*(7524), 561-562.

Lee, W. H. (2008). How to identify emerging research fields using scientometrics: An example in the field of Information Security. *Scientometrics 76*(3), 503.

Liu, S., & Chen, C. (2013). The differences between latent topics in abstracts and citation contexts of citing papers. *Journal of the American Society for Information Science and Technology, 64*(3), 627-639.

Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., & Lin, Y. (2014). Literature retrieval based on citation context. *Scientometrics, 101*(2), 1293-1307.

McCain, K. W., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics, 17*(1), 127-163.

Mei, Q. & Zhai, C. (2008). Generating impact-based summaries for scientific literature. *Proceedings of ACL '08*, Columbus.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D. & Zajic, D. (2009). Using citations to generate surveys of scientific paradigms. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Boulder.

Nakov, P. I., Schwartz, A.S. & Hearst, M.A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. *SIGIR 2004 Workshop on Search and Discovery in Bioinformatics,* Sheffield.

Nanba, H. & Okumura, M. (1999). Towards multi-paper summarization using reference information. *16th International Joint Conference on Artificial Intelligence*, Stockholm.

Nanba, H., & Okumura, M. (2005). Automatic detection of survey articles. *The Research and Advanced Technology for Digital Libraries*, Berlin.

Siddharthan, A. & Teufel, S. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. *Proceedings of NAACL/HLT-07,* Rochester.

Small, H. (1979). Co-citation context analysis: The relationship between bibliometric structure and knowledge. *Proceedings of the ASIS Annual Meeting,* Medford.

Small, H. (1982). Citation context analysis. *Progress in communication sciences, 3*, 287-310.

Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investgation. *Scientometrics, 87*(2), 373-388.

Spiegel-Rösing, I. (1977). Science studies: Bibliometric and content analysis. *Social Studies of Science*, 97-113.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press (Cambridge England and New York).

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*.

Toutanova, K., & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature, 514*(7524), 550-553.

Wang, B., Liu, S., Ding, K., Liu, Z., & Xu, J. (2014). Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. *Scientometrics, 101*(1), 685-704.

Zhang, J., Xie, J., Hou, W., Tu, X., Xu, J., Song, F., Wang, Z. & Lu, Z. (2012). Mapping the knowledge structure of research on patient adherence: Knowledge domain visualization based co-word analysis and social network analysis. *PloS One, 7*(4), e34497.

Time to First Citation Estimation in the Presence of Additional Information

Tina Nane

g.f.nane@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX, Leiden (The Netherlands)

Abstract
We are interested in modelling the time to first citation, that is how long does it take for a publication to be cited for the first time after it has been published in a journal. We argue that both cited and uncited publications should contribute to the distribution of the time to first citation. Moreover, our objective is to model the time to first citation nonparametrically, hence under no parametric assumption. Due to the similarities with the observed data in survival analysis, we employ the techniques based on censored data and describe the distribution of the time to first citation in terms of the hazard rate, that is the instantaneous rate of being firstly cited. We find that publications receive their first citation at increasing rates in the first 24 months after their publication date and at decreasing rates afterwards. Moreover, we observe that the hazard rate and hence the time to first citation is influenced by the document type, number of authors and collaboration type and field. We also investigate the difference in the time to first citations for publications grouped by their collaborative status or the assigned field.

Conference Topic
Citation and co-citation analysis

Introduction
The first citation a publication receives is an important event in the bibliometric data, as it is not only a simple citation count, but also marks a change in the status of the publication, i.e. from being uncited the publication becomes cited. Certainly, observing the first citation of a publication depends on the considered time frame. Regardless the period of analysis, certain publications will never receive their first citation, in other words we will not observe the first citation received by some publications for any finite time period we consider.

Another important aspect concerns the time it takes for a publication to receive its first citation. For some publications it takes a small amount of time, such as 1-2 months, while for others it can even take more than 10 years. Due to overlong reviewing and publication procedures, some publications might even have negative times to first citation, meaning that the publication has been cited before it has been published.

The event that a publication received its first citation, as well as the time to the first citation received considerable attention over the years, starting with Schubert and Glänzel (1986), Glänzel (1992), Rousseau (1994), Glänzel and Schoepflin (1995). Since 2000, Egghe (2000), Egghe and Rao (2001), Burrell (2001), and Glänzel et al. (2012) continued to model the first citation data. Additionally, we acknowledge the work of van Dalen and Hekens (2005) and Bornmann and Daniel (2010), that is specifically close to the present research and will be referred to later on. Most of the previous work relies on the parametric modelling of the time to first citation distribution, such as the double exponential model (Rousseau, 1994), mixtures of non-homogeneous Poisson process

(Burrell, 2001), etc. The modelling in the existing literature focuses only on publications in certain journals or fields and the uncited publications do not always contribute to the time to first citation distribution, yet they emerge as a consequence of the model (Burell, 2001). Additionally, in Egghe (2000), the proportion of the uncited documents emerges from the model.

It should be stressed however that the time to first citation distribution derived from a set of publications that contains both uncited and cited documents does not coincide with the time to first citation distribution of the publications that receive a citation. From a probabilistic perspective, the first distribution is the sub-distribution of the latter. Furthermore, not accounting for the uncited publications can lead to biases in the estimation of the distribution of the time to first citation.

Our present study aims to continue and extend the research on the time to first citation analysis. We consider all the publications, regardless the document type and field, that appeared in Web of Science (WoS) in 2000 and their first citations received until the end of 2013. The time to first citation is registered in months. Additional data is recorded for each publication, such as document type, the number of authors, institutions and countries, and information on collaboration.

We propose an approach that aims to model the time to first citation distribution by accounting for all observations (both uncited and cited publications). Our approach assumes that the event of interest is the first citation, which is time dependent and we are interesting in modelling the time to this event of interest, namely the time to first citation. The time to event analysis has been employed in many fields. In sociology, it is known as event history analysis, in economy as duration analysis and in engineering is called reliability theory. Nevertheless, it is best known in biostatistics, where most research has been performed and where it is called survival analysis.

Consequently, the terminology employed in survival analysis is ubiquitous. In biostatistics, a frequent event of interest is death and the time to the event is then expectedly called survival time. Different functionals of the distribution of the time to the event of interest are successively termed survival function, hazard or cumulative hazard function. We will employ this unfortunate terminology in the analysis of the time to first citation.

A typical feature of the data in survival analysis is that not all events of interest are observed within the period of analysis. These observations are referred to as censored observations. The uncited publications are therefore regarded as censored observations. The uncited publications are in fact right censored observations, since their first citation is conditioned to take place after the period of analysis ended, i.e. at the right of the period of analysis. This approach circumvents the issue of not having a time to first citation for the uncited publications.

In survival analysis, the distribution of the time to event data is usually characterized by its survival function, as well as its hazard rate. The hazard rate provides information on the evolution in time of the event rate, in our case first citation rate. An attractive feature of the hazard rate compared to the density function, for example, is that the hazard rate accounts for the aging effect, while the density does not. Based on our data, we provide the time to first citation distribution and investigate its behaviour via the hazard rate.

Another important aspect in survival analysis is how additional information on observations, referred to as covariates or explanatory variables influence the time to the

event of interest. The Cox model (Cox, 1972) is probably the most popular method to model the influence of covariates on the time to the event of interest. In this study, we aim to infer on the effect of different characteristics of publications on the time to first citation. In other words, is the document type, number of authors, collaboration type or the field of a publication influencing the time it takes for that publication to receive the first citation? To our best knowledge, the influence of the explanatory variables document type, collaboration or field have not been accounted so far in the time to first citation analysis.

These methods in survival analysis have been previously used to model the time to first citation distribution by van Dalen and Henkens (2005) and Bornmann and Daniel (2010). Both studies restrict themselves to publications in a specific area of research, i.e. demography and chemistry. van Dalen and Henkens (2005) propose to model the hazard rate of the time to first citation distribution under the parametric assumption of a Gompertz distribution, which, in turn, lead to hazard rate which are decreasing over time. This restriction is unintuitive and in particular, it does not fit the data of the present study. Bornmann and Daniel (2010) are very brief in explaining the methods and, more importantly, the results of the analysis are not consistent in presenting their results, as they first refer to the differences in the survival curves and later on to the differences in the hazard rate. It is not very clear, for example, if the publication characteristics have an effect on the hazard rate.

Time to first citation distribution

We consider all the publications in Web of Science (WoS) that appeared in 2000 and their first citations up until 2013. That accounts for 1,202,371 publications, from which 62.62% received their first citation until the end of 2013. The first citation of publication A is defined as the publication date (month) of a publication B that cites firstly publication A, that is the minimum publication date of all publications that cite publication A. Needless to say that since the study is restricted to WoS, we refer to the first citation covered by WoS. Moreover, we exclude self-citations, hence we condition on publication B having no common authors with publication A.

The time to first citation of publication A is the time period (in months) between the publication date of publication A and the publication date of a publication B that cites firstly publication A. The time to first citation can sometimes be negative, but this is mostly an artefact due to the slow reviewing or publication process in different journals, etc. We exclude such observation from our study.

We chose the publication date to be registered in months given the availability of the data, but also for a better insight in the first citation process. Moreover, this avoids the issue of highly discrete data. Nonetheless, it is noteworthy that the publication date in months is not available for all data. For these cases, the first month of the year (January) or the middle one (July) is usually reported.

The histogram of the time to first citation for the publications in WoS that appeared in 2000 and received their first citation within the period 2000-2013 is presented below.

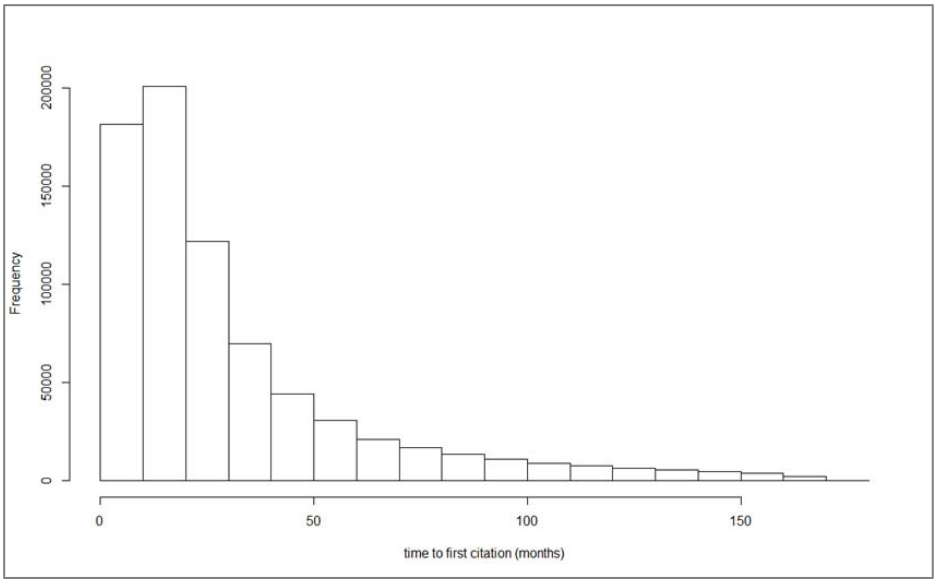


Figure 1. Histogram of the time to first citation for publications in 2010.

Most of the publications received their first citation shortly after publication. As expected, the proportion of publications that receive citations decreases over time. There are however publications that receive their first citation 13 years after their publication.

The histogram provides information on the time to first citation distribution of publications that received at least a citation until 2013. As mentioned beforehand, there is however no information on the publications that have not received any citation, apart from the percentage of the uncited publications.

Censored observations

It would be desirable though that the uncited publications also contribute to the distribution of the time to first citation, as they influence the probability of being firstly cited. Within this framework, the uncited publications did not experience the event of interest (first citation) by the duration of the study. What it is known is that their first citation occurs after the analysis ended.

In survival analysis, these observations are referred to as right censored observation. The publications that received their first citation within the period of analysis are called uncensored observations. Modelling time to event data requires that observations, both censored and uncensored have an observed time of interest, denoted as the follow-up time. For the uncensored observations, the follow-up time is the time to their first citations. For the censored observations, the follow-up time is the time period (in months) between their publication date and the end of analysis, that is December 2013, and it is referred to as the censored time.

For example, the censored time of a publication that appeared in January 2000 is 168 moths, whereas the censored time of a publication from June 2000 is 163 months. It needs to be distinguished between a publication with its time to first citation 163 months, for example it appeared in January 2000 and was firstly cited in December 2013, and a publication with its censored time 163 months. For this, we use an indicator Δ that is 1 if the publication has been cited and 0 if the publication remains uncited for the period of analysis.

The hazard rate

We are now interested in modelling the first citation rate on small units of time and its evolution in time. For this we will make use of the hazard rate, a functional of the time to first citation distribution. The hazard rate is referred to as the force of mortality in sociology, or the failure rate, in reliability. All these terms adhere to the pessimistic tone consistently used in survival analysis.

The hazard rate quantifies the rate at which first citations occur per unit of time relative to the proportion of publications that have not been yet cited. For a continuous random variable X, the hazard function is defined as

$$\lambda(t) = \lim_{\Delta t \searrow 0} \frac{P(t \leq X < t+\Delta t | X \geq t)}{\Delta t}.$$

In our case X denotes the time to first citation. We assume that the underlying time to first citation is continuous, while the observed data is discretized by measurement.

In order to compute the hazard rate at a given time point t, one needs to calculate the conditional probability in the numerator. In the present study, this is the probability of being firstly cited in the time interval [t,t+Δt), given that the publication has not been cited before time t. The conditioning ensures that at each time point t, only the publications that have not been cited up until time t are considered, therefore also the publications that are not cited throughout the entire period of analysis, i.e. the censored observations. Dividing this conditional probability by Δt, that is the width of the interval [t,t+Δt), we obtain the rate of the first citation occurrence per unit of time. By taking the limit Δt↘0 gives the instantaneous rate of occurrence of first citation. Note that, by definition, the hazard rate is not a (conditional) probability, or a density.

The hazard rate is a functional of the time to first citation distribution and can be derived for any parametric distribution and also estimated for a nonparametric distribution. The most straightforward example is the exponential distribution, for which the hazard rate is a constant function.

The hazard rate for the publications in the study is depicted in Figure 2 below.

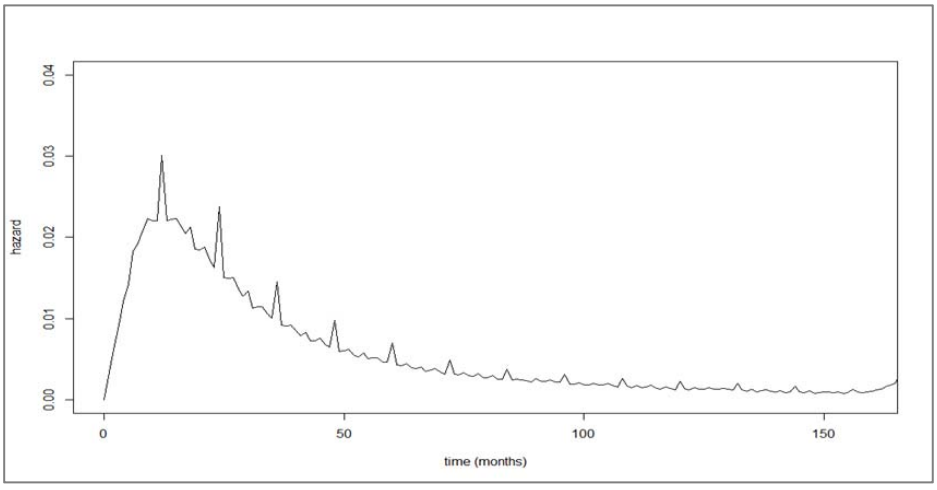


Figure 2. Hazard rate of publications in 2010.

First of all, we notice some spikes in the hazard function, which occur at the beginning and in the middle of each year in the citation window. This is due to the fact that certain journals publish once or twice a year. Moreover, when the publication date of certain

journal issues is unknown, the publication date is typically assigned to the beginning or middle of the year.

It seems that, per unit of time, publications receive their first citation at an increasing instantaneous rate up until a given time, that we refer to as the first citation peak, and despite the spikes, at decreasing instantaneous rates after the first citation peak. This shape suggests an unimodal hazard rate.

The first citation peak is for this dataset 24 months. In terms of conditional probabilities, the results can be interpreted as follows. Given that publications have not been cited before, on small unit intervals, they get cited for the first time with higher probability in the first 2 years after publications and with lower probability afterwards. The conditional probability decreases with time, but flattens after a while. That is, the decrease of the hazard is rather steep until 50 months and flattens afterwards. It can be inferred that first citation instantaneous rate is low and does not change significantly for documents that have not been cited for 4-5 years after publication.

Additional information – covariates

We are now interested in what can possibly influence the time to first citation and its hazard rate. This additional information is recorded as explanatory variables that are typically referred to as covariates in survival analysis, or as control variables in econometrics.

We consider the following covariates: document type, number of authors, collaboration type and field. By field we refer to the 250 subject categories to which journals are assigned in WoS. Surely, other covariates might be included, such as number of institutions or countries, number of pages, journal impact, etc.

Assume that covariates do not change over time, that they have a fixed value at the publication date. There can be however, covariates that change over time (time dependent covariates), such as journal impact, authors' visibility or performance.

The Cox model

The most famous model that incorporates the information on certain covariates in survival analysis is the Cox model (Cox, 1972). Regardless the fact that the model is more than 40 years old, it has been widely used and numerous versions, for particular issues with the data, have been proposed and investigated ever since.

The Cox model specifies the hazard rate at time t of a publication with a given covariate vector z as

$$\lambda(t|z) = \lambda_0(t)\exp(\beta' z),$$

where λ_0 is the underlying baseline hazard and β' is the transpose of the vector of underlying regression coefficients. Notice that if we take all covariates to be zero, we obtain the baseline hazard.

Within the Cox model, the hazard has two components. The first one, the baseline hazard, is the nonparametric part and it indicates how the hazard varies in time. The second term specifies parametrically, via an exponential function, the dependence on the covariates. It is then obvious why the Cox model is considered a semi-parametric model. Moreover, it is worth mentioning that the baseline hazard can be left unspecified when one want to estimate the regression coefficients and this flexibility has been particularly attractive for researchers.

Ever since the model was proposed, there was a great interest in estimating the regression coefficients β, that reflect how changes in the covariates produce a change in the hazard rate. The estimates were obtained via a partial likelihood method that avoided the bothersome issue of estimating the baseline hazard λ_0.
We have fitted the Cox model with the following covariates

- Document type
- Collaboration type
- Number of authors.

We will focus on estimating the (baseline) hazard and not on the regression coefficient estimation. We need to stress that conditioning on the covariates to be at a baseline value, i.e. z=0, is not the same thing as not accounting for covariates. This can be determined from the equation specifying the Cox model, but also from the figure below.

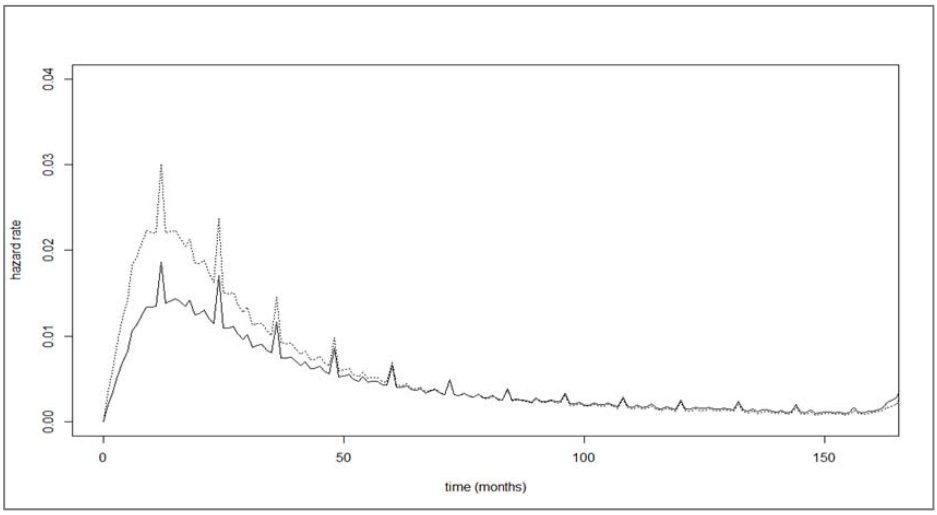


Figure 3. Hazard rate in the presence of no covariates (dotted) and baseline hazard (solid line).

Apparently, accounting for covariates shifts the hazard down in the first 60 months after the publication date and has no effect afterwards. The baseline hazard follows the same trend as the hazard rate in the presence of no covariates that is increasing until 24 months after the publication date and decreasing afterwards. Therefore, we can conclude that the covariates have a scale effect rather than a shape effect on the hazard. Furthermore, it seems that there is a proportional effect of the covariates on the baseline hazard, at least in the first 50 months. This represent a visualization of the goodness of fit of the Cox model and additionally, several tests suggest that the model fits the data well.
We want to investigate now whether certain characteristics of the publication, such as the collaborative status or the field have an impact on the instantaneous first citation rates.

Collaboration

It is commonly thought that publications that have resulted from an international collaboration are more visible to the academic community and hence receive more citations than national collaborative publications or publications that do not result from any inter institutional collaboration. It would be interesting to see if the collaboration type also influences how fast a publication receives its first citation.

As mentioned beforehand, we have fitted a Cox model with document type, collaboration type and number of authors as covariates. All the covariates have a (statistical) significant influence on the time to first citation.

To show the difference in the hazard rates among the different types of collaboration, we compute the hazard rate for publications with international, national and no collaborations. All the other covariates are set to their baseline level. Figure 4 depicts these differences.

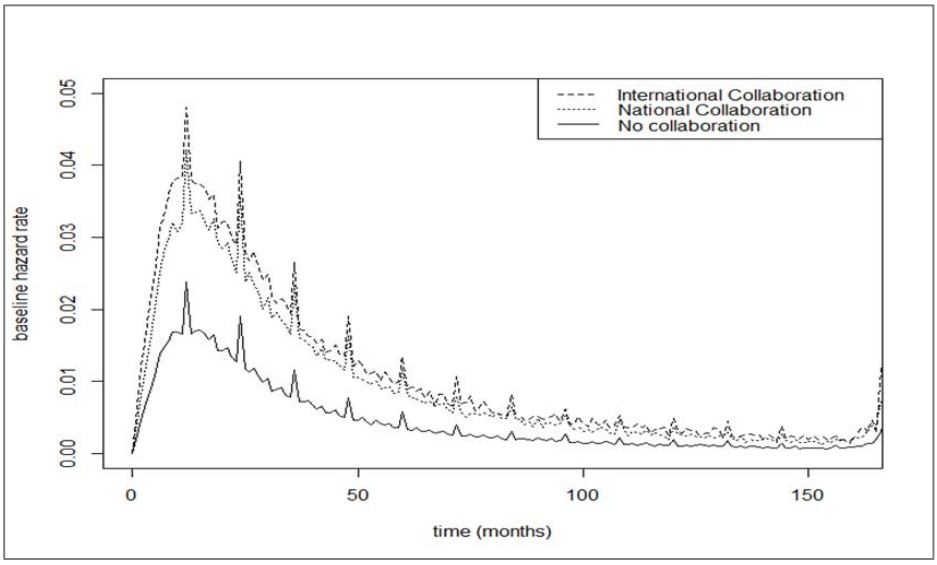


Figure 4. Baseline hazard rates in terms of collaboration type: international collaboration (dashed), national collaboration (dotted) and no collaboration (solid line).

It seems that there is a significant scale difference in the instantaneous first citation rate among publications that represent international and international collaborations and those that do not result from such collaborations. There are however small differences between baseline hazard of the international and national collaborative publications. Nonetheless, the publications that resulted from an international collaboration register higher instantaneous first citation rates than publications that represent national collaborations and these publications have, in turn, higher instantaneous first citation rates than publications whose authors are affiliated to a single institution. Similar to the overall (baseline) hazard rates, there are less and less differences in the hazard rates of different collaboration types 100 months after publication.

Contrary to the popular belief however, it seems that, apart from a scaling factor, publications receive their first citation at similar rates irrespective their collaboration type. The maximum hazard function is attained by publications of all collaboration types at the same time point, which is 24 months after the publication date. This is not different from the overall baseline hazard.

To condition further on specific values of the other covariates, we have considered the document type ‘Article’ and assume the publications has 3 authors, which is close to the overall average of the entire dataset, that is 3.31.

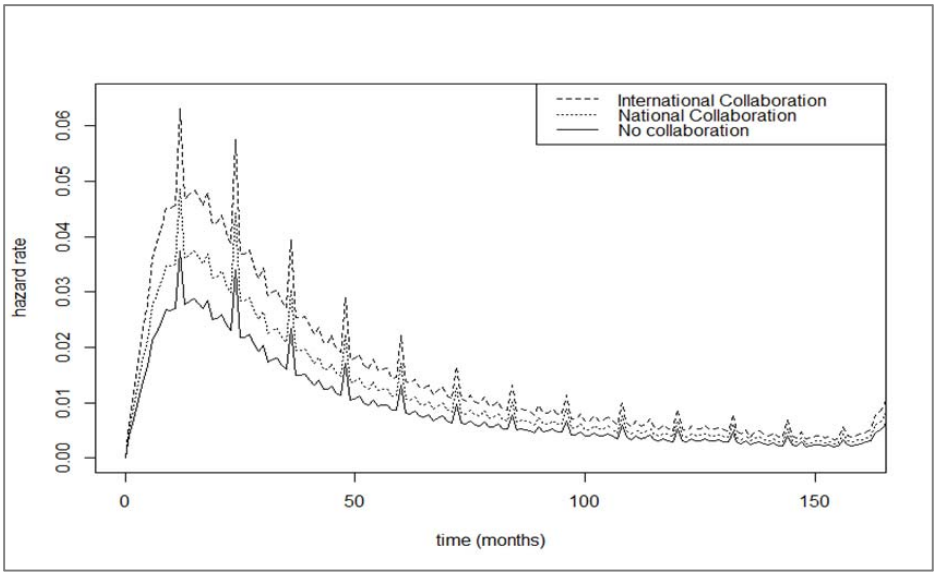


Figure 5. Hazard rates for articles with mean number of authors. International collaboration (dashed), national collaboration (dotted) and no collaboration (solid line).

Figure 5 depicts the hazard rates of articles that result from different collaborations and are written by three researches. We notice that the differences in the hazard rates have decreased. Despite similar behaviour over time, international collaborations still achieve the highest hazard rates over time, followed by national collaborations and articles produced by the same institution (no collaboration).

Field

We are also interested to see whether the field assigned to a certain publication affects the rate of being firstly cited. Nonetheless, more than half of the journals in WoS are assigned to at least two fields and some journals are assigned to six fields. This means that the field covariate cannot be uniquely defined for each publication. This difficulty cannot be overcome by using the WoS subject category assignment and hence the field cannot be included as a covariate in the Cox model. A solution is to adopt the publication-level classification system proposed by Waltman and van Eck (2012). Within this approach each publication is assigned to an unique cluster. Employing the publication-level classification system is deferred to future research.

In order to still assess the influence of the field on the time to first citation distribution, we have limited the data of all publication from 2000 to three fields: Biochemistry & Molecular Biology, Economics and Mathematics. We have now a number of 80,745 publications that have been published in 2000 and are assigned to the three fields.

We have fitted the Cox model with the following covariates

- Document type
- Collaboration type
- Number of authors
- Field

All four covariates have a (statistical) significant effect on the hazard rate. We are interested in the baseline hazard rates for the data grouped by the field. The differences

between the three baseline hazards can be observed in Figure 6. Once again, the other covariates have been set to zero.

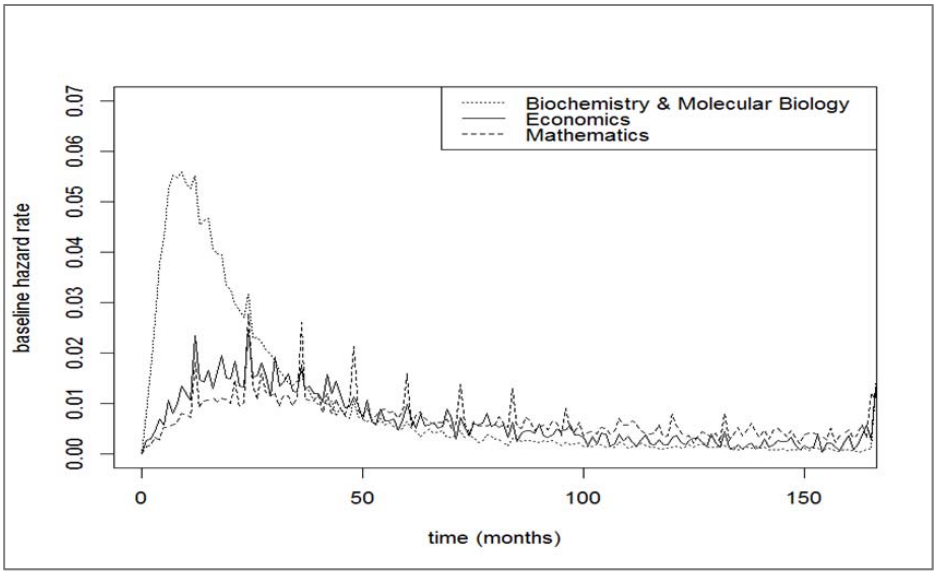


Figure 6. Baseline hazard rates in terms of field: Biochemistry and Molecular Biology (dotted), Mathematics (dashed) and Economics (solid line).

The three baseline hazard rates differ in both shape and scale. Firstly, it seems that the publications that appeared in 2000 in Biochemistry and Molecular Biology achieve their maximum first citation rate earlier than publications in Economics or Mathematics. The peak in Biochemistry and Molecular Biology is registered at 12 months, whereas the publications in Economics and Mathematics have a baseline hazard rate peak around 24 months.

We observe that there are large changes over time in the baseline hazard rate of publication in Biochemistry and Molecular Biology. Moreover, during the first part of the citation window, publications in Biochemistry and Molecular Biology have an instantaneous first citation rate three times as higher than the instantaneous first citation rates in Economics and Mathematics. The publications in Economics and Mathematics exhibit similar hazard rate behaviour.

It is noteworthy and interesting that after 60 months, the order of the three baseline hazard rates completely reverse, that is publications in Mathematics have higher baseline hazard rates than publications in Economics, that have higher baseline hazard rates than the publications in Biochemistry & Molecular Biology.

Discussion and conclusions

The first citation is probably the most important citation a publication receives. It can determine entirely the number or speed of further citations. Besides a simple citation count, it also changes the status of a publication, from being uncited to being cited. In some fields, being cited is even sufficient to become frequently cited.

The time to first citation also contributes to the number or speed of further citations. Apart from the famous sleeping beauties (van Raan, 2004), it is obvious that the more it takes for a publication to receive its first citation, the lower the probability of receiving further citations.

Time to first citation is the first step in modelling how publications accumulate citations in general over time. It is still unknown whether the time to first citation differs significantly from the time to second citation, etc.

We aimed to model the time to first citation and used a set of publications that appeared in 2000 and are included in the WoS database. Probably the most important aspect of our approach is that we employed nonparametric or semi-parametric methods of estimation. In other words, we let the data speak for itself. This ensures a greater flexibility and avoids the bothersome issue that a given model fits a particular data well, say publications that appear in a certain year and within a specific field, but fails to fit another particular data appropriately. While this is not a problem specific only to the first citation analysis, for an example on this matter in the first citation analysis, see Rousseau (1994). Another important drawback of the parametric approach is that certain employed parametric models cannot incorporate specific shapes of the first citation data. Van Dalen and Hekens (2005) for example make use of a Gompertz hazard model that cannot incorporate an unimodal hazard, as we obtained in the present study.

Apart from the nonparametric choice of estimation, we have also incorporated the uncited publications in the distribution of the time to first citation by using methods developed in survival analysis. We stress the fact that the information on uncited publications should be accounted for in modelling the time to first citation distribution, otherwise the results of the estimation can be seriously biased, especially given the high percentage of uncited publications.

We have investigated the time to first citation distribution through its hazard rate, the instantaneous rate of being firstly cited. We observe that the hazard rate increase over the first 24 months and decreases afterwards. This is somehow expected, that publications receive their first citations at higher rates until a maximum and afterwards at lower and lower rates. What is surprising is the relative short period of time over which the hazard rate is increasing. It means that the probability of a publications being cited for the first time is increasing over the first 24 months, and decrease afterwards.

Furthermore, it is of high interest to investigate whether certain characteristics of publications influence their time to first citation. We included the document type, number of authors, collaboration type and the field. We have found that all these explanatory variables (covariates) influence the time to first citation and investigated the differences between the hazard rates of publications grouped by collaboration type. The hazard rates of the three collaboration types differ in scale and not in shape and attain the maximum at the same time point. Hence, it seems that publications receive their first citations at an increasing rate up to the same time point, namely 24 months regardless their collaboration type.

A different dataset has been chosen to investigate the influence of the field on the time to first citation. It seems that, for the three selected fields, the hazard rate of the publications differ not only in scale but also in shape. The publications in Biochemistry and Molecular Biology register higher rates than publications from Economics and Mathematics, but also they have increasing first citation rates over a shorter period of time than the publications from the other two fields. The order of the three hazard rates reverse after 60 months.

As mentioned in the previous section, the problem of the overlapping fields in WoS needs to be addressed in future research and this can be overcome by considering the

publication-level classification system proposed by Waltman and Van Eck (2012). Numerous investigations are further required and desired. For example it would be very interesting to investigate whether the time to first citation distribution, and in particular the hazard rate including self citations differs from the time to first citation excluding self citations. Other covariates can be included in the analysis, such as the impact of the journal, the performance or visibility of authors, etc. Of course, it is very interesting to see whether the shape of the hazard rate changes over the time of publication, not only through the citation window. The author expects that the hazard would have the same unimodal shape, but the maximum point would be attained at different time points that is the first citation peak would be time dependent.

In terms of estimation, it is highly desirable to account for the monotonicity of the (baseline) hazard that is to provide estimates of the baseline hazard rate under the assumption of monotonicity. This is in line with the research of Lopuhaä and Nane (2013), but needs some refinement to incorporate the estimation of a unimodal baseline hazard.

References

Bornmann, L. & Daniel, H.-D. (2010). Citation speed as a measure to predict the attention an article receives: An investigation of the validity of editorial decisions at *Angewandte Chemie International Edition*. *Journal of Informetrics, 4*, 83-88.

Burrel, Q.L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics,* 52, 3-12.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*. *Series B (Methodological)*, *45*, 187-220.

van Dalen, H.P. & Henkens, K. (2005). Signals in science – On the importance of signaling in gaining attention in science. *Scientometrics*. *64*, 209-233.

Egghe, L. (2000). A heuristic study of the first-citation distribution. *Scientometrics,* 48, 343-359.

Egghe, L & Rao, I.K.R. (2001). Theory of first-citation distributions and applications. *Mathematical and Computer Modelling*, *34*, 81-90.

Glänzel, W. (1992). On some stopping times of citation processes. From theory to indicators. *Information Processing & Management, 28*, 53-60.

Glänzel, W., Rousseau, R. & Zhang, L. (2012). A visual representation of relative first-citation times. *Journal of the American Society for Information Science and Technology*, *63*, 1420-1425.

Glänzel, W. & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature, *Journal of Information Science*, *21*, 37-53.

Lopuhaä, H.P. & Nane, G.F. (2013). Shape constrained nonparametric estimators of the baseline distribution in Cox proportional hazards model. *Scandinavian Journal of Statistics, 40*, 619-646.

van Raan, A.F.J. (2004). Sleeping beauties in science (short communication). *Scientometrics*

Rousseau, R. (1994). Double exponential models for first-citation processes. *Scientometrics, 30*, 213-227.

Schubert, A. & Glänzel, W. (1986), Mean response time – a new indicator of journal citation speed with application to physics journals. *Czechoslovak Journal of Physics (B), 36*, 121-125.

Waltman, L. & van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system in science. *Journal of the American Society for Information Science and Technology*, *63*, 2378-2392.

Author Relationship Mining based on Tripartite Citation Analysis

Feifei Wang[1], Junwan Liu[2], Siluo Yang[3]

[1] *feifeiwang@bjut.edu.cn,* [2] *liujunwan@bjut.edu.cn*
School of Economics and Management, Beijing University of Technology, 100024 Beijing (China)

[3] *58605025@qq.com*
School of Information Management, Wuhan University, 430072 Wuhan (China)

Abstract
This study scrutinizes potential author relationships according to the findings based on the tripartite citation analysis. It focuses on Author co-citation analysis (ACA), author bibliographic-coupling analysis (ABCA) and author direct citation analysis (ADCA). By algorithm design and empirical analysis, the deduction from results of ACA, ABCA and ADCA to potential author relationships mining could be probable, and the empirical process would be practicable.

Conference Topic
Citation and co-citation analysis

Introduction
Citation analysis is a mature quantitative research method in Bibliometrics and Scientometrics. It is widely used in scientific evaluation, scholarly communications, academic behavior analysis, and information retrieval. Author citation analysis mainly includes three types: author co-citation, author coupling, and author direct citation.

Author co-citation analysis (ACA) is the most widely used method for the empirical analysis of disciplinary paradigm, and is frequently studied and improved upon. Many ACA studies have been conducted since Small (1973) introduced document co-citation analysis and White and Griffith (1981) introduced ACA. Bibliographic coupling was proposed as early as 1963 (M. M. Kessler, 1963). However, author bibliographic-coupling analysis (ABCA), i.e. author-coupling relationships, did not get much attention until it is formally put forward and empirically studied by Zhao (Zhao & Strotmann, 2008).

Direct citation is sometimes also called inter-citation or cross citation (Zhang et al., 2009). Compared with co-citation and bibliographic coupling, direct citation is a direct citing relationship without a third party paper. Although researchers are aware of direct citation analysis and employed from time to time (Shibata et al., 2008), it was ignored because of the unavailability of data, difficulty of implementation, and long time windows to obtain a sufficient linking signal for clustering. However, scholars are gradually paying attention to this topic (Boyack & Klavans, 2010). A number of studies have focused on journal direct citation or comparative analysis of methods. The author direct citation analysis was more clearly explored by Wang et al. (2012). Wang used this method to reveal the knowledge communication and disciplinary structure in Scientometrics. This process is named "author direct citation analysis" (ADCA) (Yang & Wang, 2015).

All of these three kinds of citation analysis methods can reveal separately the author relationship in a field. Then, how about the similarities or diversity among the tripartite citation relationships at author level? And, how could the tripartite relationships be synthetically presented to the readers or the result users? We have tried to answer these two questions in previous studies (Wang, 2014), even though the effort is still limited. Persson (2010) and Gómez-Núñez et al. (2011, 2014, 2015) tried to combine these citation measures

in a normalized way to weight existing direct citation relationships between articles or journals.

The following question is worthy of investigation as well: Could we discover potential author relationships according to the findings based on the tripartite citation analysis? To give an example: in a field, author A's paper and author B's paper both are cited by the same paper C, then A and B have co-citation relationship, which can be marked as (A, co-citation, B). Author C and author D, when citing the same paper in their respective articles, have bibliographic-coupling relationship, marked as (C, bibliographic-coupling, D). In addition, if C and A cite each other, then C and A have direct-citation or cross-citation relationship, marked as (C, directly citing, A) or (A, directly citing, C) or (A, cross citation, C). According to these primary relationships, could we deduce an integrated relationship between A and D, or B and C, even B and D? And, what will be the association strength in these potential relationships? These are the key problems that we answer in this study.

Data and methodology

Basic Data

Since the journal *Scientometrics* is one of the most representative communication channels in the field of Scientometrics, it reflects the characteristic trends and patterns of the past decades in scientometric research (Schubert A 2002). This study is based on bibliographic data based on all types of documents published in *Scientometrics* from 1978 to 2011, retrieved from the Web of Science. Author names including the cited authors were normalized because some authors may report their names differently in different papers. We identified each author by his or her surname and first initial only; the same applies to cited authors.

Methodologies

In this study, bibliometrics method is applied to identify the core authors (94 first authors who have published 5 or more papers and simultaneously have a cited frequency over 10) in Scientometrics filed. Author co-citation analysis (ACA), author bibliographic-coupling analysis (ABCA) and author direct citation analysis (ADCA) are respectively used to discover author relationships with co-citation, bibliographic-coupling and direct-citation. Co-occurrence analysis and deductive reasoning methods are used to mine potential author relationships on the findings of the tripartite citation analysis. VBA program processes all kinds of citation analysis data. The final results of author relationship are visualized with Pajek.

Results and discussion

According to the tripartite citation analysis, we obtain three original relation matrixes and their corresponding normalized matrixes (Fig. 1). The normalization method is based on Salton's Cosine similarity measures, which returns similarity values ranging between 0 and 1. In order to describe the directivity of citing behaviour and achieve vectorial deducing, the direct citation matrix is unsymmetrical.

Core author co-citaion matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.7535	0.8359	0.5426
Glanzel W	0.7535	1	0.8916	0.7579
Braun T	0.8359	0.8916	1	0.5736
Egghe L	0.5426	0.7579	0.5736	1

Core author bibliographic-coupling matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.4249	0.3612	0.2881
Glanzel W	0.4249	1	0.9171	0.4069
Braun T	0.3612	0.9171	1	0.26
Egghe L	0.2881	0.4069	0.26	1

Core author direct citation matrix

	Garfield E	Glanzel W	Braun T	Egghe L
Garfield E	1	0.0022	0.0312	0.0027
Glanzel W	0.2844	1	0.414	0.1365
Braun T	0.2511	0.173	1	0.0073
Egghe L	0.0974	0.221	0.1058	1

Figure 1. Normalized matrixes of tripartite citation analysis.

The following five steps could help us realize author relationship mining based on tripartite citation analysis, such as "A→C, B→D, B→C". These steps can also be seen as an algorithm in relation mining.

First step: Obtaining the fundamental citation relationship with strength(>0) among core authors from original matrixes

Tripartite adjacency matrixes are transformed into corresponding adjacency lists. ACA list $\{L_{1i},Q_{1i}\}$ versus matrix $\{O_{1i}, P_{1j}\}$, and relational degree X_i (i stands for the ID of author pair) in list can replace X_{ij} (i/j stand for different authors in the matrix) . ABCA list $\{L_{2i},Q_{2i}\}$ versus matrix $\{O_{2i}, P_{2j}\}$, and relational degree Y_i versus Y_{ij}. ADCA list $\{L_{3i},Q_{3i}\}$ and $\{L_{3j},Q_{3j}\}$ versus matrix $\{O_{3i}, P_{3j}\}$, and relational degree Z_i and Z_j versus Z_{ij} (the order between i and j denotes the citing direction). We used the Adjacency list in calculation process.

Second step: Filtering no-explicit-relationship author pairs

The no-relationship author pairs (X_i=0, Y_i=0, Z_i=0, and no cooperation), are filtered as $\{O_{4i},P_{4j}\}$ in the Adjacency matrix, and $\{L_{4i},Q_{4i}\}$ in the Adjacency list, which forms the basic object in subsequent analysis.

Third step: Mining the relationship of A →C from$\{L_{1i},Q_{1i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the $\{L_{4i},Q_{4i}\}$ as $\{A_k,C_k\}$ (k stands for the number of author pairs), the goal is finding the Dk with the relations $\{A_k→D_k, C_k\text{-}D_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, C_k and D_k, from $\{L_{1i},Q_{1i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{A_k,C_k\}$. The pseudo code is as follows:

If one author in the pair of $\{A_k,C_k\}$= one author in a pair of $\{L_{1i},Q_{1i}\}$, and another one in the pair of $\{A_k,C_k\}$= one author in a pair of $\{L_{3i},Q_{3i}\}$, and another one in the pair of $\{L_{1i},Q_{1i}\}$= another one in the pair of $\{L_{3i},Q_{3i}\}$

Then mark the "one author in the pair of $\{A_k,C_k\}$" (so as the "one author in a pair of $\{L_{1i},Q_{1i}\}$") as Cα, the "one author in a pair of $\{L_{3i},Q_{3i}\}$" (so as the "another one in the pair of $\{A_k,C_k\}$") as Aα, the "another one in the pair of $\{L_{1i},Q_{1i}\}$" (so as the "another one in the pair of $\{L_{3i},Q_{3i}\}$") as Dα

End with the relation between A_α and C_α according to D_α, and their relation strength equaling to the product of X_α and Y_α. If the order of author pair in $\{L_{4\alpha},Q_{4\alpha}\}$(i.e., $\{A_k,C_k\}$) is in reverse of the order of author pair in $\{L_{3\alpha},Q_{3\alpha}\}$(i.e., $\{A_k,D_k\}$), then the relation strength between Aα and Cα will be the negative value.

Finally, choose the top value (Take the absolute value of the negative value) as the final relation strength of A_α and C_α.

Fourth step: Mining the relationship of B →D from$\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the $\{L_{4i},Q_{4i}\}$ as $\{B_k,D_k\}$ (k stands for the number of author pairs), the goal is to find the A_k with the relations $\{A_k→D_k, A_k\text{-}B_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, A_k and B_k, from $\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{A_k,C_k\}$.This process is similar with the process of A→C, so the pseudo code is omitted.

Fifth step: Mining the relationship of B →C from$\{L_{1i},Q_{1i}\}\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}\{L_{4i},Q_{4i}\}$

Remark the rest (no relationship like A→C and B→D) of $\{L_{4i},Q_{4i}\}$ as $\{B_k, C_k\}$ (k stands for the number of author pairs), the goal is to find the A_k and D_k with the relations $\{A_k→ D_k, A_k\text{-}B_k, C_k\text{-}D_k\}$. We looked for the synchronous relations with strengh between A_k and D_k, A_k and B_k, Ck and D_k, from $\{L_{1i},Q_{1i}\}\{L_{2i},Q_{2i}\}\{L_{3i},Q_{3i}\}$, and matched the author pairs in $\{B_k,C_k\}$. The pseudo code as follows:

If one author in the pair of $\{B_k,C_k\}$= one author in a pair of $\{L_{2i},Q_{2i}\}$, and another one in the pair of $\{B_k,C_k\}$= one author in a pair of $\{L_{1i},Q_{1i}\}$, and another one in the pair of $\{L_{2i},Q_{2i}\}$=one author in the pair of $\{L_{3i},Q_{3i}\}$, and another one in the pair of $\{L_{1i},Q_{1i}\}$= another one in the pair of $\{L_{3i},Q_{3i}\}$

Then mark the "one author in the pair of $\{B_k,C_k\}$" (so as the "one author in a pair of $\{L_{2i},Q_{2i}\}$") as B_χ, "another one in the pair of $\{B_k,C_k\}$" (so as "the one author in a pair of $\{L_{1i},Q_{1i}\}$") as C_χ, one author in the pair of $\{L_{3i},Q_{3i}\}$(so as the "another one in the pair of $\{L_{2i},Q_{2i}\}$") as A_χ, another one in the pair of $\{L_{1i},Q_{1i}\}$(so as the "another one in the pair of $\{L_{3i},Q_{3i}\}$) as D_χ

End with the relation between B_χ and C_χ according to A_χ and D_χ, and their relation strength equaling to the product of X_χ and Y_χ and Z_χ. If the order of author pair in $\{L_{4\chi},Q_{4\chi}\}$(i.e., $\{B_k,C_k\}$) is in reverse of the order of author pair in $\{L_{3\chi},Q_{3\chi}\}$(i.e., $\{A_k,D_k\}$), then the relation strength between B_χ and C_χ will be the negative value.

Finally, choose the top value (take the absolute value of the negative value) as the final relation strength of B_χ and C_χ.

So far, all relationship among author pairs in $\{L_{4i},Q_{4i}\}$ have been built.

According to the above algorithm, potential relationships among not-directly-related core author set could be discovered by VBA programme and Access databases. The final results among A→C, B→D and B→C are visulized by Pajek as Figure 2 and 3.

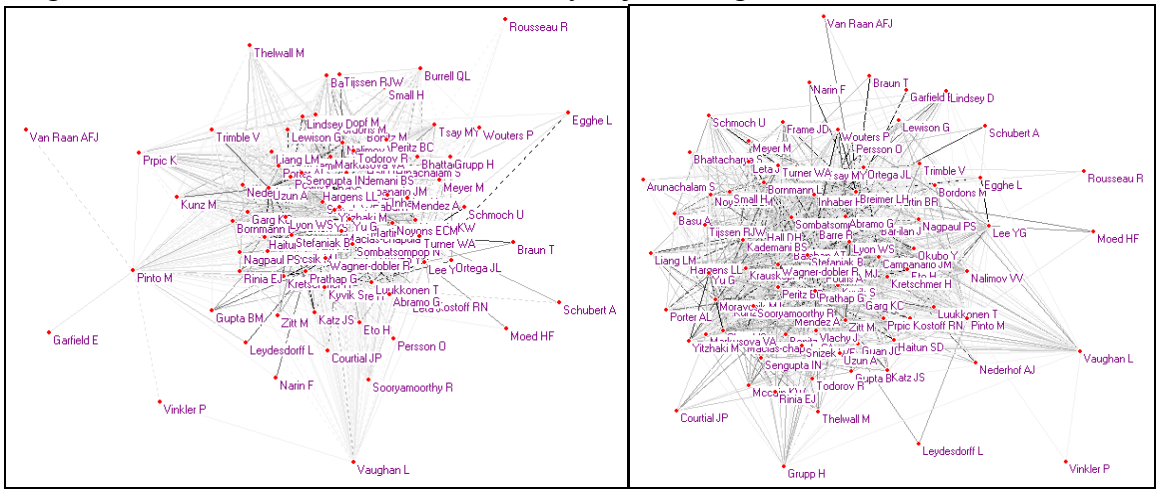


Figure 2. (a) Author relationship network of A→C. (b)Author relationship network of B→D.

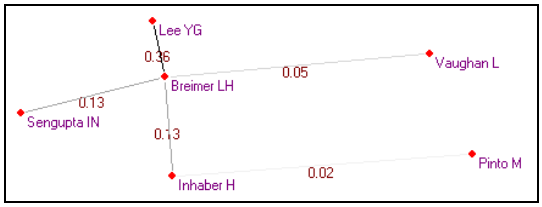


Figure 3. Author relationship network of B→C.

In Figure 3, the labels in the lines denote the value of the relationship similarity for authors in pairs. According to the results, there are different levels of potential relationship between Breimer LH and other authors, such as Inhaber H、Lee YG、Sengupta IN、Vaughan L.

Conclusions

Based on the algorithm design and empirical analysis, the deduction from results of ACA, ABCA and ADCA to potential author relationships mining could be probable, and the

empirical process would be practicable. The findings in Scientometrics field can help scholars discover more research fellows, which can promote scientific research cooperation and broader knowledge communication.

References

Boyack, K. W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389-2404.

Kessler, M.M. (1963). Bibliographic coupling between scientific papers, *Journal of the American Society for Information Science and Technology*, *14*(1), 10-25.

Gómez-Núñez, A. J., Batagelj, V., Vargas-Quesada, B., Moya-Anegón, F., & Chinchilla-Rodríguez, Z. (2014). Optimizing SCImago Journal & Country Rank classification by community detection. *Journal of Informetrics*, *8*(2), 369-383.

Gómez-Núñez, A. J., Vargas-Quesada, B., de Moya-Anegón, F. & Glänzel, W.(2011). Improving SCImago Journal & Country Rank (SJR) subject classification through reference analysis. *Scientometrics*, *89*(3), 741-758.

Gómez-Núñez, A. J., Vargas-Quesada, B. & Moya-Anegón, F. (2015). Updating the SCImago journal and country rank classification: A new approach using Ward's clustering and alternative combination of citation measures. *Journal of the Association for Information Science and Technology*, published online. http://dx.doi.org/10.1002/asi.23370.

Persson,O.(2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, 4(3), 415-422.

Schubert, A. (2002). The web of Scientometrics: A statistical overview of the first 50 volumes of the journal. Scientometrics, *53*(1), 3-20.

Shibata, N., Kajikawa, Y., Takeda, Y. & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, *28*(11), 758-775.

Small, H. (1973). Cocitation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265-269.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820-1833.

Wang, F. (2014). Influence Analysis of Core Authors in Scientometrics from an Integrated Perspective of Publication and Citation. *Science of Science and Management of S.&T.(China)*, *35*(12): 45-55.

Wang, F., Qiu, J. & Yu, H. (2012). Research on the cross-citation relationship of core authors in scientometrics. *Scientometrics*, *91*(3), 1011-1033.

White, H.D. & Griffith, B. (1981). Author cocitation: A literature mesaure of intellectual structures. *Journal of the American Society for Information Science*, *32*(3), 163-171.

Yang, S. & Wang, F. (2015). Visualizing Information Science: Author Direct Citation Analysis in China and around the World. *Journal of Informetrics*, *9*(1), 208-225.

Zhang, L., Glänzel, W. & Liang, L. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. *Scientometrics*, *81*(3), 821-838.

Charles Dotter and the Birth of Interventional Radiology: A "Sleeping-Beauty" with a Restless Sleep

Philippe Gorry[1] and Pascal Ragouet[2]

[1] *philippe.gorry@u-bordeaux.fr*
GREThA UMR CNRS 5113, University of Bordeaux, Av. Leon Duguit, 33608, Pessac (France)

[2] *pascal.ragouet@u-bordeaux.fr*
Centre E. Durkheim CNRS UMR 5116, University of Bordeaux, 3ter Pl. de la Victoire, 33076, Bordeaux (France)

Abstract
Charles Dotter is described as the father of interventional radiology, a medical specialty born at the cross-border of radiology and cardiology. Dotter's landmark paper published in 1964 was poorly cited until 1979 and can be considered from a scientometric point of view as a sleeping beauty. Sleeping-beauties are article that suffer of a delayed recognition. This paper, will explore the bibliometric characteristics of this case study and the accuracy of Van Raan's criteria to define "sleeping beauty" in science will be discussed. "The prince" is identified through citation network analysis, and the sleeping period has been documented as a restless sleep period with science and social controversy that could be documented in publications databases by differentiating bibliographic references. Therefore, a category of "sleeping beauty" –like paper should be introduced. By the end, these observations should open new avenues in identifying "sleeping beauties".

Conference Topic
Citation and co-citation analysis

Introduction

Charles Dotter, father of interventional radiology

Charles Theodore Dotter (1920–1985) was a pioneering US vascular radiologist, credited with developing interventional radiology (IR): he invented the angioplasty and the catheter-delivered stent. On January 16, 1964, he percutaneously dilated a tight, localized stenosis of the superficial femoral artery in an 82-year-old woman with painful leg ischemia and gangrene who refused leg amputation. Percutenous transluminal angioplasty (PTA) was born, and Dotter with his trainee Dr. Melvin P. Judkins, described their technique in a landmark paper published in the medical journal "Circulation" (Dotter, 1964).

Today, Charles Dotter is described as the father of interventional radiology (IR), a sub-specialty of radiology using minimally invasive image-guided procedure to diagnose, as well as to treat diseases in every organ. The Oregon Health Sciences University (OHSU), where he spent his entire medical career, boasts the Dotter Interventional Institute. Furthermore, the Society of Interventional Radiology named a Dr. C.T. Dotter lecture to honor annually extraordinary contributions to the IR field (Rösch, 2003).

However, at first, the relationship between surgeons and radiologists was adversarial because the Dotter technique was a paradigmatic revolution, inviting radiologists to transgress medical specialty boundaries. It can be summed up by Dotter's formula at that time: "The angiographic catheter can be more than a tool for passive means for diagnostic observations; used with imagination, it can become an important surgical instrument". (Payne, 2001).

Therefore, as we found out, Dotter's landmark paper was poorly cited until 1979 and can be considered from a scientometric point of view as a sleeping beauty paper.

Sleeping beauty in scientific literature

In Scientometrics, the phenomenon of delayed recognition has been well described since the pioneering observations of Garfield, and referred to as premature discoveries, resisted discoveries, delayed recognition or sleeping beauties (Burrell, 2005; Braun, 2010). Van Raan (2004) defined "sleeping beauties" as articles that go unnoticed ("sleeps") for a long period of time and then, suddenly, receives a lot of citations by a "prince" (another article). Three variables were defined for such papers: depth of sleep, length of the sleep and awaking intensity. Some publication had heaping before sleeping, and are described as "all-element-sleeping beauties" (Li, 2012).

Objectives

In the present work, we explore the bibliometric characteristics of this case study, question the sleeping-beauty definition, explore the diffusion of Dotter concept during the sleeping period, and document the awaking phase and identify "the prince" through citation network analysis.

Method

A literature search on Dotter C.T. scientific production was conducted both in PubMed and Scopus databases. Citations of Dotter work were extracted from the Web of Science database until 12/31/2013. Then, a descriptive statistics analysis was led on the corpus (219 publications; 7866 citations). Scientific collaborations of C.T. Dotter was explored with Intellixir® to draw co-publications graph. Citations network pattern during time of the landmark paper was drawn using CitNetExplorer software tool (Van Eck, 2014). Complementary queries were run using Dotter or PTA as a keyword in different search fields for different types of documents.

Result

The scientific production of Charles Dotter

Dotter published his first paper in 1948 in a top medical journal, the New England Journal of Medicine (Jan 13; 239(2):51-4). During his 33 years at OHSU, he issued 219 publications; a quarter of his scientific production was disclose in high quality journals, and split between 2 main medical disciplines: radiology and cardiology (Table 1).

Table 1. Journal distribution of C.T. Dotter scientific production.

Source title	*Publications number*	*Impact factor*
Radiology	46	5,561
Am. J. Roentgenol. Radium Ther. Nucl. Med.	27	na
Circulation	19	12,755
New England J Medicine	8	52,589
Am. J. Roetgenol.	6	2,47

Dotter had many relations in the academic community: all along his career he co-published with 140 different authors, mainly with J. Rosch, F. Keller & J. Melvin (340, 215 & 68 co-publications respectively; Fig.1 and Table 2).

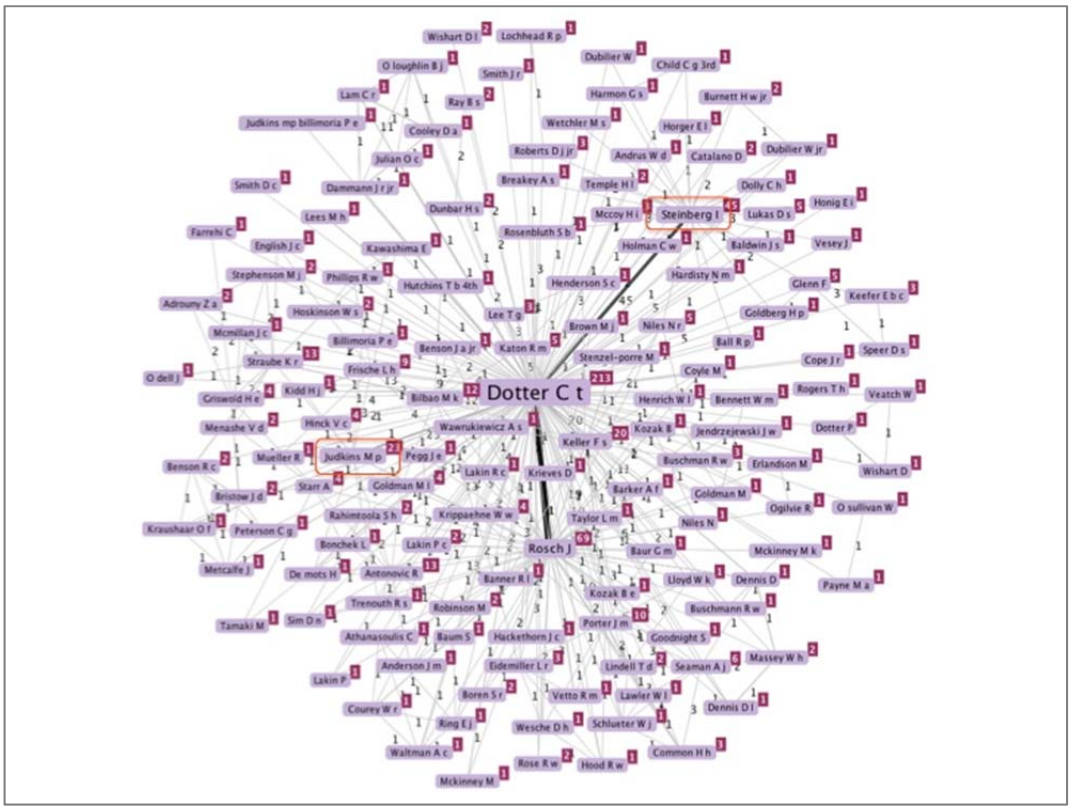


Figure 1. Network of C.T. Dotter co-publications.

Table 2. C.T. Dotter main scientific collaborators.

Author	Lab. / Dpt.	Institution	Publi.
Rösch, Johannes	Center of Cardiac Surgery	Friedrich Alexander University (DE)	340
Keller, Frederick S.	Dotter Interventional Inst.	Orgeon Health & Sciences Medical Center (USA)	215
Steinberg, Israel	Dpt. of Surgery, Medicine & Radiology	New Loma Linda Univ. (USA)	174
Judkins, Melvin P.	Coordinating Center for Collaborative studies in Coronary Artery Surgery	New York Hospital – Cornell Univ. (USA)	68
Bilbao, Marcia K.	Dpt. of Radiology	University of Oregon Mecial School (USA)	22

He published his last paper in 1981, four years before his death. By the end of his career, his scientific work totalized more than 4500 citations and reached 7866 citations at the end of 2013 (Fig. 2).

Dotter successfully diffused his results and obtained recognition from his academic community with an average of 52-251 citations every year.

It is interesting to point out that before his landmark paper was published in 1964, he was already an active researcher with 100 publications, well recognized by his academic community with 1068 citations at that time.

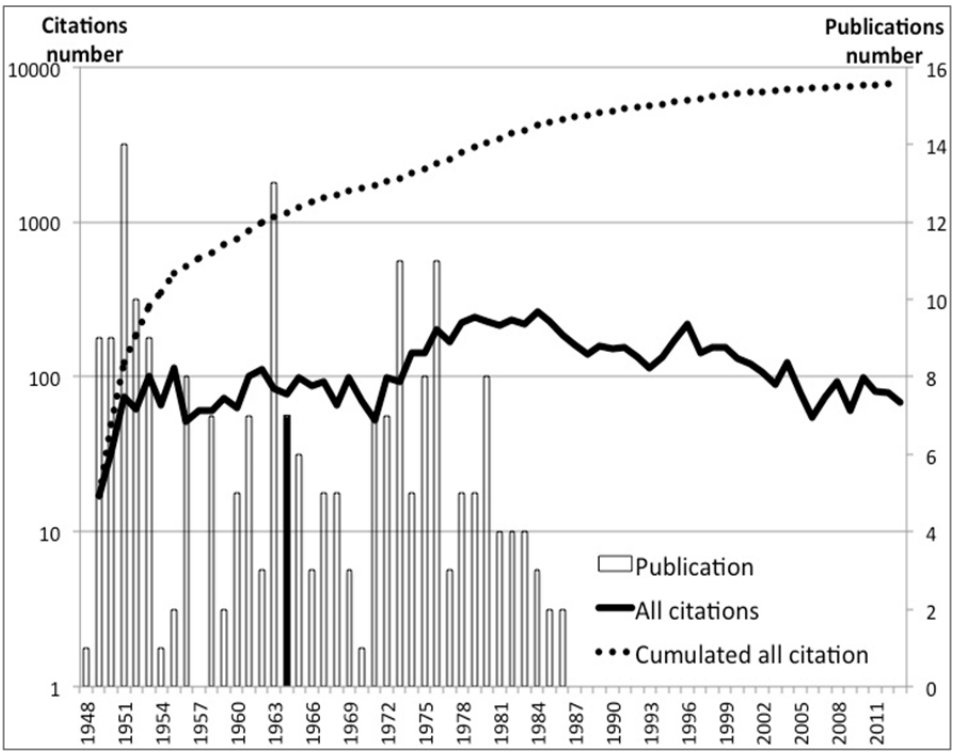


Figure 2. Dotter's publications and citations.

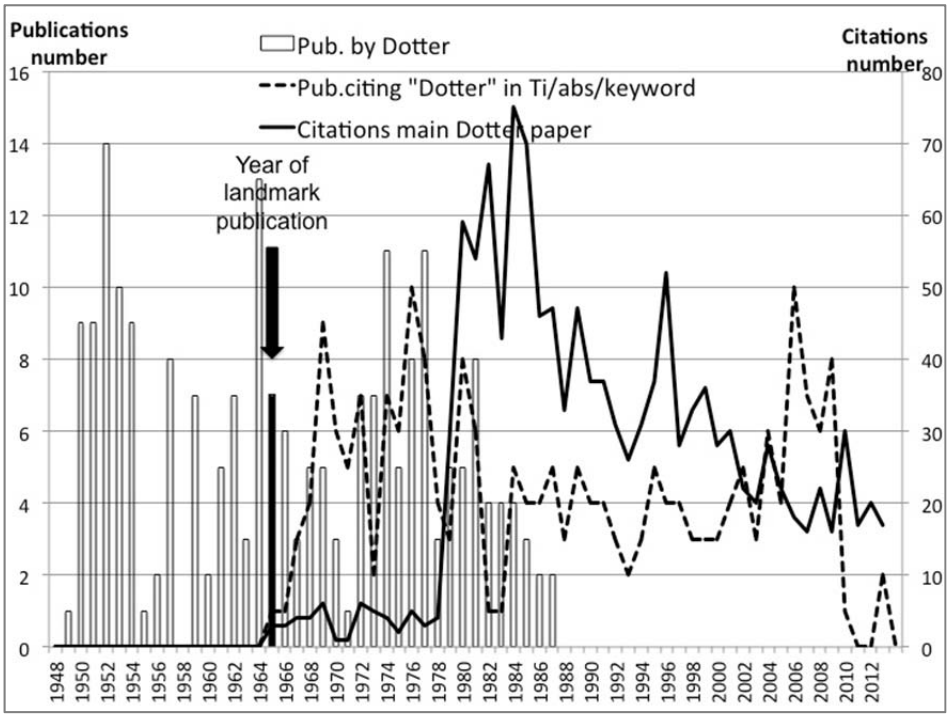


Figure 3. Dotter's main paper citations and Dotter's name apparition in the literature.

Dotter's landmark paper: a sleeping-beauty?

Dotter's landmark paper published in 1964 (Figure 2; black box) was cited with an average of 19.31 citations per year, totalizing 1275 citations today. However, during the first 14 years, his paper was cited only 51 times (Figure 3; full line) before suddenly gaining 29 citations in 1979 and more than 50 citations per year in the latter period.

Therefore, Dotter's main paper has the characteristics of a "sleeping beauty" despite the fact that it does not exactly fit Van Raan's definition (depth of sleep: 3.64 citations/year length of sleep period: 14 years; awake intensity: 52.25 citations/year).

During the delayed recognition period, Dotter was frequently named (n=76) within medical literature (Figure 3: dotted line), as well as his technique, percutaneous transluminal angioplasty (data not show) attesting that the "sleeping period" was traversed by a medical controversy.

The corresponding "Prince" was identified by visualizing the pattern of citations (Fig.4). A German cardiologist, A. Gruntzig, inventor of the coronary balloon angioplasty, was the first to referred to Dotter's previous work. He first did so in a paper published in German in the journal Deutsche Medizinische Wochenschrift in 1974, which had however only very little echo at that time until it was published in English in a well established journal in radiology (American J. of Roentgenology, 132:547-552, April 1979).

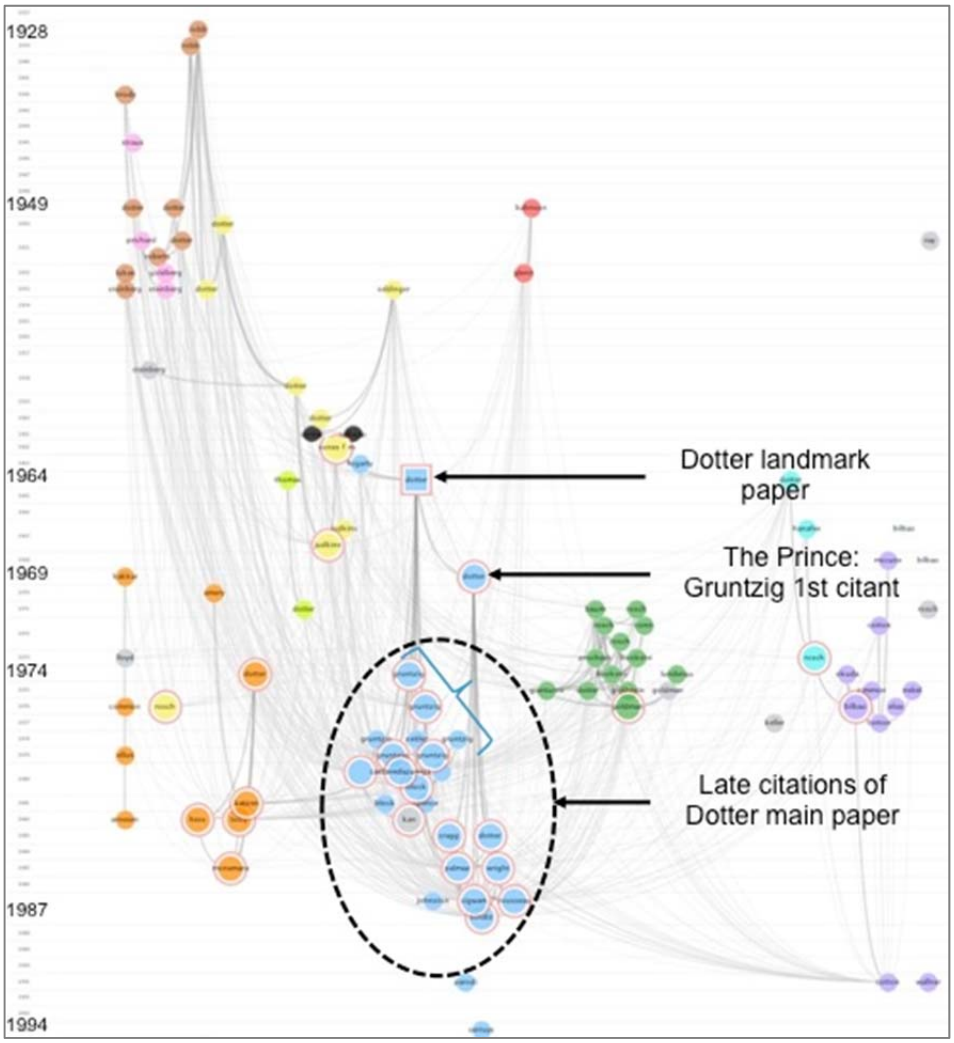


Figure 4. Citation network of CT Dotter paper and its direct and indirect successors.

Later on, Gruntzig's paper, citing Dotter pioneering work, was quickly cited in the medical literature (n=23, year +1) and its peak of citations coincided with the awaking of Dotter landmark paper citations (Figure 5).

Discussions

Dotter landmark paper has the characteristics of a sleeping-beauty but does not fit Van Raan's criteria. Therefore, this case study will discuss the accuracy of Van Raan's criteria to define

"sleeping beauty" in science, and introduce the category of "sleeping beauty" – like as a paper. Beside it is necessary to pinpoint that the sleeping period might indeed be a restless sleep period traversed by scientific controversy that could be traced back in publications databases by differentiating bibliographic references from citations in the text, or by analyzing the nature of the documents, especially article versus editorial, letter or review. These observations should open new avenues in identifying "sleeping beauties" in the literature, and nurture science resistance or controversy study in sociology of science.

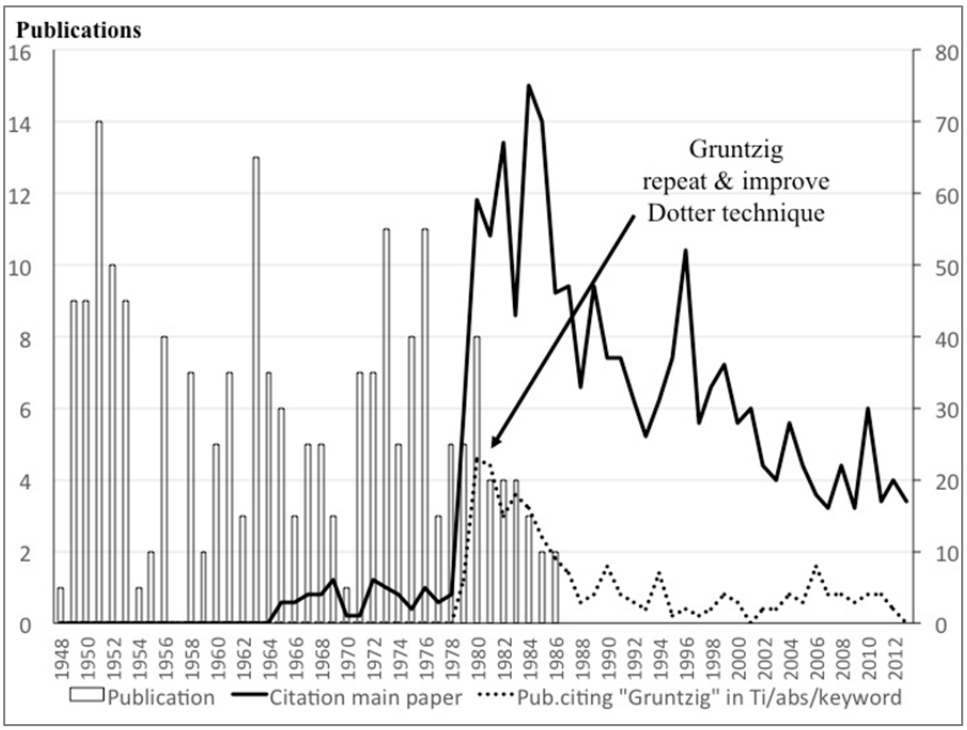


Figure 5. Citations curves of Dotter's paper & Gruntzing refering paper.

Acknowledgments

This work is supported by a grant from the French National Cancer Institute (INCA#6165).

References

Braun, T., Glänzel, W., & Schubert, A. (2010). On Sleeping Beauties, Princes and other tales of citation distribution. *Research Evaluation*, *19*(3), 195-202.

Burrell, Q.L. (2005). Are "sleeping beauties" to be expected?" *Scientometrics* 65, 381-389.

Dotter, C. & Judkins, M. (1964). Transluminal treatment of arteriosclerotic obstruction. Description of a new technic & a preliminary report of its applications. *Circulation 30*, 654-70.

Gruntzig, A. & Hopff. H. (1974), Perkutane Rekanalisation chronischer arterieller Verschlüsse mit einem neuen Dilatationskatheter. *Deutsche Medizinische Wochenschrift, 99*, 2502-2505.

Li, J. & Ye, F.Y. (2012). The phenomenon of all-elements-sleeping-beauties in scientific literature. *Scientometrics,* 92, 795-799.

Payne, M. (2001). C T Dotter : the father of Intervention. *Texas Heart Institute J. 28*, 28-38.

Rösch, J., Keller, F. S., & Kaufman, J. A. (2003). The birth, early years, and future of interventional radiology. *Journal of Vascular and Interventional Radiology*, *14*(7), 841-853.

Van Dalen, H.P. & Henkens, K. (2005). Signals in science - On the importance of signaling in gaining attention in science. *Scientometrics, 64*, 209-233.

Van Eck, N.J. & Waltman L. (2014), CitNeExplorer: a new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8: 802-

Van Raan, A.F.J. (2004). Sleeping Beauties in science. *Scientometrics, 59*, 467-472.

Citation Distribution of Individual Scientist: Approximations of Stretch Exponential Distribution with Power Law Tails

O.S. Garanina and M.Yu. Romanovsky[1]

[1]*slon@kapella.gpi.ru*
A.M.Prokhorov General Physics Institute of RAS, Vavilov str., 38, 119991 Moscow (Russia)

Abstract
A multi-parametric family of stretch exponential distributions with various power law tails is introduced and is shown to describe adequately the empirical distributions of scientific citation of individual authors. The four-parametric families are characterized by a normalization coefficient in the exponential part, the power exponent in the power-law asymptotic part, and the coefficient for the transition between the above two parts. The distribution of papers of individual scientist over citations of these papers is studied. Scientists are selected via total number of citations in three ranges: 10^2-10^3, 10^3-10^4, and 10^4-10^5 of total citations. We study these intervals for physicists in ISI Web of Knowledge. The scientists who started their scientific publications after 1980 were taken into consideration only. It is detected that the power coefficient in the stretch exponent starts from one for low-cited authors and has to trend to smaller values for scientists with large number of citation. At the same time, the power coefficient in tail drops for large-cited authors.
One possible explanation for the origin of the stretch-exponential distribution for citation of individual author is done.

Conference topic
Citation and co-citation analysis

Introduction
The discussion of how citations of individual authors are distributed has a long history going back even to E. Garfield (1955). In general, there are two points of view on this: the distribution of papers of each scientist is a so-called stretch exponent $W \sim \exp(-x^{\alpha}/T)$, where x is the number of citations, T is some normalization, α is the power exponent coefficient (Redner, 1998; Laherrere & Sornette, 1998). Usually α is considered as 0,3-0,5 (Redner, 1998, Iglesias & Pecharroman, 2006). A slightly more complicated distribution was introduced by (Tsallis & de Albuquerque, 2000).
The second point is that the above distribution has power-law (Pareto, Zipf) character, i.e. $W \sim x^{-\beta}$ where β is the power (Silagadze, 1999; Vazquez 2001; Lehmann et al., 2003). Often, this dependence is treated as the asymptote (tail) of distribution for comparably large x. In this case, the main body is considered as log-normal (Redner, 2005; Stanley, 2010). It should be noted that there are more complicated models of citation distribution.
The idea of our work is to consider the citation distribution of individual scientists taking into account that the distributions for "various-ranking" scientists can be different. Also, it is interesting to join the above stretch-exponential distributions and power-law distributions: observation of tails of citation distributions of individual scientists often demonstrates a presence of small number of extremely-high cited articles, while other articles of considered scientists can be cited much more moderately. From this point of view, the consideration of citation data of a large set of authors (like in (Redner, 1998) etc.) provides rough enough results. Thus, we concentrate on analysis of citation distributions of individual scientists, taking into account some differences in the total number of citations of each. The cumulative distribution of the number of articles with some or larger number of citations will be analyzed.
Of course, the proposed approach is rough enough, since it does not take into account the co-authoring of cited articles. The authors think that it should be considered in further studies in case of wide scientific interest.

The descriptive model is based on our previous works for tailed distributions: Gauss for stock return distributions (Romanovsky & Vidov, 2011), and exponential Boltzmann distribution for new car sells, incomes and weights (Romanovsky & Garanina, 2015). The authors do not know consistently introduced mathematical formulae for distributions with exponential main part and power law asymptote.

Multi-parametric family of curves with stretch exponential main part and power law tail

To define the general form of the desired distribution, one may proceed from the results presented in (Romanovsky & Vidov, 2011) as a starting point. According to (Romanovsky & Vidov, 2011), the sum of a large quantity N of random values similarly distributed with the probability density function (PDF) of the Student's (generally, non-integer) type $\sim z_0^{2\beta}/(z_0^2+f^2)^{2\beta}$ has the distribution of the Gaussian form for comparably small values of fluctuations f:

$$W_G(f) \approx \frac{1}{\sqrt{\pi}} \exp(-f^2)$$

and $\sim 1/f^{2\beta}$ for large f (z_0 being a normalization constant, the sum is treated as random walks in (Romanovsky & Vidov, 2011)). The obvious mathematical generalization to get the exponential part with power-law tail is to perform the transformation $f^2 \to R/T$ (here T can be interpreted as an effective "temperature"). Upon switching from parameters N, z_0, β to parameters θ, T, $\sigma\beta$, the transformation yields the curve with the stretch exponential main part and a transition to power law at the tail in an explicit form of a PDF (Romanovsky & Garanina, 2015):

$$W_{T(\sigma\beta)\theta}(R) = \frac{1}{\sqrt{\pi T}} \int_0^\infty \cos(xR^\sigma) \left\{ \frac{2}{\Gamma(\beta - 1/2)} \left[(\beta - 3/2) \frac{xT}{4\theta} \right]^{\beta/2 - 1/4} K_{\beta - 1/2} \left[\sqrt{(\beta - 3/2) \frac{xT}{4\theta}} \right] \right\}^\theta dx \quad (1)$$

Here R is variable, Γ is the gamma-function, $K_{\beta-1/2}$ is the modified Bessel function of the 2nd kind (also known as ''McDonald function'').

The approximation of Eq. (1) for comparably small R (up to several units of $T^{1/2\sigma}$) is easily reduced to only a dependence on parameter T

$$W_T(R) \cong \frac{1}{T} \exp\left(-\frac{R^{2\sigma}}{T}\right) \quad (2)$$

The general drop off law for $W_{T\beta\theta}$ in the case of large R is $R^{-\beta\sigma}$. The parameter θ describes transition among (stretch) exponential and power-law part of (1). This transition goes under larger R (and smaller values of $W_{T(\sigma\beta)\theta}$) under larger values of θ.

To obtain a general form of W, note that

$$I_\beta(x) = \frac{2}{\Gamma(\beta - 1/2)} \left[(\beta - 3/2) \frac{xT}{4\theta} \right]^{\beta/2 - 1/4} K_{\beta - 1/2} \left[x \sqrt{(\beta - 3/2) \frac{T}{\theta}} \right], \quad (3)$$

It is easy to see that it is a monotonic function of β. Indeed, if $\nu = \mu + 1$, one finds, considering the rule for modified Bessel functions of the 2nd kind, that the ratio $I_\mu(x)/I_\nu(x)$ becomes

$$\frac{I_\mu(y)}{I_\nu(y)} = \frac{K_{\mu+1/2}(y) - K_{\mu-3/2}(y)}{K_{\mu+1/2}(y)} = 1 - \frac{K_{\mu-3/2}(y)}{K_{\mu+1/2}(y)} < 1$$

Furthermore, $\forall \eta : \nu > \eta > \mu$, and one finds that $I_\nu > I_\eta > I_\mu$. Thus, it is not necessary to investigate (1,3) with an arbitrary β. It is enough to consider the integer $\beta = 2, 3, \ldots$, while integrals with intermediate β will be ''locked'' among integrals with neighboring integers β that are expressed by means of elementary functions. Then $n = \beta - 1$,

$$K_{\beta - 1/2} \left[x \sqrt{(\beta - 3/2) \frac{T}{\theta}} \right] = K_{n+1/2} = \sqrt{\frac{\pi}{2x \sqrt{(\beta - 3/2) \frac{T}{\theta}}}} \sum_{k=0}^{n} \frac{(n+k)!}{k!(n-k)! \left[2x \sqrt{(\beta - 3/2) \frac{T}{\theta}} \right]^k} \quad (4)$$

The three functions $W_{T(\sigma\beta)\theta}$ for $\sigma\beta = 2, 1, 0.8$ are:

$$W_{T(\sigma\beta)\theta}(R) = \frac{1}{\sqrt{\pi T}} \int_0^\infty \cos(xR^{\sigma\beta}) \exp\left(-x\sqrt{\frac{\theta T}{2}}\right)\left(1 + x\sqrt{\frac{T}{2\theta}}\right)^{\theta} dx \qquad (5)$$

We used here the simplest form of the function (1) for β=2 for the following approximations of empirical data. The functions $W_{T(\sigma\beta)\theta}$ for σ= 0.5, 0.25, 0.2 are shown in Fig.1. It is seen as a well-coincidence of general functions with corresponding approximation exponents for comparably small values of variable R.

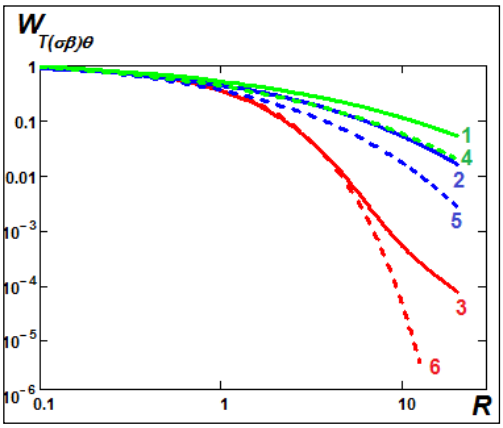


Figure 1. Functions $W_{T(\sigma\beta)\theta}$ for β=2 and σ=0.5 (curve 3), σ=0.25 (curve 2), σ=0.2 (curve 1) for comparably small R. The straight lines (4-6) are exponents $exp(-R^{2\sigma}/T)$ for σ=1,0.5,0.4, respectively. Here T=1, θ=300.

For large R, these functions drop off as R^{-2}, R^{-1}, $R^{-0.8}$, respectively (see Fig.2):

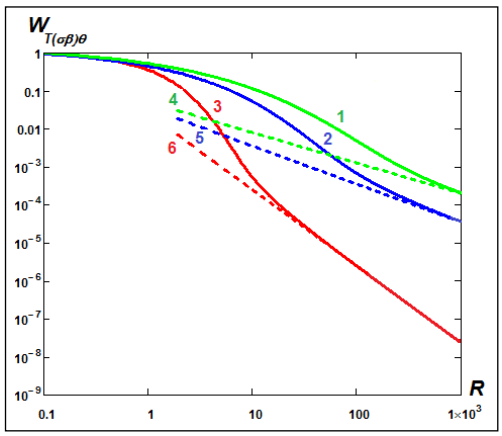


Figure 2. Functions $W_{T(\sigma\beta)\theta}$ for the same β and σ (curves 3-1) as on Fig.1. Hyperboles $R^{-\beta\sigma}$ (straight lines 6-4 on double-logarithmic plot) have σ=0.5, 0.25, and 0.2 (curve 4), respectively. Parameters T, θ are the same as on Figure 1.

Thus the introduced function (1) well-describes the stretch exponent for small (and moderate) values of argument, and provides power-law asymptotes for large R. We used these functions in the next section.

Distribution of citation of individual authors

It was found that the distributions of citations of individual authors are different. It can be expected due to, for example "Matthew effect" (see Bonitz et al., 1997; Bonitz & Scharnhorst, 2001; Stanley, 2010). One may expect that scientists with total number of citation in range 10^2-10^3, 10^3-10^4, and 10^4-10^5 have different distributions of citations. Let us call the scientists with total number of citations in these ranges as the "first-type scientist", etc. We study these intervals for physicists in the ISI Web of Knowledge. The scientists who started their scientific publications

after 1980 were taken into consideration only. We took 20 scientists for the first two ranges, and several scientists for the third. Typical examples of citation distributions are presented below on Figs. 3-5.

On Fig. 3, the cumulative citation distribution (i.e. the number of articles with citations larger than the value R) for experienced scientists with total number of citations in the first range 10^2-10^3 is presented:

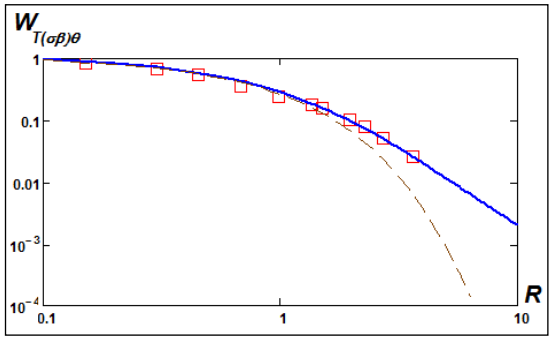


Figure 3. The distribution of articles over citations for the first-type scientist. Open squares are empirical points, the solid curve is $W_{T(\sigma\beta)\theta}$ (5) for β=2, σ=0.5, T=6.5, θ=10, dashed line is an exponent (2) with σ=0.5, T=6.5.

The function $W_{T(\sigma\beta)\theta}$ on Fig.3 is normalized on total number of articles of the first-type scientists in ISI Web of Knowledge. The variable R is the number of citations normalized on T that is the mean citation of this author. It is seen that the function $W_{T(\sigma\beta)\theta}$ (5) well describes the empirical data, the clear difference from the exponent (2) is on-site. At the same time, the total exit on the asymptotic curve $\sim R^{-2}$ does not realize. The last was observed for other-types scientists.

The citation distribution of the second-type scientist (this is a range of world well-known person) is demonstrated on Fig. 4:

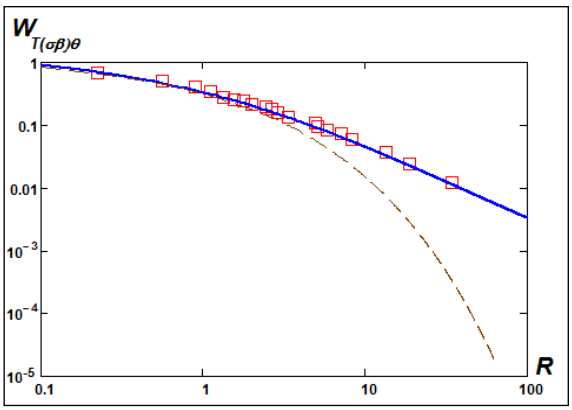


Figure 4. The distribution of articles over citations for the second-type scientist. Open squares are empirical points, the solid curve is $W_{T(\sigma\beta)\theta}$ (5) for β=2, σ=0.25, T=47.4, θ=5, dashed line is an exponent (2) with σ=0.25, T=46.

The normalization of $W_{T(\sigma\beta)\theta}$ on Fig.4 was on total number of articles also. Indeed, the variable R is normalized now on $T^{2\sigma} = (47.4)^{2\sigma} = 6.9$. The "difference" between empirical data as well as function (5) with pure stretch exponent $exp(-R^{1/2}/T)$ is larger than on Fig.3 for the first-type scientist. The total exit on the asymptotic curve $\sim R^{-1}$ is also not realized.

The citation distribution of the third-type scientist (this is a range of Nobel Prize winners) is demonstrated on Fig. 5:

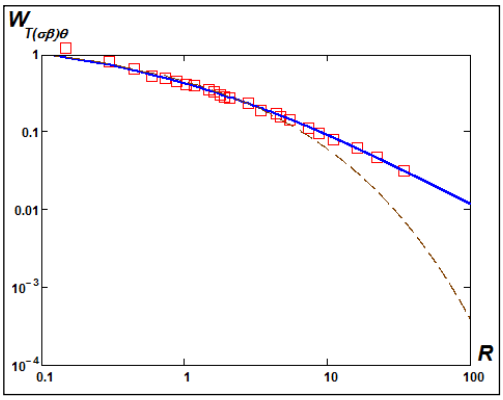


Figure 5. The distribution of articles over citations for the third-type scientist. Open squares are empirical points, the solid curve is $W_{T(\sigma\beta)\theta}$ (5) for β=2, σ=0.2, T=340, θ=5, dashed line is an exponent (2) with σ=0.2, T=340.

The normalization of $W_{T(\sigma\beta)\theta}$ on Fig.5 is the same, the variable R is normalized now on $T^{2\sigma} = 340^{2\sigma} = 10.3$. It is interesting that all values $T^{2\sigma}$ for all three-types scientists are close to each other and may characterize the citation distribution of individual scientists.

Explanation attempt

Let us try to explain the appearance of stretch exponents in cumulative distribution of such random values like citations. We start from the standard exponential distribution

$$W_1 = \exp(-x) \tag{6}$$

where we used normalization T=1 to simplify the following expressions.

How can these calculations be "translated" into the language of citations? The first cause of a citation of some article is the scientific results of this article. Since the author who can potentially cite the above article may find or not find this article, the process of citation due to the scientific significance looks like the two-body exchange (of information in this case) and is provided by distribution (6). Thus it may be that the basic cumulative distribution of citations arises due to the scientific significance of the article and looks like (6).

There are clear additional independent causes for citations. One of them is the name of author (or one of authors in case of co-authoring) of a potentially cited article. It may be the name of scientist in the group that works in the same area of science studied with the author of the cited paper, there arises another causes to cite some scientist. Since this scientist may also be chosen randomly in the process of information exchange, the probability distribution to cite this scientist looks like (6) as well. If now the citation is realized due to two causes: by scientific significance and cited article author, the random value of such citation is the factor of two random values characterized by distribution (6).

Since the causes for citation are independent, they can be considered as some coordinates. For two cases, they are above "scientific significance" and "author's name". The variation of these coordinates here are from small to large scientific significance and from large to small reputation of cited scientist. At the same time, we observe citation as being a principally one-dimensional value: the citation either exists or does not exist. Therefore, all distributions (6) reduce to one dimension. The transformation of coordinates in (7) $x^2 \rightarrow y$ provides than for cumulative distribution function

$$W_2(y) = \exp\left(-\sqrt{y}\right) \tag{7}$$

i.e. the main part of stretch exponent (2) with σ=0.25. These stretch-exponents distributions were observed by us and described in the chapter of this paper "Distribution of citation of individual authors".

The same procedure in case of three clearly existed "coordinates" provides cumulative distribution

$$W_3(y) = \exp\left(-\sqrt[3]{y}\right) \quad (8)$$

The same conduction for power-law tailed stretch exponential distributions should take into consideration the power exponents in tails for original distributions of "scientific significance" etc., and needs the volumetric calculations.

Conclusion

The 4-parametric family of functions representing the stretch exponential distribution for small and medium values of the argument combined with a power-law asymptotic tail, along with various transitions between these two parts, is introduced. These functions are demonstrated as good fits of the available empirical data for the cumulative distribution of citations to individual scientists.

Abstracting from the co-authoring of a cited paper, one may conclude that these cumulative distributions of papers of individual authors versus their citations have character of stretch exponent for small and moderate values of citations, and power-law form for asymptotic part. It looks that the "power of stretch", i.e. the introduced coefficient σ depends on the total number of citations, moreover, this coefficient starts from ½ (i.e. distributions start from normal exponent) and becomes smaller with an increase of the total number of citations. The power-law force becomes smaller in return.

The first attempt to explain the "main body" of distributions (stretch exponents) is provided.

Acknowledgements

The paper is support by RFBR grant 13-07-00672.

References

Bonitz, M., Brukner, E., & Scharnhorst, A. (1997). Characteristics and impact of Matthew effect for countries. *Scientometrics*, *40*, 407-422.

Bonitz, M., & Scharnhorst, A. (2001). Competition in science and the Matthew core journals. *Scientometrics*, *51*, 37-54.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*. 108–111.

Iglesias, J.E., & Pecharroman, C. (2006). Scaling the h-Index for Different Scientific ISI Fields. Online: http://arxiv.org/ftp/physics/papers/0607/0607224.pdf

Laherrere, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *The European Physical Journal B*, *2*, 525-539.

Lehmann, S., Lautrup, B., & Jackson, A.D. (2003). Citation networks in high energy physics. *Physics Review E,* 68. 026113.

Petersen, A.M., Fengzhong Wang, & Stanley, H.E. (2010). Methods for measuring the citations and productivity of scientists across time and discipline. *Physics Review E, 81*, 036114.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B, 4*, 131-134.

Redner, S. (2005). Citation statistics from 110 years of Physical Review. *Physics Today,* 58. 49–54.

Romanovsky, M.Yu., & Vidov, P.V. (2011). Analytical representation of stock and stock-indexes returns: Non-Gaussian random walks with various jump laws. *Physica A*, *390*, 3794–3805.

Romanovsky, M.Yu., & Garanina, O.S. (2015). New multi-parametric analytical approximations of exponential distribution with power law tails for new cars sells and other applications. *Physica A*, 427, 1-9.

Silagadze, Z.K. (1999). Citations and the Zipf-Mandelbrot's law, http://arXiv.org/abs/physics/9901035v2

Tsallis, C., & de Albuquerque, M.P. (2000). Are citations of scientific papers a case of nonextensivity? *The European Physical Journal B*, 13, 777-780.

Vazquez, A. (2001). Statistics of citation networks. *E-prints* arXiv:condmat/0105031.

Influence of International Collaboration on the Research Citation Impact of Young Universities

K. A. Khor and L.-G. Yu

mkakhor@ntu.edu.sg; mlgyu@ntu.edu.sg
Research Support Office and Bibliometrics Analysis, Nanyang Technological University, #B4-01, Block N2.1, 76, Nanyang Drive, Singapore 637331 (Singapore)

Introduction

It is widely presumed that international collaboration benefits the researchers and the organisations involved, and enhances the quality of research (Persson, 2010). However, research also suggests that the effects of international collaboration may vary across disciplines and the authors' countries (Moed, 2005).

In this study, we investigated the effect of international collaboration on the impact of publications of selected young universities, and compared to that of renowned old universities. The 5-year citations per paper (CPP) data, the international collaboration rate, the CPP differential between publications with and without international collaborations, and the difference between the percentages of international collaborated publications falling in the global top 10% highly cited publications and the percentage of overall publications falling in the global top 10% highly cited publications (Δ%Top10%) are used as the impact indications. These data are extracted from the Thomson Reuters Web of Science (WoS) database and Essential Science Indicator (ESI) based on papers published from 2004 to 2013. Young institutions ranked by the 2014 Times Higher Education (THE)'s 100 under 50 Universities are selected in this study, and some renowned universities (> 100 years old) are selected as references for "old universities".

To eliminate the discipline difference effect, the increment of 5-year (2010-2014) field weighted citation impact (FWCI) of internationally collaborated papers over the 5-year overall FWCI of the institutions in SciVal® of Elsevier is used as another indicator. The collaboration among 8 old institutions and 8 young institutions are investigated.

Results and Discussion

Correlation between International Collaboration rate and CPP in 5-year interval

Figure 1 shows the 5-year ESI CPP trends as a function of 5-year international collaborations rate trends for selected young and old universities. While old universities have higher CPP in general, there are strong correlation between international collaboration rate trends and 5-year CPP trends. For example, for old universities, the CPP increased 4.12 for every 10% increase in international collaboration rate for MIT, 3.42 for Univ Oxford, and 3.01 for Stanford Univ. Among young universities, for Nanyang Technol Univ (NTU), it is 2.24 CPP per 10% Intl Collab increment, and that for Plymouth Univ is 3.02, and 0.73 for King Fahd Univ of Petr and Min.

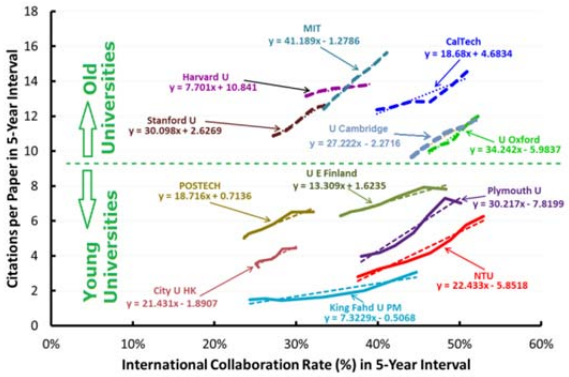


Figure 1. 5-Year CPP Trends vs. 5-Year International Collaborations Rate Trends for Selected Young and Old Universities.

The ΔCPP trends for publications with and without international collaborations for selected institutions are examined, and listed in Table 1.

Table 1. 5-Year Citations per Paper Differential between Publications with and without International Collaborations.

5-Year Period	Citations per Paper Difference between Publications with and without International Collaboration											
	Caltech	U E Finland	Univ Florence	Univ Tsukuba	Univ Melbourne	Univ Waikato	Kyushu Univ	MIT	NUS	HKUST	NTU	USM
2004-2008	5.5	3.04	3.59	5.19	4.5	2.68	2.26	3.24	0.78	1.03	0.2	1.08
2005-2009	5	3.38	3.68	5.65	4.06	1.68	2.62	3.25	0.66	1.44	0.51	0.97
2006-2010	4.2	3.42	3.79	4.87	4.3	2	2.55	3.38	0.63	0.55	0.43	0.65
2007-2011	4.2	4.1	3.91	4.85	4.42	2.1	1.85	2.68	0.82	1.33	0.47	0.11
2008-2012	4.8	4.44	4.38	4.65	4.77	2.86	1.75	2.29	1.28	1.44	0.05	-0.3
2009-2013	6.1	5.28	5.3	5.2	4.87	3.61	2.4	2.16	1.67	0.87	0.02	-0.7
ESI 2009-2013 CPP	15	8.53	7.68	6.48	8.66	5.43	5.28	15.7	7.83	6.7	6.92	3.47

From Table 1, we can find that in the case of Caltech, U Melbourne and U Tsukuba, the CPP difference between their international collaborated

publications and their publications without international collaboration is roughly 4 to 5. This explains the typical 5-year ESI CPP vs. international collaboration rate trends of these institutions: with the increase of international collaboration rate in their publications, the overall CPP of their papers has more weight from their international collaborated publications, and the overall CPP of their publications increased. Yet, for Hong Kong Univ of Sci & Techn (HKUST), Natl Univ Singapore (NUS) and NTU, the CPP gaps between publications with and without international collaboration are relatively small (around 0 to 1 CPP). This is because the fact that these institutions have attracted a lot of researchers with international background to work in these institutions, which makes the difference between their national research and international collaborated research relatively small.

Trends of difference between percentage of international collaborated publications falling in global top 10% highly cited publications and that for all publications (Δ%Top10%)

The study on difference between the percentage of international collaborated publications for an institution falling in the ESI global top 10% highly cited publications and the percentage of all publications of the same institution falling in the ESI global to top10% highly cited publications (Δ%Top10%) shows that, for all the selected young and old institutions, this difference is generally positive, means that internationally collaborated publications generally have a higher rate of high citation publications among all publications. Yet, this difference varies from one institution to another institution. For some renowned top universities like Caltech, Stanford University and University of Cambridge, although their overall CPP for their publications is already very high, the Δ%Top10% is still higher than the percentage of their overall publications falling in the global top 10%. Further investigation is needed to have an adequate explanation for this phenomenon.

Increment of field weighted citation impact (FWCI) of internationally collaborated papers over the FWCI of the involved institutions

Figure 2 shows the increment of FWCI for internationally collaborated papers over the overall FWCI of the two collaborating institutions among the selected 8 old institutions and 8 young institutions. 57 bilateral collaboration couples with 50 and more collaborating publications are identified among these 16 institutions, and the FWCI increment data for these collaboration couples are include in the plot. It can be seen that, international collaboration benefits both the young and the old institution, with the old institution to old institution collaboration provides the highest FWCI increment, followed by the old institution to young institution collaboration. Among the 57 bilateral collaborations, only 3 involved young institution to young institution collaboration, indicating that there are untapped potential for enhancement on bilateral collaboration among young institutions.

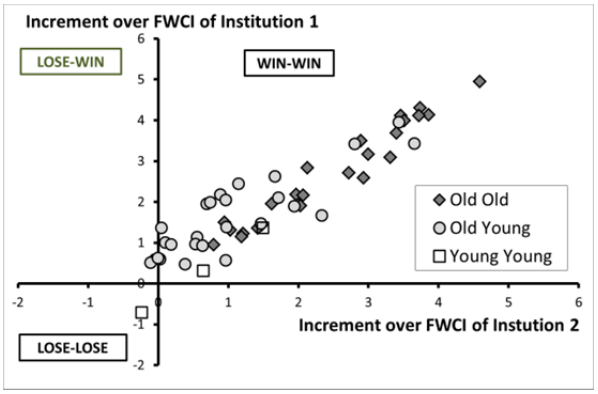


Figure 2. Increment of 5-year FWCI of internationally collaborated papers over the overall FWCI of the involved Institutions.

Conclusions

The investigation on the effect of international collaboration on the impact of publication of selected young universities and well established renowned universities show that, both young and old institutions received benefit from international collaboration using citation impact of their publications as indicator. For example, for old universities, the CPP increased 4.12 for every 10% increase in international collaboration rate for MIT and 3.42 for U Oxford. Among young universities, for NTU, it is 2.24 CPP per 10% Intl Collab increment, and that for Plymouth U is 3.02 CPP per 10% Intl Collab increment.

The percentage of publications fall in the ESI global top 10% highly cited publications for international collaborated publications is generally higher than that for all journal publications of the same institution. Yet, this difference varies from one institution to another institution.

The international collaboration also increases the FWCI of the institution, yet there are untapped potential to enhance the collaboration among young institutions.

References

Persson, O. (2010). Are highly cited papers more international? *Scientometrics*, *83*(2), 397-401.

Moed, H.F. (2005). *Citation analysis in research evaluation.* Dordrecht, Netherlands: Springer.

[redacted]

[redacted]

b.lancho@csic.es

[redacted]

Introduction

[redacted]

Research questions

[redacted]

Data and Methods

[redacted]

• Citations per paper: Average citations received by [redacted]

• References per paper: [redacted]

• Citation Rate Increment from the Collaborator [redacted]

Results

[redacted]

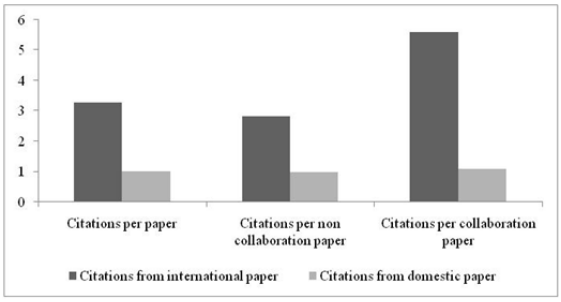


Figure 1. Comparison among the different averages in terms of citations made to Turkey, distinguishing in both cases between domestic and international articles.

[redacted]

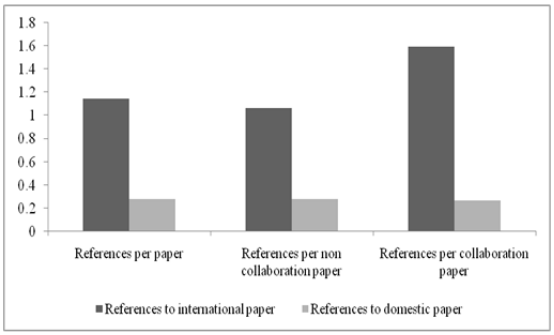


Figure 2. Comparison among the different averages in terms of references made by Turkey, distinguishing in both cases between domestic and international articles.

Table 1. This chart is referred to the total production in collaboration with Turkey and the total citations made to documents in collaboration with Turkey.

Country	Papers with different countries	Citation to collaboration documents	Citations from collaborators

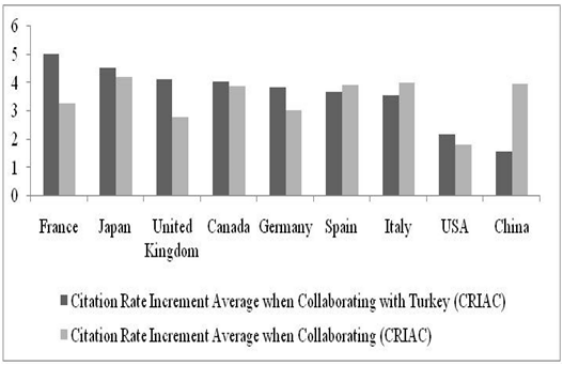


Figure 3. Comparison between CRIAC in general and CRIAC with Turkey.

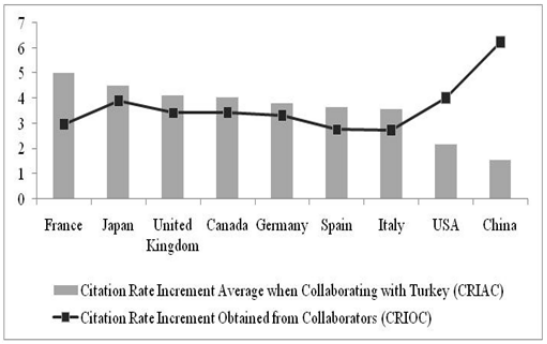


Figure 4. Comparison between the CRIAC with Turkey and the CRIOC among the nine countries with the largest production in 2004.

Interpretation

References

Scientometrics, 86

JASIST, 63

Scientometrics, 94

Bio Science, 55

Scientometrics 87

[illegible]

[illegible]

[1] s.r.goldberg@qmul.ac.uk
[illegible]

[2] t.evans@imperial.ac.uk
[illegible]

Introduction

[illegible]

Methods

[illegible] N [illegible] C [illegible] $c>0.1C$ [illegible] c [illegible]

$$\frac{n(c)}{N} = \int_{c-0.5}^{c+0.5} \frac{dx}{\sqrt{2\pi}\sigma x} exp\left\{-\frac{(\ln(x/C)+\sigma^2)^2}{2\sigma^2}\right\}$$

[illegible] σ [illegible]

[illegible] p [illegible] $\exp((N-t)/\tau)$ [illegible] $(N+1)$ [illegible]

[illegible]

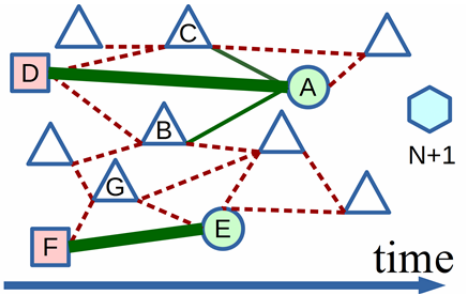


Figure 1. Illustration of Model C. A new paper (hexagon, *N+1*) is set to have four references. The first 'core' paper is chosen, A, using the global process of model B. Then with probability *q*, papers cited by A are also added to the new bibliography. Here B and C are considered (thin solid lines) but only D is added (thick line). The process continues until the required bibliography is complete. Here a second core paper E is chosen and one of its citations, F, is copied. At that point the process stops, paper G is never considered. The new bibliography is A, D, E and F.

[illegible] q [illegible] t [illegible]

Results

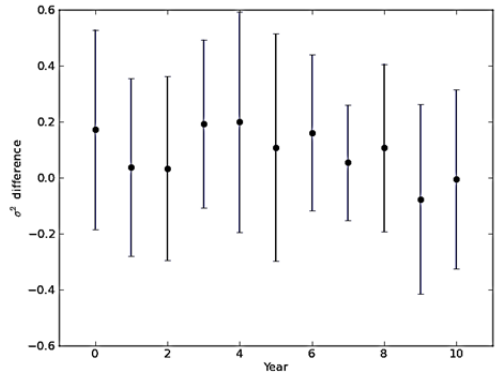


Figure 1. The difference between the width σ of the hep-th data and that found in our Model C for final fitted parameters.

Discussion

Acknowledgments

References

Complex Networks

Scientometrics *93*

Phys.Rev. E *72*

PNAS *105*

arXiv:

Complex Systems *14*

Scientometrics *62*

Citation analysis as an auxiliary decision-making tool in library collection development

Iva Vrkić

ivavrkic@gfz.hr
University of Zagreb, Faculty of Science, Department of Geophysics, Geophysical Library, Horvatovac 95, 10000 Zagreb (Croatia)

Introduction

Academic libraries in Croatia are facing constant budget cuts, making it difficult to obtain access to current scientific and professional journals (Krajna & Markulin, 2011). At the end of 2008 the Croatian economy had plummeted into recession and the *Ministry of Science, Education and Sports* ceased the funding of scientific literature acquisition (Krznar, 2011).

Parallel to budget cuts, the prices of scientific journals increased. The period from 2009 to 2014 showed a threefold increase in prices of the journals acquired by the Geophysical library in Zagreb (Figure 1), making it necessary to review the need for the purchase of each journal.

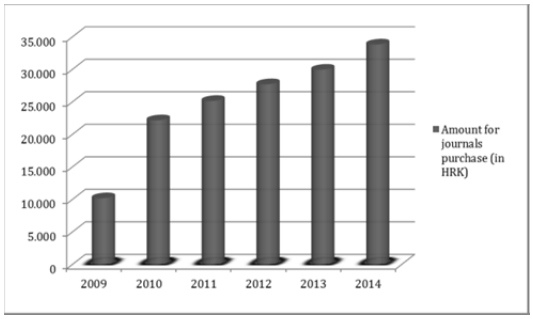


Figure 1. Threefold increase in prices of the journals acquired by the Geophysical library in Zagreb.

Quantitative and qualitative methods can be used to make optimal decisions regarding the purchase of journals (Gomez, 2002). The qualitative method is based upon interviewing lecturers and other competent scientific staff and taking their suggestions on which journals are essential. Their assessment of the journals' relevance is the most important guideline in creating an acquisitions policy. The quantitative method, on the other hand, provides the much-needed objectivity in the acquisitions process, but can only be used as an additional guideline to the qualitative method. This method can come in the form of usage statistics or the assessment of the journal's importance through citation analysis. Such an assessment is described in this paper. Although the quantitative method is objective, its results (list of most used/most relevant journals) cannot replace subject-matter experts' opinion, only inform them.

Methodology

The goal of this study is to determine the importance of certain journals for the geophysical community at the Faculty of Science in Zagreb. This will be done by compiling a list of journals most cited by the scientific staff at the Geophysical department from 2000 to 2014. References from all scientific papers published by the staff at the Department of Geophysics in the last 14 years were collected, and 6120 references were selected from journals cited by our geophysicists. The citation frequency was analysed, and references were listed for each journal.

Results and discussion

Assuming the citation frequency of articles from a certain journal confirms its importance for the scientists, the journals were listed by relevance after the data had been processed. The result is a list of 512 journals ranked by the number of citations. A "Top 15" list has also been created – 15 most cited journals by the members of the Department of Geophysics from 2000 to 2014 (Table 1).

Table 1. *Top 15* – most cited journals by the members of the Department of Geophysics form 2000 to 2014.

	Journal title	∑ citation
1	Journal of Geophysical Research	448
2	Journal of the Atmospheric Sciences	349
3	Quarterly Journal of the Royal Meteorological Society	335
4	Monthly Weather Review	222
5	Boundary-layer meteorology	213
6	Journal of Climate	202
7	Atmospheric Environment	195
8	Journal of Applied Meteorology and Climatology	185
9	Geofizika	162
10	Geophysical Research Letters	115
11	Tellus	114
12	International Journal of Climatology	108
13	The Astrophysical Journal	97
14	Bulletin of the Seismological Society of America	95
15	Annales Geophysicae	93

Data on the age of journal citations (cited by the members of the Department of Geophysics in a 14-

year period) was processed. Citation age is determined as the discrepancy between the publishing years of both the cited and the citing paper.

The citation age median for the whole set is 9 years. The histogram (Figure 2) shows that half of the citations are 0 to 9 years old, and rest of them are 10 to 133 years old. Citation frequency in 1st quartile shows statistically significant greater representation of citations in relation to the 2nd quartile ($\chi^2 = 9{,}86$; $P<0{,}0017$).

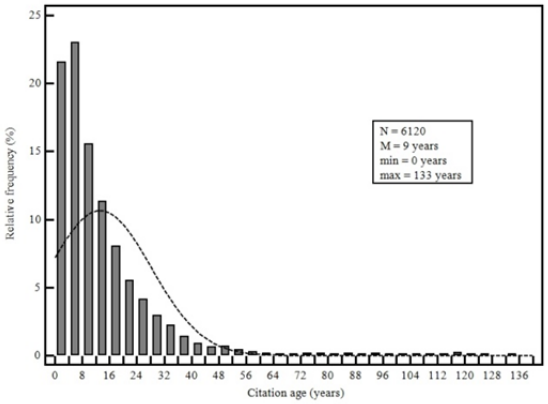


Figure 2. Citation frequency relative to citation age.

Therefore, recent scientific papers are the most cited.

Instead of a conclusion

Why is optimizing the library's acquisitions policy so important? The answer is, of course, because optimization is crucial in creating a list of the most relevant journals to be acquired, which can also be illustrated using the *Pareto principle*.

The Pareto principle is, amongst other thing, used to evaluate periodicals collections. It was named after Vilfredo Pareto, an Italian sociologist, who first used it to explain the distribution of land in Italy, where 80% of the land was owned by 20% of the population.

As previously mentioned, the principle applies to many different areas, so if applied to a periodicals collection, it will show that 20% of the periodicals in the collection will cover 80% of information needs. Also, 80% of the citations will be found in 20% of the periodicals (Dewland & Minihan, 2011).

This analysis further establishes the Pareto principle: 85,87% of the citations were found in the upper 20% of the periodical list. As a relatively low number of periodicals (20%) generates the most citations (85%), it's possible to conclude that, if an academic library strives to acquire the right periodicals and makes an optimal selection, it can provide good coverage of relevant information for its patrons, even if the quantity of said periodicals is low. In other words, a small but optimal selection of periodicals can cover the most of an institution's information needs.

References

Dewland, J. & Minihan, J. (2011). Collective serials analysis : the relevance of a journal in supporting teaching and research. *Technical Services Quarterly*, 28, 265-282.

Gomez, M. (2002). A bibliometric study to manage a journal collection in an astronomical library: some results. *Library and information services in astronomy*, 4, 214-222.

Krajna, T. & Markulin, H. (2011). Nabava knjižnične građe u visokoškolskim knjižnicama. *Vjesnik bibliotekara Hrvatske*, 54, 21-42.

Krznar, I. (2011). Identifikacija razdoblja recesija i ekspanzija u Hrvatskoj. *Istraživanja (Hrvatska narodna banka)*, 32, 1-17.

[illegible]

[illegible]

[1] clement.arsenault@umontreal.ca
[illegible]

[2] vincent.lariviere@umontreal.ca
[illegible]

Introduction

[illegible]

Literature Review

[illegible]

Data and Methodology

Results and Discussion

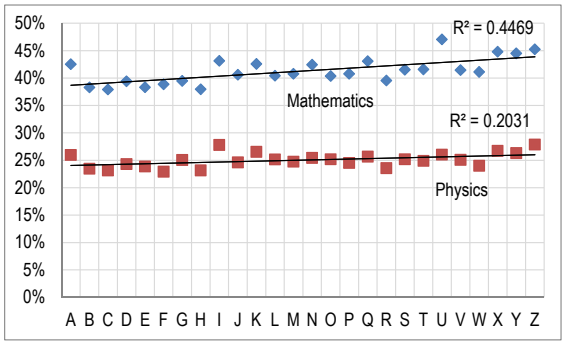


Figure 1. Uncitedness rates of Mathematics and Physics papers by initial letter of first author's surname.

References

PLoS ONE, 7

BMJ,

Information Storage and Retrieval,

The Journal of Socio-Economics,

Journal of Economic Perspectives, 20

Journal of Informetrics,

Journal of the American Society for Information Science and Technology,

Journal of Academic Librarianship, 26

Economic Inquiry,

Proceedings of the 17th International Conference on Science and Technology Indicators Conference

Review of Agricultural Economics,

Economica, 75

Journal of Informetrics, 6

[illegible]

[illegible]

[1]katranid@uom.gr, [2] tpanag@uom.gr
[illegible]

Introduction

[illegible]

Literature Review

[illegible]

Data

[illegible]

[illegible]

Methodology

[illegible]

$$P_i=p(y_i=1)=\beta_1+\beta_2(Belgium*PhD^{US})+\beta_3(Denmark*PhD^{US})+\beta_4(Germany*PhD^{US})+\beta_5(Greece*PhD^{US})+\beta_6(Italy*PhD^{US})+\beta_7(Portugal*PhD^{US})+\beta_8(Belgium*PhD^{Belgium})+\beta_9(Denmark*PhD^{Denmark})+\beta_{10}(Germany*PhD^{Germany})+\beta_{11}(Greece*PhD^{Greece})+\beta_{12}(Italy*PhD^{Italy})+\beta_{13}(Portugal*PhD^{Portugal})+\beta_{14}(Belgium*Female)+\beta_{15}(Denmark*Female)+\beta_{16}(Germany*Female)+\beta_{17}(Greece*Female)+\beta_{18}(Italy*PhD^{Italy})+\beta_{19}(Portugal*Female)$$

Results

Conclusions

References

Recherches économiques de Louvain

Scientometrics, 82,

Bulletin of Economic Research, 58,

Scientometrics

Economics of Education Review

Bulletin of Economic Research.

Do First-Articles in a Journal Issue Get More Cited?

Tian Ruiqiang, Yao Changqing, Pan Yuntao, Wu Yishan, Su Cheng and Yuan Junpeng

trq2011@sina.com

Institute of Scientific and Technical Information of China, 15 Fuxing Road, 100038, Beijing (China)

Introduction

As the advice of peers on the quality of a submitted paper prior to publication, peer review can be regarded as the pre-publication evaluation. Bibliographic citations of scientific papers used as indicators of the visibility, impact, and quality of scientific publications, could be regarded as the post-publication evaluation.

Intentionally or not, journal editors often put the accepted manuscript with nice comments by peer reviewers at the top of all papers in an issue. The First-Articles of journal issues are generally regarded with higher importance, intense creativity or superior quality through peer review process. Judge A, Cable M, Colbert E (2007) deemed that journal editors placed the best paper in the "pole position", and they confirmed this anecdotal evidence further in their study. Specifically, 75% of 16 journals indicated that quality played some primary role in selection of the first articles. Wang (2015) also admitted that journals would choose the very best paper of an issue on the cover, "a paper that in 20 year's time might win a Nobel Prize", according to the opinion of Stang, the EIC of *Journal of the American Chemical society* (Ritter 2006).

Since there are evidences that peer reviewers can successfully discriminate between manuscripts that have a greater chance to be cited in future. Further, in this sense, we made a hypothesis that the best articles selected by peer reviews—usually the First-Articles, will be superior in receiving higher citations after publication. In this paper we will illustrate how peer review and the performance of journal papers measured by bibliometric indicators could concordance with each other. In particular, we examined whether there were obvious citation differences between First-Articles and non-First-Articles published in the same issue of a journal.

Data and Methodology

Twins data, a sampling method used in labour economics, reaches "other things being equal" to a certain extent. Twin studies are often employed to evaluate the inheritance of a trait by dissecting the genetic and environmental contributions to the trait.

In this study, we regard the First-Articles and non-First-Articles in the same issue as twins. They were published in the same time and have similar disciplinary backgrounds.

We select First-Articles from Scopus and Web of Science (WoS). First, we choose journals which publish research articles on their first pages rather than other types of documents, such as editorial, letters et al. And we find that most mathematic journals satisfy this criterion well. Thus we select top100 mathematical journals by their Impact Factors from JCR 2013. Then, we acquire twins data by retrieving articles published in those 100 journals between1995-1999 in Scopus and WoS. As a result, we obtained 19,411 articles in 62 journals in WoS on December 25, 2014 and 18,524 articles in 67 journals in Scopus on January 13, 2015 respectively. The difference of journal numbers is resulted that some journals were not indexed as early as 1995-1999 while included in 2013 JCR. And we identified 2050 out of WoS and 2229 out of Scopus First-Articles, excluding those articles published on supplementary issues, special issues. Table 1 provides an overview of the samples.

Table 1. Descriptive statistics of the samples

	Scopus		WoS	
	Fr	**Non -Fr**	**Fr**	**Non-Fr**
Articles	2229	16295	2050	17361
		67 journals		62 journals

Results

First-Articles receive higher CPP&CTC

The indicator CPP (the average number of citations received per article) and CTC (the contributions to total journal citations) were taken as the criterion to assess the citation position of First-Articles and non-First-Articles in their own disciplinary citation environment. It revealed obvious differences in citations between the First-Articles and non-First-Articles. As shown in Table 2, in WoS, the First-Articles received higher average citation (AC) (16.56) since publishing, while the non-First-Articles got 13.69. In Scopus, the First-Articles accumulated 17.00 of AC, those non-First-Articles of 14.00. In WoS, the First-Articles contribute 12.5% to total citations (TC) of the journal when their proportions in total documents remain only 10.6%. Though the non-First-Articles got 89.4% share of total documents, their contributions of TC remain 87.5%. And the case is almost the same in Scopus: the First-Articles contribute 14.2% to TC when the proportions of articles remain only 12%. Though

the non-First-Articles got 88% of articles, their contributions of TC remain 85.8%.

Based on ANOVA test, we found significant difference between TC of 2050 First-Articles and 17361 non-First-Articles in WoS at the 0.05 significance level. Similarly, in Scopus there is also significantly different between 2229 First-Articles and 16295 non-First-Articles. Specifically, TC of First-Articles is significantly higher than non-First-Articles. From WoS, the non-First-Articles received mean TC of 13.69. While under same circumstance, First-Articles received clearly higher mean TC of 16.56. In terms of Scopus, the non-First-Articles reached at 14.00 of mean TC. And this time, the similar backgrounds, First-Articles performed more excellent, reaching notably higher mean TC of 17.00. Therefore, First-Articles are higher impact than non-First-Articles both in WoS and Scopus.

Table 2. TC difference in ANOVA test

	WoS			Scopus		
	Num	**Mean**	**SD**	**Num**	**Mean**	**SD**
Fr	2050	16.56	30.13	2229	17.00	27.08
N-Fr	17361	13.69	24.03	16295	14.00	24.51
P			0.000			0.000

Nearly 24% First-Articles are most highly cited, while non-cited articles account for only 10%

It shows 22.6% First-Articles in average are also the papers with highest TC among papers published in the same journal issues in WoS. And the proportion keeps stable in the observe window. In Scopus, the percentage of the most highly cited papers in First-Articles goes to almost 25%. In 1997, it even reached a peak of 27%.

Table3. Citation difference of First-Articles and non-First-Articles in WoS& Scopus

	WoS	**Scopus**
CPP-Fr	16.56	17.00
CPP-Non-Fr	13.69	14.00
CTC-Fr	0.125	0.142
CTC-NFr	0.875	0.858
Num highC	463	552
Num zeroC	228	179
highC %	0.226	0.248
ZeroC%	0.111	0.080
ZeroC Total %	0.124	0.107

As shown in Table 3, the percentage of non-cited papers in 62 mathematics journals in WoS is 12.4%. While it is much lower for First-Articles, the uncitedness rate drops to 11.1% in a whole through a period of nearly two decades. As for Scopus database, the share of papers never cited in 67 journals in mathematics decline to10.7%. In addition, the proportion of uncitedness for First-Articles stays to 8.0% on average.

Conclusion

To verify the hypothesis that the best articles selected by peer reviewers, usually the First-Articles, will be superior in receiving higher citations after publication compared with non-First-Articles published in the same journal issue, we first obtained twins data of First-Articles and non-First-Articles by retrieving articles published in top 100 (in terms of JCR 2013 JIF) mathematic journals in Scopus and WoS. Then we employed indicators CPP, CTC and TC, based on which we applied ANOVA to contrast citation bias of First-Articles and non-First-Articles in both Scopus and WoS. Results showed that there existed significant difference between First-Articles and non-First-Articles in receiving citations after publication. On the basis of these empirical grounds, we suggested that the First-Articles are biased in citations compared with non-First-Articles. We also found that it revealed a higher proportion of First-Articles to be most highly cited and comparatively lower proportion to be uncited. Furthermore, it presented a good consistency in conclusion in Scopus and WoS.

The results suggest that the peer reviewer's best recommendation go accordance with highest bibliometric indicator performance. Deliberately or not, papers received best recommendations in pre-publication evaluation process often are arranged as the First-Articles in a journal issue. The First-Articles are generally regarded as ones of high importance intense creativity or superior quality judged by peer reviewers; therefore they are expected to have a greater chance to get highly cited in the future. In fact, such understanding is supported by our analysis in this paper. After publication, those First-Articles are more likely to receive higher citations. Accordingly, peer reviewers' best recommendations and the excellent performance of journal papers measured by bibliometric indicators concordance with each other in the case of First-Articles.

Acknowledgments

We acknowledge the National Natural Science Foundation of China (NSFC Grant No.71373252) for financial support.

References

Judge A, Cable M, Colbert E. (2007). What causes a management article to be cited—article, author, or journal? *Academy of Management Journal*, 50(3), 491-506.

Wang, X., Liu, C., & Mao, W. (2014). Does a paper being featured on the cover of a journal guarantee more attention and greater impact? *Scientometrics*, 102(2), 1815-1821.

Ritter, S. K. (2006). Making The Cover. *Chemical & Engineering News*, 84(45), 24-27.

ProQuest Dissertation Analysis

Kishor Patel,[1] Sergio Govoni,[1] Ashwini Athavale,[1] Robert P. Light,[2] Katy Börner[2]

[1]*Kishor.Patel@proquest.com, Sergio.Govoni@proquest.com, Ashwini.Athavale@proquest.com*
ProQuest LLC, 7500 Old Georgetown Road, Suite 1400, Bethesda, MD 20814 (USA)

[2] *katy@indiana.edu, lightr@indiana.edu*
CNS, SOIC, Indiana University, 1320 E. Tenth Street, Bloomington, IN 47405 (USA)

Introduction

Productivity measurement has become a major issue for university leaders. Federal and state governments support teaching and research with significant investments. When university leaders are seeking new funding, it is not uncommon that they need to justify their request with productivity measurement metrics and equally important research output consumption metrics. However, it is often very difficult for university leaders to generate these metrics as they lack access to relevant data and tools to analyse and visualize large amounts of data.

Interested to address the diverse needs of university leaders, ProQuest and Indiana University analysed the ProQuest Dissertation & Theses Global (PQDT Global) database, an extensive and trusted collection of 3.8 million graduate study dissertations with 1.7 million full text records and editorially assigned metadata created by subject area experts. The database offers comprehensive North American and significant international coverage. Worldwide access to the database is logged at the dissertation level by ProQuest. Usage data mining is important for understanding user behaviour (Srivastava et al., 2000). The ProQuest Dissertations Dashboard released in 2014 provides easy access to dissertations, metadata, and usage data. It is available for free to leaders of any university that shares dissertation data with ProQuest.

ProQuest Data Analysis and Visualization

Analyses were conducted and results visualized to answer questions that seemed of particular interest to university leaders and those seeking to assess the performance of a school as a whole.

Study 1: How much attention are my school's dissertations getting?

A school's ability to generate interest in their students' dissertations may not only reflect the reputation of the school, but have long-term effects on those students' marketability and also in attracting future generations of students to join the school.

Figure 1 plots the production and access data for computer science dissertations for a selected institution given in red and labelled 'Subject University' and two groups of peer institutions rendered in green and blue. Other institutions that have published computer science dissertations are given in grey. The three institutions in the top-right corner of the plot—publishing many theses that attract many views—include both well-regarded private research institutions as well as for-profit colleges with practically open admissions. This implies that while thesis production and usage are important, they should not be used as a sole indicator for the quality of a program.

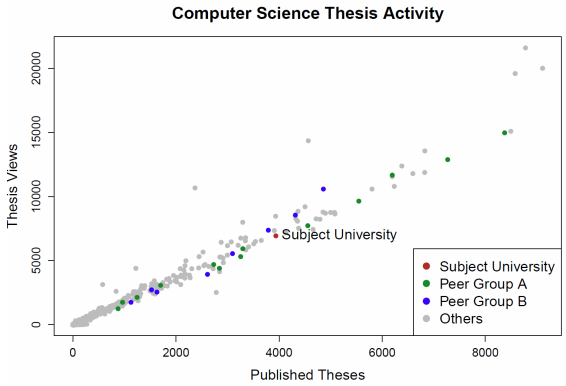


Figure 1. Comparing Subject-Area Specific Thesis Access Activity with Peer Groups.

Study 2: How can I quickly compare the number of dissertations and associated download activity for a large number of universities?

Given all dissertations or dissertations in a certain subject area, university leaders might like to understand the "market share" of an institution within a comparison or peer group.

In Figure 2, two peer groups of institutions are compared. Each institution is represented by a rectangle. Each rectangle is sized based on the total corpus of computer science dissertations available in the ProQuest dataset for that institution.

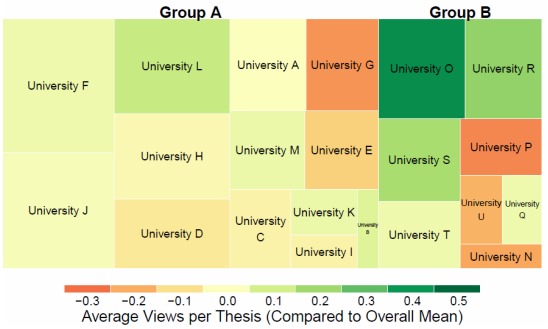


Figure 2. Treemap Comparing Thesis Production and Usage in Computer Science.

Colours tell how frequently the average dissertation at that institution is accessed in comparison to the group average. Computer science dissertations written at Universities L, O, and R are accessed more frequently than the group average, while those published at Universities G or P are accessed less.

Study 3: How is dissertation information flowing in and out of my university?

Universities are both producers and consumers of information (Mazloumian et al., 2013). Administrators are interested to understand which dissertations from which universities are used at their own institution but they also want to know who is accessing their own institution's dissertations. Plus, they might need to compare this in-flow and out-flow of information with the flows calculated for other universities.

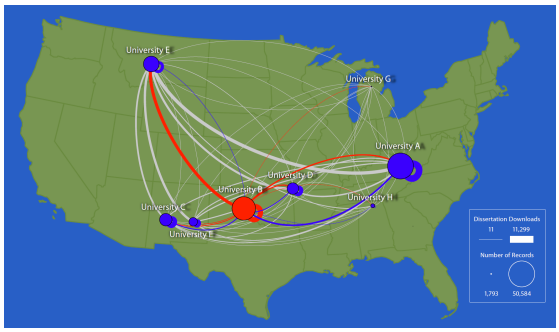


Figure 3. Information Flows within Peer Group

The example in Figure 3 looks at information flow between a group of peer schools. One institution, labelled University B, is highlighted. Red edges depict information flowing out of that institution, while blue flows show information flowing into that institution. The thicker the line, the greater is the number of dissertations. (Information always flows clockwise on the curved lines).

Future Directions

Currently, ProQuest dissertation data is not linked to publication, funding or other data. However, there is much interest in being able to study career trajectories in a more comprehensive manner (Ni & Sugimoto, 2012; Ostriker, Kuh, & Voytuk, 2011) and to examine the reputation and funding of dissertation advisors and the success (in terms of funding and publication records) of their advisees in more detail. Citation counts for dissertations, user ratings and altmetrics data, e.g., social media data, are valuable indicators of impact that we would like to explore. We also think that productivity and usage datasets can be leveraged to study the emergence of new disciplines and cross-disciplinary subject areas (Sugimoto, Li, Russell, Finlay, & Ding, 2011).

Acknowledgments

This work was partially funded by the National Institutes of Health under awards P01AG039347, U01GM098959, and U01CA198934. The authors would like to thank and acknowledge the assistance of Samuel Mills in preparing graphics for this text, Mike Gallant for information technology support as well as the ProQuest dissertations product management, development, and technical teams for their support during this research work.

References

Mazloumian, A., Helbing, D., Lozano, S., Light, R. P., & Börner, K. (2013). Global multi-level analysis of the 'Scientific Food Web'. *Scientific reports*, 3.

Ni, C., & Sugimoto, C.R. (2012). Using doctoral dissertations for a new understanding of disciplinarity and interdisciplinarity. *Proceedings of the Annual Meeting of the American Society for Information Science and Technology*. Baltimore, MD. October 26-30, 2012: ASIST.

Ostriker, J., Kuh, C., & J. Voytuk (Eds.), (2011) A Data-Based Assessment of Research-Doctorate Programs in the United States. Retrieved from: http://www.nap.edu/rdp/

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. Graph.* 11, 1 (pp. 92-99). Retrieved from http://doi.acm.org/10.1145/102377.115768

Srivastava, J., Cooley, R., Deshpande, M, & Tan, P., (2000). Web usage mining: discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations Newsletter*, 1(2), 12-23. http://doi.acm.org/10.1145/846183.846188

Sugimoto, C. R., Li, D., Russell, T. G., Finlay, S. C., & Ding, Y. (2011). The shifting sands of disciplinary development: Analyzing North American library and information science dissertations using Latent Dirichlet Allocation. *Journal of the American Society for Information Science and Technology, 62* (1), 185-204. http://onlinelibrary.wiley.com/doi/10.1002/asi.21435/abstract

ISSI

2015

INDICATORS

An Alternative to Field-normalization in the Aggregation of Heterogeneous Scientific Fields

Antonio Perianes-Rodriguez[1] and Javier Ruiz-Castillo[2]

[1] *antonio.perianes@uc3m.es*
Universidad Carlos III, Department of Library and Information Science, SCImago Research Group, C/ Madrid, 128, 28903 Getafe, Madrid (Spain)

[2] *jrc@eco.uc3m.es*
Universidad Carlos III, Departamento de Economía, C/ Madrid, 126, 28903 Getafe, Madrid (Spain)

Abstract
A possible solution to the problem of aggregating heterogeneous fields in the all-sciences case relies on the normalization of the raw citations received by all publications. In this paper, we study an alternative solution that does not require any citation normalization. Provided one uses size- and scale-independent indicators, the citation impact of any research unit can be calculated as the average (weighted by the publication output) of the citation impact that the unit achieves in all fields. The two alternatives are confronted when the research output of the 500 universities in the 2013 edition of the CWTS Leiden Ranking is evaluated using two citation impact indicators with very different properties. We use a large Web of Science dataset consisting of 3.6 million articles published in the 2005-2008 period, and a classification system distinguishing between 5,119 clusters. The main two findings are as follows. Firstly, differences in production and citation practices between the 3,332 clusters with more than 250 publications account for 22.5% of the overall citation inequality. After the standard field-normalization procedure where cluster mean citations are used as normalization factors, this figure is reduced to 4.3%. Secondly, the differences between the university rankings according to the two solutions for the all-sciences aggregation problem are of a small order of magnitude for both citation impact indicators.

Conference Topic
Indicators; Citation and co-citation analysis

Introduction
As is well known, the comparison of the citation impact of research units is plagued with obstacles of all sorts. For our purposes in this paper, it is useful to distinguish between the following three basic difficulties. (i) How can we compare the citation distributions of research units of different sizes even if they work in the same homogeneous scientific field? For example, how can we compare the output of the large Economics department at Harvard University with the output of the relatively small Economics department at Johns Hopkins? The next two difficulties have to do with the heterogeneity of scientific fields: the well-known differences in production and citation practices makes it impossible to directly compare the raw citations received by articles belonging to different fields. Given a classification system, that is, a rule for assigning any set of articles to a number of scientific fields, field heterogeneity presents the following classic hindrances in the evaluation of research units' performance. (ii) How can we compare the citation impact of two research units working in different fields? For example, how can we compare the citation impact of MIT in Organic Chemistry with the citation impact of Oxford University in Statistics and Probability? Finally, (iii) how can we compare the citation impact of two research units taking into account their

output in all fields? For example, how can we compare the citation impact of MIT and Oxford University in what we call the *all-sciences* case?

As is well known, the solution to the first two problems requires size- and scale-independent citation impact indicators. We will refer to indicators with these two properties as *admissible* indicators. Given an admissible indicator, in this paper we are concerned with the two types of solutions that the third problem admits. Firstly, the problem can be solved in two steps. One first uses some sort of normalization procedure to make the citations of articles in all fields at least approximately comparable. Then, one applies the citation indicator to each unit's normalized citation distribution. Secondly, consider the Top 10% indicator used in the construction of the influential Leiden and *SCImago* rankings. In the Leiden Ranking this indicator is defined as "*The proportion of publications of a university that, compared with other similar publications, belong to the top 10% most frequently cited…Publications are considered similar if they were published in the same field and the same publication and if they have the same document type*" (Waltman et al., 2012a). A similar definition is applied in the SCImago ranking (Bornmann et al., 2012) Note that this way of computing this particular indicator in the all-sciences case does not require any kind of prior citation normalization. For our purposes, it is useful to view this procedure as the average (weighted by the publication output) of the unit's Top 10% performance in each field. We note that this important precedent can be extended to *any* admissible indicator. Thus, given a classification system and an admissible citation indicator, we can compute the citation impact of a research unit in the all-sciences case as the appropriate weighted average of the unit's citation impact in each field. Independently of the conceptual interest of this proposal, we must compare the consequences of adopting it versus the possibility of following a normalization procedure.

Intuitively, the better the performance of the normalization procedure in eliminating the comparability difficulties across fields, the smaller will be the differences between the two approaches. Consider, for example, what we call the standard field-normalization procedure in which the normalized citations of articles in any field are equal to the articles' original raw citations divided by the field mean citation. Under the universality condition, that is, if field citation distributions were identical except for a scale factor, then the standard field-normalization procedure would completely eliminate all comparability difficulties. However, the universality condition, once claimed to be the case (Radicchi et al., 2008), is not usually satisfied in practice: even appropriately normalized, field citation distributions are seen to be significantly different from a statistical point of view (Albarrán et al., 2011a; and Waltman et al., 2012a). Therefore, at best, normalization procedures provide an approximate solution to the original comparability problem.

Using a measuring framework introduced in Crespo et al. (2013), recent research has established that different normalization procedures perform quite well in eliminating most of the effect in overall citation inequality that can be attributed to differences in production and citation practices between fields. This is the case for large Web of Science (WoS hereafter) datasets, classification systems at different aggregation levels, and different citation windows (Crespo et al., 2013, 2014; Li et al., 2013; Waltman & Van Eck, 2013; Ruiz-Castillo, 2014). The reason for the good performance of target (or cited-side) normalization procedures is that field citation distributions, although not universal, are extremely similar (Glänzel, 2007; Radicchi et al., 2008; Albarrán & Ruiz-Castillo, 2011; Albarrán et al., 2012; Waltman et al., 2012a; Radicci & Castellano, 2012; Li et al., 2013). It should be noted that this research on target normalization procedures uses WoS classification systems distinguishing at most between 235 sub-fields.

In principle, given the good performance of normalization procedures, we expect that the differences between the two approaches would be of a small order of magnitude.

Nevertheless, this is an empirical question that has never been investigated before. To confront this question, in this paper we conduct the following exercise.

- Ruiz-Castillo & Waltman (2015) apply the publication-level algorithmic methodology introduced by Waltman and Van Eck (2012) to a WoS hereafter dataset consisting of 9.4 million publications from the 2003-2012 period. This is done along a sequence of twelve independent classification systems in each of which the same set of publications is assigned to an increasing number of clusters. In this paper, we use the classification system recommended in Ruiz-Castillo and Waltman (2015), consisting of 5,119 clusters, of which 4,161 are referred to as significant clusters because they have more than 100 publications over this period. For the evaluation of research units' citation impact, we focus on the 3.6 million publications in the 2005-2008 period, and the citations they receive during a five-year citation window for each year in that period. It should be noted that, using the size- and scale-independent technique known as Characteristic Scores and Scales, Ruiz-Castillo and Waltman (2015) show that, as in previous research, significant clusters are highly skewed and similarly distributed.
- Our research units are the 500 universities in the 2013 edition of the CWTS Leiden Ranking (Waltman et al., 2012b). We analyze the approximately 2.4 million articles – about 67% of the total– for which at least one author belongs to one of these universities. We use a fractional counting approach to solve the problem –present in all classification systems– of the assignment of responsibility for publications with several co-authors working in different institutions. The total number of articles corresponding to the 500 universities is approximately 1.9 million articles –about 50% of the total.
- We evaluate the citation impact of each university using two size- and scale-independent indicators. Firstly, we use the Top 10% indicator, already mentioned. Secondly, one characteristic of this indicator is that it is not monotonic in the sense that it is invariant to any additional citation that a high-impact article might receive. Consequently, we believe that it is interesting to use a second indicator possessing this property. In particular, we select a member of the Foster, Greer, and Thorbecke (FGT hereafter) family, introduced in Albarrán et al. (2011b). We apply this indicator to the set of high-impact articles mentioned before. As will be seen below, the fact that both of our indicators are additively decomposable facilitates the comparability of the two solutions to the all-sciences aggregation problem.
- Using Crespo et al.'s (2013) measurement framework, Li et al. (2013) indicate that the best alternative among a wide set of target normalization procedures is the two-parameter system developed in Radicci and Castellano (2012). However, recent results indicate that the standard, one-parameter field-normalization procedure exhibits a good performance in reducing the effects on overall citation inequality attributed to differences in production and citation practices between fields (Radicchi et al., 2008; Crespo et al., 2013, 2014; Li et al., 2013; and Ruiz-Castillo, 2014). Consequently, in this paper we adopt this procedure in the usual solution to the all-sciences aggregation problem.
- We present two types of results. Firstly, we assess the performance of the standard normalization procedure in facilitating the comparability of the citations received by articles belonging to different clusters. Secondly, we assess the consequences of adopting the two solutions to the all-sciences aggregation problem by comparing the corresponding university rankings according to the two citation impact indicators.

The rest of the paper is organized into three sections. Section II presents the citation impact indicators, as well as the two solutions to the all-sciences aggregation problem. Section III describes the data, and includes the empirical results, while Section IV concludes.

The aggregation of heterogeneous scientific fields in the all-sciences case

Notation and citation indicators

It is convenient to introduce some notation. Given a set of articles S, and J scientific fields indexed by $j = 1,\ldots, J$, a *classification system* is an assignment of articles in S to the J fields. Let I be the number of research units, indexed by $i = 1,\ldots, I$. In this Section, the assignment of articles in S to the I research units is taken as given. Let $\boldsymbol{c_{ij}} = \{c_{ijk}\}$ be the *citation distribution of unit i in field j*, where c_{ijk} is the number of citations received by the k-th article, and let $\boldsymbol{c_j}$ be the *citation distribution of field j*, that is, the union of all research units' citation distributions in that field: $\boldsymbol{c_j} = \cup_i \{\boldsymbol{c_{ij}}\}$. Finally, let $\boldsymbol{C} = \cup_i \cup_j \{\boldsymbol{c_{ij}}\}$ be the *overall citation distribution*, or the citation distribution in the all-sciences case. For later reference, let N_{ij} be the number of articles in distribution $\boldsymbol{c_{ij}}$, let $N_i = S_j N_{ij}$ be the total number of articles published by unit i, let $N_j = S_i N_{ij}$ be the total number of articles in field j, and let $N = S_i S_j N_{ij}$ be the total number of articles in the all-sciences case.

A *citation impact indicator* is a function F defined in the set of all citation distributions, where $F(\boldsymbol{c})$ is the citation impact of distribution $\boldsymbol{c}$. Let $\boldsymbol{c^r}$ be the r-th replica of distribution $\boldsymbol{c}$. An indicator F is said to be *size-independent* if, for any citation distribution $\boldsymbol{c}$, $F(\boldsymbol{c^r}) = F(\boldsymbol{c})$ for all r. An indicator F is said to be *scale-independent* if for any $\lambda > 0$, and any citation distribution $\boldsymbol{c}$, $F(\lambda\boldsymbol{c}) = F(\boldsymbol{c})$. An indicator F is said to be *additively decomposable* if for any partition of a citation distribution $\boldsymbol{c}$ into G sub-groups, indexed by $g = 1,..., G$, the citation impact of distribution $\boldsymbol{c}$ can be expressed as follows:

$$F(\boldsymbol{c}) = S_g\, (M_g/M) F(\boldsymbol{c_g}),$$

where M_g is the number of publications in sub-group g, and $M = \Sigma_g\, M$ is the number of publications in distribution $\boldsymbol{c}$.

Consider the following two difficulties for comparing the citation impact of any pair of research units: the two units may be of different sizes, and if they work in different fields, then their raw citations are not directly comparable. As it is well known, these two difficulties can be overcome using a size- and scale-independent indicator. The following two indicators are good examples of size- and scale-independent indicators that, in addition, are additively decomposable.

1. Let $\boldsymbol{X_j}$ be the set of the 10% most cited articles in citation distribution $\boldsymbol{c_j}$, and let $\boldsymbol{x_{ij}}$ be the sub-set of articles in $\boldsymbol{X_j}$ corresponding to unit i, so that $\boldsymbol{X_j} = \cup_i\{\boldsymbol{x_{ij}}\}$ with $\boldsymbol{x_{ij}}$ non-empty for some i. If n_{ij} is the number of articles in $\boldsymbol{x_{ij}}$, then the *Top 10% indicator for unit i in field j*, T_{ij}, is defined as

$$T_{ij} = n_{ij}/N_{ij}. \qquad (1)$$

Of course, for field j as a whole, if $n_j = \Sigma_i\, n_{ij}$ is the number of articles in $\boldsymbol{X_j}$, then $T_j = n_j/N_j = 0.10$.

2. Let z_j be the *Critical Citation Line* –CCL hereafter– for citation distribution $\boldsymbol{c_j}$, and denote the articles in $\boldsymbol{c_j}$ with citations equal to or greater than z_j as *high-impact articles*. For any high impact article with citations c_{il}, define the *CCL normalized high-impact gap* as $(c_{il} - z_j)/z_j$.

Consider the family of FGT indicators introduced in Albarrán et al. (2011b) as functions of normalized high-impact gaps. The second member of this family, A_{ij}, is defined as

$$A_{ij} = (1/N_{ij})[S_l(c_{il} - z_j)/z_j], \quad (2)$$

where the sum is over the high-impact articles in citation distribution $\boldsymbol{c_j}$ that belong to unit i. We refer to this indicator as the *Average of high-impact gaps for unit i in field j*. For the entire field j as a whole, the average of high-impact gaps is defined as

$$A_j = (1/N_j)[S_k (c_k - z_j)/z_j],$$

where the sum is over the high-impact articles in citation distribution $\boldsymbol{c_j}$.

To facilitate the comparison with T_{ij}, in the sequel we will always fix z_j as the number of citations of the article in the 90th percentile of citation distribution $\boldsymbol{c_j}$. In that case, the set of high-impact articles coincides with the set of the 10% most cited articles in citation distribution $\boldsymbol{c_j}$. The two main differences between the two indicators are the following. Firstly, one or more citations received by a high-impact article increases A_{ij} but does not change T_{ij}. In other words, A_{ij} is monotonic but T_{ij} is not. Secondly, T_{ij} is more robust to extreme observations than A_{ij}.

The solution to the all-sciences aggregation problem using the standard field-normalization procedure

For any i, let $\boldsymbol{c_i} = (\boldsymbol{c_{i1}},\ldots, \boldsymbol{c_{ij}},\ldots, \boldsymbol{c_{iJ}})$ be the *raw citation distribution of unit i in the all-sciences case*. Differences in production and citation practices across fields make impossible the direct comparison of the raw citations received by articles in different fields. In order to achieve some comparability, one possibility is to use some normalization procedure. For any article k in citation distribution $\boldsymbol{c_{ij}}$, the normalized number of citations c^*_{ijk} according to the *standard field-normalization procedure* is defined as

$$c^*_{ijk} = c_{ijk}/\mu_j.$$

For any i, let $\boldsymbol{c^*_i} = \cup_j \cup_k \{c^*_{ijk}\} = (\boldsymbol{c^*_{i1}},\ldots, \boldsymbol{c^*_{ij}},\ldots, \boldsymbol{c^*_{iJ}})$ be the *normalized citation distribution of unit i in the all-sciences case*. Since normalized citations are now comparable, it makes sense to apply any indicator to citation distribution $\boldsymbol{c^*_i}$. For any i, let $\boldsymbol{F^*_i} = F(\boldsymbol{c^*_i})$ be the citation impact of distribution $\boldsymbol{c^*_i}$ according to the indicator F. For any pair of research units u and v in the all-sciences case, the citation impact values $\boldsymbol{F^*_u}$ and $\boldsymbol{F^*_v}$ are now comparable, and can be used to rank the two units in question.

Note that, since $\boldsymbol{c^*_i}$ for $i = 1,\ldots, I$ forms a partition of $\boldsymbol{C^*}$ and F is assumed to be additively decomposable, we can write

$$F^* = F(\boldsymbol{C^*}) = S_i (N_i/N)\boldsymbol{F^*_i}.$$

Thus, if we rank universities by the ratio $\boldsymbol{F^*_i}/F^*$, $i = 1,\ldots, I$, then the value one can serve as a benchmark for evaluating the research units in the usual way. For later reference, since $\boldsymbol{c^*_{ij}}$ for $j = 1,\ldots, J$ forms a partition of $\boldsymbol{c^*_i}$, for each i we can write

$$\boldsymbol{F^*_i} = F(\boldsymbol{c^*_i}) = S_j (N_{ij}/N_i)\boldsymbol{F^*_{ij}}, \quad (3)$$

where $F^*_{ij} = F(\boldsymbol{c}^*_{ij})$ for all j, that is, F^*_{ij} is simply the citation impact of citation distribution $\boldsymbol{c}^*_{ij}$ according to F.

A solution to the all-sciences aggregation problem without field-normalization

For any i and any j, denote by $F_{ij} = F(\boldsymbol{c}_{ij})$ the citation impact of distribution $\boldsymbol{c}_{ij}$ according to F. A convenient measure of citation impact for unit i in the all-sciences case, F_i, can be defined as the weighted average of the values F_{ij} achieved in all fields, with weights equal to the relative importance of each field in the total production of unit i:

$$F_i = S_j\,(N_{ij}/N_i)F_{ij} \qquad (4)$$

The comparison of expressions (4) and (5) illustrate the differences between the two solutions to the all-sciences aggregation problem when the evaluation of the units' citation impact is made with additively decomposable indicators. Finally, it is convenient to compute the weighted average of these quantities as follows:

$$F = S_i\,(N_i/N)F_i.$$

Thus, as before, if we rank universities by the ratio F_i/F, $i = 1,\ldots, I$, then the value one can serve as a benchmark for evaluating the research units in the usual way. In practice, we have information concerning some but not all research units. Therefore, we compute F as the following weighted average: $F = \mathrm{S}_j\,(N_j/N)F_j$, where $F_j = \mathrm{F}(\boldsymbol{c}_j)$.

The aim of the paper

The main aim of this paper is the comparison between the rankings of research units obtained with and without the standard field-normalization procedure, $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$, respectively.

To understand the way the results will be presented, we need to review the connection between the performance of the normalization procedure and the relationship between the solutions to the all-sciences aggregation problem. For that purpose, we need to introduce some more notation. For any j, let $\boldsymbol{x}_j$ be the set of high-impact articles in distribution $\boldsymbol{c}_j$, that is, the set of articles in $\boldsymbol{c}_j$ with citations equal to or greater than z_j, or the set of the 10% most cited articles in $\boldsymbol{c}_j$. Let us denote by $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots, \boldsymbol{x}_j,\ldots, \boldsymbol{x}_J)$ the set of high-impact articles in the all-sciences case. On the other hand, let $\boldsymbol{Y}$ be the set of the 10% most cited articles in the overall normalized citation distribution $\boldsymbol{C}^* = \cup_j \{\boldsymbol{c}^*_j\}$. Let $\boldsymbol{y}_j$ be the sub-set of articles in $\boldsymbol{Y}$ belonging to field j, so that $\boldsymbol{Y} = (\boldsymbol{y}_1,\ldots, \boldsymbol{y}_j,\ldots, \boldsymbol{y}_J)$. Note that, in practice, the sets $\boldsymbol{y}_j$ might be empty for some j.

Under the universality condition, that is, if all fields are equally distributed except for a scale factor then, at every percentile of field citation distributions, normalized citations will be the same for all fields. In other words, the normalization procedure will work perfectly. In particular, in this situation we would have $z_j/\mu_j = z^*$ for all j. Consequently, we would have $\boldsymbol{y}_j = \boldsymbol{x}_j$ for all j, and $\boldsymbol{Y} = \boldsymbol{X}$. Since citation distributions $\boldsymbol{c}^*_{ij}$ and $\boldsymbol{c}_{ij}$ have the same number of articles and our indicators are a function solely of high-impact articles, we would have $F^*_{ij} = F(\boldsymbol{c}^*_{ij}) = F_{ij} = F(\boldsymbol{c}_{ij})$ for all i and j. In view of equations (4) and (5), we would have $F^*_i = F_i$ for all i. In other words, the rankings $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$ will be identical.

As we know, in practice the universality condition is not satisfied. However, the better the performance of the normalization procedure, that is, the closer is the set $\boldsymbol{Y}$ to set $\boldsymbol{X}$, the more similar the rankings $(F^*_1, \ldots, F^*_I)$ and $(F_1, \ldots, F_I)$ are expected to be for any F. Note that this

conjecture has to be verified in practice. In any case, the empirical section begins by assessing the performance of the normalization procedure.

On the other hand, independently of the normalization procedure's performance, we should measure the consequences of adopting the two solutions to the all-sciences aggregation problem using indicators with different properties. The reason, of course is that whenever $\boldsymbol{Y}$ and $\boldsymbol{X}$ differ, that is, when the set of high-impact articles under the two solutions differ, the consequences for the university rankings might be of a different order of magnitude depending on the citation impact indicator we use. This is the reason why we will study the situation using the Top 10% and the Average of high-impact gaps.

Empirical results

The data and descriptive statistics

As indicated in the Introduction, our dataset results from the application of a publication-level methodology to 9,446,622 distinct articles published in 2003-2012 (see Ruiz-Castillo & Waltman, 2015). Publications in local journals, as well as popular magazines and trade journals have been excluded (see Ruiz-Castillo & Waltman, 2015 for details). We work with journals in the sciences, the social sciences, and the arts and humanities, although many arts and humanities journals are excluded because they are of a local nature. The classification system consists of 5,119 clusters, and citation distributions refer to the citations received by these articles during a five-year citation window for each year in that period. In this paper, we focus on the set of 3,614,447 distinct articles published in 2005-2008. In terms of the notation introduced in Section II.1, we have $\boldsymbol{C} = \cup_j \{\boldsymbol{c_j}\} = (c_1,\ldots, c_N)$ with J = 5,119, and N = 3,614,447.

The research units are universities. Publications are assigned to universities using the fractional counting method that takes into account the address lines appearing in each publication. An article is fully assigned to a university only if all addresses mentioned in the publication belong to the university in question. If a publication is co-authored by two or more universities, then it is assigned fractionally to all of them in proportion to the number of address lines. For example, if the address list of an article contains five addresses and two of them belong to a particular university, then 0.4 of the article is assigned to this university, and only 0.2 of the article is assigned to each of the other three universities.

We know the total number of address lines of every publication, but we have information about the number of address lines of specific institutions only for the 500 LR universities. This number is well below I, the total number of research units in the notation introduced in Section II.1. There are 2,420,054 distinct articles, or 67% of the total, with at least one address line belonging to a LR university. The total number of articles in the LR universities according to the fractional counting method is 1,886,106.1, or 52.2% of the total. The distribution of this total among the 500 universities is available in Perianes-Rodriguez & Ruiz-Castillo, 2014a.

The performance of the normalization procedure

We assess the performance of the normalization procedure using the measurement framework introduced in Crespo et al. (2013), we first estimate the effect on overall citation inequality attributable to differences in production and citation practices between clusters, and then the reduction in this effect after applying the standard field-normalization procedure. Given the many clusters with very few publications (see Ruiz-Castillo & Waltman, 2015), we apply this method to the 3,332 clusters with more than 250 publications. These clusters include 3,441,666 million publications, or 95.2% of the total.

We begin with the partition of, say, each cluster citation distribution into P quantiles, indexed by $p = 1,.., P$. In practice, in this paper we use the partition into percentiles, that is, we choose $P = 100$. Assume for a moment that, in any cluster i, we disregard the citation inequality within every percentile by assigning to every article in that percentile the mean citation of the percentile itself, μ_i^p. The interpretation of the fact that, for example, $\mu_i^p = 2\ \mu_j^p$ is that, on average, the citation impact of cluster i is twice as large as the citation impact of cluster j in spite of the fact that both quantities represent a common underlying phenomenon, namely, the same *degree of citation impact* in both clusters. In other words, for any π, the distance between μ^p and μ_j^p is entirely attributable to the difference in the production and citation practices that prevail in the two clusters for publications with the same degree of excellence in each of them. Thus, the citation inequality between clusters at each percentile, denoted by $I(p)$, is entirely attributable to the differences in citation practices between the 3,332 clusters holding constant the degree of excellence in all clusters at quantile π. Hence, any weighted average of these quantities, denoted by *IDCC* (*I*nequality due to *D*ifferences in *C*itation impact between *C*lusters), provides a good measure of the total impact on overall citation inequality that can be attributed to such differences. Let $\boldsymbol{C'}$ be the union of the clusters citation distributions, $\boldsymbol{C'} = \cup \{\boldsymbol{c_j}\}$ for $j = 1,\ldots, 3{,}332$. We use the ratio

$$IDCC/I(\boldsymbol{C'}) \qquad (6)$$

to assess the relative effect on overall citation inequality, $I(\boldsymbol{C'})$, attributed to the differences in citation practices between clusters (for details, see Crespo et al., 2013).

Finally, we are interested in estimating how important scale differences between cluster citation distributions are in accounting for the effect measured by expression (6). For that purpose, we use the relative change in the *IDPC* term, that is, the ratio

$$[IDCC - IDCC^*]/IDCC, \qquad (7)$$

where *IDCC** is the term that measures the effect on overall citation inequality attributed to the differences in cluster distributions after applying the standard field-normalization procedure (for details, see again Crespo et al., 2013). The estimates of expressions (6) and (7) are as follows:

Table 1. The effect on overall citation inequality, I('C), of the differences in citation impact between clusters before and after standard field-normalization, and the impact of normalization on this effect.

		Normalization impact =100 [*IDCC – IDCC*/IDCC*]
Before MNCS normalization, 100 [*IDCC/I(C')*]	22.5 %	-
After MNCS normalization, 100 [*IDCC*/I(C')*]	4.3 %	84.3 %

It can be observed that the effect of the differences in citation practices between such a large number of clusters represents 22.5% of overall citation inequality, a figure much larger than what has been found in the previous literature for at most 235 sub-fields. Nevertheless, the standard field-normalization procedure reduces this effect down to 4.3%, quite an achievement.

Differences in university rankings under the two solutions to all-sciences aggregation problem

The university rankings without and with normalization according to the Top 10% indicator, T_i and T^*_i, and according to the Average of high-impact gaps, A_i and A^*_i can be found in Perianes-Rodriguez & Ruiz-Castillo (2014a). We begin with the comparison of university rankings according to T_i and T^*_i. The Pearson correlation coefficient between university values is 0.995, while the Spearman correlation coefficient between ranks is 0.992. However, high correlations between university values and ranks do not preclude important differences for individual universities. In analyzing the consequences of going from T_i to T^*_i, we must take two aspects into account. Firstly, we should analyze the re-rankings that take place in such a move. Secondly, we should compare the differences between the university values themselves. Fortunately, we have a relevant instance with which to compare our results: the differences found in Ruiz-Castillo and Waltman (2015) in going from the university rankings according to T_i using the Web of Science classification system with 236 journal subject categories, or sub-fields, and the classification system we are using in this paper with 5,119 clusters.

As much as 38.4% of universities experience very small re-rankings of less than or equal to five positions, while 67 universities, or 13.4% of the total, experience re-rankings greater than 25 positions. These figures are 20.2% and 39.0% when going from the WoS classification system to our dataset. Among the first 100 universities, 61 experience small re-rankings in going from T_i to T^*_i, while only 44 are in this situation in the change between classification systems. As far as the cardinal changes is concerned, 78.4% of universities have changes in top 10% indicator values smaller than or equal to 0.05 when going from T_i to T^*_i. This percentage is 71% among the first 100 universities. These figures are 50.1% and 60.0% in the change between classification systems. For most universities, the differences are more or less negligible. Although for some universities more significant differences can be observed, the conclusion is clear. The differences observed in university rankings according to the top 10% indicator when we adopt the two solutions for solving the all-sciences aggregation problem are considerably less than according to the same indicator when we move from the WoS classification system to our dataset (Perianes-Rodriguez & Ruiz-Castillo, 2014a).

The Pearson correlation coefficient between the university rankings according to the average of high-impact gaps, A_i and A^*_i, is 0.596, while the Spearman correlation coefficient between ranks is 0.984. However, the low Pearson correlation coefficient is due to the presence of the well-known extreme observation of the University of Göttingen (Waltman et al., 2012b; Ruiz-Castillo & Waltman, 2015). Without this university, this correlation coefficient becomes 0.986. In any case, as before, high correlations between university values and ranks do not preclude important differences for individual universities. The ordinal differences in university rankings according to this indicator with and without field-normalization are of a similar order of magnitude as those obtained with the top 10% indicator. For example, 33.0% of universities experience very small re-rankings of less than or equal to five positions, while 80 universities, or 16.0% of the total, experience re-rankings greater than 25 positions. Among the first 100 universities, only 44 experience small re-rankings in going from A_i to A^*_i (in comparison with 61 when going from T_i to T^*_i). As far as the cardinal changes is concerned, 64.2% of universities have changes in indicator values smaller than or equal to 0.05 when going from A_i to A^*_i –a comparable figure with 78.4% when going from T_i to T^*_I (Perianes-Rodriguez & Ruiz-Castillo, 2014a).

The conclusion is inescapable. In spite of the fact of the limitations of the standard normalization procedure in the presence of so many clusters, the differences observed in university rankings when we adopt the two solutions for solving the all-sciences aggregation problem are of a relatively small order of magnitude regardless of which of then two rather different citation impact indicators is used in obtaining the university rankings.

Conclusions

The heterogeneity of the fields distinguished in any classification system poses a severe aggregation problem when one is interested in evaluating the citation impact of a set of research units in the all-sciences case. In this paper, we have analyzed two possible solutions to this problem. The first solution relies on prior normalization of the raw citations received by all publications. In particular, we focus on the standard field-normalization procedure in which field mean citations are used as normalization factors. The second solution extends the approach adopted in the Leiden and SCImago rankings for computing the Top 10% indicator in the all-sciences case to any admissible indicator. This solution does not require any prior field-normalization: the citation impact of any research unit in the all-sciences case is calculated as the appropriately weighted sum of the citation impact that the unit achieves in each field.

Using a large WoS dataset consisting of 3.6 million publications in the 2005-2008 period and an algorithmically constructed publication-level classification system that distinguishes between 5,119 clusters, this simple alternative has been confronted with the usual one when the citation impact of the 500 LR universities are evaluated using two indicators with very different properties: the top 10% indicator, and the average of high-impact gaps.

The shape of the citation distributions of 4,161 significant clusters with more than 100 publications in our dataset has been previously shown to be highly skewed and reasonable similar (Ruiz-Castillo & Waltman, 2015). Previous results with WoS classification systems that distinguishes at most between 235 sub-fields indicate that, when this is the case, the standard field-normalization procedure performs well in reducing the overall citation inequality attributed to the differences in production and citation practices between fields. In this paper we have shown that this is not exactly the case, even when we restrict the attention to 3,332 clusters with more than 250 publications. Therefore, a priori it was not obvious what to expect when confronting the solutions to the all-sciences aggregation problem with and without prior field-normalization.

Interestingly enough, the differences between the university rankings obtained with both solutions is of a relatively small order of magnitude independently of the citation impact indicator used in the construction of the university rankings. In particular, these differences are considerably smaller than the ones obtained in Ruiz-Castillo and Waltman (2015) for the move from the WoS classification system with 236 sub-fields to the one used in this paper with 5,119 clusters.

In principle, it seems preferable to evaluate the citation impact of research units in the all-sciences case avoiding any kind of prior normalization operation. However, the empirical evidence presented in this paper indicates that that the use of the traditional methodology does not lead to very different results. This is a convenient conclusion, since there are instances when normalization is strongly advisable. For example, when one is interested in studying the research units citation distributions in the all-sciences case –as we do in the companion paper Perianes-Rodriguez and Ruiz-Castillo (2014b).

It should be noted that, before being accepted, it would be convenient to replicate the results of this paper for other datasets, other classification systems, other types of research units, and other ways of assigning responsibility between research units in the case of co-authored publications.

References

Albarrán, P., & Ruiz-Castillo, J. (2011). References-made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology*, *62*, 40–49.

Albarrán, P., Crespo, J., Ortuño, I., & Ruiz-Castillo, J. (2011a). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*, 385–397.

Albarrán, P., Ortuño, I., & Ruiz-Castillo, J. (2011b). The measurement of low- and high-impact in citation distributions: technical results. *Journal of Informetrics*, *5*, 48–63.

Bornmann, L., De Moya Anegón, F., & Leydesdorff, L. (2012) The new excellence indicator in the world report of the SCImago Institutions Rankings 2011. *Journal of Informetrics*, *6*, 333-335.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields. *PLoS ONE*, *8*, e58727.

Crespo, J. A., Herranz, N., Li, Y., & Ruiz-Castillo, J. (2014). The effect on citation inequality of differences in citation practices at the Web of Science subject category level. *Journal of the Association for Information Science and Technology*, *65*, 1244–1256.

Glänzel, W. (2007). Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, *1*, 92–102.

Li, Y., Castellano, C., Radicchi, F., & Ruiz-Castillo, J. (2013). Quantitative evaluation of alternative field normalization procedures. *Journal of Informetrics*, *7*, 746–755.

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014a). *An alternative to field-normalization in the aggregation of heterogeneous scientific fields*. Working Paper, Economic Series 14-25, Universidad Carlos III (http://hdl.handle.net/10016/19812).

Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2014b). *University citation distributions*. Working Paper, Economic Series 14-26, Universidad Carlos III (http://hdl.handle.net/10016/19811).

Radicchi, F., & Castellano, C. (2012). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*, e33833.

Radicchi, F., Fortunato, S., & Castellano, C. (2008), "Universality of citation distributions: Toward an objective measure of scientific impact", *PNAS*, *105*, 17268-17272.

Ruiz-Castillo, J. (2014). The comparison of classification-system-based normalization procedures with source normalization alternatives in Waltman and Van Eck. *Journal of Informetrics*, *8*, 25–28.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*, 102-117. (DOI: 10.1016/j.joi.2014.11.010).

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*, 2378–2392.

Waltman, L., Van Eck, N. J., & Van Raan, A. F. J. (2012a). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*, 72–77.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., Van Eck, N. J., Van Leeuwen, T. N., Van Raan, A. F. J., Visser, M. S., & Wouters, P. (2012b). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*, 2419–2432.

Waltman, L., & Van Eck, N. J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, *7*, 833–849.

Correlating Libcitations and Citations in the Humanities with WorldCat and Scopus Data

Alesia Zuccala[1] and Howard D. White[2]

[1] *spl465@iva.ku.dk*
Royal School of Library and Information Science, University of Copenhagen
Birketinget 6, DK-2300 Copenhagen S (Denmark)

[2] *whitehd@drexel.edu*
College of Computing and Informatics, Drexel University
32nd and Chestnut Streets, Philadelphia, PA, 19104 (USA)

Abstract
The term *libcitations* was introduced by White et al. (2009) as a name for counts of libraries that have acquired a given book. Somewhat like citations, these library holdings counts, which vary greatly, can be taken as indicators of the book's cultural impact. Torres-Salinas and Moed (2009) independently proposed the same measure under the name *catalog inclusions*. Both articles sought an altmetric for authors of books in, e.g., the humanities, since the major citation indexes, oriented toward scientific papers, have not served them well. Here, using very large samples, we explore the libcitation-citation relationship for the same books by correlating their holdings counts from OCLC's WorldCat with their citation counts from Elsevier's Scopus. For books cited in two broad fields of the humanities during 1996-2000 and 2007-2011, we obtain positive, weak, but highly significant correlations. These largely persist when books are divided by main Dewey class. The overall results are inconclusive, however, because the Scopus citation counts for the books tend to be very low. Further correlational research should probably use the much higher book citation counts from Google Scholar. Nevertheless, a qualitative analysis of widely held and widely cited books clarifies the libcitation measure and helps to justify it.

Conference Topic
Indicators

Introduction
Journal-oriented scientists have long had citation counts as an indicator of the impact of their articles, and journal-based citation indexes cater to them. But the same indexes cover citations to books less well, and book-oriented scholars in the humanities and softer social sciences feel themselves at a disadvantage, especially if citation measures are going to be used in performance evaluations and funding decisions (see Kousha, Thelwall, & Rezaie 2011 for a review). White et al. (2009) responded to this lack by proposing that one measure of a book's cultural impact could be the number of libraries that hold it. The idea behind this altmetric was that librarians who acquire a book are somewhat like scholars who cite it, in that both acts involve assessment and choice on behalf of communities of readers. To bring out the parallel, White et al. called the librarians' formal act of acquisition a *libcitation* (first syllable as in "library"). They wrote that the libcitation count (also known as a library holdings count) for a particular book "increases by 1 every time a different library reports acquiring that book in a national or an international union catalog. Readers are invited to think of union catalogs in a new way: as 'librarians' citation indexes'" (p. 1084). OCLC's WorldCat was mentioned as a prime example of a union catalog—that is, one that pools the cataloging records of OCLC member libraries and reports how many of them hold each cataloged item.

At the same time and wholly independently, Torres-Salinas and Moed (2009) made an identical proposal. Their name for libcitations (our term here) was *catalog inclusions*, and they, too, stressed the parallel between such inclusions and citations to journal articles (p. 11). They, too, named WorldCat as a potential source of library holdings data. Moreover, both

they and White et al. raised the possibility of empirically testing the relationship between libcitation counts and citation counts for the same set of books: are the two correlated?

The question is important because citation counts, when scrupulously used, have become a standard performance indicator in many disciplines, and, given the inadequacies of citation data for books, it would be very interesting if libcitations could serve a similar purpose. Torres-Salinas and Moed (2009, p. 24) saw correlation research of this sort in terms of validating the holdings-count idea:

> One way of doing this is to examine...the degree of correlation between the number of times book titles are cited in the serial literature on the one hand, and the number of library catalogs in which they are included on the other.

That is just what the present paper does for books (aka titles) in two broad fields in the humanities: *History* and *Literature & Literary Theory*. It draws on a special database of book citation data from Elsevier's Scopus and libcitation data for the same books from WorldCat, as described in Zuccala and Guns (2013), a research-in-progress paper. White et al. (2009, p. 1094) had anticipated what would be found:

> It is an open question whether libcitation counts for books and book chapters will correlate significantly with citation counts for the same works. Indeed, they may not. Our exploratory trials have shown some books to be high in both citation and libcitation counts. Occasionally, a book turns up that is well cited despite being held by relatively few libraries. More common are books that are meagerly cited, but relatively widely held. This overall mix produces low correlations.

These remarks were occasioned by spot-checking citation counts in the Web of Science. Using Scopus instead, Zuccala and Guns (2013) provided the first empirical answer to the open question: they found low but significant correlations.

The present paper continues this line of analysis (also described in Sieber and Gradmann, 2011). We do not hypothesize that libcitations *cause* citations (or the reverse)—merely that the two variables may positively co-vary.

Our database covers more than 100,000 books, and it now allows correlations to be obtained in the 10 main Dewey subject classes. As before, it has a total libcitation count for each book, but also disaggregates that total into counts for members of the Association of Research Libraries (ARL) and counts for non-members. The non-members include thousands of academic and public libraries whose collections are not primarily intended to support advanced research. In contrast, the 125 ARL institutions own very large subject collections that support graduate degree programs and specialized faculty research in many disciplines. (When multiple libraries in ARL institutions buy the same book, its count can go well beyond 125.) The books with the greatest cultural impact achieve libcitation counts in the thousands by appealing to ARL members and non-members alike. Plum Analytics, a commercial firm specializing in altmetrics, now includes a book's holding count in WorldCat as one of its indicators of "usage."

The results of our analyses, while interesting and suggestive, return us to a common criticism of both the Web of Science and Scopus: within the time frame of our study, they pick up too few citations to books to correlate those citations with libcitations on a firm basis. Both WoS and Scopus have recently expanded their efforts to capture citations to books, but it is too early to assess the full effect of these new data on bibliometrics. Kousha, Thelwall and Rezaie (2011) demonstrate that Google Books and Google Scholar give considerably higher citation counts for books than Scopus does. Our findings point to the same conclusion.

Overview of the database

Here we re-present several details about our database from Zuccala and Guns (2013) and add some new ones. The Scopus database from Elsevier supplied our citation data, which was

granted through the Elsevier Bibliometrics Research Program. Having requested separate datasets in *History* and *Literature & Literary Theory*, we further limited them to citations that appeared in journal articles during two periods, 1996-2000 and 2007-2011. We examined the Scopus data to determine the overall frequency with which various types of publications were cited: books, research articles, conference proceedings, review papers, notes, and other materials. Cited materials that were "non-sourced"—that is, that did not have a Scopus identification number linking them to a source journal—were classified as books, the unit of analysis in which we were interested.

Table 1 shows the number of journals in each field (as classified by Scopus) from which we drew citing articles. The lower part of Table 1 gives the numbers (N's) of books cited in the journal articles in each field and period. It will be seen that, in both fields, the N's of books cited in the earlier period are much smaller than those in the later, because Scopus covered fewer humanities articles in the 1990s.

Table 1. Journals and journal citation data in Scopus (April 2011).

Journal counts and classification codes	
History (N=494 source journals)	ASJC 1202 (Scopus Classification Code)
Literature & Literary Theory (N=419 source journals)	ASJC 1208 (Scopus Classification Code)
Both History and Literature (N=110 source journals)	Both ASJC 1202 and ASJC 1208
Counts of books cited during 1996-2000	*Counts of books cited during 2007-2011*
History (N=20,996)	History (N=50,466)
Literature & Literary Theory (N=7,541)	Literature & Literary Theory (N=35,929)

We searched the apparent books in WorldCat, using an API developer key granted to us by the Online Computer Library Center (OCLC). The key allowed us to match titles cited *at least once* in Scopus with titles held by *at least one ARL and one non-ARL library* covered by WorldCat. (These libraries, while mostly North American, include participants worldwide.) For every matched title (confirming that it was a book), we retrieved the OCLC accession number, ISBN number, publisher's name, publisher's location, and library count data. These were added to the book's citation data from Scopus to create a unique Scopus-WorldCat relational database.

Once a book has been published, it takes time for it to be acquired and cataloged by a library. A book published in a given year could have been acquired by a library no earlier than that year, but might have been acquired up to and including November 2012. Our holdings counts were current as of that cut-off date.

To improve publication-date accuracy, we analyzed only books published in the six years immediately preceding our two five-year citation windows. Thus, the books cited in 1996-2000 were limited (by filtering their Scopus records) to those published during 1990-1995. The books cited in 2007-2011 were likewise limited to those published during 2001-2006.

Converted to the four files at the bottom of Table 1, our book data come to 114,932 cases in all, 81 percent of which are unique titles. The remaining 19 percent are titles that appear more than once. Some were cited in both our earlier and later periods. Others were cited in both the History and the Literature journals, or in the journals that Scopus has assigned to both fields jointly, as shown in Table 1. We did not attempt to re-assign these latter titles to a single field, but allowed them to enter into the counts for both fields. There seems no easy way to avoid double counting, because that is the way in which Scopus has structured the data. Even so, a trial analysis with duplicates removed does not greatly affect the correlations.

Data analyses and results

Our data analyses were conducted with SPSS, the Statistical Package for the Social Sciences. Table 2 gives summary statistics for the titles in History and Literature. Means and standard deviations have been rounded to whole numbers. As noted in Zuccala and Guns (2013, p. 357), both citations and libcitations exhibit the highly skewed distributions typical of bibliometrics. However, the subsets of ARL libcitations for both History and Literature have bimodal distributions, with peaks at 1-4 and 100-104 holding libraries, and a low point at 45-54 libraries. In other words, the ARL libraries tend to acquire large numbers of rarely held titles, large numbers of widely held titles, and markedly fewer titles with holdings counts in between. This saddle-shaped distribution may reflect the opposing needs of specialized researchers: on their behalf, ARL libraries acquire many books held by few other members, but also many books that almost every member *must* have. The titles with the maximum counts in Table 2 (e.g., 92 citations; 4,725 libcitations) will be named in Tables 6 through 9.

Table 2. Summary statistics for two fields in combined time periods.

History combined periods N=71462					
	Minimum	Maximum	Mean	Std. Dev.	Median
Citations	1	92	2	3	1
ARL libcitations	1	212	59	40	63
Non-ARL libcitations	1	4603	278	351	178
Total libcitations	2	4725	338	372	250
Literature combined periods N=43470					
	Minimum	Maximum	Mean	Std. Dev.	Median
Citations	1	91	2	3	1
ARL libcitations	1	215	62	38	67
Non-ARL libcitations	1	4603	305	395	189
Total libcitations	2	4725	367	412	267

In Table 3, *citation* counts for every book are correlated with total *libcitation* counts for every book in major subsets of the database. Citation counts are also separately correlated with the libcitation counts for ARL members and non-members. (Only the libcitation variables are labeled, but the unlabeled citation variable is present in all the cells.) These are Spearman rho correlations, calculated with ranks of the count values rather than the counts themselves. Unlike Pearson r's, rho's do not require the assumption of normally distributed populations and so accommodate bibliometric skew (Zuccala & Guns, 2013: 357).

Table 3. Total, ARL, and non-ARL libcitations to books correlated with citations to the same books in two fields, two periods, and combined periods.

History 1996-2000			**History 2007-2011**			**History combined**		
Total	ARL	Non-ARL	Total	ARL	Non-ARL	Total	ARL	Non-ARL
0.26	0.29	0.25	0.25	0.28	0.24	0.24	0.26	0.23
N=20996			N=50466			N=71462		
Literature 1996-2000			**Literature 2007-2011**			**Literature combined**		
Total	ARL	Non-ARL	Total	ARL	Non-ARL	Total	ARL	Non-ARL
0.23	0.28	0.22	0.18	0.24	0.17	0.20	0.24	0.19
N=7541			N=35929			N=43470		

The rho's are all positive and weak, with values much like those in Zuccala and Guns (2013, p. 357). Because of the large numbers of books involved, all are significant at $p < .001$ by

one-tailed test. The hypothesis of no relationship can thus be safely rejected: citations and libcitations do capture a certain amount of scholarly impact in common. A sign of this in Table 3 is that citations, which are essentially a researchers' practice, always correlate a bit more highly with libcitations from research libraries—that is, ARL members. However, none of the rho's are strong enough to indicate that libcitations can substitute for citations as a measure. Libraries, especially ARL members, do buy many books that turn out to be well cited, but they buy even more books that are not highly cited in the journals covered by Scopus. This raises questions about the citation-libcitation relationship that we will return to later with specific examples.

Table 4 may clarify the situation in our two subject fields. The total libcitation counts for books have been divided at their medians. Citation counts for the same books have been collapsed into three groups, as shown in the column labels. In both History and Literature, the two variables are directly related: as citation counts rise, the percentage of books with above-median libcitation counts also rises sharply. For example, in History, only 43% of books cited once have libcitation counts in the top half, whereas for books cited two to four times the comparable figure is 59%, and for books with five or more citations, 79%. The percentages in the Literature table are almost identical.

Table 4. Libcitations and citations cross-tabulated in two fields for combined periods.

History combined periods				
	Citations			
Libcitations	1	2-4	5 or more	
GT Median	43%	59%	79%	50%
LE Median	57%	41%	21%	50%
	100%	100%	100%	100%
N =	46578	19165	5719	71462
Literature combined periods				
	Citations			
Libcitations	1	2-4	5 or more	
GT Median	44%	59%	78%	50%
LE Median	56%	41%	22%	50%
	100%	100%	100%	100%
N =	29876	10668	2926	43470

However, this effect must be viewed in light of the extreme skew of the citation counts seen in the column marginals. Roughly two-thirds of all books in our samples have only one citation each, and roughly another quarter have only two to four citations. The fraction of titles with five or more citations is relatively small. Thus, the Spearman rho's for these grouped variables, though highly significant ($p < .001$), are even lower than when the variables are ungrouped in Table 3—only 0.22 for History and 0.19 for Literature.

We turn to a finer breakdown of the data. As mentioned in Zuccala and Guns (2013, p. 358), historians who publish in History journals do not exclusively cite works of history, nor do literary scholars who publish in Literature journals exclusively cite works of literature or literary theory. Instead, both groups cite books across the full range of subjects covered by the Dewey Decimal Classification. We were able to get the Dewey class numbers for most of our book titles from WorldCat. (Some books do not receive Dewey classifications.) In Table 5 we subdivide the books cited in History and Literature journals in our two time periods by their main Dewey classes.

Class 000 in Dewey is formally "Computer science, information, general works." This class is traditionally used for general reference books and books in trans-disciplinary fields such as librarianship, journalism, publishing, and reading. Historians and literary scholars mainly cite

books in areas like these, rather than in computer science. Hence, we have shortened the long label here to "General works."

The Table 5 cells contain 120 replications of our correlational study in subsets of the data. We are again correlating each book's total citations with its total libcitations, as well as the libcitation counts from ARL members and ARL non-members. In making comparisons, be aware that non-ARL libcitations make up by far the larger share of total libcitations. The two categories thus tend to produce correlations that are similar or identical, and so the non-ARL results will not be separately discussed here.

Table 5. Libcitations correlated with citations to books by field, period, and main Dewey classes.

	History 1996-2000				**History 2007-2011**			
Main Dewey Classes	Libcites	ARL	Non-ARL	N =	Libcites	ARL	Non-ARL	N =
000 General works	0.20	0.21	0.20	350	0.23	0.28	0.22	794
100 Philosophy and psychology	0.20	0.21	0.19	1055	0.18	0.20	0.17	2041
200 Religion	0.27	0.27	0.26	1766	0.27	0.29	0.25	4186
300 Social sciences	0.26	0.28	0.26	8067	0.23	0.25	0.21	16585
400 Language	0.11	0.11	0.12	247	0.17	0.16	0.17	672
500 Science	0.20	0.27	0.19	914	0.13	0.23	0.11	1543
600 Technology	0.25	0.35	0.23	824	0.12	0.24	0.09	1990
700 Arts and recreation	0.21	0.24	0.20	1056	0.19	0.26	0.18	3788
800 Literature	0.17	0.26	0.15	1620	0.20	0.26	0.19	4725
900 History and geography	0.28	0.31	0.27	4388	0.27	0.29	0.25	10439
	Literature 1996-2000				**Literature 2007-2011**			
Main Dewey Classes	Libcites	ARL	Non-ARL	N =	Libcites	ARL	Non-ARL	N =
000 General works	0.09	0.08	0.09	155	0.17	0.36	0.14	548
100 Philosophy and psychology	0.19	0.22	0.18	585	0.23	0.27	0.22	1919
200 Religion	0.13	0.19	0.12	398	0.25	0.29	0.23	2221
300 Social sciences	0.14	0.16	0.14	1344	0.19	0.22	0.18	6322
400 Language	0.22	0.24	0.21	505	0.22	0.24	0.20	1218
500 Science	0.04	0.09	0.04	115	0.06	0.12	0.06	516
600 Technology	0.13	0.28	0.11	130	0.09	0.24	0.07	703
700 Arts and recreation	0.18	0.21	0.17	591	0.22	0.26	0.20	3268
800 Literature	0.23	0.31	0.21	2616	0.26	0.31	0.25	11171
900 History and geography	0.14	0.25	0.12	742	0.21	0.26	0.20	3963

Even with Table 5's extensive partitioning, the N's underlying the correlations are large enough that most of the rho's remain highly significant ($p < .001$ by one-tail test). Of the correlations between citations and total libcitations, 21 out of 40 remain at or above 0.20. Large N's can cause correlations that are statistically but not substantively significant (Babbie 2015, p. 469). Nevertheless, certain patterns do lend substance to the overall analysis:

- Some 33 of the 40 ARL correlations remain in the 0.20s or higher.
- Some 37 of the 40 ARL correlations are higher than those for the non-ARL libraries in their row. This reinforces the supposed connection between citations and libcitations in research environments.
- As examples of subject accord, the ARL correlation for books classed in *900 History and geography* is second-highest (0.31) in History 1996-2000, and tied-highest (0.29) in History 2007-2011.
- As further examples of subject accord, the ARL correlation for books classed in *800*

Literature is highest (0.31) in Literature 1996-2000, and second-highest (0.31) in Literature 2007-2011.

- In both our History periods, the lowest correlations occur for books classed in *400 Language*. The N's for books in this class, which is historically Dewey's smallest, are likewise small. While historians make use of research from all fields, it is unsurprising that books on language are not their chief resource.
- In both our Literature periods, the lowest correlations occur for books classed in *500 Science,* and the N's for books in this class are small as well. One would not expect literary scholars to cite numerous science books. However, one might expect them to cite more books in *400 Language* than historians, and that is what the data show.
- Table 5 in fact shows wide variation in the number of books that Scopus authors have cited in each class. In both History periods, books classed in *300 Social Sciences* are most numerous. This makes sense because of the close interplay between historical and social scientific topics. Books classed in *900 History and geography* are the second-most numerous, and books in *800 Literature* are third. In both Literature periods, the same three classes dominate but in another order: *800 Literature* first, as seems fitting, then *300 Social Sciences* and *900 History and geography*. For our two broad fields in the humanities, these are reassuringly reasonable outcomes.

Since libcitations are a new altmetric, we think it informative to display the titles that have top-ranked libcitation counts in particular contexts (as do both Torres-Salinas and Moed, 2009 and White et al., 2009). This allows a qualitative as well as a quantitative analysis. White (2005) proposed the label *bibliograms* for bibliometric distributions in which not only the ranked counts but also the terms associated with them are analyzed as communications. "Bibliograms," he wrote (p. 443), "consist of (1) at least one seed term that sets a context, (2) terms that co-occur with the seed across some set of records, and (3) counts of how frequently terms co-occur with the seed by which they can be ordered high to low." Here, we use main Dewey class names as seed terms. We then rank the books that co-occur with them (as OCLC accession numbers) by their libcitation or citation counts. Lastly, the OCLC numbers are used to retrieve full bibliographic data from WorldCat so that we can comment on the authors, titles, and nature of the top-ranked books.

Table 6 comprises extracts from 40 bibliograms. We display, for our two fields and two time periods, the titles with the highest *total* libcitation counts in each of the 10 main Dewey classes. Many of these books have subtitles, but they have been omitted in favor of authors' surnames (or those of first authors in collaborations). We also display their ARL libcitation counts and their citation counts in Scopus.

The books in Table 6 do not resemble typical scientific articles. They are the sort of titles that present readers, like everyone else, may have purchased for reasons having nothing to do with bibliometrics. They exemplify the broad cultural impact of the humanities—for example, standard reference works on language, music, religion; biographies of famous men (Peter Gay's *Freud*, David McCullough's *Truman* and *John Adams*); novels (Toni Morrison's *Paradise*, Dan Brown's *The Da Vinci Code*); popularizations of science (Dava Sobel's *Longitude*, Malcolm Gladwell's *Blink*, Carl Sagan's *Cosmos*); best-selling social critiques (Susan Faludi's *Backlash*, Robert Hughes's *Culture of Complaint*); advice for business executives (James Collins's *Good to Great*, Thomas Peters and Robert Waterman's *In Search of Excellence*). While some exemplify high scholarship, others are not scholarly at all (Ernest Hemingway's *A Moveable Feast*); some are even children's books (David Wiesner's *Flotsam*, Peter Spier's *Noah's Ark*, both Caldecott Medal winners). They come to the fore here because they were bought by thousands of libraries, and they had citation counts of at least one in Scopus. Persons at research universities who specialize in manifestations of popular culture are legion.

Table 6. Books with highest libcitation counts by field, period, and main Dewey class.

History 1996-2000					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
1	160	2592	General works	The Oxford dictionary of modern quotations	Augarde
1	143	2936	Philosophy and psychology	Freud	Gay
1	101	2789	Religion	Crossing the threshold of hope	John Paul II
1	124	4233	Social sciences	My American journey	Powell
1	105	3433	Language	The story of English	McCrum
2	108	2572	Science	Longitude	Sobel
1	112	3204	Technology	Healing and the mind	Moyers
1	130	2133	Arts and recreation	Culture of complaint	Hughes
1	122	4132	Literature	Paradise	Morrison
4	137	4724	History and geography	Truman	McCullough
History 2007-2011					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
1	160	2592	General works	The Oxford dictionary of modern quotations	Augarde
2	145	4059	Philosophy and psychology	Blink	Gladwell
1	93	2931	Religion	Under the banner of heaven	Krakauer
4	152	3967	Social sciences	Freakonomics	Levitt
5	182	2760	Language	The Oxford English dictionary	Simpson
4	104	3284	Science	A short history of nearly everything	Bryson
2	148	4496	Technology	In search of excellence	Peters
4	123	2596	Arts and recreation	New Grove dictionary of music	Grove
6	122	4725	Literature	The Da Vinci code	Brown
5	140	4655	History and geography	John Adams	McCullough
Literature 1996-2000					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
2	155	2076	General works	Double fold	Baker
3	145	4059	Philosophy and psychology	Blink	Gladwell
1	87	3511	Religion	Noah's ark	Spier
3	152	3967	Social sciences	Freakonomics	Levitt
1	105	3433	Language	The story of English	McCrum
1	125	3884	Science	Cosmos	Sagan
1	141	4195	Technology	Good to great	Collins
1	86	4133	Arts and recreation	Flotsam	Wiesner
13	122	4725	Literature	The Da Vinci code	Brown
1	140	4655	History and geography	John Adams	McCullough
Literature 2007-2011					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
1	115	3342	General works	The road ahead	Gates
1	75	2455	Philosophy and psychology	Care of the soul	Moore
1	128	3083	Religion	The Oxford companion to the Bible	Metzger
2	154	3169	Social sciences	Backlash	Faludi
2	148	3119	Language	The Oxford companion to the English language	McArthur
1	112	2068	Science	Black holes and time warps	Thorne
1	93	4314	Technology	Men are from Mars, women are from Venus	Gray
1	130	2133	Arts and recreation	Culture of complaint	Hughes
1	125	3455	Literature	A moveable feast	Hemingway
1	121	3600	History and geography	The fifties	Halberstam

Thus, even the most pop-cultural books in Table 6 are widely held by ARL members. It is a misconception that these libraries acquire only works of rarified scientific or scholarly status. In fact, they buy innumerable works that would also be found in public and school libraries. The best example is the single most widely held item in our database—*The Da Vinci Code*, owned by 122 (of 125) ARL members. Whatever one may think of this novel, it had a huge impact for several years, and scholars in the humanities will want copies on hand, if only to attack Dan Brown's transgressions. Nevertheless, the citation counts for these books in Table 6's leftmost column are very low. Brown's novel has the most, and these may include book reviews.

By contrast, Table 7 displays the titles that are *most* highly cited in our categories. As implied earlier, relatively high citation counts tend to signal a research orientation, and these 40 books, which have the top counts in their respective Dewey classes, are almost all distinctly more academic than those in Table 6. Their *total* libcitation counts tend to be lower than those in Table 6, suggesting more specialized readerships. (The exception is *The Guardian*, a Nicholas Sparks novel.) A fair number of them address themes prominent in the humanities (race, class, gender, imperialism), and their authors include names famous to postmodern scholars, if not to the general public (e.g., Edward Said, Gilles Deleuze, Judith Butler, Donna Haraway, Gayatri Spivak, and, with two books, Giorgio Agamben).

Three-fourths of these books are held by a hundred or more ARL libraries. Of those that are not, some may reflect genuinely narrower acquisition by ARL members. Others (if not errors) may reflect delayed or incomplete reporting of an acquired book that makes its libcitation count deceptively small. That may have happened, for instance, with Spivak's *Death of a Discipline*, whose ARL count in Table 7 is only 22, but whose count as an e-book in WorldCat is 1,246 at this writing.

In any event, ARL libcitation counts range unbrokenly over values from 1 to 215. Given this variation, why are the correlations of ARL counts with citations not higher? We have already noted that they tend to be higher than correlations of *total* libcitations with citations, but only slightly. In both cases the problem is the same: the great majority of books in our database have only one citation (or at most a few). Thus, a key variable in our study has little variability. As one illustration, Table 8 lists the five books with the highest ARL libcitation counts in our two fields (time periods combined, and omitting the *Oxford English Dictionary*, already shown). These books are best-sellers not only among ARL members but in libraries of all kinds. Yet their citation counts in Scopus are minuscule and much the same, just as they were for the books in Table 6. To anyone familiar with these titles, it is incredible that Table 8 reflects their full citation records. Rather, their true counts are not being captured.

Not too long ago, this assumption could only have been checked with data from the Web of Science, but now we can spot-check citations to books in Google Scholar. When that is done, the results are very different from what Scopus shows, whether the Scopus figures are as low as one or as high as 92. Table 9 suggests the nature of the problem. The counts there reflect our judgment calls, such as to include only those for the 2000 edition of *DSM-IV-TR* or the 2007 edition of *The Elements of Style*. Google Scholar itself does not break down by edition the many citations to the feminist classic *In a Different Voice*. Nor does it allow us to extract citations to books in our two periods of study. Nevertheless, the Google Scholar counts indicate where further correlational research should be directed (see also Prins et al., 2014).

Table 7. Books with highest citation counts by field, period, and main Dewey class.

History 1996-2000					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
14	117	573	General works	The letters of the Republic	Warner
30	115	798	Philosophy and psychology	The production of space	Lefebvre
19	111	689	Religion	Ritual theory, ritual practice	Bell
75	129	1195	Social sciences	Imagined communities	Anderson
11	76	509	Language	Biblical Hebrew syntax	Waltke
29	107	450	Science	Bayes or bust?	Earman
25	84	364	Technology	Curing their ills	Vaughan
13	108	650	Arts and recreation	Orientalism	MacKenzie
56	119	1381	Literature	Culture and imperialism	Said
71	119	1406	History and geography	Britons	Colley
History 2007-2011					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
24	114	546	General works	"The tyranny of printers"	Pasley
39	26	413	Philosophy and psychology	The navigation of feeling	Reddy
37	109	478	Religion	Formations of the secular	Asad
92	114	602	Social sciences	Carnal knowledge and imperial power	Stoler
22	12	481	Language	Bilingualism and the Latin language	Adams
31	115	556	Science	The body of the artisan	Smith
32	100	342	Technology	Contagious divides	Shah
17	92	412	Arts and recreation	The reformation of the image	Koerner
26	32	2802	Literature	The guardian	Sparks
83	116	813	History and geography	The birth of the modern world, 1780-1914	Bayly
Literature 1996-2000					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
71	110	415	General works	The reading nation in the Romantic period	St. Clair
79	102	391	Philosophy and psychology	The open	Agamben
36	87	404	Religion	Saint Paul	Badiou
91	117	545	Social sciences	State of exception	Agamben
37	101	377	Language	The translation zone	Apter
12	95	294	Science	The spacious word	Padron
37	71	259	Technology	The companion species manifesto	Haraway
27	104	348	Arts and recreation	In the break	Moten
85	22	559	Literature	Death of a discipline	Spivak
87	106	462	History and geography	Writing history, writing trauma	LaCapra
Literature 2007-2011					
Cites	**ARL**	**Libcites**	**Dewey class**	**Title**	**Author**
6	117	573	General works	The letters of the Republic	Warner
17	108	632	Philosophy and psychology	Difference and repetition	Deleuze
6	114	771	Religion	Fragmentation and redemption	Bynum
41	131	1049	Social sciences	Gender trouble	Butler
19	84	301	Language	Discourse and social change	Fairclough
9	117	1034	Science	The origins of order	Kauffman
5	112	475	Technology	The commodity culture of Victorian England	Richards
11	122	983	Arts and recreation	Gone primitive	Torgovnick
38	120	843	Literature	The location of culture	Bhabha
23	125	891	History and geography	Imperial eyes	Pratt

Table 8. Books with the top five ARL libcitation counts in two fields.

History combined				
Cites	**ARL**	**Libcites**	**Title**	**Author**
2	212	4101	Diagnostic and statistical manual of mental disorders: DSM-IV-TR	
2	194	2478	In a different voice	Gilligan
3	180	1282	The alchemy of race and rights	Williams
2	176	1348	On the law of nations	Moynihan
1	176	1136	Theoretical perspectives on sexual difference	Rhode
Literature combined				
Cites	**ARL**	**Libcites**	**Title**	**Author**
1	215	3792	Publication manual of the American Psychological Association	
1	204	3436	The elements of style	Strunk, White
1	203	2046	A theory of justice	Rawls
3	178	1466	There's no such thing as free speech, and it's a good thing, too	Fish
1	175	995	Sex and reason	Posner

Table 9. Same data, but with citations in Scopus replaced by citations in Google Scholar.

History combined				
Cites	**ARL**	**Libcites**	**Title**	**Author**
5364	212	4101	Diagnostic and statistical manual of mental disorders: DSM-IV-TR	
30044	194	2478	In a different voice	Gilligan
2431	180	1282	The alchemy of race and rights	Williams
146	176	1348	On the law of nations	Moynihan
102	176	1136	Theoretical perspectives on sexual difference	Rhode
Literature combined				
Cites	**ARL**	**Libcites**	**Title**	**Author**
1393	215	3792	Publication manual of the American Psychological Association	
2988	204	3436	The elements of style	Strunk, White
782	203	2046	A theory of justice	Rawls
616	178	1466	There's no such thing as free speech, and it's a good thing, too	Fish
1546	175	995	Sex and reason	Posner

Discussion

The correlations in this paper suggest that libcitations and citations are not entirely different measures of impact. However, we are left wanting citation counts for books that do not have so many low, tied values. It is possible that better data would again produce low or even negligible correlations. It is also possible that the correlations would be much higher than those seen here. The libcitation measure draws on a varied mix of assessments, and they are not necessarily the same as those that go into scholars' acts of citation. But, as our data make plain, they indicate major intellectual achievements no less forcefully than citations. In fact, one can argue that many of the humanities titles in Table 6 are *truly* major achievements, in that they have reached large publics beyond academe.

What, then, do libcitations measure? Briefly, they estimate the potential readerships, or users, of a given book. Citations, in contrast, measure actual uses to which the book has been put within research-oriented communities. It is therefore not surprising that citations and libcitations are associated, especially if the latter come from libraries that serve researchers,

such as those in ARL. But libcitations also measure broad cultural impacts that citations may miss, because libcitations rest on chains of judgments within the world of publishing, and this world, which subsumes the scholarly one, extends into every part of life. The chains include authors, agents, past editors who have built publishers' reputations, present-day editors of various kinds, referee-readers, marketers, and wholesalers. Librarians are only the last link.

This speaks to the common objection that librarians do not evaluate individual titles, but put their acquisitions on automatic pilot through approval plans and the like; how, then, can libcitations reflect genuine worth? On the contrary, librarians are highly attuned to potential demand in their communities, and it is they who approve the approval plans and buy into the pre-formed collections. It is quite true that such moves favor some publishers over others, but that is because librarians trust the chains of judgment those publishers represent. And so do their communities, who routinely expect librarians to have acquired certain books they learn about and are displeased if they have not.

Libcitations are sales figures—a market measure. They reflect virtual unanimity on the worth of some titles, but they vary enormously. In our database, although the counts run to the high values seen in our tables, many titles are held by only one ARL and one non-ARL library, just as many papers have only a citation or two. Research on libcitation-citation correlations should continue, but even if they remain low, that does not invalidate the libcitation measure. It is better thought of as a free-standing gauge of authors' cultural impact. Having published a book, what author would not prefer a thousand libraries to hold it rather than 10?

Acknowledgments

The authors are grateful to the Elsevier Bibliometrics Research Programme (http://ebrp.elsevier.com/) and OCLC WorldCat for granting access to the data used to build the unique database for this study. We also thank Dr. Roberto Cornacchia for helping to develop the database, as well as Maurits van Bellen and Robert Iepsma for their data cleaning and standardisation work. Dr. Stefanie Haustein of the École de bibliothéconomie et des sciences de l'information (EBSI), Université de Montréal, kindly provided information on the library holdings-count measure at Plum Analytics.

References

Babbie, E. R. (2013). *The practice of social research*. 14th edition. Boston, MA: Cengage Learning.

Kousha, K., Thelwall, M., & Rezaie, S. (2011) Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology, 62*(11), 2147-2164.

Prins, A., Costas, R., van Leeuwen, T., & Wouters, P. (2014). Using Google Scholar in research evaluation of social science programs, with a comparison with Web of Science data. *Proceedings of the Science and Technology Indicators Conference 2014 Leiden*, 434-443.

Sieber, J., & Gradmann, S. (2011). How to best assess monographs? An attempt to assess the impact of monographs using library infrastructure and Web 2.0 tools. European Educational Research Quality Indicators. Retrieved December 17, 2014 from

http://edoc.hu-berlin.de/docviews/abstract.php?id=38002

Torres-Salinas, D., & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in economics. *Journal of Informetrics, 3*(1), 9-26.

White, H. D. (2005). On extending informetrics: An opinion paper. Proceedings of the 10th International Society for Scientometrics and Informetrics Conference, 2, 442-449.

White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology, 60*(6)*,* 1083-1096.

Zuccala, A., & Guns, R. (2013). Comparing book citations in humanities journals to library holdings: Scholarly use versus 'perceived cultural benefit' (RIP). *Proceedings of the 14th International Society for Scientometrics and Informetrics Conference, 1*, 353-360.

A Vector for Measuring Obsolescence of Scientific Articles

Jianjun Sun[1], Chao Min[1] and Jiang Li[2]

[1] *sjj@nju.edu.cn*
School of Information Management, Nanjing University, Nanjing (China)

[1] *marlonmassine@yeah.net*
School of Information Management, Nanjing University, Nanjing (China)

[2] *li-jiang@zju.edu.cn*
Department of Information Resource Management, Zhejiang University, Hangzhou (China)

Abstract
Diachronous studies of obsolescence categorized articles into three general types: "flashes in the pan", "sleeping beauties" and "normal articles", by using quartiles to identify first 25% and last 75% articles reaching 50% of their total citations, or by using averages to define threshold values of sleeping and awakening periods. However, the average-based and quartile-based criteria, sometimes, less effectively distinguished "flashes in the pan" and "sleeping beauties" from normal articles. In this research, we proposed a vector for measuring obsolescence of scientific articles, as an alternative to these criteria. The obsolescence vector is designed as $O = (G_s, A^-, n)$, where n is the age of an article, G_s and A^- are parameters for revealing the shape of citation curves. Among Nobel laureates' 28,340 articles, each of which received over 20 citations, we identified 265 flashes in the pan (approximately 1%) and 40 sleeping beauties (approximately 0.1%) by the obsolescence vector. By a few case studies, it is verified that obsolescence vector yielded more reasonable classifications than did the average-based and quartile-based criteria.

Conference Topic:
Indicators

Introduction
In a previous study (Li et al., 2014), we introduced G_s index, an adjustment of Gini coefficient, for measuring the inequality of "heartbeat spectrum" of "sleeping beauties". "Sleeping beauty" in science was first proposed by van Raan (2004), in order to describe a phenomenon where papers did not achieve recognition in citations until many years after their original publication. As in the fairy tale, a princess (an article) sleeps (goes unnoticed) for a long time and then, almost suddenly, is awakened (receives a lot of citations) by a prince (another article). "Heartbeat spectrum" was defined as a vector of a sleeping beauty's annual citation(s) received in the sleeping period.

How to categorize recognition to a paper as "early", "delayed" or "normal"? Diachronous studies of obsolescence answered this question, by using quartiles to identify first 25% and last 75% articles reaching 50% of their total citations (Costas et al., 2010), or by using averages to define threshold values of sleeping and awakening periods (van Raan, 2004; van Dalen & Henkens, 2005). In this research, we propose an obsolescence vector based on the G_s index, as an alternative to both approaches.

Literature review
"Obsolescence" (or "ageing") studies, in the field of bibliometrics, attempt to answer the question how long does the information in a research paper remain current, by measuring the number of citations the paper received since publication (Cunningham & Bocock, 1995). There are two approaches to measure obsolescence: "synchronous" and "diachronous" distribution (Nakamoto, 1988). They are also referred to as "citations from" and "citations to" approaches (Redner, 2005), or "retrospective citation" and "prospective citation" approaches

(Burrell, 2002; Glänzel, 2004). The former considers the age distribution of references of a paper in a particular year, while the latter analyzes the distribution of citations over time.

A number of metrics has been proposed, from a synchronous perspective, to measure obsolescence of scientific literature. "Half-life" was described (Burton & Kebler, 1960) as "half the active life", which means the time during which one-half of the currently active literature was published. Price (1970) suggested the percentage of references (from all articles) up to five years old as an index to reveal obsolescence of scientific documents, which is also named "Price Index".

From a diachronous perspective, a citation curve (Garfield, 1989; Avramescu, 1979; Li et al., 2014) is the time distribution of citations a paper received. It is also referred to as "life-cycle" (Cunningham & Bocock, 1995), "citation patterns" (Li & Ye, 2014; Wang, Song, & Barabási, 2013; Guo & Suo, 2014; Redner, 2005), or "citation history" (Redner, 2005; ABT, 1981; Persson, 2005; Vlachý, 1985; Costas et al., 2010). A "typical citation curve" describes the history of an article which received a few citations in the first following years after publication, then rose to a citation peak, but afterwards was gradually less cited with time. It is identified that lognormal function best fits typical citation curves (Egghe & Rao, 1992). For most scientific papers, death (no longer being cited by other papers) comes within ten years after publication (Price, 1976). Nevertheless, the minority appears exponential increase in citations in a long time, whose citation curves fit exponential function (Li & Ye, 2014).

The peaking time of citations features the shape of citation curves, reflecting the immediacy of publications. Some articles were noticed immediately after publication but ignored very soon, and hence were named as "flashes in the pan" (van Dalen & Henkens 2005; Costas et al., 2010). Their citations peaked much earlier than typical citation curves. Some went unnoticed for a long time and then, almost suddenly, received a lot of citations, and hence were referred to as "sleeping beauties" (van Raan, 2004), "premature discoveries" (Stent, 1972; Wyatt, 1975), "resisted discoveries" (Barher, 1961) or "delayed recognition" (Cole, 1970). Their citations peaked much later than typical citation curves. Van Raan (2004) suggested three criteria for distinguishing sleeping beauties: (1) they deeply slept (receive at most 1 citation per year on average), or less deeply slept (between 1 and 2 citations per year on average) for a few years after publication; (2) they slept at least five years; and (3) they were awakened by over 20 citations during the four years following the sleeping period. However, the criteria are not always applicable to answer Garfield (1980)'s question how abrupt a citation boost must be to suggest delayed recognition. Moreover, the criteria ignored the citations received after the awakening period (Li, 2014; Li & Ye, 2012).

Different from van Raan's average-based criteria, Costas et al. (2010) used quartiles. They identified the year after publication in which the document received for the first time at least 50% of its citations ("Year 50%"), then calculated, for all documents of the same year of publication in the same field, the percentiles 25 and 75 of the distribution function of the value of "Year 50%", and recorded them as "P25" and "P75". As a result, the articles were categorized into "flashes in the pan" ("Year 50%" <"P25"), "delayed recognition" ("Year 50%" >"P75") and the rest as "normal publications" ("P25"≤"Year 50%"≤"P75"). These criteria considered the whole citation history of articles rather than only sleeping and awakening periods, and avoided the deficiency of van Raan's definitions. However, the excessive percentages of early and delayed recognition identified by these criteria caused the originally rare phenomena normal.

Methodology

Design of the obsolescence vector

Suppose there are seven ten-year old articles whose citation curves are drawn in Figure 1. P_1 is a sleeping beauty who deeply slept for six years (received no citations) but was suddenly awakened by 40 citations in the following four years. P_2 is a flash in the pan, which immediately received 32 citations within the first two years after publication, but was ignored afterwards and rarely received citations. P_3 is a typical citation curve, which reached citation-peak in the fourth year. It was successfully fitted by the lognormal function in the program OriginPro 8 ($R^2 = 0.972$). P_4 is an article whose citations increase exponentially. Exponential function successfully fits the curve with $R^2 = 0.983$. Both P_5 and P_6 are waveform curves, but they have different initial values, hence have distinct normalized curves in Figure 1. P_7 is a horizontal line, and coincides with the 45 degree diagonal in the right side of Figure 1, which is called "the line of equality" and indicates absolutely even distribution.

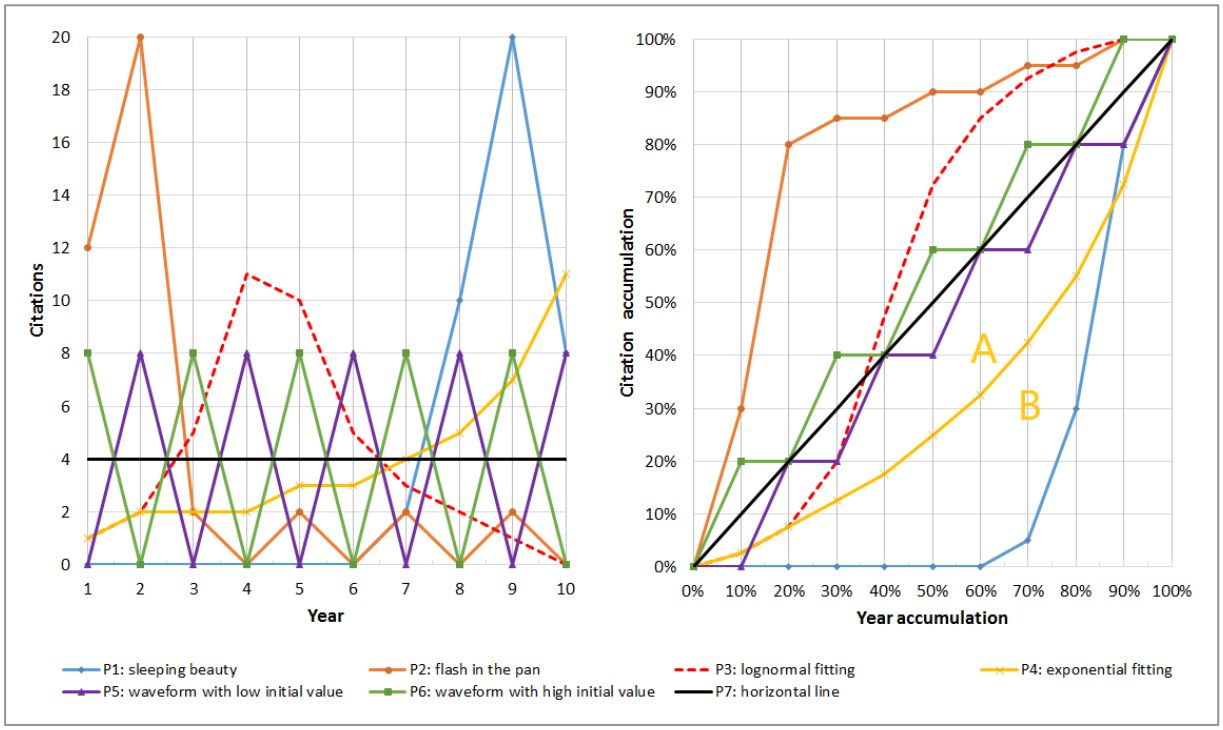


Figure 1. From citation curves to normalized cumulative citation curves of P1-P7 (left: citation curves; right: normalized cumulative citation curves).

The value of G_s, taking P4 as an example, equals to the ratio of the area that lies between the line of equality and the normalized cumulative citation curve (marked A in Figure 1) over the total area under the line of equality (sum of A and B), i.e.,

$$G_s = \frac{A}{A+B}. \quad (1)$$

The normalized cumulative citation curve (hereafter "normalized curve") of P4 is a "Lorenz curve", because the sequence of citations is in an ascending order. Since the areas A and B form an isosceles right triangle, we have

$$A + B = \frac{1}{2}. \quad (2)$$

Thus, putting Eq. (2) into Eq. (1), we have

$$G_s = 2A. \quad (3)$$

The calculation of G_s is determined by the calculation of the area B which can be divided into several trapeziums and a triangle. In this study, we remain the expression of the segment function of G_s in our previous study (Li et al., 2014),

$$G_s = \begin{cases} 1 - \frac{2\times[n\times c_1+(n-1)\times c_2+\cdots+c_n]-C}{C\times n}, & C > 0 \\ 1, & C = 0 \end{cases} \quad (4)$$

but redefine the parameters. In the new definition, n is the age of a paper, C is the total number of citations the paper received during the n years, and $c_i (i \in \{1,2,\cdots,n\})$ is the number of citations the paper received in the i^{th} year after publication. Here, $Gs \in (-1,1]$ and depends on the age (n) of articles. The value of G_s gradually approaches to -1, if the article no longer receives citations.

The value of G_s, to certain extent, characterizes the shape of citation curves:

(1) large G_s indicates delayed recognition, while small G_s denotes early recognition, as P_1 and P_2 shown in Table 1;

(2) $G_s < 0$ implies that there exists leaping early in citation curves, for example, both P_2 and P_6 received a large number of citations immediately after publication, while P_3 has a fast rising period although it does not have immediacy; and

(3) $G_s = 0$ suggests a horizontal citation curve (as P_7), or a citation curve including at least one high-citation period (to guarantee $A^- < 0$) which is offset by at least one low-citation period.

The value of A is not always positive. For P_2, $A<0$, since its normalized curve in Figure 1 is above the line of equality. Since

$$A = A^+ + A^- , \quad (5)$$

putting Eq. (5) into Eq. (3), we have

$$A^- = \frac{1}{2}G_s - A^+. \quad (6)$$

A^+ is the area between the line of equality and the normalized curve under the line of equality. Similar to the calculation of G_s, we calculate A^+, and accordingly have the value of A^-. In case of P_3, the intersection of the normalized curve and the line of equality in Figure 1 exists in between the accumulation year 30% and 40%. Therefore, there is a minor error (a difference) between the output and target of A^+ values of P_3. In cases of P_1, P_4 and P_5, there is no error in the calculation of A^+.

The fast rising period of a citation curve is hidden from the value of G_s if $A^- < 0 < A^+$. In case of $A^+ = 0$, we have

$$A^- = A = \frac{1}{2}G_s. \quad (7)$$

Hence, the value of A^- provides complementary explanation to the shape of citation curves:

(1) recognition to the article is normal or delayed rather than early if $A^-=0$;

(2) there exists leaping in the citation curve of the article if $A^-<0$; and

(3) citation leaping appears early if $A^- = \frac{1}{2}G_s$.

We propose a vector for measuring obsolescence of scientific articles: $O=(G_s, A^-, n)$, where G_s is an index revealing the history of citations, A^- is a parameter uncovering citation leaping and age n is an adjusting parameter. We calculated the obsolescence vectors for P_1-P_7 as shown in Table 1.

Table 1. Obsolescence vectors for P1-P7.

Article	*Citation curve*	*Citations*	*A*	A^+	*Obsolescence vector*		
					G_s	A^-	*n*
P1	Sleeping beauty	40	0.335	0.335	0.670	0.000	10
P2	Flash in the pan	40	-0.300	0.000	-0.600	-0.300	10
P3	Lognormal fitting	40	-0.075	0.028	-0.150	-0.103	10
P4	Exponential fitting	40	0.183	0.183	0.365	0.000	10
P5	Waveform with low initial value	40	0.050	0.050	0.100	0.000	10
P6	Waveform with high initial value	40	-0.050	0.000	-0.100	-0.050	10
P7	Horizontal line	40	0.000	0.000	0.000	0.000	10

Criteria for categorizing the patterns of obsolescence

In this research, we use the terms "flashes in the pan", "sleeping beauties" and "normal articles" as the patterns of obsolescence, but provide three different approaches for measurement, in order to characterize obsolescence vector. We remain van Raan's average-based criteria in the first approach. By following the criteria, we define variables for "flashes in the pan": "noticed" (van Dalen and Henkens, 2005) as receiving over 10 citations, "ignored" as receiving less than two citations per year on average and "immediately" as within two years since publication. We also define the duration of light disappearing for at least five years, since a flash is likely to reappear. Then, we suggest average-based criteria as follows:

flashes in the pan (F_1): articles which received more than 10 citations in the first two years since publication, and then in the next five years received no more than 2 citations per year on average;

sleeping beauties (S_1): articles which received no more than 2 citations per year on average in the first five years since publication, and then in the next four years received more than 20 citations; and

normal articles (N_1): which neither satisfy the criteria for F_1 nor for S_1.

The second approach uses quartiles. We adjust "relative ranking in a field" in Costas et al. (2010) to "relative age", since the former requires the population of articles in a filed which involves a huge dataset. Thus, for a single article, we record the percentiles 25 and 75 of its age as "A25" and "A75". Then, we define quartile-based criteria for the patterns of obsolescence as follows:

flashes in the pan (F_2): articles that reached "Year 50%" within 25% of its age, i.e., "Year 50%" <"A25";

sleeping beauties (S_2): articles that reached "Year 50%" with the time exceeding 75% of its age, i.e., "Year 50%" >"A75"; and

normal articles (N_2): which neither satisfy the criteria for F_1 nor for S_1, i.e., "A25"≤"Year 50%"≤ "A75".

Based on the obsolescence vectors of the seven cases in Table 1, we propose new criteria for categorizing the patterns of obsolescence as follows,

flashes in the pan (F_3): $G_s \leq$ -0.6 and $A^- = \frac{1}{2} G_s$;

sleeping beauties (S_3): $G_s \geq 0.6$ and $A^- = 0$; and

normal articles (N_3): which neither satisfy the criteria for F_3 nor for S_3.

Data

A dataset was prepared to make comparisons of the above three sets of criteria, and to verify the efficiency of the proposed obsolescence vector. From the Web of Science, we collected 58,963 articles of 629 Nobel Prize winners during the period of 1901-2012, in the fields of Chemistry, Physics, Physiology or Medicine, and Economic Sciences. The definition S_2 requires that a sleeping beauty should have more than 20 citations. For the purpose of comparisons, we eliminated articles, which received no more than 20 citations, and remained a collection of 28,340 articles published between 1900 and 2000. Then, we searched the number of annual citations to these articles up to 2011 in the Web of Science. Thus, every article in this collection aged at least eleven, which is sufficient for a sleeping beauty with the shortest sleeping period to be awakened.

Results

Obsolescence vector as an alternative to average-based and quartile-based criteria

The life-cycles of most articles in the dataset have already drawn to their close. As shown in Table 2, the peak of G_s distribution appears in the interval (-0.4,-0.2] and the values of G_s for 84.3% articles are negative. Moreover, 95.0% of the articles have $A^-<0$. Small G_s values (minus) indicate the end of cife-cycles, as shown by article P_2 in Figure 1. It is calculated that 68.4% of the articles with $G_s > 0$ have $A^- < 0$. Thus, there are only a small fraction of citation curves having the shape of P_1, P_4 and P_5 in Figure 1. What they have in common is that there is no citation rise and fall in the initial stage of citation curves. The rise and fall of citations must be a citation leaping or like a lognormal shape. Articles with the largest and smallest G_s values are categorized into sleeping beauties (S_3) and flashes in the pan (F_3), respectively. The obsolescence vector for the former (Rayleigh, 1914) is O = (0.892, 0, 98). Although published as early as in 1914, it received no citations until 1992. It does not satisfy S1, since it was not awakened by more than 20 citations within four years after sleeping period. However, it satisfies S_2, since recognition to it was delayed to the last four years of its age. This example reveals the deficiency of S_1. The latter (Ryle & Bailey, 1968) has an obsolescence vector O = (-0.960, -0.480, 44). The article received 26 citations immediately in the publication year, but the number rapidly fell to zero four years later and it was never cited till the end. It satisfies both F_1 and F_2.

Table 2. Comparisons of the three approaches to measuring obsolescence.

G_s	N	$N(A^-<0)$	F_1	S_1	F_2	S_2	F_3	S_3	$F_1\&F_3$	$F_2\&F_3$	$S_1\&S_3$	$S_2\&S_3$
(-1,-0.8]	494	494	41	0	489	0	265	0	34	262	0	0
(-0.8,-0.6]	3,897	3,897	62	6	3,856	0	1,734	0	57	1,704	0	0
(-0.6,-0.4]	6,808	6,808	30	16	5,250	0	0	0	0	0	0	0
(-0.4,-0.2]	7,213	7,213	21	22	985	0	0	0	0	0	0	0
(-0.2,0]	5,477	5,477	7	25	25	0	0	0	0	0	0	0
(0,0.2]	2,894	2,344	7	27	0	15	0	0	0	0	0	0
(0.2,0.4]	1,140	543	5	26	0	228	0	0	0	0	0	0
(0.4,0.6]	348	141	2	7	0	304	0	0	0	0	0	0
(0.6,0.8]	65	17	1	1	0	65	0	37	0	0	1	37
(0.8, 1)	4	0	0	0	0	4	0	3	0	0	0	3
Total	**28,340**	**26,934**	**176**	**130**	**10,605**	**616**	**1,999**	**40**	**91**	**1,966**	**1**	**40**

It seems that the condition $G_s \leq$ -0.6 and $A^- = \frac{1}{2} G_s$ for flashes in the pan is a loose condition, since it yields 1,999 flashes in the pan in the dataset. If it is intensified to be $G_s \leq$ -0.8 and $A^- = \frac{1}{2} G_s$, the number of flashes in the pan shrinks to 262, closer to the result of criterion F_1. Considering that 81.6% of the articles aged over 20, we suggest the criterion for flashes in the pan be $G_s \leq$ -0.8 and $A^- = \frac{1}{2} G_s$ on condition that n≥20.

The criterion S_3 for sleeping beauties is more stringent than S_1 and S_2, and selected only 40 qualified articles from the dataset. The 40 articles is a subset of the collection by S_2, but covers 39 articles out of the collection by S_1. In Table 2, there are six articles satisfying S_1 whose G_s values exist in the interval (-0.8, -0.6]. For example, the article in Figure 2 received only nine citations within the first five years after publication, but suddenly received 25 citation in the following four years. It also satisfies S_2, since it reached "Year 50%" within ten years (13.9% of its age) after publication. Nevertheless, this article is more like a "typical citation curve" which spent seven years to gradually reach citation-peak and slowly declined to death afterwards. The obsolescence vector of this article is O = (-0.648, -0.324, 72) which does not satisfy S_3. Moreover, we identified 3,897 articles of its kind, which have $G_s \in$ (-0.8, -0.6]. Therefore, it is more reasonable to categorize it as a "normal article" rather than a "sleeping beauty".

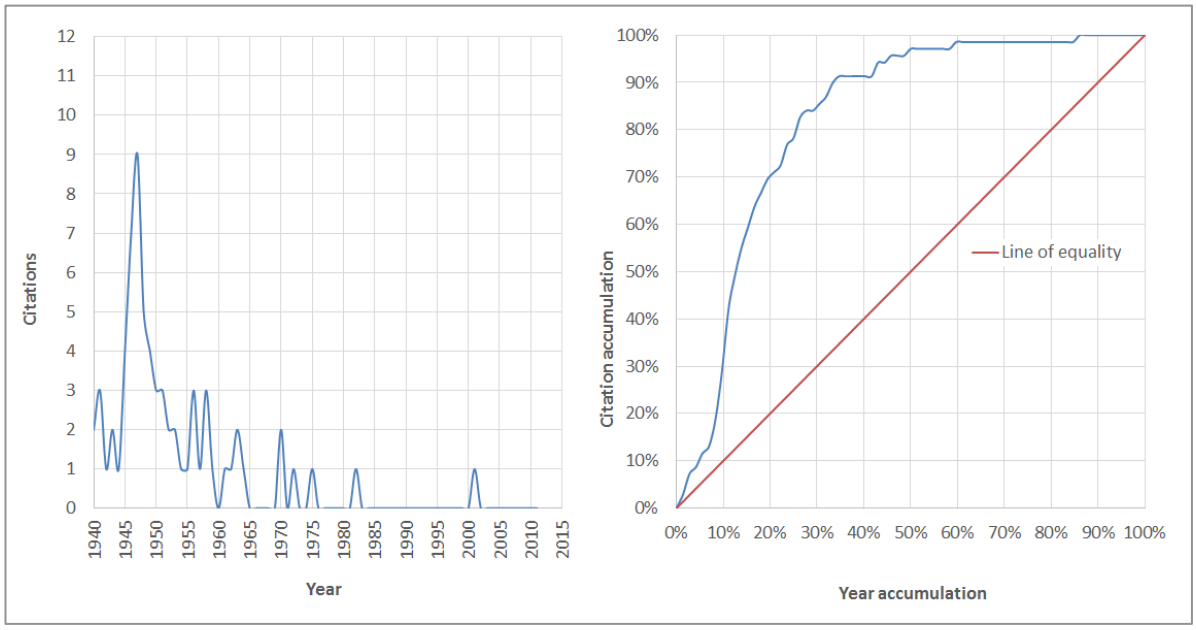


Figure 2. A sleeping beauty by average-based and quartile-based criteria, but a normal article by obsolescence vector (Landsteiner, 1940).

Citation-curve differences of obsolescence

The calculation of G_s values, sometimes, remains citation leaping under cover. As shown in Figures 3, Zewail's and Corey's articles were published in the same year of 2000, and have the same G_s values 0.083. However, they received different citations and have different citation curves. The obsolescence vector of the two articles are O=(0.083, 0, 12) and O=(0.083, -0.004, 12), respectively. Due to the citation leaping since 2007, the normalized curve of Corey's article in Figure 3 surpassed the line of equality in 2010 and yielded $A^- < 0$ which does not appear in the normalized curve of Zewail's article. Therefore, it is a sign of citation leaping to have A^-<0.

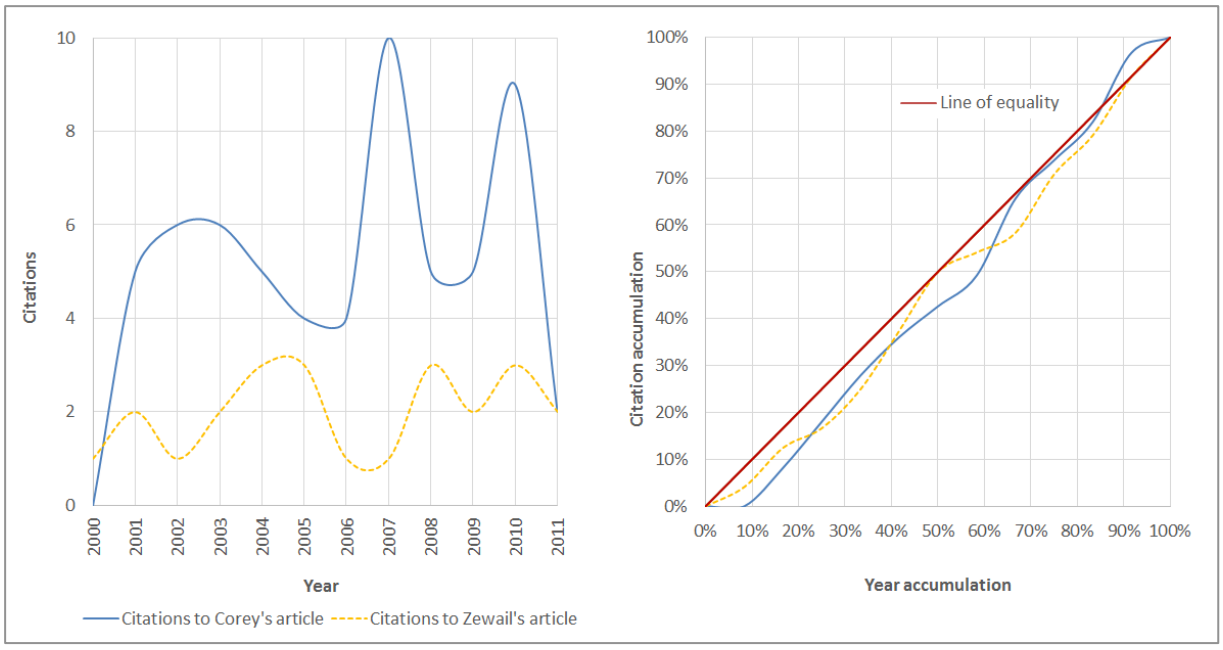


Figure 3. Zewail's article with O = (12, 0.083, 0) and Corey's article with O = (12, 0.083, -0.004).

Age differences of obsolescence

The years of 1950, 1990 and 2000 were selected for the publication years for sampling articles, in order to explore age differences of obsolescence. They were aged 62, 22 and 12, respectively. It appears that older articles have smaller G_s values while younger ones have larger G_s values. It is clear in Table 3 that the peak of G_s distribution among the intervals shifted from (-0.6, -0.4] in 1950, to (-0.4,-0.2] in 1990, even to (-0.2, 0] in 2000. Most of the old articles have been ignored and receive rare or no citations after recognition, similar to the example in Figure 2. Therefore, their G_s values gradually decline. It is hence identified that age exerts significant influence on the values of G_s.

Table 3. Age differences of obsolescence.

G_s	*Year* 1950		*Year* 1990		*Year* 2000	
	N	*N*(A^-<0)	*N*	*N*(A^-<0)	*N*	*N*(A^-<0)
[-1,-0.8]	11	11	12	12	0	0
(-0.8,-0.6]	65	65	45	45	8	8
(-0.6,-0.4]	66	66	190	190	31	31
(-0.4,-0.2]	42	42	250	250	81	81
(-0.2,0]	28	28	148	148	216	216
(0,0.2]	22	16	80	68	173	117
(0.2,0.4]	8	3	27	9	46	10
(0.4,0.6]	6	0	5	2	8	1
(0.6,0.8]	0	0	0	0	0	0
(0.8, 1]	0	0	0	0	0	0
Total	**248**	**231**	**757**	**724**	**563**	**464**

Disciplinary differences of obsolescence

The obsolescence of economic sciences is slower than that of fundamental sciences, including chemistry, physics and physiology & medicine. It is a sign of slow obsolescence to have more positive G_s values and less $A^- < 0$. In Table 4, the distribution of G_s values of economic sciences peaked in the interval (0, 0.2], while in other disciplines, it peaked in the interval (-0.4,-0.2] or (-0.6,-0.4]. The percentage of $A^- < 0$ in positive G_s values is only 50.4%, far less

than 69.8-75.8% in fundamental sciences. Moreover, older articles tend to have higher absolute G_s values, in each of the four disciplines.

Table 4. Disciplinary differences of obsolescence

G_s	*Chemistry*			*Physics*			*Physiology & Medicine*			*Economic sciences*		
	N	*N(A-<0)*	*Age*	*N*	*N(A-<0)*	*Age*	*N*	*N(A-<0)*	*Age*	*N*	*N(A-<0)*	*Age*
[-1,-0.8]	34	34	56.1	124	124	36.4	336	336	51.0	0	0	0.0
(-0.8,-0.6]	625	625	49.8	653	653	35.1	2,615	2,615	45.9	4	4	38.3
(-0.6,-0.4]	1,727	1,727	41.4	1,185	1,185	33.2	3,850	3,850	41.0	44	44	36.2
(-0.4,-0.2]	2,690	2,690	37.5	1,212	1,212	35.0	3,193	3,193	36.2	118	118	36.8
(-0.2,0]	2,236	2,236	35.3	1,008	1,008	34.6	1,972	1,972	30.7	263	263	35.6
(0,0.2]	1,099	926	39.3	576	483	42.2	730	594	34.5	489	341	30.0
(0.2,0.4]	307	161	53.9	289	180	58.9	155	78	49.8	389	124	28.2
(0.4,0.6]	67	34	71.1	147	63	71.9	33	13	60.4	101	31	37.2
(0.6,0.8]	10	3	90.5	38	10	86.9	5	0	47.2	12	4	52.3
(0.8, 1]	0	0	0.0	4	0	90.0	0	0	0.0	0	0	0.0
Total	**8,795**	**8,436**		**5,236**	**4,918**		**12,889**	**12,651**		**1,420**	**929**	

Discussion

Further discussion on $A^- < 0$

Significant citation leaping is likely to result in recurring appearance of $A^-<0$ area. For example of Hsu et al.'s article (1997), citation leaping appeared twice in the citation curve. The first citation peak appeared in 1998, the second year after publication, which led the normalized curve to reach the line of equality. In 1999, the article received six citations. The normalized curve hence surpassed the line of equality. However, the citation leaping disappeared afterwards, and the normalized curve dropped under the line of equality. Nevertheless, the second citation peak, higher than the first one, appeared in 2005 and boosted the normalized curve above the line of equality again. Comparing this example with the supposed waveform citation curves, i.e., P_5 and P_6 in Figure 1, it is identified that the appearance of $A^-<0$ area is originated by citation leaping. Furthermore, double appearance of $A^-<0$ area indicates double citation leaping in which the first one happened immediately after publication and the second one is higher. However, the characteristics of double or multiple appearance of $A^-<0$ area are not in consideration of the new designed obsolescence vector, since the number of this kind is rare.

Limitations

The obsolescence vector cannot differentiate two citation curves if there is multiplier relationship between their annual citations. For example, both (0, 8, 0, 8, 0, 8, 0, 8, 0, 8) and (0, 4, 0, 4, 0, 4, 0, 4, 0, 4) have the same obsolescence vector O=(0.1, 0, 10). The obsolescence vector is applicable to categorize articles into "flashes in the pan", "sleeping beauties" or "normal articles", by distinguishing citation leaping in citation curves. It does not characterize citation history of "normal" articles, which account for a large percent. As normal articles, P_3-P_6 in Figure 1 have entirely different obsolescence patterns. However, they cannot be uncovered by obsolescence vector.

It is controversial whether someone who won a major prize has received increased citations on all his/her work (Hugget, 2013; Mazloumian et al., 2011). However, the results are generalized from articles of Nobel laureates rather than randomly sampled authors, and hence are potentially biased. In addition, "recognition" is referred to as a large number of citations,

e.g., 20. Thus, whether the obsolescence vector is applicable to articles receiving less than 20 citations requires further research.

Conclusions

We proposed a vector for measuring obsolescence of scientific articles, $O = (G_s, A^-, n)$, where n is the age of an article, G_s and A^- are parameters for the shape of the article's citation curves. By distinguishing inequality of citation distribution, obsolescence vector is applicable to categorize articles into three general types:

flashes in the pan: $G_s \leq$ -0.8 and $A^- = \frac{1}{2} G_s$ for n≥20 or $G_s \leq$ -0.6 and $A^- = \frac{1}{2} G_s$ for n<20;
sleeping beauties: $G_s \geq 0.6$ and $A^- = 0$; and
normal articles: which neither satisfy the criteria for F_3 nor for S_3.

The age, subject category and citation curve of articles exert significant influence on G_s values. Older articles tend to have higher absolute G_s values. The criterion for "flashes in the pan" is adjustable in terms of the age of articles. In case of articles younger than, e.g., ten years old, as shown in Figure 1, it is feasible to mildly adjust the criterion as $G_s \leq$ -0.6. Disciplinary differences exist in the proposed obsolescence vector. Articles in economic sciences appear higher G_s values than those in fundamental sciences, including chemistry, physics and physiology & medicine. In case of articles receiving no more citations, their G_s values tend to decline, till to -1.

As an alternative to average-based and quartile-based criteria, the obsolescence vector avoided overlooking the period after sleeping beauties being awakened, and tightened the loose conditions by using quartiles. By obsolescence vectors, we identified 265 flashes in the pan (approximately 1%) and 40 sleeping beauties (approximately 0.1%), among 28,340 articles of Nobel laureates, which receive more than 20 citations by the year of 2011. The low percentages of flashes in the pan and sleeping beauties remained them rare phenomena.

Acknowledgement

This research was financially supported by the National Natural Science Foundation of China (NSFC No. 71203193 and 71273125).

References

ABT, H. A. (1981). Long-term citation histories of astronomical papers. *Publications of the Astronomical Society of the Pacific*, *93*, 207-210.

Avramescu, A. (1979). Actuality and obsolescence of scientific literature. *Journal of the American Society for Information Science, 30*(5), 296-303.

Burrell, Q. L. (2002). The nth-citation distribution and obsolescence. *Scientometrics*, *53*(3), 309-323.

Burton, R. E., & Kebler, R. W. (1960). The "half-life" of some scientific and technical literatures. *American Documentation*, *11*(1), 18-22.

Cole, S. (1970). Professional standing and the reception of scientific discoveries. *American Journal of Sociology*, *76*, 286–306.

Costas, R., van Leeuwen, T. N., & van Raan, A. F. J. (2010). Is scientific literature subject to a "sell-by-date"? A general methodology to analyze the "durability" of scientific documents. *Journal of the American Society for Information Science and Technology*, *61*(2), 329–339.

Cunningham, S. J., & Bocock, D. (1995). Obsolescence of computing literature. *Scientometrics*, *34*(2), 255-262.

Egghe, L.,, & Rao, I. K. R. (1992). Citation age data and the obsolescence function: Fits and explanations. *Information and Processing Management, 28*(2), 201-217.

Garfield, E. (1980). Premature discovery or delayed recognition-why? *Current Contents*, *4*, 488-493.

Garfield, E. (1989). More delayed recognition. Part 1. Examples from the genetics of color blindness, the entropy of short-term memory, phosphoinositides, and polymer rheology. *Current Contents*, *38*, 3-8.

Guo, J. L., & Suo, Q. (2014). Comment on" Quantifying Long-term Scientific Impact". *Science*, *345*(6193), 149.

Hugget, S. (2010). Does a Nobel Prize lead to more citations. *Research Trends*, Retrieved April 7, 2015 from http: //www.researchtrends.com/issue20-november-2010/does-a-nobel-prize-lead-to-more-citations.

Hsu, C. P., Song, X.,, & Marcus, R. A. (1997). Time-dependent Stokes shift and its calculation from solvent dielectric dispersion data. T*he Journal of Physical Chemistry B*, *101*(14), 2546-2551.

Li, J. (2014). Citation Curves of "All-elements-sleeping-beauties": "Flash in the Pan" first and then "Delayed Recognition". *Scientometrics*, *100*(2), 595-601.

Li, J., & Ye, F. Y. (2012). The phenomenon of all-elements-sleeping-beauties in scientific literature. *Scientometrics*, *92*(3), 795–799.

Li, J., & Ye, F. Y. (2014). A Probe into the Citation Patterns of High-quality and High-impact Publications. *Malaysian Journal of Library and Information Science*, *19*(2), 31-47.

Li, J., Shi, D., Zhao, S. X., & Ye, F. Y. (2014). A study of the "heartbeat spectra" for "sleeping beauties". *Journal of Informetrics*, *8*(3), 493-502.

Mazloumian, A., Eom, Y., Helbing, D., Lozano, S., & Fortunato, S. (2011). How citation boosts promote scientific paradigm shifts and nobel prizes. *Plos One*, *6*(5), e18975.

Nakamoto, H. (1988). Synchronous and dyachronous citation distributions. In L. Egghe, & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 157–163). Amsterdam: Elsevier Science Publishers.

Persson, O. (2005). "Citation Indexes for Science"-A 50 year citation history. *Current Science*, *89*(9), 1503-1504.

Price, D. (1970). Citation measures of hard science, soft science, technology, and non-science. In C. E. Nelson & D. K. Pollock (Eds.), *Communication among Scientists and Engineers* (pp. 3-22). Lexington, MA: Heath.

Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306.

Rayleigh, L. (1914). On the theory of long waves and bores. Proceedings of the royal society of London series A- Containing papers of a mathematical and physical character, 90(619), 324-328.

Redner, S. (2005). Citation statistics from more than a century of physical review. *Physics Today*, *58*(1), 49–54.

Stent, G. S. (1972). Prematurity and uniqueness in scientific discovery. *Scientific American*, *227*(6), 84–93.

van Dalen, H. P., & Henkens, K. (2005). Signals in science – On the importance of signaling in gaining attention in Science. *Scientometrics*, *64*(2), 209–233.

van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472.

Vlachý, J. (1985). Citation histories of scientific publications. The data sources. *Scientometrics*, *7*(3), 505-528.

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127-132.

Wyatt, H. V. (1961). Knowledge and prematurity-journey from transformation to DNA. *Perspectives in Biology and Medicine, 18*(2), 149-156.

Field-Normalized Citation Impact Indicators and the Choice of an Appropriate Counting Method

Ludo Waltman and Nees Jan van Eck

{waltmanlr, ecknjpvan}@cwts.leidenuniv.nl
Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)

Abstract
Bibliometric studies often rely on field-normalized citation impact indicators in order to make comparisons between scientific fields. We discuss the connection between field normalization and the choice of a counting method for handling publications with multiple co-authors. Our focus is on the choice between full counting and fractional counting. Based on an extensive theoretical and empirical analysis, we argue that properly field-normalized results cannot be obtained when full counting is used. Fractional counting does provide results that are properly field normalized. We therefore recommend the use of fractional counting in bibliometric studies that require field normalization, especially in studies at the level of countries and research organizations.

Conference Topic
Citation and co-citation analysis; Indicators

Introduction
In discussions on bibliometric indicators, two topics that receive a considerable amount of attention are field normalization and counting methods. Field normalization is about the problem of correcting for differences in citation practices between scientific fields. The challenge is to develop citation-based indicators that allow for valid between-field comparisons. Counting methods are about the way in which co-authored publications are handled. For instance, if a publication is co-authored by two countries, should the publication be counted as a full publication for each country or should it be counted as half a publication for each country?

The topics of field normalization and counting methods are usually discussed separately from each other. However, we argue that there is a close connection between the two topics. Our argument is that proper field normalization is possible only if a suitable counting method is used. In particular, we claim that properly field-normalized results cannot be obtained when one uses the popular full counting method, in which co-authored publications are fully assigned to each co-author. The fractional counting method, which assigns co-authored publications fractionally to each co-author, does provide properly field-normalized results. The problem of full counting basically is that co-authored publications are counted multiple times, once for each co-author, which creates a bias in favor of fields in which there is a lot of co-authorship and in which co-authorship correlates with additional citations. This is the essence of the argument that we present in this paper. Our argument builds on an earlier paper (Waltman et al., 2012), but in the present paper we elaborate the argument in more detail and we also present an extensive empirical analysis.

This paper is a shortened version of a more extensive working paper (Waltman & Van Eck, 2015). The working paper includes additional empirical analyses comparing different counting methods at the level of institutions and countries. Furthermore, the working paper considers different variants of fractional counting and also studies first author and corresponding author counting methods.

Counting methods

Our focus is on the comparison between full counting and fractional counting. In the case of full counting, a publication is fully assigned to each co-author. For instance, a publication co-authored by four countries counts as a full publication for each of the four countries. In the fractional counting case, a publication is fractionally assigned to each co-author. The weight with which a publication is assigned to a co-author indicates the share of the publication allocated to that co-author. The sum of the weights of all co-authors of a publication equals one. An example of fractional counting is the situation in which a publication co-authored by four countries is assigned to each country with a weight of 1 / 4 = 0.25.

There is a quite extensive literature on counting methods. Because of space limitations, we mention only a few selected studies. A systematic terminology for counting methods is proposed by Gauffriau, Larsen, Maye, Roulin-Perriard, and Von Ins (2007). They refer to full counting as whole counting and to fractional counting as normalized counting. Gauffriau, Larsen, Maye, Roulin-Perriard, and Von Ins (2008) present a comparison of counting methods at the country level. They also provide an overview of earlier literature on counting methods. Another country-level comparison is reported by Aksnes, Schneider, and Gunnarsson (2012). At the institution level, Waltman et al. (2012) present a comparison between full and fractional counting. Interesting work on counting methods can also be found in various papers by Ruiz-Castillo and colleagues, who propose the idea of a so-called multiplicative counting method (e.g. Albarrán, Crespo, Ortuño, & Ruiz-Castillo, 2010).

Relation between counting methods and field normalization

Our aim in this section is to demonstrate the close connection between counting methods and field normalization. In particular, we aim to make clear that full counting is fundamentally inconsistent with the idea of field normalization. We argue that full counting yields results that suffer from a bias in favor of fields in which there is a lot of co-authorship and in which co-authorship correlates with additional citations. This bias is caused by the fact that co-authored publications are counted multiple times in the case of full counting, once for each co-author.

We present our argument by providing two simple examples. Both examples take countries as the unit of analysis and focus on the mean normalized citation score (MNCS) indicator (Waltman, Van Eck, Van Leeuwen, Visser, & Van Raan, 2011). However, the underlying ideas of the two examples are more general, and similar examples can be given with authors or organizations as the unit of analysis and with other field-normalized indicators.

Table 1. Example involving a single field.

	Authors	No. of cit.	Norm. cit. score
Publication 1	Country A	3	0.6
Publication 2	Country A	6	1.2
Publication 3	Country B	1	0.2
Publication 4	Country A; Country B	10	2.0

Example involving a single field

We consider a world in which there are just four publications. These publications have been produced by two countries, labeled as country A and country B. Table 1 shows for each publication the countries by which the publication is authored and the number of citations the publication has received. The table also shows the normalized citation score of each publication. For simplicity, it is assumed that all four publications are in the same field. The normalized citation score of a publication is therefore obtained simply by dividing the number of citations of the publication by the average number of citations of all four publications. The

average number of citations of the four publications equals (3 + 6 + 1 + 10) / 4 = 5, and therefore the normalized citation score of for instance publication 1 equals 3 / 5 = 0.6. Of course, the average of the normalized citation scores of the four publications equals one.
We now calculate both for country A and for country B the MNCS. Using full counting, we obtain

$$\mathrm{MNCS}_\mathrm{A} = \frac{0.6 + 1.2 + 2.0}{3} = 1.27 \text{ and } \mathrm{MNCS}_\mathrm{B} = \frac{0.2 + 2.0}{2} = 1.10.$$

On the other hand, using fractional counting, we get

$$\mathrm{MNCS}_\mathrm{A} = \frac{1.0 \times 0.6 + 1.0 \times 1.2 + 0.5 \times 2.0}{1.0 + 1.0 + 0.5} = 1.12 \text{ and } \mathrm{MNCS}_\mathrm{B} = \frac{1.0 \times 0.2 + 0.5 \times 2.0}{1.0 + 0.5} = 0.80,$$

where publication 4 has been assigned with a weight of 0.5 to country A and with a weight of 0.5 to country B.
The important thing to observe in this example is that in the case of full counting country A and country B both have an MNCS above one. One of the main ideas of field-normalized indicators such as the MNCS indicator is that the value of one can be interpreted as the world average. Under this interpretation, country A and country B both perform above the world average. Since there are no other countries in our example, the conclusion would be that all countries in the world perform above the world average. There are no countries with a below-average performance. In our opinion, the conclusion that everyone is above average does not make much sense. Moreover, this conclusion is fundamentally different from the conclusion that is reached in the case of fractional counting. Using fractional counting, country A has a performance above the world average while the performance of country B is below the world average.
Looking a bit more in detail at our example, we observe that in the fractional counting case we have

$$\frac{2.5 \times \mathrm{MNCS}_\mathrm{A} + 1.5 \times \mathrm{MNCS}_\mathrm{B}}{2.5 + 1.5} = \frac{2.5 \times 1.12 + 1.5 \times 0.80}{2.5 + 1.5} = 1.$$

Hence, the weighted average of the MNCS of country A and the MNCS of country B, with weights given by each country's fractional number of publications, equals exactly one. This is a general property of fractional counting. The weighted average of the MNCSs of all countries in the world will always be equal to exactly one.
In the full counting case, the weighted average of the MNCS of country A and the MNCS of country B equals

$$\frac{3 \times \mathrm{MNCS}_\mathrm{A} + 2 \times \mathrm{MNCS}_\mathrm{B}}{3 + 2} = \frac{3 \times 1.27 + 2 \times 1.10}{3 + 2} = 1.20,$$

where the weight of each country is given by the number of publications of the country obtained using full counting. So in the full counting case the world average at the country level does not equal one but instead equals 1.20. Taking 1.20 as the world average, we conclude that country A, with an MNCS of 1.27, has an above-average performance while

country B, with an MNCS of 1.10, performs below average. This is in agreement with the conclusion reached using fractional counting.

So in our example there is a difference of 1.20 – 1 = 0.20 between the world average obtained using full counting and the world average obtained using fractional counting. We refer to this difference as the full counting bonus. In principle, the full counting bonus can be either positive or negative, but we will see that in practice the bonus is usually positive. The full counting bonus is caused by the fact that publications co-authored by multiple countries are counted multiple times in the case of full counting, and therefore the citation impact of multi-country publications relative to single-country publications determines whether the full counting bonus is positive or negative. The bonus will be positive if publications co-authored by multiple countries receive more citations than publications authored by a single country. Conversely, a negative bonus will be obtained if multi-country publications are cited less frequently than single-country publications. As can be seen in Table 1, in our example the only publication co-authored by multiple countries is publication 4, and this is also the most highly cited publication. In the full counting case, publication 4 is fully assigned both to country A and to country B. Hence, the most highly cited publication in our example is counted two times, once for country A and once for country B. This double counting of publication 4 explains why both countries have an MNCS above one and why the full counting bonus is positive.

Example involving multiple fields

In the example discussed above, all publications are in the same field. We now consider an example that involves more than one field. This example is presented in Table 2. There are six publications, three in field X and three in field Y, and there are four countries. Countries A and B are active only in field X, while countries C and D are active only in field Y. The three publications in field X have all received the same number of citations, and therefore these publications all have a normalized citation score of one. This is not the case in field Y, in which publication 6, co-authored by countries C and D, has received more citations than publications 4 and 5, which are single-country publications. Of course, the average normalized citation score of the publications in field Y equals one, just like in field X.

Table 2. Example involving multiple fields.

	Field	Authors	No. of cit.	Norm. cit. score
Publication 1	Field X	Country A	10	1.0
Publication 2	Field X	Country B	10	1.0
Publication 3	Field X	Country A; Country B	10	1.0
Publication 4	Field Y	Country C	4	0.8
Publication 5	Field Y	Country D	4	0.8
Publication 6	Field Y	Country C; Country D	7	1.4

Using fractional counting, the four countries all have an MNCS of exactly one. For countries A and B this is immediately clear. In the case of countries C and D, the MNCS is calculated as $(1.0 \times 0.8 + 0.5 \times 1.4) / (1.0 + 0.5) = 1$. So fractional counting tells us that all four countries perform at the world average. This is indeed the outcome that we would expect to obtain. The publications of countries A and B have all been cited equally frequently as the average of their field, so countries A and B obviously perform at the world average. In the case of countries C and D, we observe that these countries have exactly the same performance and that they are the only countries active in field Y. Based on these two observations, it is natural to conclude that the performance of countries C and D is at the world average.

We now consider the full counting case. Using full counting, countries A and B have an MNCS of one, while countries C and D have an MNCS of $(0.8 + 1.4) / 2 = 1.10$. The full

counting results seem to suggest that countries C and D have a better performance than countries A and B. However, a more careful analysis shows that this is not a correct interpretation of the results. To see this, we calculate both for field X and for field Y the average of the MNCSs of the countries active in the field. The average MNCS of the countries active in field X equals one, while the average MNCS of the countries active in field Y equals 1.10. Hence, both countries A and B active in field X and countries C and D active in field Y perform at the world average of their field. Like in the fractional counting case, we conclude that all four countries have an average performance. Countries C and D have a higher MNCS than countries A and B only because they are active in a field with a higher full counting bonus. Field Y has a full counting bonus of $1.10 - 1 = 0.10$, while the full counting bonus in field X equals zero.

Conclusions based on the examples

Based on the above examples, two important conclusions can be drawn. The first conclusion is that there is a need to carefully distinguish between two field normalization concepts. We refer to these concepts as weak field normalization and strong field normalization. Weak field normalization requires the average of the normalized citation scores of all publications in a field to be equal to one. Strong field normalization is more demanding. It requires the weighted average of the MNCSs of all countries active in a field to be equal to one, where the weight of a country is given by its number of publications in the field.

As shown in the above examples, full counting yields results that are in agreement with the idea of weak field normalization, but these results may violate the idea of strong field normalization. For instance, in the first example discussed above, the average normalized citation score of the four publications equals one (weak field normalization), but the average MNCS of the two countries does not equal one (no strong field normalization). Fractional counting results, on the other hand, satisfy not only the idea of weak field normalization but also the idea of strong field normalization. Using fractional counting, the weighted average of the MNCSs of all countries active in a field will always be equal to one.

When citation-based indicators are calculated using full counting, there is a risk of misinterpretation. People may confuse the concepts of weak and strong field normalization, and they may fail to understand that the idea of strong field normalization does not apply in the case of full counting. In the second example presented above, they may for instance draw the incorrect conclusion that countries C and D perform above the world average. In the fractional counting case, people will not draw such an incorrect conclusion, because fractional counting results are in agreement with the idea of strong field normalization.

We now turn to the second conclusion that follows from our examples. The fact that full counting yields results that are incompatible with the idea of strong field normalization may in itself be regarded as just a minor issue. Instead of having a world average of one, the average of all countries in the world may for instance be equal to 1.10 or 1.20. Although a world average of one might be somewhat more convenient, the exact value of the world average may in the end seem to be of limited importance.

However, our second conclusion is that deviations of the world average from one actually do have serious consequences, at least when making comparisons between fields. This is what is shown in the second example given above. Using full counting, the average MNCS of the countries active in field X equals one, while the average MNCS of the countries active in field Y equals 1.10. So in field X the world average equals one, while in field Y we have a world average of 1.10. Direct comparisons of the MNCSs of the countries active in field X and the countries active in field Y therefore do not yield valid conclusions. Based on their MNCSs, the countries active in field Y seem to perform better than the countries active in field X, but

taking into account the fact that field Y has a higher world average than field X, it actually should be concluded that all countries perform at the same level.

Essentially, the second conclusion that we draw based on our examples is that full counting is fundamentally inconsistent with the idea of field normalization. Citation-based indicators calculated using full counting yield results that do not allow for valid comparisons between fields, and this is the case even when field-normalized indicators, such as the MNCS indicator, are used. When full counting is used in the calculation of field-normalized indicators, countries that focus their activity on fields with a high full counting bonus have an advantage over countries that are active mainly in fields with a low full counting bonus. Fractional counting does not suffer from this problem. Fractional counting results are compatible with the idea of strong field normalization, and these results therefore do allow for proper between-field comparisons.

Empirical analysis of the full counting bonus

In the previous section, we have introduced the idea of the full counting bonus and we have illustrated this idea using theoretical examples. In this section, we present a large-scale empirical analysis of the full counting bonus. This analysis for instance makes clear which fields benefit most from the full counting bonus, and the analysis shows the differences between fields caused by the bonus.

Calculation of the full counting bonus

We first explain in more detail the way in which we calculate the full counting bonus. For simplicity, we assume that our interest is in the full counting bonus at the level of countries. However, the full counting bonus can be calculated in a similar way at the level of for instance authors or organizations.

Suppose we have a set of n publications. This could be for instance the set of all publications in a specific field and in a specific year. For each publication i, we have a citation score c_i. The citation score of a publication can be defined in different ways. It may be simply the number of times a publication has been cited, but it may also be something more advanced, for instance a field-normalized citation score. We also know for each publication the countries by which the publication has been co-authored. We use m_i to denote the number of countries that have co-authored publication i.

In order to obtain the full counting bonus, we first calculate for each country the average citation score of its publications. We perform this calculation both using full counting and using fractional counting. Next, we calculate a weighted average of the average citation scores of all countries. In the case of full counting, we use the number of publications of a country obtained using full counting as the weight of the country. In the case of fractional counting, we use a country's number of publications obtained using fractional counting as the country's weight. Finally, we calculate the full counting bonus as the difference between the weighted average in the full counting case and the weighted average in the fractional counting case.

The above approach to calculating the full counting bonus is somewhat complicated. However, a mathematically equivalent but much simpler approach is available. In this approach, the full counting bonus is calculated as

$$\mathrm{FCB} = \frac{\sum_{i=1}^{n} m_i c_i}{\sum_{i=1}^{n} m_i} - \frac{\sum_{i=1}^{n} c_i}{n},$$

where the first term equals the above-mentioned weighted average in the full counting case while the second term equals the weighted average in the fractional counting case. In the first term, the citation score c_i of publication i co-authored by m_i countries is counted m_i times. This is because in the full counting case publication i is fully assigned to each of the m_i countries. In the second term, the citation score c_i of publication i is counted only once, regardless of the number of countries m_i by which publication i has been co-authored. This is because in the fractional counting case the total weight with which publication i is assigned to the m_i countries equals one.

In our empirical analysis, we consider two definitions of the citation score of a publication. Both definitions include a normalization for field. In the first definition, the citation score of a publication is obtained by dividing the number of citations of the publication by the average number of citations of all publications in the same field and in the same year. Averaging the citation scores of multiple publications then gives us the MNCS indicator. This indicator was also used in the theoretical examples presented in the previous section. In the second definition of the citation score of a publication, we determine whether a publication belongs to the top 10% most frequently cited publications of its field and publication year. A publication belonging to the top 10% has a citation score of one, while a publication belonging to the bottom 90% has a citation score of zero. When this second definition is used, averaging the citation scores of multiple publications yields the $PP_{top\ 10\%}$ indicator, where $PP_{top\ 10\%}$ stands for the proportion of top 10% publications (Waltman et al., 2012; Waltman & Schreiber, 2013). When the full counting bonus is calculated for the set of all publications in a specific field and in a specific year, the second term in the above equation for the full counting bonus will be equal to one in the case of our first definition of the citation score of a publication. This term will be equal to 0.1 (or 10%) in the case of our second definition.

Empirical results

We perform our analysis using the Web of Science (WoS) database. The analysis is based on publications in the period 2009–2010. Only publications of the WoS document types 'article' and 'review' are taken into account. A four-year citation window is used, including the year in which a publication appeared. For the purpose of the calculation of the field-normalized citation scores of publications, fields are defined by the WoS journal subject categories.

We consider three units of analysis: Authors, organizations, and countries. To determine the number of organizations and the number of countries by which a publication has been co-authored, we take into account both the regular addresses of the publication and the reprint address. The number of organizations and the number of countries of a publication is obtained by counting the number of distinct organization names and the number of distinct country names mentioned in the addresses of the publication.

The full counting bonus depends on two factors. On the one hand, it depends on the variation among publications in the number of authors, organizations, or countries. For instance, if all publications have the same number of authors, there can be no full counting bonus at the level of authors. On the other hand, the full counting bonus also depends on the relation between the number of authors, organizations, or countries of a publication and the citation score of the publication. There can for instance be no author-level full counting bonus if publications with different numbers of authors on average all have the same citation score.

Figure 1 presents the distribution of publications based on their number of authors, organizations, and countries. Not surprisingly, the figure shows that the variation among publications in the number of authors is largest while the variation among publications in the number of countries is smallest. Figure 2 presents the relation between the number of authors, organizations, and countries of a publication and the average citation score given by the MNCS indicator. In general, an increasing relation can be observed between the number of

authors, organizations, and countries of a publication and the average citation score. The relation is strongest for countries and weakest for authors. In fact, when the number of authors is between two and five, there is hardly any dependence of the average citation score of a publication on the number of authors. Publications with three or four authors on average even have a slightly lower citation score than publications with two authors. Results for the $PP_{top\ 10\%}$ are not shown, but are similar to the results for the MNCS indicator.

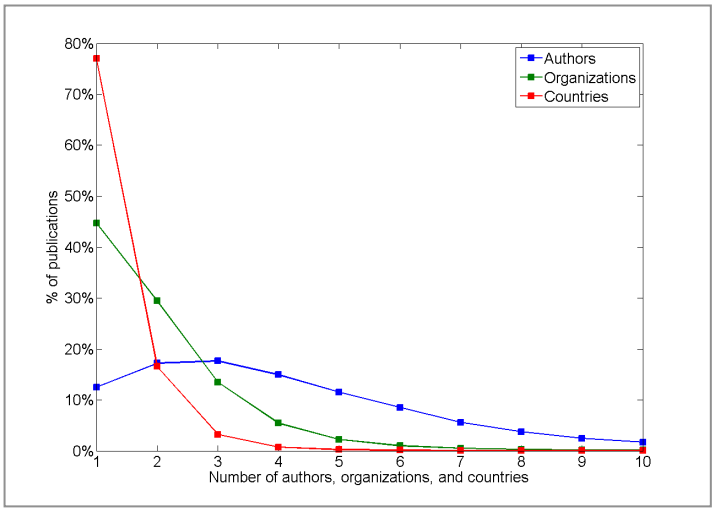


Figure 1. Distribution of publications based on their number of authors, organizations, and countries.

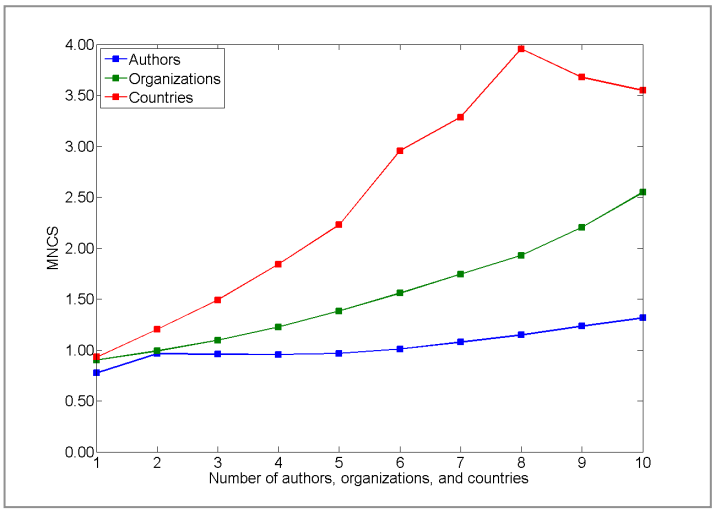


Figure 2. Relation between the number of authors, organizations, and countries of a publication and the MNCS indicator.

Figures 1 and 2 make clear that publications often have multiple co-authors and that the citation impact of a publication tends to increase with the number of co-authors. Co-authored publications are counted multiple times in the case of full counting, and our expectation based on Figures 1 and 2 therefore is to observe full counting bonuses that are positive and of significant size. This is indeed what is reported in Tables 3 and 4. The tables show the full counting bonus at the level of authors, organizations, and countries for five broad fields of science and also for all fields of science taken together. Table 3 relates to the MNCS indicator, while Table 4 relates to the $PP_{top\ 10\%}$ indicator. In order to facilitate comparison between the results obtained for the two indicators, the full counting bonus is presented as a percentage of the average value of the indicator. For instance, in the case of the MNCS

indicator, we obtain a full counting bonus of 0.248 at the level of authors for all fields of science. The average value of the MNCS indicator equals one, and therefore the full counting bonus is reported as 0.248 / 1 = 24.8% in Table 3. Likewise, the $PP_{top\ 10\%}$ indicator has an average value of 0.1 (or 10%), and therefore a full counting bonus of 0.0304 (or 3.04%) is reported as 0.0304 / 0.1 = 30.4% in Table 4.

Table 3. Full counting bonus for the MNCS indicator at the level of authors, organizations, and countries, including a breakdown into five broad fields of science.

	Authors	Organizations	Countries
All fields	24.8%	21.1%	12.6%
Biomedical and health sciences	20.9%	26.8%	16.7%
Life and earth sciences	14.7%	16.2%	12.7%
Mathematics and computer science	8.2%	8.0%	6.9%
Natural sciences and engineering	35.2%	19.3%	10.8%
Social sciences and humanities	14.7%	11.2%	5.6%

Table 4. Full counting bonus for the PPtop 10% indicator at the level of authors, organizations, and countries, including a breakdown into five broad fields of science.

	Authors	Organizations	Countries
All fields	30.4%	26.5%	17.1%
Biomedical and health sciences	24.9%	34.5%	22.6%
Life and earth sciences	22.8%	24.3%	19.7%
Mathematics and computer science	11.3%	11.3%	9.7%
Natural sciences and engineering	43.3%	20.6%	13.0%
Social sciences and humanities	21.3%	17.2%	8.3%

Based on the results for the MNCS indicator presented in Table 3, a number of conclusions can be drawn. At all three analysis levels (i.e., authors, organizations, and countries), there turns out to be a full counting bonus that is positive and of significant size. In general, the bonus is highest at the level of authors and lowest at the level of countries. We have seen in Figure 2 that the number of countries of a publication has a much stronger effect on a publication's citation score than the number of authors, but apparently this is offset by the fact that publications with a large number of countries occur much less frequently than publications with a large number of authors, as shown in Figure 1. The full counting bonus at the level of organizations is generally in between the country-level and author-level bonuses, although there are two main fields (i.e., 'Biomedical and health sciences' and 'Life and earth sciences') in which the organization-level bonus is higher than the author-level one.

The results reported in Table 3 also indicate that at the levels of authors and organizations the full counting bonus is lowest in the 'Mathematics and computer science' main field. At the country level, 'Social sciences and humanities' is the main field with the lowest bonus. The 'Natural sciences and engineering' main field has the highest bonus at the level of authors, while the highest bonus at the organization and country level can be found in the 'Biomedical and health sciences' main field.

The results for the $PP_{top\ 10\%}$ indicator reported in Table 4 are quite similar to the MNCS results presented in Table 3. However, full counting bonuses turn out to be consistently higher for the $PP_{top\ 10\%}$ indicator than for the MNCS indicator.

More detailed results at the level of 250 WoS journal subject categories can be found in an Excel file that is available at www.ludowaltman.nl/counting_methods/. The Excel file also indicates how the five main fields listed in Tables 3 and 4 are defined in terms of the WoS journal subject categories. There turn out to be rather large differences between subject categories in the full counting bonus. For instance, the subject categories with the highest MNCS full counting bonus at the level of organizations and countries are 'Medicine, general

& internal' and 'Physics, nuclear'. The subject categories have bonuses of, respectively, 148% and 176% at the organization level and 89% and 70% at the country level. Other subject categories have bonuses that are close to zero or even negative. Examples of such subject categories include 'Chemistry, organic' and 'Ergonomics'.
It is important to be aware of the consequences of the large differences between subject categories in the full counting bonus. Consider a university that has a full counting MNCS of 2.50 in the 'Medicine, general & internal' subject category and a full counting MNCS of 1.00 in the 'Chemistry, organic' subject category. What should we conclude based on these values? The obvious conclusion may seem to be that in terms of citation impact our university is performing much better in the 'Medicine, general & internal' subject category than in the 'Chemistry, organic' subject category. However, this conclusion does not take into account the effect of the full counting bonus. As mentioned above, the 'Medicine, general & internal' subject category has an organization-level full counting bonus of almost 150%, while the full counting bonus for the 'Chemistry, organic' subject category is close to zero. Taking into account the effect of the full counting bonus, we need to conclude that in both subject categories our university performs around the average level of all organizations worldwide.

Commonly used arguments in favor of full counting

In practice, most bibliometric analyses use full counting instead of fractional counting. Below we list three arguments that are often given to argue against the use of fractional counting and to justify the use of full counting. We also provide a response to each argument.

Argument 1: The different co-authors of a publication usually have not contributed equally. By giving equal weight to each co-author, fractional counting fails to properly represent the contributions made by the different co-authors. Hence, giving equal weight to each co-author is arbitrary and lacks a sound justification.

It is true that there can be large differences between co-authors in the contribution they have made to a publication. At the level of an individual publication, fractional counting may therefore significantly misrepresent the contributions made by individual co-authors. However, at the level of a large set of publications, for instance all publications of an organization or a country, we believe that it is reasonable to assume that the error will be within an acceptable margin. This is because errors at the level of individual publications are likely to cancel out. The contribution of an organization or a country to certain publications may be overestimated, but most probably there will then be other publications for which the contribution of this organization or this country is underestimated.
Furthermore, the argument that giving equal weight to each co-author of a publication is arbitrary may equally well be used as an argument against full counting. Like fractional counting, full counting gives the same weight to each co-author of a publication.

Argument 2: Fractional counting provides an incentive against collaboration, which is often considered undesirable.

We believe that citation impact and collaboration represent different dimensions of scientific performance and that in general these dimensions can best be measured separately from each other. Citation-based indicators should be assessed based on the degree to which they measure citation impact in an accurate way. In this respect, we believe that for many purposes fractional counting performs better than full counting. If in addition to citation impact one also considers collaboration to be a relevant dimension of scientific performance, then additional indicators should be used to measure this dimension. If one desires to do so, these indicators can then be used to provide an incentive to collaboration. By assessing citation-

based indicators based on the effect they may have on collaboration, one fails to make a proper distinction between the citation impact dimension of scientific performance and the collaboration dimension.

Argument 3: Fractional counting is more difficult to understand and less intuitive than full counting.

To a certain degree, we agree with this argument. Fractional counting yields non-integer publication and citation counts. These non-integer counts are more difficult to understand and require more explanation than the integer publication and citation counts provided by full counting. Fractional counting may also be less intuitive than full counting. For instance, consider a researcher who has produced some of his publications on his own while he has produced other publications with one or two co-authors. The researcher may feel that his co-authored publications are of similar importance to his oeuvre as his single-author publications. However, fractional counting gives less weight to the co-authored publications of the researcher than to his single-author publications. This is not in agreement with the feelings the researcher has about the importance of the different publications in his oeuvre, and therefore from the point of view of the researcher fractional counting can be regarded as less intuitive than full counting.

On the other hand, from a different point of view, it can also be argued that fractional counting is actually more intuitive than full counting. Earlier in this paper, we have given two examples showing that field-normalized citation impact indicators calculated using full counting can easily be misinterpreted. Field-normalized indicators calculated using fractional counting are much more easy to interpret in a correct way. As we have explained, this is because indicators based on fractional counting yield results that are compatible with the idea of strong field normalization. Unlike full counting indicators, fractional counting indicators therefore allow comparisons between fields to be performed in an easy and intuitive way. So from this point of view indicators based on fractional counting can be considered more intuitive than their full counting counterparts.

Conclusions

In this paper, we have presented a new perspective on the choice between different counting methods, leading to an important new argument in favor of fractional counting. Building on our earlier work (Waltman et al., 2012), this argument is based on the observation that the problem of choosing an appropriate counting method is closely connected to the problem of field normalization of citation-based indicators.

We have argued that from a field normalization point of view fractional counting is preferable over full counting. As we have shown, properly field-normalized results cannot be obtained using full counting, and field-normalized indicators calculated using full counting can easily be misinterpreted. Fractional counting does provide properly field-normalized results, and these results can be interpreted in a much more straightforward way than results obtained using full counting. Essentially, the problem of full counting is that co-authored publications are counted multiple times, once for each co-author, which creates an unfair advantage to fields with a lot of co-authorship and with a strong correlation between co-authorship and citations. For instance, the average full counting MNCS of all organizations or all countries active in these fields is significantly higher than one. On the other hand, fields in which co-authorship is less common or in which co-authorship does not correlate with citations are disadvantaged. Full counting yields results that are biased against organizations and countries whose activity is focused on these fields. Fractional counting does not suffer from this problem. In the case of fractional counting, each publication is counted only once, regardless

of its number of co-authors, and this ensures that comparisons between fields can be made in an unbiased way.

What are the practical implications of the analysis presented in this paper? In our view, this depends on the level of aggregation at which a bibliometric study is performed. In the case of a study at a high aggregation level, such as the level of countries or organizations (e.g., university rankings), we consider it absolutely essential to use fractional counting instead of full counting. At this level, there is a serious risk of misinterpretation of full counting results. Moreover, we believe that arguments in favor of full counting, such as the ones discussed in the previous section, are of limited relevance at a high aggregation level.

The situation is more difficult at a low level of aggregation, for instance at the level of researchers or research groups. At this level, we believe that reasonable arguments can be given in favor of both full and fractional counting. Especially the third argument discussed in the previous section plays an important role at this level. As pointed out in this argument, full counting is in agreement with the intuitive idea that all publications of a researcher or a research group should be considered of equal importance.

However, there is a more fundamental reason why the argument presented in this paper in favor of fractional counting is less relevant at a low level of aggregation. The argument depends on the connection between counting methods and field normalization, but the entire idea of field normalization may be seen as problematic at a low aggregation level. Field-normalized indicators have a limited accuracy (e.g., Van Eck, Waltman, Van Raan, Klautz, & Peul, 2013), and it is questionable whether these indicators are sufficiently accurate for applications at a low aggregation level. If the accuracy of field-normalized indicators at a low aggregation level is considered insufficient, the argument presented in this paper in favor of fractional counting has no relevance at this level.

In this paper, we have not shown how results obtained using full and fractional counting differ in practice. We refer to our working paper (Waltman & Van Eck, 2015) for an extensive comparison of full and fractional counting in bibliometric studies at the level of institutions and countries. The working paper also considers different variants of fractional counting, and it studies first author and corresponding author counting methods.

References

Aksnes, D.W., Schneider, J.W., & Gunnarsson, M. (2012). Ranking national research systems by citation indicators. A comparative analysis using whole and fractionalised counting methods. *Journal of Informetrics*, *6*(1), 36-43.

Albarrán, P., Crespo, J.A., Ortuño, I., & Ruiz-Castillo, J. (2010). A comparison of the scientific performance of the U.S. and the European Union at the turn of the 21st century. *Scientometrics*, *85*(1), 329-344.

Gauffriau, M., Larsen, P.O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2007). Publication, cooperation and productivity measures in scientific research. *Scientometrics*, *73*(2), 175-214.

Gauffriau, M., Larsen, P.O., Maye, I., Roulin-Perriard, A., & Von Ins, M. (2008). Comparisons of results of publication counting using different methods. *Scientometrics*, *77*(1), 147-176.

Van Eck, N.J., Waltman, L., Van Raan, A.F.J., Klautz, R.J.M., & Peul, W.C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE*, *8*(4), e62395.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., Van Eck, N.J., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419-2432.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *64*(2), 372-379.

Waltman, L., & Van Eck, N.J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. arXiv:1501.04431.

Waltman, L., Van Eck, N.J., Van Leeuwen, T.N., Visser, M.S., & Van Raan, A.F.J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37-47.

Forecasting Technology Emergence from Metadata and Language of Scientific Publications and Patents[1]

Olga Babko-Malaya, Andy Seidel, Daniel Hunter, Jason HandUber, Michelle Torrelli and Fotis Barlos

{olga.babko-malaya, andy.seidel, daniel.hunter, jason.handuber, michelle.torrelli, fotis.barlos}@baesystems.com
BAE Systems, Burlington, MA 01803

Abstract
This paper describes a multidisciplinary study and development effort to analyze full text and metadata of scientific articles and patents for indicators of new disruptive and game-changing technical breakthroughs. The system we are developing can scan millions of documents in two languages, English and Chinese, and extract meaningful trends and predictions. Whereas traditional approaches to innovation analytics rely on citation analysis to analyze impact or identify the most influential patents or researchers in the field, our system takes a step further and combines these methods with an analysis of text in order to identify and characterize emerging technologies. The paper describes the indicators and forecasting models, as well as presents the results of applying these indicators to forecast levels of interest in a particular technology based on the analysis of English and Chinese patents. It further shows how the indicators we developed can provide insights into the nature and the lifecycle of emerging technologies.

Conference Topic
Indicators

Introduction
This paper describes Abductive Reasoning Based on Indicators and Topics of EmeRgence, or ARBITER, an automated system whose purpose is to identify and characterize emerging technologies and emerging fields in science. It does so by processing very large collections of scientific publications and patents in multiple languages and identifies trends, associations, and predictions more rapidly than with current methods. Unlike previous approaches to detecting emergence, which are based on the citation analysis of papers and patents (e.g. Bettencourt et al., 2008; Shiebel et al., 2010; Roche et al., 2010), we are extracting information from the text of publications and patents, identifying authors, their affiliations, addresses, as well as classifying types of organizations and publications. Moreover, we apply natural language processing technologies to extract scientific terminology from the full text of the documents, to identify different types of relationships between citations, authors, terms, and organizations, including contrast, opinion, and related work, and to characterize maturity and other properties of terms based on their contextual patterns. This diverse set of features enables us to efficiently process multiple collections and various types of data without dependency on the presence of a specific feature in a collection. For example, our approach is not hampered by the lack of prior art references in Chinese patents, which is a problem for a standard, citation-based analysis of innovative technologies.

To define indicators of emergent technologies and scientific fields, we have developed a pragmatic theory of technoscientific emergence, described in Brock et al. (2012), which builds on Actant Network Theory (Latour, 2005). An Actant Network is a heterogeneous network of human and non-human elements, including people, institutions, funders, meetings, documents, and scientific terminology, interconnected by disparate relationships. The membership of elements within such a network, and the nature and extent of the relationships

[1] Approved for public release; unlimited distribution.

between these elements, is dynamic and constantly changing. To model emergence, we have developed indicators that measure the character and evolution of Actant Networks, including

- Extent of different types of elements in a network, including prolific and prominent entities
- Number of relationships and the volume of traffic in a network
- Growth of entities and relationships, including average growth rate and slope measures
- Novelty of elements and relationships
- Prevalence of the marketplace actant
- Extent of patenting activities
- Amount of disagreements and uncertainties.

In our previous work, we have shown how these indicators can be applied to characterize communities of practice (Babko-Malaya et al., 2013a), identify the presence of the debate in the community (Babko-Malaya et al., 2013b), as well as determine whether practical applications exist for research fields (Thomas et al., 2013). This paper presents the results of applying these indicators to forecast prominence of technology terms, as measured by a significant increase in term frequency. Whereas ARBITER processes both scientific articles and patents, the results presented in this paper are limited to the analysis of patents.

This paper contains three further sections. First, we give an overview of metadata and full text features, describe different categories of indicators designed to identify emerging technologies, as well as demonstrate how the indicators are combined via Bayesian networks into a forecasting model. The next section presents the results of the correlation analysis of indicators with future term prominence for English and Chinese patents, which measures the ability of our indicators to forecast a significant increase in term usage. The final section outlines how the system can be applied to characterize the nature and the lifecycle of the technology.

System Description

Feature Extraction

ARBITER extracts features from the metadata and full text of scientific papers and patents, including Lexis-Nexis Patent data, which includes granted patents and published patent applications from United States and Chinese national patent offices, and Thomson Reuters Web of ScienceTM (abstracts of journals and conference proceedings for the same time period, ~40M records). The features we extract from these sources include metadata features (such as title, author, author affiliation, patent assignees, etc.), as well as features that are based on the analysis of text. All feature extraction capabilities, including language features, are developed for two languages: English and Chinese. A summary of our features is shown in Figure 1. The entities we extract include people, organizations, documents, and scientific terminology, interconnected by different types of relationships.

To analyze persons, we extract authors from scientific articles and inventors from patents. In order to be able to count unique mentions of researchers, we developed a disambiguation component, which groups them into equivalence classes. Our analysis of researchers builds on features such as researcher impact, including Hirsch index and prolificness (measured by patent/paper productivity), as well as co-authorship and citation graphs.

To identify organizations, we extract author affiliations and patent assignees from metadata, as well as funding organizations from the text of acknowledgements and footnotes of scientific papers. All organizations are classified into three classes: Commercial, Academic, and Government/Nonprofit. The organization classification component allows us to evaluate

the extent and changes in the Academic vs. Commercial involvement in a certain field, as well as the diversity of researchers and organizations.

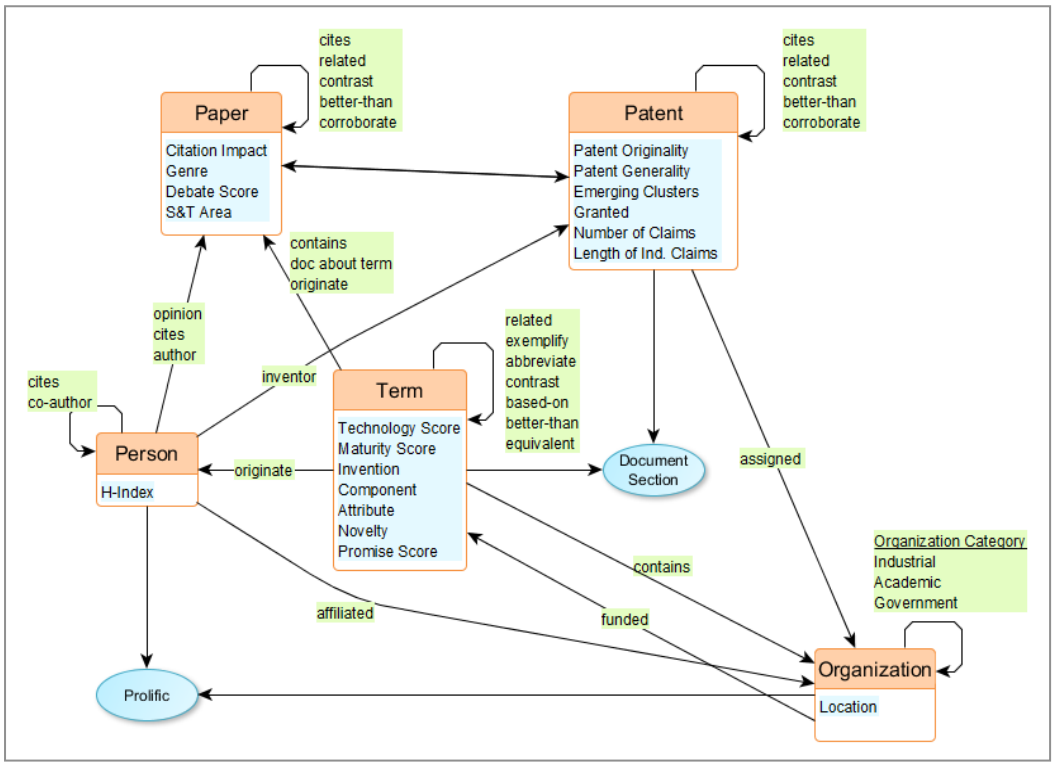


Figure 1. Actant Network extracted from metadata and text.

Our analysis of documents uses citation-based metrics developed by one of our team partners to measure generality, originality, and membership in "emerging clusters" (Breitzman & Thomas, 2015). We further measure mean citation impact of papers and patents, and analyze the structure and length of patent claims.

Our other partners have developed several modules for linguistic processing of text in English and Chinese. For example, to identify scientific terminology, we apply a technology described in Meyers et al. (2010) that extracts scientific noun phrases from the text of papers and patents. The extracted terms are noun phrases that tend to occur frequently in a set of articles from a specific field, but rarely occur in more general or popular articles.

In order to characterize these terms, we score terms based on the extent to which the term behaves like a technology (Anick et al., 2014), as well as assign a maturity score based on how often the term is mentioned in text as being used.

To analyse documents, we apply a genre classifier to evaluate the types of documents that are being published in a certain field, such as review articles or product reviews, as well as to classify documents based on the extent of the debate in the community (Babko-Malaya et al., 2013b). Using the document structure parser, we further identify different sections of documents and categorize claims in patents. To support Chinese extraction, we have adapted a tool to support word segmentation and part of speech tagging to scientific literature and patents (Li & Xue, 2014).

All entities we extract are linked by various types of relations. Whereas some relations are extracted from metadata (e.g. affiliated, invented, assigned, cites, co-author), many relations are extracted from text using information extraction techniques. These relations include opinion relations as well as relations like abbreviate, exemplify, and related work (based on,

better than, contrast, etc), which are described in more detail in Meyers (2013) and Meyers et al. (2014) and are illustrated below.

All entities and relations extracted from full text were evaluated against manually created gold standard corpora. Performance of extraction components is generally comparable across English and Chinese with the f-score above 70-75% in both languages.[2]

Indicators

Using this network, we have developed over 200 indicators that measure different characteristics and changes in the network associated with particular technologies and concepts. The indicators we developed are driven by our pragmatic theory, which defines emergence as the growth in the robustness of actant networks (Brock et al., 2012). The indicators we apply to identify potential disruptive technologies are therefore designed to analyze the relationships between the target entity and other elements in the actant network, including the extent and nature of these relationships, their novelty, dynamic changes, as well as impact, prominence and diversity. Other indicators we explore relate technology emergence to their practicality, as well as the presence of the debate in a community.[3]

Term Momentum Indicators. Our first set of indicators measures momentum in the usage of a particular term. These indicators are time series of annual counts, such as counts of term usage by inventors and organizations, with a further focus on prolific inventors and organizations. In addition, our 'section-based' indicators analyze term usage in independent claims, summary of invention, and abstract sections of patents. The rationale behind an analysis of term usage in specific sections is that these indicators can better measure the extent of the acceptance of the term by the community. For example, if a term occurs in independent claims of patents, it means that it has been legally accepted.

Term Characterization. Beyond indicators based on the momentum associated with individual terms, we also developed indicators that examine different characteristics of these terms. These characteristics include (1) the likelihood that the term describes a technology, (2) the maturity of the technology described by the term, (3) the degree to which the term functions as a description of an invention, and (4) the degree to which a term refers to a component of another technology.

Term characterization scores are calculated by collecting and aggregating evidence from the term's context. For example, to compute maturity scores, we define a set of 'usage' patterns, i.e. patterns that indicate that a term was used or applied: *We used [term] for …, [term] was used for …, employ [term], …* The maturity score is then derived from the number of times these 'usage' patterns are applied to the term. Likewise, the degree to which the term is used as a component is computed based on term usage in 'component'-specific contexts, as illustrated by the sentence *"A typical RFID tag consists of/contains an RFID antenna and RFID chip"*. The terms *RFID antenna* and *RFID chip* are tagged as components in this context, given that they occur as the objects of verbs *consist of* or *contains*. Our expectation is that a time series analysis of maturity of technologies, including their usage as an invention or a component, might be indicative of a change in the lifecycle of a technology, and therefore can be used to identify potentially disruptive technologies (Arthur, 2009).

Semantic Relations. Another class of language-based indicators is based on semantic relations we extract from text. These relations include Opinion, Abbreviate, Exemplify,

[2] Although performance is comparable, there is some variation in the frequency and the type of relations that we extract in the two languages. Some relations are very sparse in Chinese (such as Abbreviations, Contrast, Exemplify (Term1 is an example of Term2). Another difference is that text processing in Chinese is significantly slower than in English due to word segmentation.

[3] The indicators described in this section are focused on the analysis of patents. Similar indicators have also been developed for the analysis of scientific articles, but their analysis is beyond the scope of this paper.

Originate, and different types of Related Work, including Contrast, Based On, and Better Than (Meyers et, 2014). For example, Practical relations represent the author's view that the technology is either being used specially or is useful in some way. Therefore, the indicator that measures the number of Practical relations attached to a term may identify an increase in interest to using a given technology, or its new application. Meanwhile, the relation Abbreviate, which links scientific terms to their abbreviations, can be used to detect the timeline of the acceptance of the term by the community. Finally, relations like Contrast may help to identify the early stages of technology development, given that scientists developing innovative concepts tend to contrast their work with existing research, whereas as the technology becomes more accepted, the number of contrast relations declines.

Document and Inventor Characteristic indicators. This class of indicators measures characteristics of the papers or patents that are using the term. Some of these indicators measure citations to papers containing a given term, or the impact factor of the journals in which the term appears. Others compute dispersion of term usage across technologies or countries, or the number of prior art references in patents.

Inventor Characteristic indicators. In addition to characteristics of documents, we also analyse the inventors and patent assignees who use the term in patents. Examples include the Hirsch index of an inventor or the impact of prior patents granted to inventors or patent assignees.

Novelty. Term Novelty indicators measure the first appearance of a term anywhere in a patent document, as well as the first appearance of a term in specific sections of a patent, such as in the independent claims. Another Novelty indicator computes the first time a term appears with an abbreviation attached. These indicators are thus designed to analyse the timeline of the acceptance of the term by the community.

Most of the indicators described above are time series of annual counts or scores, such as a "number of prominent inventors per year using term in patents." To simplify the modelling process, we reduced each time series to a single value by applying three different methods:

(1) Find the slope of the regression line of indicator values against time (a measure of how fast the indicator is increasing over time);

(2) Calculate the average growth rate for the indicator value over the period selected for the time series;

(3) Compute the sum of indicator values for three years prior to the reference period.

We also experimented with (a) the x2 coefficient of the best-fitting, second-order polynomial for indicator value as a function of year (a measure of curvature, or rate of acceleration), and (b) the two-year prediction of this best-fitting polynomial. These indicators, while sometimes informative, were usually redundant with slope.

Forecasting Models

Our models are tree-augmented Naive Bayes networks (Friedman et al., 1997). Such networks have a structure like that of the network shown in Figure 2. For clarity, we display only a fragment of the model; a complete model may contain 30 to 50 indicator variables.

Bayesian networks provide a factorized representation of a joint probability distribution over a set of variables, and efficiently update the distribution, given evidence in the form of values for variables. In our models, there is a unique root node that represents the unobserved future prominence of an entity. In the above model, this is the node labeled "Prominence3." Prominence is normalized to be between 0 and 1, with a special value of -1 for cases in which the usage of the term decreases. As evidence is entered into the net, the probability distribution over the possible values of prominence is updated.

Bayesian Networks have shown good performance as classifiers (Friedman et al., 1997). We use a version of a Bayesian classifier in which links between indicator variables capture

synergistic effects among those variables – i.e. information about two or more variables tells us more about prominence than the sum of the information value of the individual variables. Capturing synergistic effects has been shown to improve classifier performance (Friedman et al., 1997).

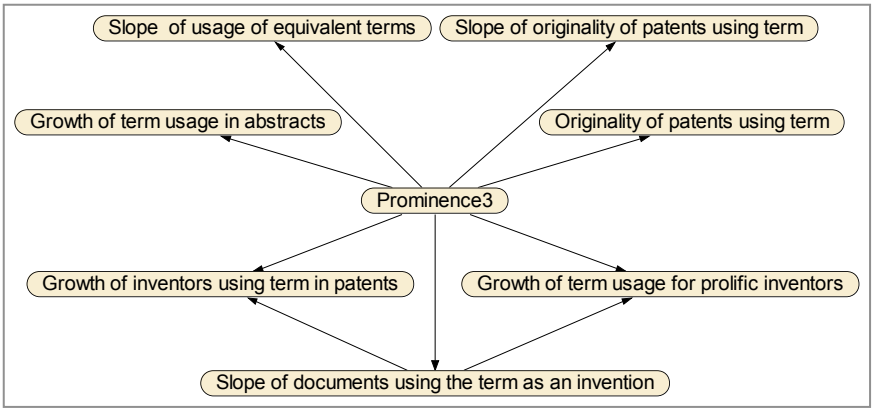


Figure 2. Fragment of model for predicting term prominence.

We chose to use Bayesian networks for several reasons. First, we executed a performance comparison between Bayesian networks (looking at common confusion matrix measurements such as the true and false positive rate, F1 score, etc.) and other classifiers such as JRip, J48, SVM, and meta-classifiers wrapping these, including Bagging and AdaBoostM1. Second, we chose Bayesian networks due to their flexibility and ease of interpretation. Finally, Bayesian networks provide insight into the contribution of indicator variables by supporting the computation of information-theoretic quantities such as mutual information and conditional mutual information.

We use a fine-grained discretization of prominence values instead of a binary prominent/not-prominent variable. This allows more precise computation of information-theoretic relations between indicator variables and prominence than does a binary variable. For example, some variables may be good at predicting very high prominence, while others merely discriminate prominent from non-prominent entities.

Although the prominence variable has a fine-grained discretization, it can be used as a binary classifier by choosing a threshold for prominence. The threshold is chosen through the multi-objective optimization process, described below.

Model Generation and Optimization

Automated model generation must answer the following questions in order to create the desired Bayes net:

- Which indicator variables should be included?
- Which indicator variables should be linked?
- How should continuous variables be discretized?
- How much weight should the training algorithm give to the training data relative to the untrained prior distribution so as to avoid over fitting?
- What threshold for predicting prominence provides the best trade off between recall, precision, and other performance goals?

All of these questions are answered by an optimization loop. This optimization loop uses a multi-objective elitist genetic algorithm (NSGA-II) to search the model parameter space (i.e. answers to the above questions) and rewards solutions that score well relative to specified recall and precision goals. The optimizer uses stratified 10-fold cross validation to compute metrics (e.g. recall and precision) for various combinations of system and ground truth

prominence thresholds. This process leverages the recall ↔ precision trade-off parameter. Finally, the optimizer promotes and further explores solutions that perform relatively well via: (1) uniform crossover, (2) Gaussian mutation for continuous variables, and (3) random flip mutation for discrete variables. The end result is an answer to the above questions that is optimized to the specified objectives.

Indicator Analysis

The analysis described below measures how well the indicators and models can forecast future term prominence, where a term is considered prominent if it has achieved a significant increase in usage.[4] To perform this analysis, we computed indicator values and generated models by processing all documents up to a given year (called the reference period), and then compared system outputs against a ground truth variable measuring an increase in term usage three years after the reference period. This analysis measures the ability of our models to forecast a significant increase in term frequency three years into the future.

By using automated model generation process described above, we generated domain-specific models for different technology areas in English and Chinese patents, including Computer Science, Communications, Biotechnology, and Semiconductors. The performance was higher for Chinese than for English, with the average recall of 0.49 and 0.52 for English patents and recall of 0.47 and precision of 0.61 for Chinese patents. The higher precision for Chinese patents is most likely due to Chinese patents containing a higher percentage of prominent terms than English patents.

To analyze individual indicators, we computed rank correlations between indicators and term prominence. Table 1 illustrates the performance of our indicators for English patents for the domain of Computer Science using Spearman's rank correlation coefficient (Rho) and three approaches to summarizing time series: slope, growth, and sum. For example, in Table 1, Rho slope for the indicator "Number of organizations per year using term in patents" shows the rank correlation for the indicator "the slope of the regression line fitted to the number of organizations using a selected term each year leading up to the reference period."

Table 1 reveals that indicators are significantly correlated with prominence for at least one computation (slope, growth, or sum), with the exception of one — the number of significant opinion relations. This is not unexpected, since opinion relations rarely occur in patents.[5] It also shows that term momentum indicators have the strongest rank correlations with prominence, i.e. measuring past momentum is particularly useful for predicting future prominence. Given that the other classes of indicators are conceptually very different from term momentum indicators, we expect that their effect on the forecasting model is additive to the momentum indicators, rather than duplicative. To test this hypothesis, we computed the partial correlations of non-momentum indicators with prominence, after the most basic term momentum has been accounted for (prior term usage in patents).

[4] One of the limitations of our system is that our analysis applies to individual terms, rather than sets of terms that are representative of technologies or research areas. This limitation is due to the problem of generation of ground truth data for training of our statistical models. In the future, we plan to extend this approach to analyse clusters of related terms, which are representative of technologies and scientific fields.

[5] Our analysis of scientific articles has shown that opinion-type relations (such as positive, standard, and negative opinion) are very infrequent in scientific literature as well, which suggests that opinion-based indicators are not particularly useful for the analysis of scientific literature and patents.

Table 1. Spearman rank correlations with future increase in term usage in English patents.

	Time Series indicators	Rho-Slope	Rho-Growth	Rho-Sum
Term Momentum Indicators	Number of unique organizations per year using term in patents	0.48	0.26	0.47
	Number of prolific organizations per year using term in patents	0.47	0.25	0.46
	Number of unique inventors per year using term in patents	0.50	0.13	0.47
	Number of prolific patenting inventors per year using term in patents	0.45	0.30	0.50
	Number of times per year term is used in patents	0.50	0.26	0.47
	Number of times per year equivalent terms are used in patents	0.48	0.25	0.45
	Number of times per year term is used in summary of invention section	0.52	0.26	0.51
	Number of times per year term is used in Independent claims	0.46	0.38	0.51
	Number of times per year term is used in Abstract section	0.47	0.33	0.52
	Number of industrial assignees using term per year	0.49	0.19	0.46
	Number of academic patent assignees using term per year	0.21	0.26	0.30
Term Characteristics	Annual technology score	N/S	N/S	0.19
	Annual maturity score	0.11	0.13	0.33
	Term usage as an invention	0.12	0.18	0.19
	Term usage as a component	0.23	0.25	0.27
Semantic relations	Annual counts of Exemplify relations	0.33	0.35	0.37
	Annual counts of Practical relations	0.33	0.33	0.37
	Annual counts of Opinion Significant relations	N/S	N/S	N/S
	Term usage with an abbreviation	0.19	0.23	0.24
	Annual counts of Contrast relations	0.20	0.26	0.26
	Annual counts of Based on relations	0.23	0.18	0.24
	Annual counts of Better than relations	0.17	0.13	0.18
Document Characteristic	Originality of patents using the term	N/S	N/S	0.19
	Average citation impact of documents about the term	N/S	N/S	0.31
	Term frequency in an emerging cluster	0.18	0.12	0.42
	Number of prior art references	0.02	-0.12	0.22
	Citations to high impact patents	N/S	N/S	0.31
	Dispersion of term usage across technologies	0.12	N/S	0.46
Inventor Char.	Number of patent inventors using the term as invention	0.12	0.17	0.19
	Hirsch index of the inventor	N/S	N/S	0.19
	Citation impact of prior patents granted to inventor(s)	N/S	N/S	0.29

Table 2 lists the indicators in the descending order of their partial correlations with prominence. An interesting finding is that the indicators that provide information over and above term momentum indicators include the ones that are based on language features, such as Practical and Exemplify relations, as well as term characterization. The indicators that have low or even negative correlations include document- and inventor-based indicators, such as the Hirsch index of the inventor, or the average citation index of document using the term. Having said that, it is important to note that document and inventor indicators are consistently selected by our forecasting models, which indicates that they are not really replaceable by other indicators.

Table 2. Partial correlation of indicators with prominence, controlling for momentum indicator.

Indicator	Partial Correlations
Annual counts of Practical relations	0.199
Term usage as an invention	0.170
Annual counts of Exemplify relations	0.169
Term usage as a component	0.159
Citations to high-impact patents	0.149
Annual maturity score	0.134
Annual technology score	0.129
Annual counts of Based_on relations	0.120
Annual counts of Contrast relations	0.114
Originality of patents using the term	0.101
Term usage with an abbreviation	0.098
Annual counts of Better_than relations	0.080
Citation impact of prior patents granted to inventor(s)	0.019
Average citation impact of documents about the term	-0.023
Number of prior art references	-0.042
Term frequency in an emerging cluster	-0.057
Hirsch index of the inventor	-0.074

Comparing indicators with different rationale, such as practicality versus discursive interest, one interesting finding is that the indicators focusing on the practicality of a field have the strongest correlations with prominence. These indicators include maturity scoring, usage as a component, Practical relations, and term usage by industrial patent assignees. Indicators focused on discursive interest in the term, such as Contrast relations, Better Than relations, and term usage by academic researchers in the field, have weaker (although still significant) correlations with prominence (as shown in Table 1 above). This suggests that, while both practicality and discursive interest are useful characteristics for the analysis of patents, the former is of particular value in forecasting the future prominence of terms.

Our further analysis of indicators focused on trying to identify indicators with complementary strengths. For example, we discovered that many of our indicators are good at predicting whether term usage will increase or decline/remain stable, but there are only a few indicators that are good at predicting different degrees of positive changes in term usage. This is illustrated by Table 3, which shows rank correlations between indicators and future changes in term usage coded as positive versus non-positive (Rho+/), as well as rank correlations considering positive values only (Rho-Pos).

As Table 3 shows, the correlations for the classification problem (Rho+/-) are generally higher, which suggests that it is more straightforward for an indicator to forecast whether or not a term will have a positive prominence, versus forecasting different degrees of positive prominence. It also reveals that some indicators might have particular strengths. For example, while momentum indicators and some document characteristic indicators perform best for delineating between positive and non-positive cases, the best indicator for distinguishing between different levels of positive prominence is “the proportion of granted patents using term relative to published documents”.

Table 3. Spearman correlations for indicators based on different conditions.

	Time Series indicators	Rho+/-	Rho-Pos
Term Momentum Indicators	Number of unique organizations per year using term in patents - Slope	0.50	0.21
	Number of prolific patenting organizations per year using term in patents - Slope	0.49	0.19
	Number of unique inventors per year using term in patents - Slope	0.52	0.22
	Number of prolific patenting inventors per year using term in patents - Slope	0.52	0.22
	Number of times per year term is used in patents - Slope	0.53	0.22
	Number of times per year equivalent terms are used in patents - Slope	0.51	0.20
	Number of times per year term is used in summary of invention section - Sum	0.54	0.24
	Number of times per year term is used in Independent claims section - Sum	0.53	0.25
	Number of times per year term is used in Abstract section - Sum	0.55	0.26
	Number of industrial assignees using term per year - Slope	0.51	0.21
	Number of academic patent assignees using term per year - Sum	0.33	0.09
Term Character ization	Annual technology score - Sum	0.21	0.05
	Annual maturity score - Sum	0.33	0.14
	Term usage as an invention - Sum	0.17	0.12
	Term usage as a component - Sum	0.27	0.13
Semantic relations	Annual counts of Exemplify relations - Sum	0.36	0.19
	Annual counts of Practical relations - Sum	0.37	0.18
	Term usage with an abbreviation - Sum	0.22	0.15
	Annual counts of Contrast relations - Sum	0.24	0.15
	Annual counts of Based_on relations - Sum	0.21	0.15
	Annual counts of Better_than relations - Sum	0.14	0.14
Document Characteristic	Originality of patents using the term - Sum	0.21	0.07
	Average citation impact of documents about the term- Sum	0.30	0.03
	Term frequency in an emerging cluster - Sum	0.46	0.15
	Number of prior art references - Sum	0.27	0.05
	Citations to high-impact patents - Sum	0.33	0.16
	Dispersion of term usage across technologies - Sum	0.50	0.18
Inv- entor Char.	Number of patent inventors using term as invention-Sum	0.18	0.10
	Hirsch index of the inventor - Sum	0.30	-0.02
	Citation impact of prior patents granted to inventor(s) - Sum	0.36	0.07
Single value	Proportion of granted documents using term relative to published documents	0.39	0.29
	The year the term first appeared in a patent	-0.15	0.01
	The year the term first appeared with an abbreviation	0.25	0.17

We further evaluated performance of indicators across one-, two- and three-year gap periods and observed a significant difference. All indicators tend to perform better in predicting longer forecasts (such as three-year gap) than shorter periods (such as one- or two-year gap). This may be because a three-year forecast smoothed out some of the year-by-year volatility in term usage.

Table 4. Spearman correlations for term prominence indicators in Chinese patents.

Time Series indicators	Rho-Slope	Rho-Growth	Rho-Sum
Number of unique inventors per year using term in patents	0.50	N/S	0.46
Number of prolific patenting inventors per year using term in patents	0.50	N/S	0.46
Number of times per year term is used in patents	0.50	0.06	0.46
Number of times per year term is used in Independent claims section	0.50	0.16	0.44
Number of unique organizations per year using term in patents	0.48	N/S	0.43
Number of prolific patenting organizations per year using term	0.48	N/S	0.44
Number of times term is used in summary of invention section	0.18	N/S	0.11
Annual maturity score	0.08	0.08	0.28

Finally, Table 4 shows correlation analysis for some of the indicators that were applied to Chinese Computer Science patents. It is important to note that citations rarely occur in Chinese patents, so indicators that are based on citation metrics cannot be used for the analysis of term prominence in Chinese. A comparison of correlations for English and Chinese (Tables 1 and 4) reveals that the general patterns across two collections are very similar, with Slope and Sum term momentum indicators performing particularly well, along with the Sum version of the Maturity Score.

Future Plans: Term Characterization

In addition to predicting future levels of interest to a technology, we expect that the indicators we developed can also provide some insights into the nature of the technology, its lifecycle, and other term characteristics. An example of this type of analysis is illustrated by 10 computer science terms, shown in Table 5.

Table 5. An analysis of 10 computer science terms.

Term	Pe	Term Characterization Analysis
RFID antenna	0.60	a device, becoming widely used in diff applications in 2007
Instant messaging	0.47	a technology or method, innovative, not a component
Robotics	0.31	a branch of technology, not a specific device, mature
XML	0.31	technology name, active area of research
Speech recognition	0.31	widely accepted technology, but best practice is being debated
Cellular telephone	0.31	a widely used standalone device, still of interest
RDF	0.31	technology name, becoming more widely used
Linux operating system	0.31	a widely accepted mature technology
GPS	0.30	a technology, widely used, mature, active area of research
Quantum computing	0	a principle or concept, innovative, no practical applications

The Pe column shows our predictions for the future changes in term usage, as described above, where zero value indicates that term usage will remain stable or decline in the future, whereas positive values predict that there will be an increased community interest in the term. The terms were analysed using 2007 as the reference period, forecasting term usage in 2010. The most interesting terms in this list include *RFID antenna* and *instant messaging*, the other terms, except for *quantum computing*, have slightly lower positive Pe values, indicating that there will be some growth in their usage between 2007 and 2010. The fact that quantum computing has zero value is not unexpected, considering that the data processed for this analysis included patent literature only, and this term has rarely been used in patents until 2007.

In addition to identifying terms with high prominence, we expect that the indicators described in the paper can also be used to characterize technologies, as illustrated in Table 5. For example, by using individual indicators or groups of indicators, we can potentially identify

widely accepted and mature technologies, terms that function as components of other technologies, active areas of research, as well as areas where best practice is being debated.

For example, Figure 3 reveals the values for the indicator that computes the average growth rate of term usage by academic institutions. This indicator can be used to identify innovative technologies that attract a growing attention from academia. Out of the 10 terms, technologies with the highest growth of academic assignees include *RFID antenna, instant messaging*, and *RDF*.

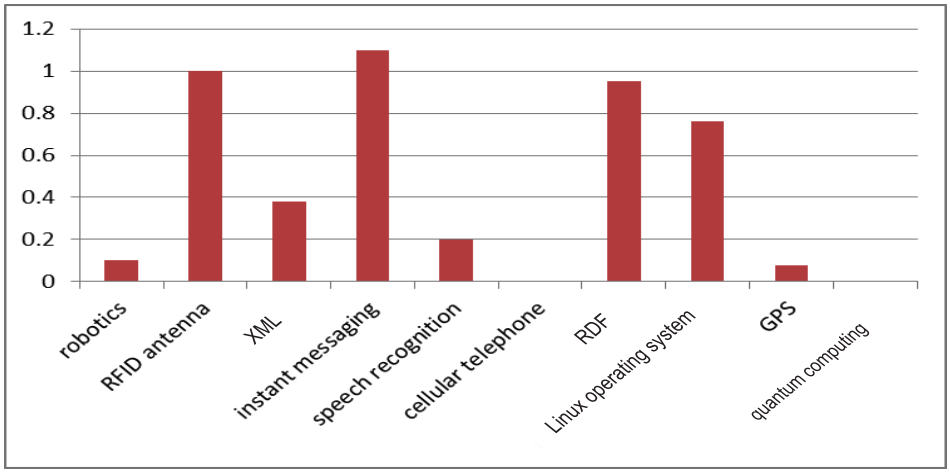


Figure 3. The average growth rate of academic assignees using term from 2002 to 2007.

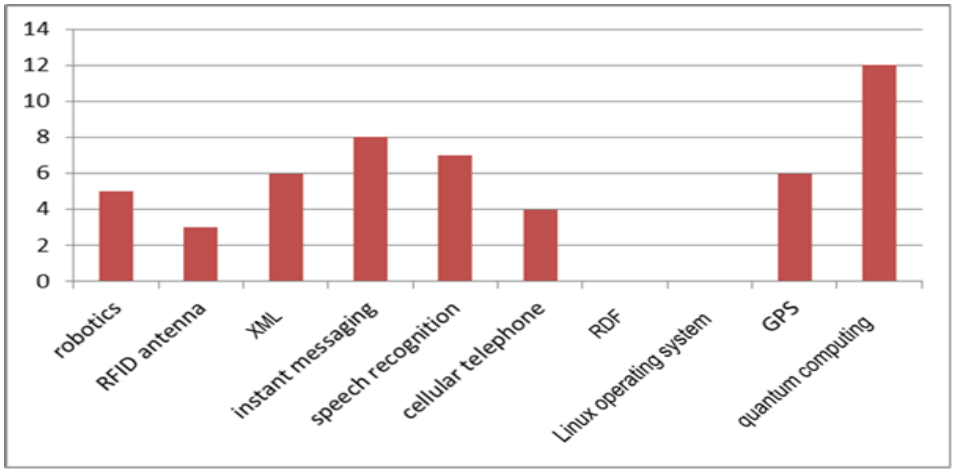


Figure 4. The number of inventors using term as an invention from 2005 to 2007.

Figure 4, on the other hand, illustrates the indicator values for "the number of inventors that were using the term as a description of an invention". Interestingly, the term that has the highest indicator value in this case is *quantum computing*. The terms with the higher values in Figure 3, *RDF and RFID antenna* have the lowest indicator values in Figure 4. This example suggests that individual indicators or groups of indicators may be used to detect different types of emerging technologies and that these differences might be related to their nature or lifecycle. It further illustrates that individual indicators can help to identify newer terms like *quantum computing*, and that high values of specific indicators may be indicative of the future potential of the term.

Conclusion

The system presented is capable of scanning millions of technical documents, extracting key indicators from both text and metadata, and forecasting meaningful trends and predictions from the extracted metrics. In particular, the extracted indicators are useful in predicting levels of interest in particular technologies. We also showed how the indicators provide insight into the nature and the lifecycle of emerging technologies, including their maturity, practicality, stages of development, and acceptance by the community.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. government.

References

Anick P, Verhagen M., & Pustejovsky J. (2014). Identification of technology terms in patents. In *Proceedings of LREC 2014*.

Arthur, B. (2009). The Nature of Technology: What It Is and How It Evolves. Free Press.

Babko-Malaya O., Thomas P., Hunter D., Meyers A., Pustejovsky P., Verhagen M., & Amis G. (2013a). Characterizing communities of practice in emerging science and technology fields, In *Proceedings of the International Conference on Social Intelligence and Technology 2013*.

Babko-Malaya O., Meyers A., Pustejovsky J., & Verhagen M. (2013b). Modeling debate within a scientific community. In *Proceedings of the International Conference on Social Intelligence and Technology 2013*.

Bettencourt, L., Kaiser, D., Kaur, J., Castillo-Chávez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, *75*(3), 495–518.

Breitzman, A., & Thomas, P. (2015). The emerging clusters model: A tool for identifying emerging technologies across multiple patent systems. *Research Policy*, 44(4), 195-205.

Brock, D.C, Babko-Malaya O., Pustejovsky, J., Thomas, P., Stromsten, S., & Barlos, F. (2012). Applied actant-network theory: Toward the automated detection of technoscientific emergence from full-text publications and patents. In *Proceedings of the AAAI Fall Symposium on Social Networks and Social Contagion 2013*.

Friedman N, Geiger, D., & Goldszmidt, M. (1997). Bayesian networks classifiers. *Machine Learning*, 29, 131-163.

Latour B. (2005). *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford University Press.

Li, S., & Xue, N. (2014). Effective document-level features for Chinese patent word segmentation, In *Proceedings of ACL 2014*.

Meyers, A., Zachary, G., Grieve-Smith, A., He, Y., Liao, S., & Grishman, R. (2014). Jargon-Term Extraction by Chunking. In *Proceedings of SADAATL 2014*.

Meyers, A. (2013). Contrasting and corroborating citations in journal articles, In *Proceedings of Recent Advances in Natural Language Processing 2013*.

Meyers, A., Lee G., Grieve-Smith A., He, Y., & Taber, H. (2014). Annotating relations in scientific articles. In *Proceedings of LREC 2014*.

Schiebel, E., Hörlesberger, M., Roche, I., François, C., & Besagni, D. (2010). An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices. *Scientometrics*, *83*(3), 765-781.

Roche, I., Besagni, D., François, C., Hörlesberger, M., & Schiebel, E. (2010). Identification and characterization of technological topics in the field of molecular biology. *Scientometrics*, *82*(3), 663-676.

Thomas P., Babko-Malaya O., Hunter D., Meyers A., & Verhagen M. (2013). Identifying emerging research fields with practical applications via analysis of scientific and technical documents. In *Proceedings of ISSI 2013*.

Understanding Relationship between Scholars' Breadth of Research and Scientific Impact

Shiyan Yan[1] and Carl Lagoze[2]

[1] *shiyansi@umich.edu*
School of Information, University of Michigan, Ann Arbor

[2] *clagoze@umich.edu*
School of Information, University of Michigan, Ann Arbor

Abstract
Many existing metrics to evaluate scholars consider their scientific impact without considering the importance of breadth of research. In this paper, we define a new metric for breadth of research based on the generalized Stirling metric that considers multiple aspects of breadth of research. We extract research topics in computer science using concept extraction and clustering from the literature in the ACM dataset. We then assign authors a distribution over these research topics, from which we calculate scores of breadth of research for each author. We design five simulation experiments that evaluate the ability of a metric to measure breadth of research and use these experiments to compare our new metric to traditional metrics. The results show how these metrics perform in different experiments, concluding that no metric consistently outperforms the others. We test the relationship between our new metric and scientific impact and find a weak correlation between them. Finally, we find that the variation of the metric over time illustrates a possible publication pattern for scholars.

Conference Topic
Indicators

Introduction
An increasing number of scholars are engaged in interdisciplinary research (Porter, Cohen, David Roessner, & Perreault, 2007; Wagner et al., 2011). Some of this is due to the emergence of new scholarly "disciplines" that are inherently multi-disciplinary such as information science, while some arises from scientific problems such as climate change that require expertise from multiple fields. Meanwhile, scholarly impact and influence continues, by and large, to be measured by indices that ignore breadth of research and may even penalize scholars who diversify their research portfolio. For example, H-index, which is used extensively to measure scholarly impact, and which has been criticized for its limited focus (Weingart, 2005), may be unfair when comparing scholars with different degrees of breadth of research. Ultimately, a metric or a set of metrics is needed that accounts for breadth of research, so that breadth of research can be measured and be included in an evaluation system of scholars' scientific influence.

In this paper we describe research that explores the area of scholarly impact metrics and breadth of research. The contributions of our work are as follows. We design a new metric to measure scholars' breadth of research that builds on traditional metrics. We develop a multi-stage method for extracting topics from a corpus (in our case computer science papers) and calculate the scores of breadth of research for authors who have published papers in computer science conferences. We design five simulation experiments that compare the relative performance of existing metrics and our new metric for measuring breadth of research. We measure the relationship of breadth of research and H-index for scholars who are authors in our corpus. Finally, we explore the variation of breadth of research for scholars over time to observe their paper publication behavior over their careers.

The structure of this paper is as follows. The next section describes related work in the areas relevant to our work. Following that, we report on the dataset we used in our research. We

then describe our process of dictionary extraction, topic extraction, paper assignment and author assignment to topics. In the subsequent section we illustrate our new metric and compare it to traditional metrics. The penultimate section describes simulation experiments to show the performance of the new metric, the relationship between the new metric and metrics of research impact, and the variation over time of breadth of research for scholars. Our conclusions and possible future work are listed in the final section.

Related Work

There is a variety of existing literature relevant to the area of breadth of research. The areas covered by this literature include topic extraction, topic relationship extraction, metrics design and the relationship between different aspects of research evaluation systems.

There are many methods to associate topics to publication. The simplest one is to use the classification codes in a dataset, such as ISI subject categories in Web of Science, as the set of topics. But these categories are too coarse-grained and hide intra-disciplinary variability. Another method is to use unsupervised learning algorithms to extract some topics according to the content of papers or the citation network of papers. Topic modelling (Blei, Ng, & Jordan, 2003) is one of the popular unsupervised learning algorithms based on content of papers. This model has been used to identify the disciplines that comprise interdisciplinary work funded by NSF (Nichols, 2014). The ACT model (author-conference-topic) (Li et al., 2010) is an adaptation of Blei's model. Another approach is to use community detection in networks as a basis for finding topics. One example is the use of two-round clustering (Rosvall & Bergstrom, 2008) over the citation network to extract topic-associated communities (Velden & Lagoze, 2013). Another method using both the citation network and the word distribution of abstracts (Jo, Hopcroft, & Lagoze, 2011) finds temporally-ordered topics from a corpus of scientific literature, such as the ACM dataset.

Understanding the relationship between topics is also an important step after topic extraction, because the calculation of the similarity of topics is necessary for understanding the breadth of research. Some researchers have extracted the relationships and used information visualization techniques to represent the relationship between different topics. For example, Yan (2013) detects the path between different disciplines to find the evolution of some areas. Another paper describes a new method to find the diversity subgraph in a multidisciplinary scientific collaboration network (He, Ding, Tang, Reguramalingam, & Bollen, 2013). An interesting visualization method leverages the circle of science to visualize the relationship between disciplines in one dimension (Boyack & Klavans, 2009).

Many metrics have been designed to measure factors related to scientific influence. The most common metrics are impact factor and H-index, which measure the number of citations of scholars' papers. Although these metrics have many problems such as lack of universality between different disciplines (Kaur, Radicchi, & Menczer, 2013), they are still widely used in systems like Google Scholar. Some alternative metrics also use the number of citations to measure the scientific influence of scholars (Ruscio, Seaman, D'Oriano, Stremlo, & Mahalchik, 2012). They offer advantages over simple metrics such as H-index, but they also focus solely on the citation count of papers. Other metrics based on the centrality of scholars in a network (e.g., co-authorship) like PageRank and betweeness centrality (Bollen, Van de Sompel, Hagberg, & Chute, 2009) are also widely used. However, the correspondence of centrality to actual influence is unknown.

As mentioned earlier, commonly used metrics of scholarly influence fail to consider breadth of scholars' research. In response a number of researchers have created some metrics for the degree of interdisplinarity and more generally breadth of research. The report of quantitative metrics and context in interdisciplinary scientific research (Wagner et al., 2011) is a good survey for metrics for interdisciplinarity. Specialization and integration (Porter et al., 2007)

are good metrics of interdisciplinarity because they consider similarity between disciplines when measuring interdisciplinarity. They can be modified easily in the context of a diversity of research topics. Some papers discuss different dimensions of interdisplinarity (Rafols & Meyer, 2010; Rafols, Leydesdorff, O'Hare, Nightingale, & Stirling, 2012): diversity, coherence and intermediation. They define diversity as a combination of variety, balance and disparity. Coherence means link strength between different disciplines. Intermediation is based on the network structure and is measured by betweenness centrality, clustering coefficient and average similarity. Other papers describe metrics based on these dimensions. Cassi, Mescheba, and de Turckheim (2014) divides the Stirling metric into "within component" and "between component" to measure the diversity of articles. Jensen & Lutkouskaya (2013) defines six indicators based on the dimensions and measure the breadth of research at two levels (article and laboratory). Karlovčec and Mladenić (2014) defines a new diversity metric based on Generalized Stirling. The metric incorporates connectedness of the citation graph into the original metric and applies it in exploratory analysis of the research community in Slovenia. Roessner, Porter, Nersessian, and Carley (2012) validates the interdisciplinarity metrics with ethnographic materials (field observations and unstructured interviews).

Finally, some research has focused on the relationship between breadth of research and other factors considered in scientometrics (not just scientific influence). One interesting paper finds that the papers with an average degree of interdisciplinarity will get higher impact than papers with too high or too low degree of interdisciplinarity (Sternitzke & Bergmann, 2008). The results are convincing but metrics used in this paper are quite simple (Jaccard similarity and cosine similarity). Two papers find that interdisciplinary papers have potentially lower impact than more focused papers. One of them finds that multidisciplinary papers are not frequently cited in contrast to the disciplinary papers (Levitt & Thelwall, 2008). The other explains how high-ranked journals suppress interdisciplinary research (I Rafols & Meyer, 2010). Other papers describe some factors that can encourage researchers to be involved in interdisciplinary research work (Carayol & Thi, 2005; van Rijnsoever & Hessels, 2011). They provide some theories to explain why scholars choose interdisciplinary projects. Some findings support that there are no correlations between citation ranks and ranked interdisciplinarity indices (Ponomarev, Lawton, Williams, & Schnell, 2014). In contrast, other researchers confirm that the degree of interdisciplinarity is strongly correlated with the impact factor (Silva, Rodrigues, Oliveira, & da F. Costa, 2013).

Dataset

We extract abstracts, full text and other metadata from the ACM digital library for proceedings of major conferences in computer science. From these proceedings we select authors whose names are unambiguous and who have published at least five papers. The standard for unambiguity is whether using the full name as the query sent to Google Scholar returns only one researcher profile with the same name. We extract the citation numbers and H-indexes by crawling over Google Scholar. Overall we crawled H-indexes and citation numbers for 8911 authors from Google Scholar in August 2014. We also used the Wikipedia dataset to extract important terms in computer science.

Topic Extraction and Assignment

Both traditional metrics and the new metric designed in this paper require a distribution over different topics or areas for authors. In order to generate topic distributions, we leverage the text data in the papers of ACM digital library and implement three steps to form distributions: dictionary extraction, topic extraction and author assignment.

Dictionary Extraction

How to define topics is the first problem to be solved in the topic extraction and assignment. In our work, we extract a dictionary of n-grams in computer science and cluster them into topics using the Affinity Propagation algorithm (Frey & Dueck, 2007). Three different sources of dictionaries are used in this paper: grams that are frequently used in papers, grams that can be matched to their abbreviations in the papers, and entries in Wikipedia.

Dictionary extraction follows these steps:

1. Extract bigrams and trigrams that occur frequently in papers using a threshold of more than 10 times for bigrams and more than 5 times for trigrams. The threshold helps to eliminate noisy grams with low frequency.
2. Extract grams from papers that conform to the pattern "n-grams (abbreviation)", e.g. machine learning (ML).
3. Intersect the results of step 1 and step 2 (3816 terms in total).
4. Build a network of entries in Wikipedia according to hyperlinks between them in the website.
5. Make use of grams in step 3 and search their neighbours in the network of Wikipedia terms. If their neighbours also occur frequently in papers (with frequency higher than the thresholds mentioned above), add the terms into the final dictionary (6100 terms)

The top 5 bigrams and top 5 trigrams in the final dictionary are shown in Table 1:

Table 1. Grams with top frequency

Grams	*Frequency*
User Interface	2372
Software development	2102
Programming language	2042
Software engineering	1988
Operating system	1761
Wireless sensor network	586
World wide web	467
Graphical user interface	305
Support vector machine	300
Discrete event simulation	287

Topic Extraction and Assignment

After extracting the dictionary, we count the co-occurrence measure for every pair of terms. We then calculate the similarity between different terms by:

$$Sim_{ij} = \log \frac{Cooccur_{ij} + 1}{Max(Cooccur_{ij}) + 2}$$

The logarithm calculation makes the distribution of similarity more uniform and avoids the influence of outliers of co-occurrence numbers. We weight co-occurrences of terms in abstracts of papers more than those in full text based on the intuition that abstracts generally have a stronger “topic signal”. Using the computed similarity matrix of terms, we then run Affinity Propagation to cluster together similar terms and choose an exemplar for every cluster. The benefits of Affinity Propagation are that there isn’t a need to parameterize the number of clusters and that the exemplars for every cluster provide a straightforward explanation of what these clusters are about. More than two hundred clusters, or topics, are generated. Here are two examples of the clustering results:

Exemplar: digital library

Terms:
citation analysis, citation index, community building, digital earth, digital library, digital library software, digital preservation, digital reference, discourse analysis, dublin core.
Exemplar: machine learning
Terms:
active learning, adaptive control, bayes classifier, belief propagation, clinical trial, computational learning theory , concept learning, conditional random field.
We then assign every paper a probabilistic assignment to the different topics according to their respective frequency of n-grams associated with the particular topic. Therefore, every paper will have a distribution over topics.

Author Assignment

Using the clusters of grams in computer science and the topic distributions for every paper, we assign authors into different topics according to their papers. Every author is represented by a distribution over topics, which are used to calculate scores of metrics. There does not exist a "gold standard" list of researchers that ranks breadth of research that we can use to evaluate how reasonable our topic assignments are. We list below some topic distributions for well-known computer scientists to demonstrate our assignment.

John Koza	
1 genetic programming	0.567
2 programming language	0.083
3 knowledge base	0.063
Peter Denning	
1 memory management	0.107
2 computer systems	0.093
3 information systems	0.050
Eric Horvitz	
1 user interface	0.082
2 information retrieval	0.067
3 machine learning	0.051
4 speech recognition	0.047

Breadth of Research Measurement

With the author distribution of topics established, the key question is how to translate this into a measure of breadth of research for authors. As mentioned in the section describing related work, many metrics have been used to measure the "degree of interdisciplinarity". Compared to previous metrics to measure breadth of research, we design a new metric that considers the topic distribution, similarity distribution and coherence within research topics.

Summary of Old Measurements

There are many measurements of diversity or interdisciplinary, like entropy (Weaver, 1949), Simpson's index (Simpsons, 1949) and generalized Stirling (Stirling, 2007). Each of these is computed as follows. Denote p_i as the probability of topic distribution for an author over topic i, d_{ij} as the distance between topic i and topic j.

$$Entropy = \sum_{i=1}^{n} -p_i \times \log_n (p_i)$$

$$Simpson = 1 - \sum_{i=1}^{n} {p_i}^2$$

$$Generalized\ Stirling = \sum_{i,j} d_{ij}^{\alpha}\ (p_i \times p_j)^{\beta}$$

Comparing them, only generalized Stirling considers not only the distribution of topics but also the similarity between topics. The further the distance between topics in which an author publishes papers, the more diverse will the author's research interest be. However, the traditional metrics do not consider the notion of differing *coherence* between different research topics. And the degrees of influence of topics with small proportions are very limited. The new measurement is a modified version of the generalized Stirling metric and it incorporates the coherence of topics and value of *minor topics* (topics with small proportions).

New Measurement

The new metric for breadth of research is defined as follows.

Denote d_{ij} as the distance between two topics, which are defined as the average distance (inverse of similarity defined above) between terms in the two topics, p_i as the probability of an author's paper belong to topic i, coh_i as the *coherence* of topic i. Coherence of each topic is the proportion of authors for whom the respective topic is their major research topic, which is an important signal to illustrate whether a research topic concentrate on some core research questions. Parameters α, β, γ are used to control the relative weights of different components.

$$Breadth\ of\ Research = \sum_{i,j} d_{ij}^{\alpha}\ (p_i + p_j)^{\beta} (Coh_i \times Coh_j)^{\gamma}$$

We modify the product of p_i and p_j in generalized Stirling to summation of p_i and p_j because the summation will give minor topics more chances to be counted into the measurement of breadth of research. We add the coherence term into the metric because different topics have different "density" within themselves. For example, some topics like digital library are less coherent topics because there are many diverse subtopics in these topics. But for topics like operation systems, researchers concentrate on several narrow subtopics. A researcher focusing on digital library should have larger breadth of research than operating systems researchers if other variables are controlled (so the gamma should have a negative value).

The new metric leverages properties of papers (topic distribution), properties of topics (coherence) and properties of relationship (topic similarity). The tunable parameters give the metric more flexibility to balance between different aspects of breadth of research.

Experiments

Simulation Experiment

There is no established standard for determining the quality of metrics of breadth of research. Furthermore, there is no ground truth to show the rankings of scholars' breadth of research with which to validate the various metrics. We propose an alternative evaluation method based on a set of axioms concerning breadth of research and then test how the metrics perform according to these axioms.

In addition to the definition of d_{ij} and coh_i defined in the previous section, the following definitions relate to the axioms.

- Denote A_i as the article i, $C=\{ A_1, A_2 ... \}$ as a *collection* of articles, and N_C as the number of articles in collection C.
- Denote t_i as the topic i, $D_A(t)$ as the topic distribution of article A over topic t. $(\sum_t D_A(t) = 1)$

- Denote $D_C(t)$ as the topic distribution of collection C over topic t. $D_C(t) = \frac{1}{N_C}\sum_{A_i \in C} D_{A_i}(t)$. $(\sum_t D_C(t) = 1)$
- Denote *score(C)* as the score of a metric over the collection of articles C

Axiom1: Publish in Old Topics

If an author publishes a paper in a topic in which she has published many papers before, her breadth of research should decrease.

Choose t, s.t. $t = arg\,max_t\, D_C(t)$, construct a new article A_{new}, s.t. $D_{A_{new}}(t) = 1$. $C' = C \cup \{A_{new}\}$. Then *score(C') < score(C).*

Axiom2: Publish in New Topics

If an author publishes a paper in a new topic in which she has never published, her breadth of research should increase.

Choose t, s.t. $D_C(t)=0$, construct a new article A_{new}, s.t. $D_{A_{new}}(t) = 1, C' = C \cup \{A_{new}\}$. Then *score(C') > score(C).*

Axiom3: Publish in New Topics Twice

If an author publishes papers in two new topics in a sequence, the increase of breadth of research in the second time should be smaller than the increase of that in the first time.

Choose t_1 and t_2, s.t. $D_C(t_1)=0$, $D_C(t_2)=0$, $t_1 \neq t_2$,construct two new articles A_{new1} and A_{new2}, s.t. $D_{A_{new1}}(t) = 1$ and $D_{A_{new2}}(t) = 1$. $C' = C \cup \{A_{new1}\}$, $C'' = C' \cup \{A_{new2}\}$. Then *score(C')-score(C) > score(C'')-score(C').*

Axiom4: Publish in Close Topics

If an author publishes a paper in a new topic close to the author's research interest, the improvement of her breadth of research should be less than that of publishing a new paper in a randomly chosen topic.

Randomly Choose t_1 s.t. $D_C(t_1)=0$, construct a new article A_{new1}, s.t. $D_{A_{new1}}(t_1) = 1$. $C' = C \cup \{A_{new1}\}$. Choose t_2 s.t. $D_C(t_2)=0$ and $arg\,min_t(inf_{t_0 \in \{t|D_C(t)>0\}} d_{t_0 t_1})$. Construct a new article A_{new2}, s.t. $D_{A_{new2}}(t_2) = 1, C'' = C' \cup \{A_{new2}\}$. Then *score(C'') < score(C')*

Axiom5: Publish in Coherent Topics

If an author publishes a paper in a new topic with high coherence, the improvement of her breadth of research should be less than that of publishing a new paper in a randomly chosen topic.

Randomly Choose t_1 s.t. $D_C(t_1)=0$, construct a new article A_{new1}, s.t. $D_{A_{new1}}(t_1) = 1$. $C' = C \cup \{A_{new1}\}$. Choose t_2 s.t. $D_C(t_2)=0$ and $t_2 = arg\,max_t\,(Cohe_t)$. Construct a new article A_{new2}, s.t. $D_{A_{new2}}(t_2) = 1, C'' = C' \cup \{A_{new2}\}$. Then *score(C'') < score(C').*

We implemented five simulation experiments based on the original dataset with 8911 authors to test how the traditional metrics and our new metric conform to the axioms. The results are shown in Table 2.

Table 2. Probability that metrics satisfy of the axioms

	Entropy	*Simpson's*	*GL Stirling* $(\alpha = 2; \beta = 0.3)$	*New Metric* $(\alpha = 1, \beta = 0.5, \gamma = -0.5)$
Axiom1	0.99	0.99	0.97	0.88
Axiom2	0.89	0.97	0.86	0.86
Axiom3	0.97	0.94	0.50	0.50
Axiom4	0	0	0.76	0.70
Axiom5	0	0	0.54	0.62

The results show that entropy and Simpson's perform well in the first three axioms because they don't consider distances between topics and introduce less noise. Because every new topic will be regarded equally for these metrics, they cannot follow Axiom4 and Axiom5. Generalized Stirling and our metric perform reasonably well in Axiom1 and Axiom2, but worse than entropy and Simpson's. They perform relatively badly in Axiom3 because relatively bad performance on publishing a paper in new topic (Axiom2) will aggregate when testing the performance of publishing two papers in two new topics. But they perform well in Axiom4 because of the consideration of distances. Also we find our metric performs better than generalized Stirling in Axiom5, which means coherences of topics and greater weights on minor topics are beneficial when we consider variation of metrics when people publish in topics with different coherence levels.

Parameter Sensitivity

The performance of new metric is influenced by the value of parameters α, β and γ. We tested the performance of the new metric with different settings. The results are shown in Table 3, Table 4 and Table 5.

Table 3. Average Prob of satisfying the axioms with different α.

	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	$\alpha = 100$
Axiom1	0.40	0.42	0.48	0.62
Axiom2	0.33	0.38	0.44	0.55
Axiom3	0.34	0.32	0.24	0.22
Axiom4	0.38	0.57	0.66	0.64
Axiom5	0.63	0.61	0.57	0.52

Table 4. Average Prob of satisfying the axioms with different β.

	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 100$
Axiom1	0.86	0.67	0.30	0.08
Axiom2	0.69	0.57	0.24	0.16
Axiom3	0.40	0.40	0.29	0.05
Axiom4	0.57	0.57	0.59	0.53
Axiom5	0.61	0.61	0.59	0.52

Table 5. Average Prob of satisfying the axioms with different γ.

	$\gamma = 0.1$	$\gamma = 1$	$\gamma = 10$	$\gamma = 100$
Axiom1	0.58	0.47	0.45	0.45
Axiom2	0.24	0.39	0.47	0.48
Axiom3	0.09	0.26	0.34	0.38
Axiom4	0.49	0.57	0.59	0.59
Axiom5	0.62	0.66	0.58	0.53

The tables show that the metric is very sensitive to the α, β and γ. In order to find the best parameter setting, we calculated the average performance over five different simulation experiments for every parameter settings. We selected the settings with highest average performance and a minimum threshold of at least 0.5 in every experiment. The best setting for Generalized Stirling is $\alpha = 2, \beta = 0.3$. The best setting for the new metric is $\alpha = 1, \beta = 0.5$ and $\gamma = -0.5$. They are used in the comparison of metrics in Table 2.

Summation Modification

One of important modifications of our metric is the replacement of product with summation in the second term of metric. We test the effect of this. If we control the distance term and coherence term in the metric to be the same for every topic and set $\beta = 1$. The metric using summation will definitely follow Axiom2 but not follow Axiom1 and Axiom3.

Let *n* represents the number of topic.

Axiom1: Publish in Old Topics

$$score(C) = \sum_{i,j} d^{\alpha}\ (p_i + p_j)(\text{coh} \times \text{coh})^{\gamma} = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$= \sum_{i,j} d^{\alpha}\ (p_i{}' + p_j{}')(\text{coh} \times \text{coh})^{\gamma} = score(C')$$

Axiom2: Publish in New Topics

$$score(C) = \sum_{i,j} d^{\alpha}\ (p_i + p_j)(\text{coh} \times \text{coh})^{\gamma} = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$< \sum_{i,j} d^{\alpha}\ (p_i{}' + p_j{}')(\text{coh} \times \text{coh})^{\gamma} = (n)d^{\alpha}(coh)^{2\gamma} = score(C')$$

Axiom3: Publish in New Topics Twice

$$score(C) = (n-1)d^{\alpha}(coh)^{2\gamma}$$

$$score(C') = (n)d^{\alpha}(coh)^{2\gamma}$$

$$score(C'') = (n+1)d^{\alpha}(coh)^{2\gamma}$$

$$score(C'') - score(C') = score(C') - score(C)$$

From the derivation above, the performance of new metric in Axiom 1 and Axiom 3 should be worse than the metric with product. The performance of Axiom 2 should be better than the metric with product. So we construct a metric using product in the second term and compare the performance of it with the new metric in different parameter settings.

$$Breadth\ of\ Research = \sum_{i,j} d_{ij}^{\alpha}\ (p_i \times p_j)^{\beta}(Coh_i \times Coh_j)^{\gamma}$$

The results in Table 6 shows that the metric using summation outperforms product in Axiom 2, and metric using product outperforms summation in Axiom1, which is consistent with the results of derivation. But the results for the other three axioms are close between the two metrics, which means the interaction between different terms in the metric (distance term, distribution term and coherence term) will influence the results of simulation.

Table 6. Comparison between metric with summation and production.

Metric	*Parameter setting*	*Axiom1*	*Axiom2*	*Axiom 3*	*Axiom4*	*Axiom5*
Production	$\alpha = 0.1\ \beta = 0.1 \gamma = -0.1$	0.99	0.85	0.45	0.22	0.59
	$\alpha = 100\ \beta = 1 \gamma = -1$	0.82	0.62	0.47	0.69	0.53
	$\alpha = 1\ \beta = 1 \gamma = -10$	0.83	0.40	0.39	0.55	0.76
Summation	$\alpha = 0.1\ \beta = 0.1 \gamma = -0.1$	0.97	0.89	0.45	0.22	0.59
	$\alpha = 100\ \beta = 1\ \gamma = -1$	0.69	0.69	0.50	0.69	0.55
	$\alpha = 1\ \beta = 1 \gamma = -1$	0.69	0.47	0.41	0.54	0.77

Relationship between breadth of research and scientific impact

We tested the Pearson correlation between metrics of breadth of research and H-indexes of scholars. Our results (Table 7) show that some metrics have a positive relationship with H-index. Others have weak negative relationship. Because publication numbers may influence

the correlation between breadth of research and scientific impact i.e. the increase of numbers of publications may bring increase of breadth of research and increase of H-index simultaneously to make them positively correlated to each other, we test the partial correlation between metrics of breadth of research to H-index controlling publication numbers (Table 7). They are weaker than Pearson correlations. And all the weak partial correlation scores don't illustrate strong correlation between metrics for breadth of research and H-index for scholars.

Table 7. Correlation between breadth of research and H-index.

	Pearson Corr.	*Partial Corr.*
Entropy v.s. H-index	-0.1722	-0.0769
Simpson's v.s. H-index	0.2102	0.0922
GL Stirling v.s. H-index	0.4283	0.1820
New Metric v.s. H-index	0.4337	0.1832

The Variation of metrics over publication years

We illustrate in Figure 1 the variation of average scores of metrics for all the scholars over publication years. Simpson's, generalized Stirling and our new metric initially increase and then level off, which explains a possible publication pattern of scholars: scholars' breadth of research may increase with the increase of publications in the early stage of their career. But because of accumulation of publications, their accumulative breadth of research will not change dramatically in the late years. For the entropy metric with base n, it is normalized by topic number. So it keeps in a stable level over year, which shows a different pattern compared to other metrics.

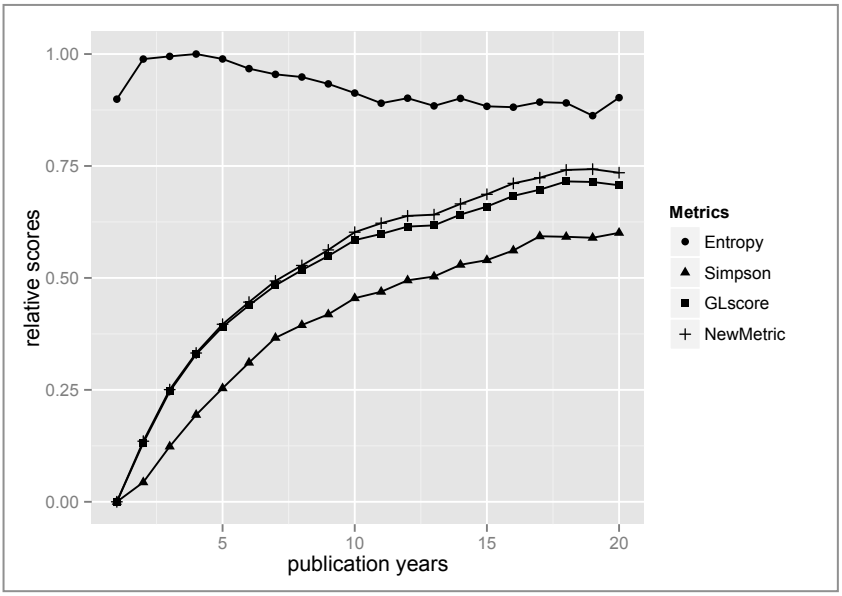


Figure 1. Variation of metrics over publication years.

Conclusion and Future Work

In this paper, we describe a new metric based on generalized Stirling to evaluate breadth of research for scholars in computer science. The metric makes use of topic distribution, similarity between topics, and coherence of topics and it can capture the diversity aspects of breadth of research. The simulation experiments show that traditional metrics can perform well in some axiom, but they don't perform well when coherence within topics and similarity between topics are considered. In contrast, generalized Stirling metric and the new metric for

breadth of research work better in the simulation related to similarity between topics and coherences but perform worse in the experiments of adding new topics. It is a trade-off between the simplicity of metrics and the concern of topic similarity and coherence.

With the new metric for breadth of research, we find the correlation between breadth of research and scientific metrics are weak, especially when we control publication numbers. From our study, there's no evidence to show whether the increase of breadth of research will influence the impact of scholars' publication. Also, after testing the variation of the new metric over years, we find a possible publication pattern of scholars: Breadth of research increases in the beginning with the increase of publications. But they increase slowly when publications have been accumulated.

There are a number of research questions that arise from the work described in this paper. The first one is finding alternative methods to generate research topics. Unsupervised learning models based on both text contents and citation information may be helpful to extract topics and show topic variation for authors. The second question is how to improve the simulation results for the new metric. The new metric performs better than general Stirling and other traditional metrics in some aspects. But if more information from co-author and citation network can be incorporated into the metric, the performance may be better and interpretable.

Acknowledgments

The research is funded by NSF 1258891 EAGER: Collaborative Research: Scientific Collaboration in Time.

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PloS One*, *4*(6), e6022. doi:10.1371/journal.pone.0006022

Boyack, K., & Klavans, R. (2009). Measuring multidisciplinarity using the circle of science. *From WRK1: Tracking and Evaluating Interdisciplinary Research, Workshop at ISSI*, *87122*.

Carayol, N., & Thi, T. (2005). Why do academic scientists engage in interdisciplinary research? *Vasa*.

Cassi, L., Mescheba, W., & de Turckheim, É. (2014). How to evaluate the degree of interdisciplinarity of an institution? *Scientometrics*. doi:10.1007/s11192-014-1280-0

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–6. doi:10.1126/science.1136800

He, B., Ding, Y., Tang, J., Reguramalingam, V., & Bollen, J. (2013). Mining diversity subgraph in multidisciplinary scientific collaboration networks: A meso perspective. *Journal of Informetrics*, 1–18.

Jensen, P., & Lutkouskaya, K. (2013). The many dimensions of laboratories' interdisciplinarity. *Scientometrics*, *98*(1), 619–631. doi:10.1007/s11192-013-1129-y

Jo, Y., Hopcroft, J., & Lagoze, C. (2011). The web of topics: discovering the topology of topic evolution in a corpus. *Conference on World Wide Web*, 257–266.

Karlovčec, M., & Mladenić, D. (2014). Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*. doi:10.1007/s11192-014-1355-y

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics*, *7*(4), 924–932. doi:10.1016/j.joi.2013.09.002

Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science*, *59*, 1973–1984. doi:10.1002/asi.20914

Li, D., He, B., Ding, Y., Tang, J., Sugimoto, C., Qin, Z., Dong, T. (2010). Community-based topic modeling for social tagging. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management - CIKM '10*, 1565. doi:10.1145/1871437.1871673

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, 741–754. doi:10.1007/s11192-014-1319-2

Ponomarev, I. V., Lawton, B. K., Williams, D. E., & Schnell, J. D. (2014). Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? *Scientometrics*, 755–765. doi:10.1007/s11192-014-1320-9

Porter, A. L., Cohen, A. S., David Roessner, J., & Perreault, M. (2007). *Measuring researcher interdisciplinarity*. *Scientometrics* (Vol. 72, pp. 117–147). doi:10.1007/s11192-007-1700-5

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, *41*(7), 1262–1282. doi:10.1016/j.respol.2012.03.015

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 1–28.

Roessner, D., Porter, A. L., Nersessian, N. J., & Carley, S. (2012). Validating indicators of interdisciplinarity: linking bibliometric measures to studies of engineering research labs. *Scientometrics*, *94*(2), 439–468. doi:10.1007/s11192-012-0872-9

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(4), 1118–23. doi:10.1073/pnas.0706851105

Ruscio, J., Seaman, F., D'Oriano, C., Stremlo, E., & Mahalchik, K. (2012). Measuring scholarly impact using modern citation-based indices. *Measurement: Interdisciplinary Research & Perspective*, *10*(3), 123–146. doi:10.1080/15366367.2012.711147

Silva, F. N., Rodrigues, F. a., Oliveira, O. N., & da F. Costa, L. (2013). Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics*, *7*(2), 469–477. doi:10.1016/j.joi.2013.01.007

Simpsons, E. H. (1949). Measurement of Diversity. Retrieved October 9, 2014, from http://www.nature.com/nature/journal/v163/n4148/abs/163688a0.html

Sternitzke, C., & Bergmann, I. (2008). Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, *78*(1), 113–130. doi:10.1007/s11192-007-1961-z

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society, Interface / the Royal Society*, *4*(15), 707–19. doi:10.1098/rsif.2007.0213

Van Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy*, *40*(3), 463–472. doi:10.1016/j.respol.2010.11.001

Velden, T., & Lagoze, C. (2013). The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology*, *64*(12), 2405–2427.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J.T., Boyack, K. W., Keyton, J., Rafols, I., & Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics*, *5*(1), 14–26. doi:10.1016/j.joi.2010.06.004

Weaver, W. (1949). Recent Contributions to the Mathematical Theory of Communication 1 Introductory Note on the General Setting of the Analytical Communication Studies.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, *62*(1), 117–131.

Yan, E. (2013). Finding knowledge paths among scientific disciplines. *arXiv Preprint arXiv:1309.2546*, (812), 1–31.

Transforming the Heterogeneity of Subject Categories into a Stability Interval of the MNCS

Marion Schmidt[1] and Daniel Sirtes[1*]

[1] *schmidt@forschungsinfo.de, sirtes@forschungsinfo.de*
iFQ Institute for Research Information and Quality Assurance, Schützenstr. 6a, D-10117 Berlin (Germany)

Abstract
The internal homogeneity of research disciplines in subject categories (SC) of the Web of Science database (WoS) regarding their publication and citation practices is an essential precondition for the field-normalization of citation indicators. This imperative of underlying homogeneity seems not to be met throughout all categories, as has been shown in former research. A keyword-based clustering method displays both the diversity of research areas included in an SC and that the clusters' mean citation rate differ substantially. This proof-of-concept paper on the basis of one country set and two SCs presents a bootstrapping method, which allows quantifying the degree of heterogeneity within subject categories as a stability interval. The MNCS 95% stability interval of our set has a range of 6.7% and 7.3% compared to its score. This kind of robustness measure could be implemented for future evaluative citation analysis in order to convey the coarseness of bibliometric point estimates.

Conference Topics
Methods and techniques; Citation and co-citation analysis; Indicators

Introduction
Field-normalized citation indicators such as the MNCS (Waltman, Eck, Leeuwen, Visser, & Raan, 2011) normalize the citation rate of a given publication corpus based on expectancy values of subject categories which correspond to the respective average citation rates within a research field (Vinkler, 1986; Mcallister, Narin, & Corrigan, 1983). Field normalization has been developed in order to neutralize the obvious diversity of publication and citation practices between field and subfields, as a corrective to otherwise unfair comparisons between the citation impact results of corpora with varying subject distributions.
Various methods for field delineation have been proposed (Glänzel & Schubert, 2003; Glänzel, Thijs, Schubert, & Debackere, 2009; Ruiz-Castillo & Waltman, 2014) including many proposals for clustering methods and arguments to determine the correct levels of aggregation. So far, however, no classification systems other than those provided by the database vendors could be established as standard throughout the bibliometrics community.
However, it is easily observable that the classification of the WoS subject categories diverges in size and specificity. Van Eck et al. (2013) provide furthermore strong evidence of heterogeneity within the medical subject categories along the characteristics of clinical and experimental research: After terms have been extracted from titles and abstracts, substructures are made visible by a term cloud procedure. These substructures can be assigned intellectually to clinical or experimental research and differ significantly in their citation rates along these dimensions. An intuitive explanation for this phenomenon would be the assumption that clinical researchers cite experimental studies, but that experimental researchers cite clinical studies only to a lesser extent.
Van Eck et al. (2013) draw the conclusion that the impact of clinical research is structurally underestimated by classical normalized citation indicators. The substructures made visible correspond to a facet that can be seen as transverse to a valid and comprehensible classification according to medical fields such as Clinical Neurology, Cardiac &

[*] The order of authorship is merely alphabetical.

Cardiovascular Fields, etc. Further theoretical issues beyond classification or clustering criteria seem to be not yet solved: If, for example, publications in so called hot topic areas are compared only with similar publications, even only with those who share not only the same topic, but also the same instruments, etc.? This could be seen as an over-normalization (Sirtes, 2012b; Sirtes, 2012a). Or is it legitimate to aggregate hot topics with less active research areas and thereby highlight the former as particularly successful? With the latter attitude the strategic decision of a researcher for a high impact research fields would be gratified while at the same time an implicit premise would be set that not all delineable areas in a functionally differentiated research landscape would be of equal value, insofar impact differences, which are effects of the functional differentiation, would not be neutralized.

By introducing finer classification systems these issues are addressed, although not answered based on theoretical reasons, as only further normalization options are created, whereas the resulting differences are not directly interpretable. Besides, in-house classifications systems are not easily compatible with a desirable trend towards greater standardization and reproducibility in the bibliometric community.

In the present paper we introduce a concept for quantifying heterogeneity differences within subject categories and thus maintain the WoS subject categories as basis for the field normalization, as they provide community-wide comparability and mutual reproducibility. Heterogeneity differences between subject categories are quantified and used to construct error or stability intervals, which can be integrated into the calculation of the total impacts of an institution or a country as before. The approach thus combines two advantages: on the one hand, we continue to work at the level of a standard classification system and on the other hand, underlying structures on a secondary level are made transparent.

Methods and Data

Keyword terms of all articles, reviews and letters published in journals of two medical subject categories (Parasitology (P), Otorhinolaryngology (O)) of the publication year 2008 have been extracted.[1] WoS keywords are not a controlled vocabulary like, e.g., Medical Subject Headings in PubMed/Medline and are therefore not per se complete and normalized. Table 1, however, shows that the amount of publications that have not assessed with keywords is relatively small. Keywords have, on the other hand, the advantage of simple accessibility; it is not necessary to exclude i.e. filler words. In order to accomplish a basic normalization, a stemming procedure is carried out which neutralizes different inflexions.

All distinct keyword terms are normalized with an Oracle Text stemming function and coupled by the *contains* function, again as provided in Oracle Text. Stemmed terms must therefore not be necessarily identical, but one term can contain the other, respectively. This also applies to keywords, which are phrases and may contain single keywords and be thus coupled with them. These keyword pairs are used for a coupling procedure of the corresponding publications; Salton's Cosine is used to neutralize differing amounts of keywords.

With the aim to reproduce the visual substructures of Van Eck et al. (2013) in a first step with our cluster procedure, these two subject categories have been chosen as they display different types of sub-structures in the discussed work. Parasitology displays quite distinct structures with three visible clusters seemingly characterized by significant differences in citation levels whereas Otorhinolaryngology displays a more fuzzy structure.[2]

[1] All calculations are processed in an Oracle database of WoS raw data (SCI, SSCI, A&HCI, CPCI-S, CPCI-SS) frozen in the 17th calender week 2013.

[2] http://www.neesjanvaneck.nl/basic_vs_clinical/

Table 1: Share of publications with keywords.

	Parasitology	*Otorhinolaryngology*
JARL 2008 (all)	3727	5122
JARL 2008 (percentage of publications with keywords)	98.0%	90.6%

The ratio of realized to theoretical possible relations between all items gives an impression about the broadness of the empirical basis of the coupling results. Table Table 2 gives the percentage of realized to theoretically possible relations of all publications (JARL = Articles, Letters and Reviews with publication type Journal Article) in 2008.

Table 2: Ratio realized relations to possible relations.

	Parasitology	*Otorhinolaryngology*
JARL 2008 (all)	18.2%	11.3%
JARL 2008 (only with keywords)	19.0%	13.8%

The resulting distance measures for publication pairs are imported into the statistical program R, converted into dissimilarity values and the clustering method Ward is used. Ward as a standard hierarchical-agglomerative clustering procedure was chosen, because it is crucial for our approach to have a clustering procedure which does not require a fixed number of clusters as parameter. Furthermore, single linkage with its well-known tendency to dilated cluster structures seems to impose to weak requirements on the clusters' homogeneity and complete linkage too strong requirements.
The usual cut-off-value of 5 was determined manually; however in future iterations of the procedure the optimal cut off value will be estimated.
As shown in Table 2 not all publications in the respective sets are actually assigned with keywords, thus we have added a non-keyword cluster with its mean citation rate in order to represent all publications in our dataset. This appears as a legitimate solution given that fact that non-keyword items have considerably smaller mean citation rates compared to the whole subject category and have to be taken into account in order to appropriately represent the SC.

Results

The visualization for the subject category parasitology as resulting from (Van Eck et al.., 2013) indicates a distribution of three discernable substructures which are clearly different in citation level. With our method, we arrive at eleven clusters. Table 3 shows four of the top keywords[3] and the respective mean citation rates, whereas Figure 1 gives the frequency distribution of the clusters (as the width of the bars) and the mean citation rates in a histogram. The topics of the clusters can only partially confirm Van Eck et al.'s conclusion. The keywords of cluster 5, 6, and 7 have all clear connection to experimental laboratory research, however only 5 (with the most distinctly molecular biology focus) has a very high citation rate compared to the rest. It is possible, that parasitology is rather a special case compared to other medical SCs, as it also encompasses topics such as classical biology (cluster 1), epidemiology (clusters 2 and the more clinical 4), a veterinary cluster (8), and clusters that are joined by common parasites (3, 9,10, and 11).

[3] All keywords were in the top 10 most frequent ones. Redundant keywords (like 'plasmodium' and 'plasmodium falciparum') and keywords that were not informative in understanding the topic of the cluster (like 'parasites') were excluded.

Table 3 - Top keywords and mean citation rate of keyword clusters in parasitology (ordered by cluster size).

Cluster	Top Keywords				Mean Citation Rate
1	Phylogeny	Evolution	Ecology	Morphology	3.91
2	Infection	epidemiology	Seroprevalence	Antibodies	5.76
3	Malaria	plasmodium falciparum	infected erythrocytes	cerebral malaria	6.25
4	Transmission	Children	Resistance	Efficacy	7.02
5	Expression	in-vitro	Protein	gene-expression	7.57
6	Mice	in-vivo	dendritic cells	immune-response	6.69
7	Identification	PCR	linked-immunosorbent-assay	Antibodies	5.50
8	Sheep	Cattle	haemonchus-contortus	Ivermectin	4.11
9	Disease	trypanosoma cruzi	chagas disease	risk-factors	6.09
10	Diptera	Culicidae	aedes-aegypti	anopheles-gambiae	5.32
11	Cryptosporidium	Parvum	Giardia	Genotypes	7.88

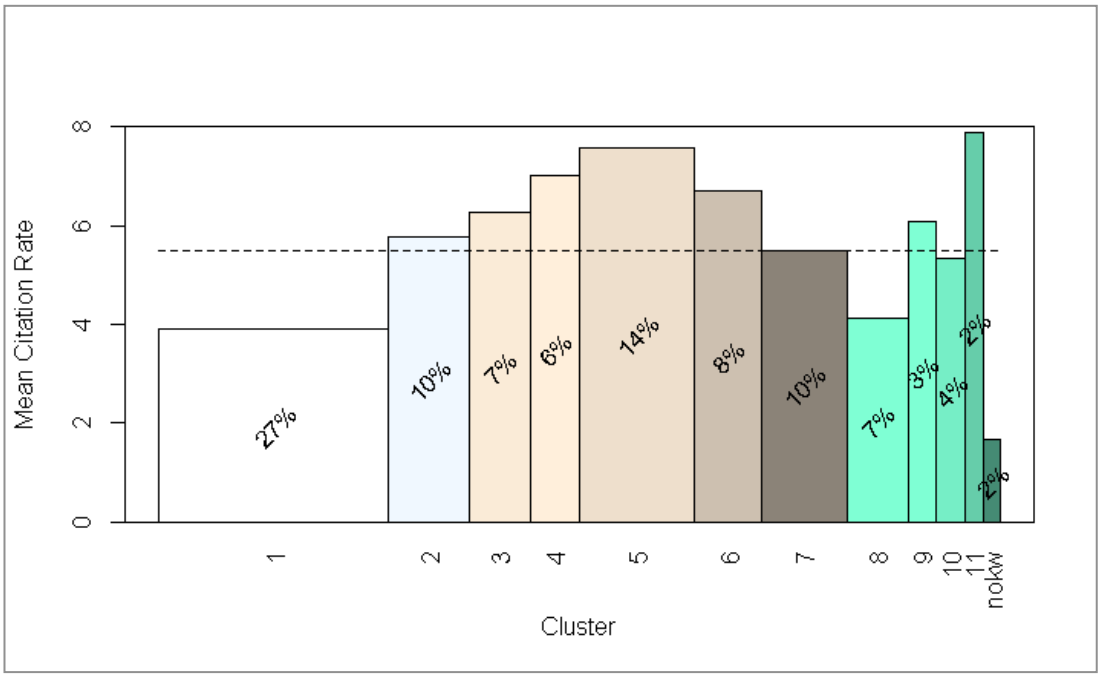


Figure 1: Share and Mean Citation Rate of Parasitology Clusters. The dotted line represents the MCR of the whole SC.

In the second case of otorhinolaryngology, the structure shown by (Van Eck u. a., 2013) is quite fuzzy and less-structured, which is mirrored by our cluster distribution. It consists of one larger and a considerable amount of very small cluster. There are also significant variations between mean citation levels ranging from around 2 to larger than 4, it is however more difficult to interpret the cluster's respective keyword frequencies.

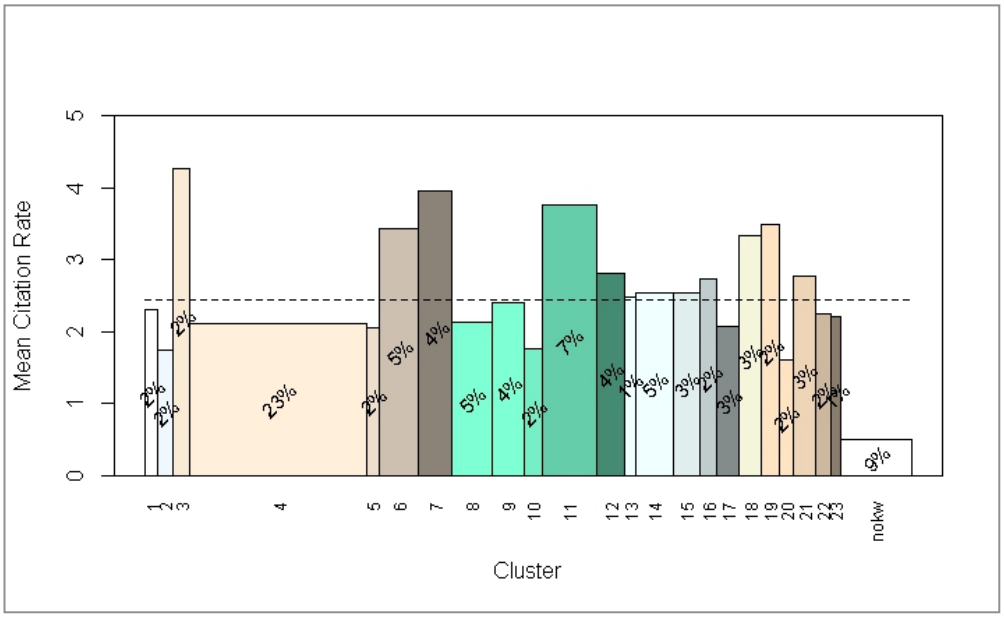


Figure 2: Share and Mean Citation Rate of Otorhinolaryngology Clusters. The dotted line represents the MCR of the whole SC.

In order to calculate the MNCS and its stability, sets of publications with an affiliation in Germany have been selected. The size of the sets were 208 (P) and 486 (O) publications respectively.

On the basis of the resulting cluster distributions, a bootstrapping approach has been utilized.

A set of MCR clusters equal to the size of the German set has been drawn with replacement from the clusters' MCRs with the probabilities equal to the clusters' share. The arithmetic mean of this combination has been calculated and served as the Expected Citation Score (ECS_i). Each raw citation score of the German papers was then divided by the ECS_i and the arithmetic mean of the results delivered the $MNCS_i$. 10'000 iterations of this procedure have been executed. The distribution of the scores are depicted in Figure 3.

Finally, the 2.5% and 97.5% quantiles of this distribution have been calculated.

The resulting MNCS 95% stability interval of the German set for parasitology ranges from 1.35 to 1.46 with an MNCS of 1.40 and for otorhinolaryngology from 0.87 to 0.93 with an MNCS of 0.9. Thus, although parasitology displays a much wider distribution, as can also be seen in Figure 3, the relative deviance of the MNCS ([95% range of $MNCS_i$]/MNCS) is quite similar with 7.3% and 6.7%, respectively.

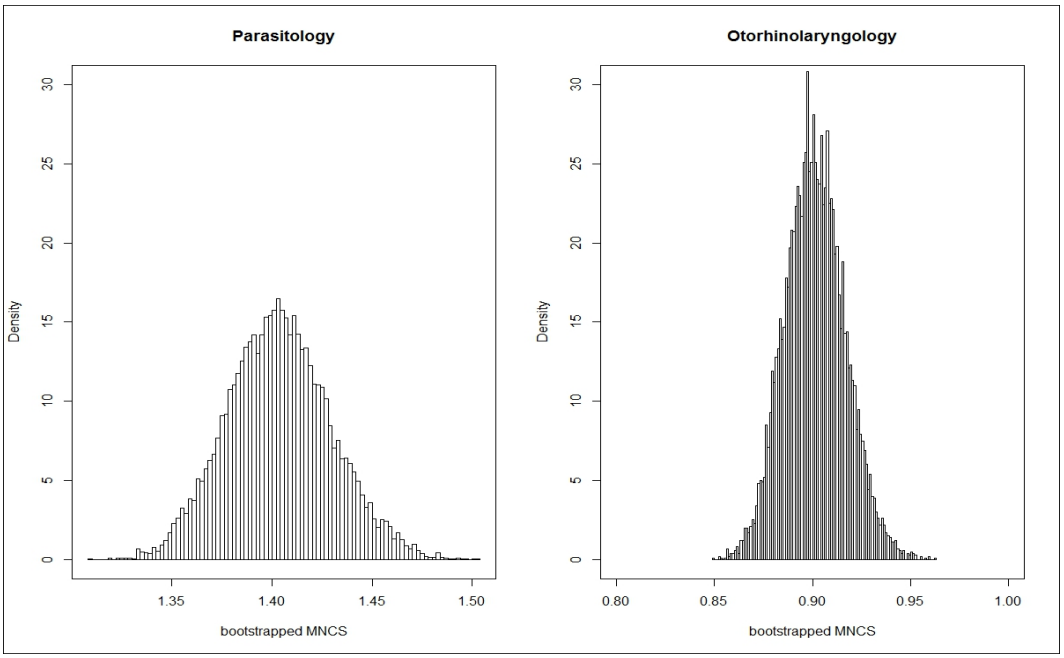


Figure 3: Distribution of $MNCS_i$ for German publications.

Discussion

These preliminary results show in the case of parasitology that clusters can be delineated and differing topical foci can be identified as well. While a dimension clinical versus experimental research is perceivable, other facets also occur: It may be the case that parasitology is a special SC as the clusters have also rather unusual topics compared to other medical disciplines such as classical biology, veterinary sciences and epidemiology. The Mean Citation Rates vary massively with a total range of MCRs of 3.97 citations per publication In the second case of otorhinolaryngology, the cluster distribution is less harmonic, more frayed out and not easily interpretable (confirming here the results of (Van Eck et al., 2013)). The coupling procedure succeeded on a relatively smaller amount of publications and many more clusters have been created. Furthermore, the citation levels are all much lower and the range of MCRs, the publications without keywords notwithstanding, have only a total range of 2.6 citations per publication.

The hitherto work was intended as a proof of concept: We were able to show that subject category substructures with different citation levels exist. Differences in citation homogeneity are however not in both cases concordantly attributable to topical structures. For the current state of this work, some simplifications have been applied: Citation rates should be processed and normalized document type-specific as articles, letters and reviews are cited differently. However, citation level differences in our results are so clear and dominant that they couldn't possibly only be caused by different document type patterns in the clusters. For a final implementation of this method, the calculations will be processed document type-specific and the expansion of the method to sets of multiple SCs, including an SC fractionalization will be developed. An exclusion of letters might be contemplated as for example about half of the publications without keywords in otorhinolaryngology are letters (about three quarters of all letters in this SC). Furthermore, parameters of the study like the clustering method and definition of cut off-values will be systematically varied and analyzed. It is even conceivable to calculate such stability intervals on the basis of percentile based indicators, which are less sensitive to outliers than the MNCS. However, already as it stands this method shows promise in circumventing to problem of calculating normalized citation scores on non-standard classification schemes while taking into account the heterogeneity of research areas in the

classical WoS SC classification. This method could also be combined with already existing bootstrapping methods of the publications sets themselves as implemented for example in the Leiden Ranking (www.leidenranking.com). Together they could account for both the robustness of the citation scores given the size and distribution of the publication sets themselves, as well as the underlying uncertainty of the expected citation rates. We believe that such methods that display the coarseness of bibliometric point estimates, which especially clients of evaluative bibliometric analyses are prone to disregard and thus revel or despair at minute changes of their scores and ranks, are an important step to the correct interpretation of bibliometric indicators and crucial for the development of bibliometrics into a mature science.

References

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367. http://doi.org/10.1023/A:1022378804087

Glänzel, W., Thijs, B., Schubert, A., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, *78*(1), 165–188. http://doi.org/10.1007/s11192-008-2109-5

Mcallister, P. R., Narin, F., & Corrigan, J. G. (1983). Programmatic evaluation and comparison based on standardized citation scores. *IEEE Transactions on Engineering Management EM*, *30*(4), 205–211. http://doi.org/10.1109/TEM.1983.6448622

Ruiz-Castillo, J., & Waltman, L. (2014). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Working Paper*. Abgerufen von http://e-archivo.uc3m.es/handle/10016/18385

Sirtes, D. (2012a). Finding the Easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics*, *6*(3), 448–450.

Sirtes, D. (2012b). How (dis-)similar are different citation normalizations and the fractional citation indicator? (And How it can be Improved). In *Proceedings of 17th International Conference on Science and Technology Indicators* (S. 894–896). Montréal: Éric Archambault, Yves Gingras, and Vincent Larivière.

Van Eck, N. J., Waltman, L., Van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS ONE*, *8*(4), e62395. http://doi.org/10.1371/journal.pone.0062395

Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, *10*(3-4), 157–177. http://doi.org/10.1007/BF02026039

Waltman, L., Eck, N. J. van, Leeuwen, T. N. van, Visser, M.S., & Raan, A. F. J. van. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics*, *87*(3), 467–481. http://doi.org/10.1007/s11192-011-0354-5

Measuring Interdisciplinarity of a Given Body of Research

Qi Wang

qi.wang@indek.kth.se
KTH Royal Institute of Technology, Lindstedsvägen 30, SE-100 44 Stockholm (Sweden)

Abstract
Identifying interdisciplinary research topics is an essential subject, not only for research policy but also research funding agencies. Previous research was constructed on measuring interdisciplinarity mainly at the macro level of research, such as Web of Science subject category and journal. However, these studies lack analysis at the micro level of the current science system. It means few studies have analyzed interdisciplinarity at the level of publications. To cover this gap, we introduce an approach for measuring interdisciplinarity at the level of micro research topics. The research topics are clustered by direct citation relations in a large scale database. According to the characteristics of boundary-crossing research, we provide an alternative approach to measure interdisciplinarity. Comparing with the widely used Rao-Stirring indicator (Integration score), we found that the results obtained by two indicators of interdisciplinarity have a strong correlation, thus we believe that this approach could effectively identify boundary-crossing research topics.

Conference Topic
Indicators

Introduction
In bibliometric and scientometric research, measuring interdisciplinarity is a difficult yet important topic. However, although it has been widely recognized that interdisciplinary research solves complex problems, promotes scientific developments and innovations, there is still no consensus on how to define and measure this type of research. Specifically, a variety of definitions on boundary-crossing research have been proposed, such as interdisciplinary multidisciplinary, transdisciplinary and cross-disciplinary; however the definitions of each term as well as discriminations among them are quite ambiguous (for more details see Huutoniemi K. et al., 2010; Wagner C.S. et al., 2011). In a broad sense, these concepts all refer to the research that cross boundaries between disciplines. We do not intend to explore the nuances among the concepts in this study. Thus, at the very beginning of this article we need to emphasis that, for the purpose of this research, the term *interdisciplinary research topics* used to refer to all type of boundary-crossing research, in other words, it covers all type of research with interdisciplinarity.

Furthermore, due to the controversy in defining research with interdisciplinarity at the conceptual level, there is no consensus on how to measure interdisciplinarity in practices. Various approaches are utilized to analyze interdisciplinarity, including both quantitative methods such as bibliometric indicators, text-mining and qualitative methods such as interviews and surveys. In particular, bibliometric approaches have been widely applied to measure and identify interdisciplinarity, such as citation-based indicators (Porter & Chubin, 1985; Leydestorff, 2007; Porter, Roessner & Heberger, 2008; Porter & Rafols, 2009; Rafols & Meyer, 2010; Leydestorff & Rafols, 2011; Rafols et al., 2012; Lariviere & Gingras, 2014), author-based indicators (Qin et al., 1997; Schummer, 2004; Abramo et al., 2012), as well as similar indicators but relying on a variety of classification systems of science (Tijssen, 1992; Morillo, Bordons, & Gomez, 2001; 2003; Braun & Schubert, 2003; Sugimoto, 2011;

Sugimoto et al., 2011). Additionally, a few studies have applied text-mining approaches, LDA for example, to explore interdisciplinarity of a given issue (Wang et al., 2013; Nichols, 2014).

In this article, we explore a citation-based measurement for identifying interdisciplinary research topics at the level of publications. We also use the Web of Science (WoS) classification system, but with a different approach. More specifically, we first construct micro research topics based on the direct citation relations among individual publications. Meanwhile, the publications are assigned into one or several subject categories on the basis of the journal where the publication has appeared and of WoS classification system. It implies that a research topic constructed might belong to one or several WoS subject categories according to publications within the cluster. In other words, WoS subject categories that attached to publications are regarded as traditional boundaries of scientific disciplines, whereas micro research topics constructed on the relatedness among publications might break the existing knowledge boundaries. We assume, then, that a cluster can be regarded as an interdisciplinary research topic if there is a considerable number of within-cluster citations spanning distant WoS subject categories. The indicator proposed in this article combines knowledge diversity with knowledge integration, in which heterogeneity and connectedness of subject categories within research topics are taken into account. It provides an alternative approach to measure interdisciplinarity and simplifies the previous citation-based approaches.

Data and Methodology

This study was based on data from the in-house WoS database of the Centre for Science and Technology Studies (CWTS) of Leiden University. The database used in this study covers the period from 2002 to 2013, a 10-year period. The total number of publications in our database is about 9 million. The methodology that we introduce for measuring interdisciplinarity of micro research topics can be divided into three steps.

Step 1 Clustering publications into micro research topics

The clustering method is mainly based on the previous studies by Waltman & van Eck (2012; 2013). First, the relatedness of publications was measured by the normalized direct citation relation among individual publications (for details see Waltman & van Eck, 2012). Furthermore, based on the relatedness matrix, an improved Louvain algorithm (Blondel et al., 2008), namely a 'Smart Local Moving algorithm' (SLM) was applied to cluster individual publications (for details see Waltman & van Eck, 2013). Labels of each cluster were selected from titles and abstracts of publications within cluster (for details see Waltman & van Eck, 2012).

Measuring interdisciplinarity on the level of micro research topics, constructed based on the citation relations, is one of the most important distinctions between this study and previous research. There are two reasons for measuring the degree of interdisciplinarity in this approach. First, WoS subject categories attached to journals cannot properly describe publication itself. For instance, although *Journal of the Association for Information Science and Technology* belongs to two categories, INFORMATION SCIENCE & LIBRARY SCIENCE and COMPUTER SCIENCE, it does not necessarily mean that all publications appeared in this journal span the two categories. More generally, some publications associated with the category of INFORMATION SCIENCE & LIBRARY SCIENCE and others related to the category of COMPUTER SCIENCE. The second reason is that WoS assigned journals such as *Nature*, *Science*, and *Plos One* as MULTIDISCIPLINARY SCIENCE. Instead of focusing on a specific scientific field, this sort of journals covers almost the full range of scientific disciplines. When measuring interdisciplinarity on the level of journals, this sort of journals may have high interdisciplinarity scores. However, although the journals are composed of publications

spanning over different scientific disciplines, it does not necessarily mean the integration of knowledge from various sources exists.

In order to avoid the problems mentioned above, we constructed micro research topics based on the relatedness of individual publications, which are expected to provide a more accurate body of research topics within the current science system.

Step 2 Calculating a similarity matrix of ISI subject categories

Porter and Rafols (2009) analyzed a sample of more than 30,000 WoS publications and their cited references, in which publications were assigned to subject categories on the basis of the WOS classification of journals the publications appeared. They constructed a matrix of subject categories using the relations of articles and their cited references, and then applied Salton's cosine (Salton & McGill, 1983) to obtain the similarity matrix of subject categories. The similarity value s_{ij} is high if subject category i and j are cited a lot by the same publications.

However, in this study, two subject categories are considered to be strongly related if they both cite a lot to the same subject categories. Specifically, the construction of a similarity matrix of subject categories is done in two steps.

In the first step, for each pair of a citing subject category i and a cited subject category j, the number of citations from publications in subject category i to publications in subject category j is counted. We use c_{ij} to denote the number of citations from publications in subject category i to publications in subject category j. Note that according to the WoS classification system, one journal might be attributed into multiple subject categories. Therefore a fractional counting strategy is adopted to handle publications belonging to more than one subject category.

The second step is to construct a similarity matrix of subject categories based on the citation matrix created in the first step. The cosine similarity measure is used for this purpose. Hence, the similarity of two subject categories i and j is given by

$$s_{ij} = \frac{\sum_k c_{ik} c_{jk}}{\sqrt{\left(\sum_k c_{ik}^2\right)\left(\sum_k c_{jk}^2\right)}}$$

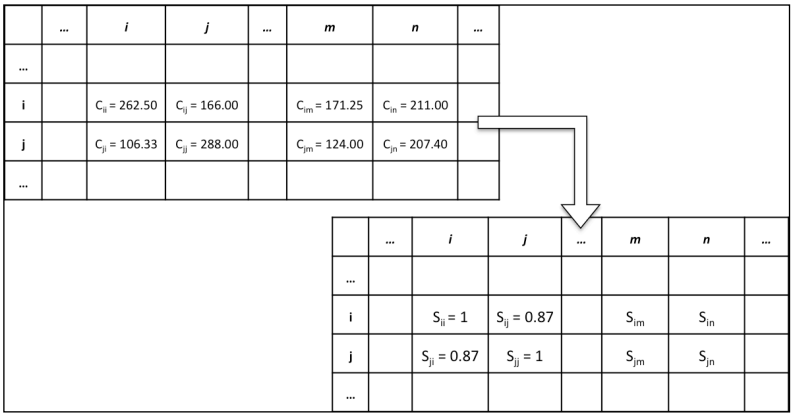


Figure 1. An example of the formula for calculating similarity.

Figure 1 can be used as an example to illustrate how the formula of similarity applied. The top left table is the matrix of citation relations among subject categories, which is not symmetric. Since a fractional counting strategy is used in this study, the numbers of citations are not always integers. As we mentioned above, c_{ij} means the number of citations from subject category i to j. Moreover, according to the above formula, we obtained the symmetric

similarity matrix of subject categories, which is shown in lower right of figure 1. In this case, subject category *i* and *j* are all cite a lot to the categories *i*, *j, m* and *n.* Therefore, the similarity between *i* and *j* is quite high, that is 0.87.

In short, using the cosine similarity measure, s_{ij} is high if publications in the two categories tend to cite the same categories. If publications in two subject categories tend to cite completely different categories, the similarity between the categories is low.

Step 3 Determining the degree of interdisciplinarity

As mentioned above, we suppose that a research topic could be regarded as an interdisciplinary research topic should satisfy two criteria; one is that it contains distant subject categories, the other is there are citation relations among different subject categories within this topic. In short, a cluster that is consisted with citation relations spanning different subject categories might be an interdisciplinary research topic.

Following the criterion discussed above, we explore the indicator to measure interdisciplinarity, whose formula is as follows:

$$Interdisciplinarity = \frac{1}{n\_cit}\sum_i^k \sum_j^k n\_cit_{ij} d_{ij},$$

where $d_{ij} = 1 - s_{ij}$. Within a cluster, n_cit_{ij} is the number of citations between subject categories i and j, and n_cit is the sum of citations obtained by $n\_cit = \sum_i^k \sum_i^k n\_cit_{ij}$. The indicator includes three attributes: *variety*, the number of subject categories within a cluster (denoted as *k*), *connectedness*, the number of cross-citations (denoted as n_cit_{ij}) and *distance*, the degree of distinctiveness between subject categories (denoted as d_{ij}). In short, a research topic can be considered to be more interdisciplinary if the citation relations within that cluster cross various WoS subject categories.

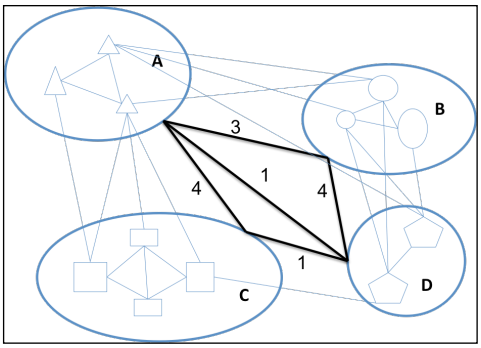


Figure 2. An example of the citation relations within a research topic.

Figure 2 shows a research topic including 12 publications that belong to 4 subject categories. The black lines represent the citation relations among different subject categories, and the blue lines are the links within the same category. In our measurement, the citations crossing subject categories (black lines in the Figure) and distances of subject categories are taken into account.

Results

Clustering analysis

Table 1 provides the basic statistic results of original and restricted database. The restricted database was constructed based on two criteria. First, we expect to analyze research topics with a relatively large number of publications only. Therefore, we set a restriction on the number of publications of each cluster so that clusters with more than 100 publications could

be advanced in the next step. Second since the accuracy of measurement is highly related to the quality of clustering results, we reviewed the clusters with the indicator, *mean citation score*. It obtained by using the total number of citations divided by the total number of publications within a cluster. If the number of citations is less than the number of publications of a cluster, publications belong to the cluster are connected loosely, resulting in the emergence of clusters with poor qualities. In this case, we found 667 clusters with low mean citation scores (defined as less than 2), which accounted for 7% of the total. Thus, it turns out that most of clusters have relatively strong interconnections. The analysis in the following sections is performed base on the restricted database.

Table 1. Basic statistic results of original and restricted database.

	# of pubs	# of topics	Average pubs	Max pubs	Min pubs	St.d pubs
Original	9,146,302	9,565	956	10744	1	1026
Restricted	8,930,360	7,864	1,135	10744	100	1040

Similarity matrix

Using Salton's cosine (Salton & McGill, 1983), we obtained a similarity matrix of WoS subject category, the range of similarity values is between 0 and 1. It implies that the similarity s_{ij} is zero if subject category i and j never cite to the same categories, whereas s_{ij} approaches one if they both cite a lot to the same categories. To test the accuracy and reliability of our similarity matrix, we have compared it with the one obtained by Porter & Rafols (2009), whose method have been introduced above. As expected, the result shows there is positive correlation between the two matrices ($r = 0.7405$). In general, we believe that the results obtained from the two approaches with slight differences are consistent.

Interdisciplinarity of research topics

The average interdisciplinarity score of each research topic is about 0.42 with a standard deviation of 0.11. The largest score is 0.72 associated with the research on respiratory system, while the lowest is close to 0.0086. The distribution of research topics over the interdisciplinarity score is shown in figure2. As can be seen, the majority of research topics have interdisciplinarity scores between 0.35 and 0.55.

In order to better interpret the results, we aggregated the WoS subject category into five main fields according to the Leiden Ranking 2013. Table 2 lists the five main fields. Specifically, a publication appearing in one or several main fields is based on the journal where it has been published. When a publication has appeared in a journal of multi-assignation and these subject categories are assigned into different main fields, the publication is expected to appear in more than one field (more details see CWTS Leiden Rank 2013, pp4). Thus, a research topic might be assigned into several main fields if the publications within this topic belong to more than one field.

Before turning to the interdisciplinarity score, we emphasize that it is quite difficult and almost impossible to define a clear cutting-off point between interdisciplinary and non-interdisciplinary research topics. Considering the difficulty, we selected the research topics with an interdisciplinarity score greater than 0.6143, which account for around 1% of the total. For the purpose of understanding the knowledge integration across main fields in the macro level, we applied following strategy. Regarding a research topic, if the number of publications in one main field is larger than 50% of the total, then the topic is assigned into this main field. Otherwise, the research topic would be assigned into its two dominant main fields. In doing so, the select topics (top 1% of the total) are tabulated in Table 3, in which each row is the main field with the most number of publications and each column is the main field holding the second number of publications. For instance, in the first row, 1 means there

is one research topic whose publications mostly appear in main fields 1 and 2, as well as main field 1 has the most number of publications.

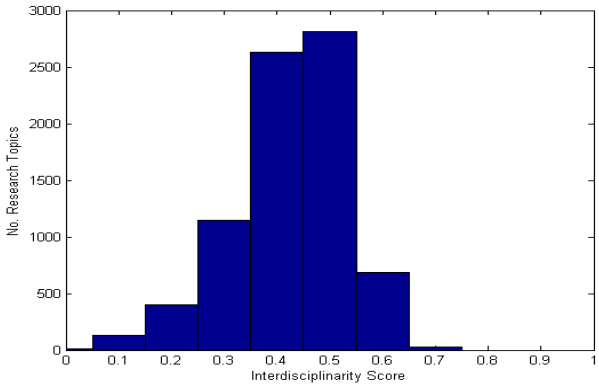


Figure 3. Distribution of research topics over interdisciplinarity score.

Table 2. Labels of main fields.

ID	*Labels of Main Fields*
Main Field -1	Social sciences & humanities
Main Field -2	Biomedical &health sciences
Main Field -3	Natural sciences & engineering
Main Field -4	Life & earth sciences
Main Field -5	Mathematics & computer science

Table 3. Distribution of research topics over the main fields.

	Main field-1	Main field-2	Main field-3	Main field-4	Main field-5	Total
Main field-1	11	1	0	0	0	12
Main field-2	1	33	6	1	2	43
Main field-3	0	2	25	1	0	28
Main field-4	0	0	1	8	0	9
Main field-5	0	2	1	0	5	8

As can be seen, most research topics in the top 1% of the total belong to the main field 2, that is BIOMEDICAL & HEALTH SCIENCES. Meanwhile, among the research topics that across two main fields, the topics whose publications mainly appear in the main field 2 contribute the largest proportion. Primarily, this is because the most number of research topics fall into this main field. In addition, the research conducted by Porter & Rafols (2009) have demonstrated that subject categories MEDICINE- RESEARCH & EXPERIMENTAL and NEUROSCIENCES have high degrees of interdisciplinary according to the Integration score (aka, Rao-Stirling's diversity) (more details see Porter & Rafols, 2009, pp723). In our classification system, the two subject categories both belong to main field 2, which is partially verified that the main field of BIOMEDICAL &HEALTH SCIENCES has relatively high interdisciplinarity. Main field 5, that is MATHEMATICS & COMPUTER SCIENCE, holds the smallest number of research topics with high interdisciplinarity, as shown in table 3. This result is also consist with the research by Porter & Rafols (2009), in which they showed subject category MATHEMATICS that is assigned into main field 5 in our study has the lowest integration score between 1975 and 2005.

For the purpose of examining the quality of the indicator, we now take a more derailed look at research topics. In doing so, we randomly select 5 research topics from the top 1%, one from

each main field. For each research topic, Table 4 gives the three most important subject categories and the two most cited publications.

Table 4. Selected research topics with high interdisciplinarity.

Cluster ID	Information of Publication	
4323	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -1 (53%); Main Field -4 (27%) 705 56 VETERINARY SCIENCES (244); SOCIOLOGY (225); PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH (47) Rijken M et al. (2005). Comorbidity of chronic diseases - Effects of disease pairs on physical and mental functioning (88) Odendaal J.S.J. & Meintjes R.A. (2003). Neurophysiological correlates of affiliative behaviour between humans and dogs (82)
3644	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -2 (54%); Main Field -3 (25%) 875 36 RADIOLOGY, NUCLEAR MEDICINE & MEDICAL IMAGING (715); NUCLEAR SCIENCE & TECHNOLOGY (533); ENVIRONMENTAL SCIENCES (464) Stabin M.G. et al. (2005). OLINDA/EXM: The second-generation personal computer software for internal dose assessment in nuclear medicine (370) Gorden A.E.V. et al. (2003). Rational design of sequestering agents for plutonium and other actinides. (227)
4083	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -3 (74%); Main Field -2 (13%) 760 63 NUCLEAR SCIENCE & TECHNOLOGY(282); INSTRUMENTS & INSTRUMENTATION (259); PHYSICS, NUCLEAR (255) Spalding K.L. et al. (2005). Retrospective birth dating of cells in humans (182) Lappin G. & Garner R.C. (2003). Big physics, small doses: the use of AMS and PET in human microdosing of development drugs (137)
7577	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -4 (50%); Main Field -3 (46%) 190 26 ASTRONOMY & ASTROPHYSICS(100); GEOSCIENCES, MULTIDISCIPLINARY (81); METEOROLOGY & ATMOSPHERIC SCIENCES (67) Rietveld M.T. et al. (2003). Ionospheric electron heating, optical emissions, and striations induced by powerful HF radio waves at high latitudes: Aspect angle dependence (91) Pedersen T.R. et al. (2003). Magnetic zenith enhancement of HF radio-induced airglow production at HAARP (45)
8434	Main field (R_pubs) T_pubs Rank Subject Categories (N_pubs) Title (Times cited)	Main Field -5 (55%); Main Field -3 (34%) 108 99 ROBOTICS (49); COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE (34); INSTRUMENTS & INSTRUMENTATION (22) Vergassola M. et al. (2007) 'Infotaxis' as a strategy for searching without gradients (103) Yoerger D.R. et al. (2007). Techniques for deep sea near bottom survey using an autonomous underwater vehicle (38)

Take two clusters as examples, cluster 3644 and cluster 4083 are randomly selected from BIOMEDICAL & HEALTH SCIENCES and NATURAL SCIENCES & ENGINEERING respectively; however, the two most frequent main fields of both clusters are the same. Apart from that, as can be concluded from table 4, most publications of both clusters belong to the

subject category of NUCLEAR SCIENCE & TECHNOLOGY. Hence we infer that the two research topics are similar at a certain degree. Observing the detailed information of publications in each cluster, we found that both clusters are related to the research on nuclear medicine, that is "a medical specialty involving the application of radioactive substances in the diagnosis and treatment of disease"[1]. However, there is a considerable difference in terms of the degree of interdisciplinary score. Cluster 3644 is much more interdisciplinary than cluster 4083 as shown from table 4. To understand the differences, we visualized the two clusters using the map of subject categories.

The map of subject categories can represent the position of a cluster in the global map of science, as well as show whether the cluster has the characteristics of interdisciplinary research. For instance, we can observe from the map of subject categories whether clusters are dispersed over many distant subject categories. The software VOSviewer (van Eck & Waltman, 2010) was used to construct the map of subject categories. In this study, the baseline map was generated by the citations between WoS subject categories using publications from 2002 to 2013. Figure 4 and 5 were generated by overlaying on the baseline map with circles, in which size of circles represents the number of publications in each WoS subject category, nodes represent subject categories, as well as links shows citations among them.

Comparing the two figures, we found that cluster 3644 are more diverse that it contains citations spanning various subject categories with larger distances (i.e. COMPUTER SCIENCE THEORY AND METHOD, ENGINEERING ELECTRICAL AND ELECTRONIC), as well as its number of publications in various subject categories are quite even. Thus, it is reasonable that cluster 3644 has a higher interdisciplinary score than cluster 4083, although they have a similar research topic. Meanwhile, it can be inferred that the two clusters have different research focuses since the subject categories with the most number of publications of the two clusters are quite different. That also explains why publications with a similar research topic were classified into two clusters.

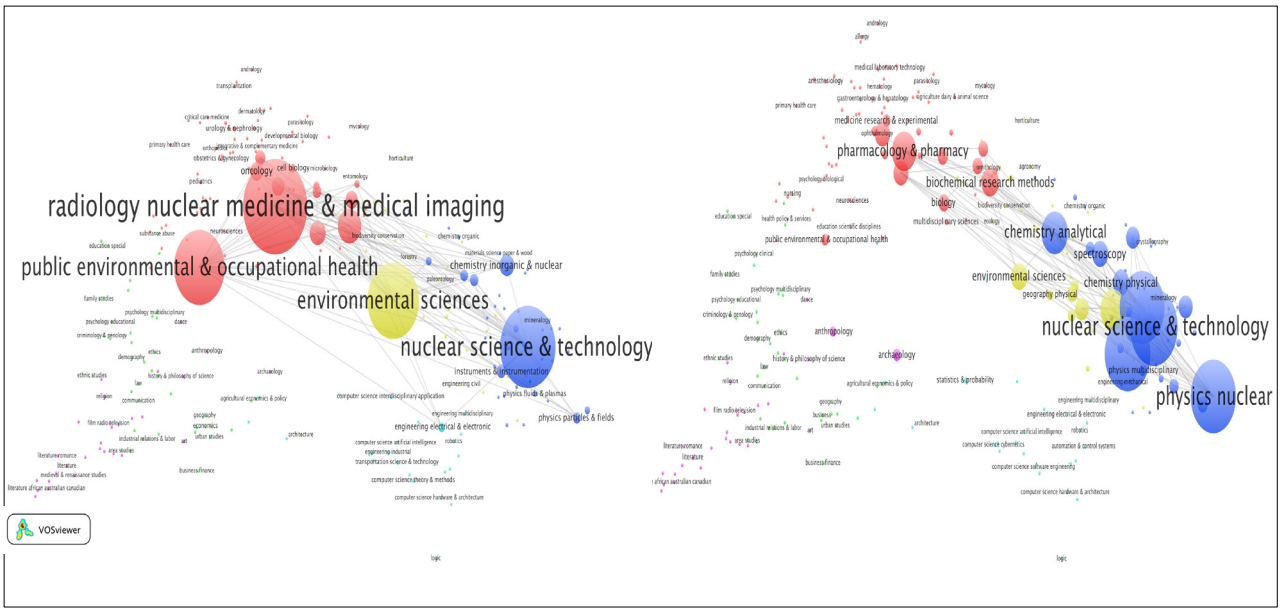


Figure. 4. A map of subject categories (note: the left panel is cluster 3644; the right panel is cluster 4083).

[1] http://en.wikipedia.org/wiki/Nuclear_medicine.

An example of Information Science and Library Science. Readers of this paper might be familiar with research in the field of information and library science; therefore, we now take a specific look at a cluster in this subject category. To give an example, we select the cluster that holds the highest interdisciplinarity value among all the clusters whose most publications belong to this subject category. In doing so, we obtained cluster 4982, which ranks 72 among the top 1% most interdisciplinary clusters. The detailed information of this cluster is shown in table 6.

As can be seen, the cluster includes 565 publications, and most of them belong to main fields of SOCIAL SCIENCES AND HUMANITIES and MATHEMATICS AND COMPUTER SCIENCE, that fit what figure 10 shows. Moreover, it also can be seen that this research topic covers various subject categories, such as computer science research, ergonomics, business, laws, and psychology. Furthermore, based on the most cited publications and the figure of citation network of this cluster, we can estimate that this research topic is rated to the research on information privacy. This is probably in line with what our cognition, that research on information privacy involves studies on either information or computer technology, or social science research such as law and psychology, or studies which overlap the two types of research.

To find more evidence, we searched the courses related to information privacy in MIT OpenCourseWare, using "information privacy" as the key words. Then, 1150 results have been obtained. The courses include from *The Economics of Information*, *Communications and Information Policy* to *Biomedical Computing*, *Information and Entropy*. That proves the research topic of information privacy is interdisciplinary in character.

Table 5. Publication information of cluster 4982.

Cluster ID	Information of Publication	
4982	Main field (R_pubs)	Main Field -1 (52%); Main Field -5 (44%)
	T_pubs	565
	Rank	72
	Subject Categories (N_pubs)	COMPUTER SCIENCE, INFORMATION SYSTEMS (141); BUSINESS (108); INFORMATION SCIENCE & LIBRARY SCIENCE (107)
	Title (Times cited)	Malhotra N.K., Kim S.S. & Agarwal J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model (169)
		Nissenbaum H. (2004). Privacy as contextual integrity (110)

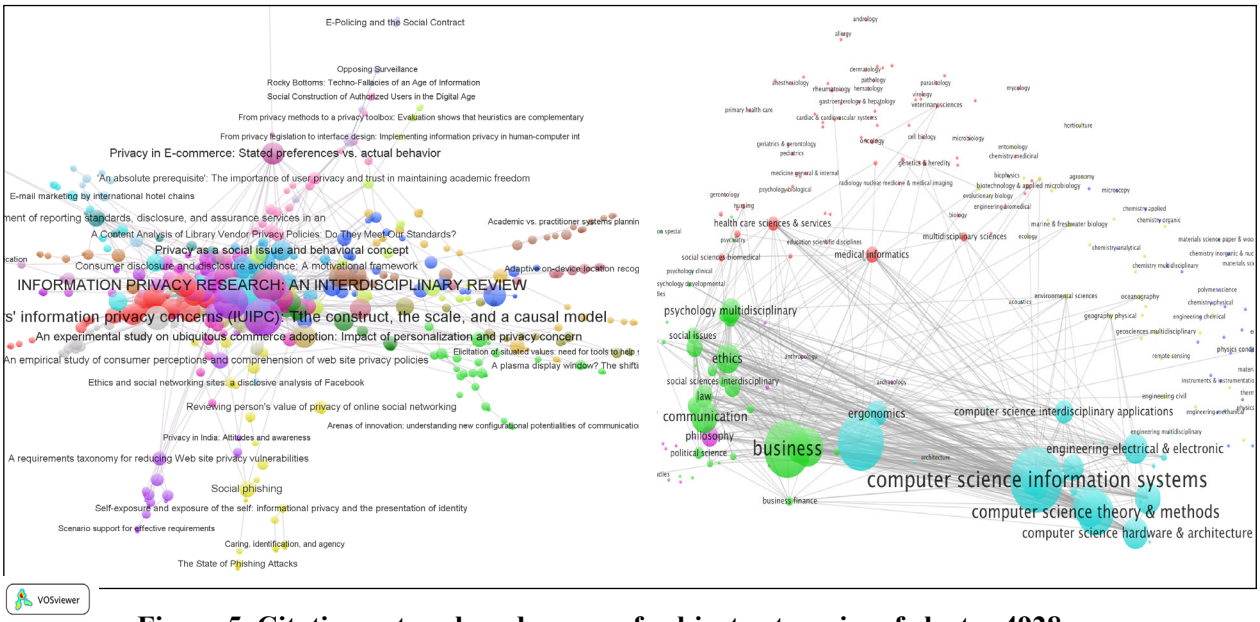


Figure 5. Citation network and a map of subject categories of cluster 4928.

Discussion and Conclusion

In this article, we proposed an alternative approach to investigate interdisciplinarity. The measurement is based on a publication-level and direct citation relations based classification system. Hence, several interdisciplinarity research topics were identified with the new interdisciplinarity score in the current science system.

The interdisciplinarity score proposed not only takes citation relations among various WoS subject categories within a cluster into consideration, but it incorporates a measure of how distant the subject categories. As mentioned above, the indicator proposed in this article is similar, to some extent, with the widely used indicator of interdisciplinarity, that is Rao-Stirling index or Integration score (Porter & Rafols, 2009). The most crucial distinction between the two indicators of interdisciplinarity is that, for each research topic, we use the number of citations among subject categories instead of the number of publications in different subject categories. We consider that the number of citations among subject categories can reflect both how diverse as well as how compact a cluster is. Furthermore, to test the robust of this approach, we estimated Pearson's correlation between the two indicators. The correlation coefficient is 0.9552, which high correlation suggests that there is no difference between the original Rao-Stirling index and the variant proposed in this article.

Another distinction with previous research is that our study is based on a publication-level and direct citation relations based classification system, in which publications were assigned into different research topics according to their citation relations. It implies the research topics constructed can more closely match the current structure of scientific research and provide more detailed information of the research content per se (Waltman & van Eck, 2012). There are 250 WoS subject categories in total, providing a coarse description of science. On the contrary, we worked on a classification with around 10,000 research topics, deriving from large-scale clustering. While the clusters in this study are small compared with WoS classification, it is important and necessary to explore interdisciplinary research topics at different level of classification system of science.

Moreover, we need to emphasis the concept of 'interdisciplinary research topic' that we used in this article again. Here, this term is related to all types of crossing boundary research topics, which can be considered as a loose standard. Since there is a gradual transition from

mono-disciplinary to interdisciplinary research, it is somewhat impossible to define a clear line to distinguish mono-disciplinary and interdisciplinary related research.

In summary, we have introduced an alternative approach for identifying interdisciplinary research topics. By in-depth analysis of some randomly selected topics, especially based on citation networks and overlay maps, we believe that they are boundary-crossing research topics. Since most research on the measurement of interdisciplinarity have conducted based on an existing classification system of science, such as journal and WoS subject category, we expect this study could provide another perspective on the current science system. The identified research topics could more accurately reveal interdisciplinary research within the current structure of scientific research.

Acknowledgments

This paper was written during a research stay at the Centre for Science and Technology Studies (CWTS) of Leiden University. I acknowledge the support of CWTS. I would like to thank Ludo Waltman for the extremely helpful discussions and suggestions on this study. I also appreciate Ulf Sandström for his comments on the early version.

References

Abramo G., D'Angelo, C.A., & Costa, F.D. (2012). Identifying interdisciplinarity through the disciplinary classification of coauthors of scientific publications. *Journal of American Society for Information Science and Technology*, *63*(11), 2206-2222.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: Theory and experiment*, P10008.

Braun, T. & Schubert, A. (2003). A quantitative view on the coming of age of interdisciplinarity in the sciences 1980-1999. *Scientometrics*, 58(1), 183-189.

Huutoniemi, K., Klein, J.T., Bruun, H., & Hukkinen, J. (2010). Analyzing interdisciplinarity: Typology and indicators. *Research Policy*, *39*, 79-88.

Leiden Ranking 2013. Retrieved June 2, 2015 from: http://www.leidenranking.com/Content/CWTS%20Leiden%20Ranking%202013.pdf

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of American Society for Information Science and Technology*, *58*(9), 1303-1319.

Leydesdorff, L. & Rafols, I. (2011). Indicators of the interdisciplinarity of journal: diversity, centrality, and citations. *Journal of Informetrics*, *5*(1), 87-100.

Lariviere, V., & Gingras, Y. (2014). Measuring interdisciplinarity. In Cronin B., & Sugimoto C.R. (Eds.) *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*, Cambridge: MIT Press.

Morillo, F., Bordons, M., & Gomez I. (2001). An approach to interdisciplinarity through bibliometric indicators. *Scientometrics*, *51*(1), 203-222.

Morillo, F., Bordons, M., & Gomez, I. (2003). Interdisciplinarity in science: A tentative typology of disciplines and research areas. *Journal of American Society for Information Science and Technology*, *54*(13), 1237-1249.

Nichols, L.G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741-754.

Porter, A.L., & Chubin, D.E. (1985). An indicator of cross-disciplinary research. Scientometrics, 8(3-4), 166-176.

Porter, A., Roessner, J.D. & Heberger, A.E. (2008). How interdisciplinary is a given body of research? *Research Evaluation*, *17*(4), 273-282.

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745.

Qin, J., Lancaster, F.W., & Allen, B. (1997). Types and levels of collaboration in interdisciplinary research in the science. *Journal of American Society for Information Science and Technology*, *48*(10), 893-916.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicator of interdisciplinarity: case studies in bionanosciene. *Scientometrics*, *82*(2), 263-287.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal ranking can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, *41*, 1262-1282.

Salton, G., & McGill, M.J. (1983). *Introduction to modern information retrieval.* Auckland, New Zealand: McGraw-Hill.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanonscience and nanotechnology. *Scientometrics*, *59*(3), 425-465.

Sugimoto, C.R., Ni, C., Russell, T.G., & Bychowski, B. (2011). Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in library and information science. *Journal of the American Society for Information Science and Technology*, *62*(9), 1808-1828.

Sugimoto, C.R. (2011). Looking across communicative genres: a call for inclusive indicators of interdisciplinary. *Scientometrics*, *86*(2), 449-461.

Tijssen, R.J.W. (1992). A quantitative assessment of interdisciplinary structures in science and technology: co-classification analysis of energy research. *Research Policy*, 21, 27-44.

Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., Rafols, I., & Borner, K. (2011) Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of literature. *Journal of Informetrics*, *165*, 14-26.

van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *847*, 523-538.

Waltman, L., & van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378-2392.

Waltman, L., & van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B*, *86*, 471.

Wang, L., Notten, A., & Surpatean, A. (2013). Interdisciplinarity of nano research fields: a keyword mining approach. *Scientometrics*, *94*, 877-892.

How often are Patients Interviewed in Health Research? An Informetric Approach

Jonathan M. Levitt[1] and Mike Thelwall[2]

[1]J.M.Levitt@wlv.ac.uk, [2]m.thelwall@wlv.ac.uk
Statistical Cybermetrics Research Group, University of Wolverhampton, Wolverhampton WV1 1LY (UK)

Abstract
In recent years research funding bodies have increased their emphasis on the engagement between researchers and the public. As part of this increased emphasis, the UK's National Institute for Health Research aims to promote a research-active population. A way in which patients can be research-active is by participating in research interviews. In order to assess the past levels of this type of contribution of patients to research, this paper investigates the extent to which health research refers to patient interviews. Co-word indicators for the interviewing and qualitative interviewing of patients are used to gauge how the levels of interviewing and qualitative interviewing in Web of Science (WoS) articles have varied over time, between science and social science and between WoS categories. The results indicate that the level of interviewing of patients, referred to in WoS articles, rose steadily between 1991 and 2013. Moreover, the amount of interviewing and qualitative interviewing varied substantially between health-related fields, with a marked tendency for more interviews in social science research and fewer in science research.

Conference Topic
Indicators

Introduction
Over the past few years research funding bodies have increased their emphasis on public involvement in health research. For example, the UK's National Institute for Health Research, in a recent strategic plan, listed as a key objective, "Citizens helping to identify and deliver research of the highest quality" (NIHR, 2014), adding that citizen participation health research "is contributing to a 'research active' nation focused on best health for all." In particular, those who are ill seem to be particularly important because they can provide first-hand understanding of the specific illness being researched. In order to understand the potential contribution of ill people to health research, it helps to understand their past contribution to health research. This paper addresses two aspects of past contribution: the extent to which this contribution has varied over time and the extent to which this contribution has varied between subjects. This paper also introduces and demonstrates a novel technique: the use of co-word metrics to gauge the levels of both interviewing and qualitative interviewing of patients, and applies it to Web of Science (WoS) articles.

Background
Informetric techniques Although the individual words in abstracts can be irrelevant to the content of the articles, analyses of the words in academic publications have been used extensively. Collections of articles have been mapped, based on the words in their titles (Leydesdorff & Zaal, 1988; Milojević et al., 2011), their titles and keywords (Whittaker, 1989), their titles and abstracts (Peters & van Raan, 1993), their titles with references used for context (van den Besselaar & Heimeriks, 2006), or their full text (Glenisson et al., 2005). However, other research with similar goals has ignored the text in articles and used subject headings instead (An & Wu, 2011). Automatic analyses of the text of articles have also been used to identify, or differentiate between, different types of methods used. For instance, this approach has been used to track the evolution, over time, of computing technologies within library and information science research and to identify articles that used specific statistical

techniques (Thelwall & Wilson, in press). One particularly relevant study searched for a set of methods-related keywords (e.g., cohort study) in the titles of health-related articles in the Web of Science, and then compared the citation impacts of the articles found for each method (Patsopoulos, Analatos, & Ioannidis, 2005).

Patient involvement in research

In addition to often being involved in decisions about their own care (Charles, Gafni, & Whelan, 1997), patients are routinely the subjects of medical research to investigate the causes of, or cures for, their maladies. Patients can also be more actively involved in research by giving their opinions in open-ended questionnaires, or in interviews, or focus groups and by participating in steering groups for the co-ordination of research. Patients may also be involved in developing or promoting informational material to fellow sufferers (Greenfield, Kaplan, & Ware, 1985) or even in developing research policies (Nilsen et al., 2006). Gaining the patient's perspective can be helpful for research, for example, to get insights into the extent to which symptoms, in practice, vary from the norm (Cotrell & Schulz, 1993) and to understand and prioritise the problems that sufferers believe to be the most important to address (Serrano-Aguilar et al., 2009). Seeking the views of patients is sufficiently widespread for systematic reviews of this practice to be published for specific ailments (Morton et al., 2010). Nevertheless, the apparently widespread knowledge of the importance of patient involvement does not ensure that it occurs for all conditions.

Research questions

This paper investigates a contribution that ill people have made to health research, namely the extent to which health research has interviewed patients. The research questions are:

1. To what extent has the level of the research interviewing (and in particular the qualitative interviewing) of patients varied over time?
2. To what extent has the level of the research of interviewing (and in particular the qualitative interviewing) of patients varied between subject categories?

Method

The main data used to address the research questions is the approximate number of articles that refer to patient interviews and approximate number of articles that refer to qualitative patient interviews. This data, obtained for different WoS databases and subject categories, must be normalised to allow comparisons between findings for different years and subjects.

A simple way of normalising is to calculate the rate of interviewing and qualitative interviewing in each subject category would be to divide by the number of articles in the dataset investigated. For some subject categories only a small proportion of articles are closely related to patients, however, and so this ratio would be flawed. For instance, less than one fifth of Pharmacology Pharmacy articles refer to ‘patient’ in the topic.

In order to normalise the interview metric, this paper divides instead by the number of articles that refer to patients. This interview metric indicates the extent to which articles that refer to also refer to interviews. This choice is based on the reasonable assumption that studies on patient interviews will in generally refer to patient in their abstracts. In order to normalise the qualitative interview metric, this paper divides by the number of articles that refer to patients and interviews. This qualitative interview metric indicates the extent to which articles that refer to patient interviews also refer to the interviews being qualitative. This metric was chosen in order to limit the metric to research that plausibly could qualitatively interview patients (i.e., where patients and interviews are mentioned).

In order to calculate the interview metric and qualitative interview metric the following data was extracted from WoS: (a) the number of articles that contain ‘patient*’ in the topic (patient frequency), (b) the number of articles that contain ‘patient*’ and ‘interview*’ in the topic

(patient interview frequency), and (c) the number of articles that contain 'patient*', 'interview*' and at least one of 'qualitative*', 'open-ended', 'in-depth', ''semi structured' and 'semistructured' in the topic (patient interview qualitative frequency). The interview metric was defined as 1000*patient interview frequency/patient frequency; the qualitative interview metric was defines as 100*patient interview qualitative frequency/patient interview frequency. The multipliers of 1000 and 100 were chosen in order for most of the findings to be expressed between 10 and 100. The definition of the qualitative interview metric was preferred to the alternative definition of 10000*patient interview qualitative frequency/patient frequency as it indicates how the proportion of interviews that are qualitative varied over time and between subjects.

A possible source of inaccuracy in the interview metric is that articles with patient and interview in the topic do not necessarily refer to patient interviews. The accuracy of the interview metric was gauged through content analysis of a random sample of 50 WoS articles containing 'patient*' and 'interview*' in the topic; 90% of the records referred to interviews of patients or people associated with their illness. A possible source of inaccuracy in the qualitative interview metric is that articles with patient, interview and an indicator of qualitative in the topic do not necessarily refer to qualitative patient interviews. The accuracy of the qualitative interview metric was gauged through a content analysis of a random sample of 50 WoS records containing 'interview*' and at least one of ''qualitative*', 'open-ended', 'in-depth', ''semi structured' and 'semistructured'; 96% of the records indicate that the interviews were qualitative. Other possible sources of inaccuracy in these metrics are false positives (e.g., 'patient' can be used in sense not related to health, i.e., not impatient) and omissions (e.g., the list of terms for qualitative research is unlikely to be exhaustive).

As a high proportion of the search terms are in the article abstracts, it is important to confine the study to periods in which a high proportion of WoS records contain abstracts. A total of 84% of the records, of a random sample of 50 WoS articles published in 1991, contain abstracts, whereas the figure for WoS articles published in 1990 is only 8% (for 2013 the figure is 100%). Consequently, this study does not investigate years prior to 1991.

Results

In this paper, ''Patient incidence' denotes the number of articles with 'patient*' in the topic, 'Interview incidence' denotes the number of articles with 'interview*' in the topic per 1,000 articles with 'patient*' in the topic, and 'Qualitative interview incidence' denotes the number of articles with the indicators of qualitative in the topic per 100 articles with 'interview*' in the topic, 'SCI only' denotes articles in the Science Citation Index (SCI) and not in the Social Sciences Citation Index (SSCI), 'SCSI only' denotes articles in the SSCI and not in the SCI, 'SCI & SSCI' denotes articles in both the SCI and SSCI, and 'A&HCI' denotes articles in the Arts & Humanities Citation Index.

Table 1: Patient, interview and qualitative interview incidences for five WoS datasets.

Datasets	*Articles containing patient* in the topic*	*Interview articles per 1000 patient articles*	*Qualitative interview articles per 100 interview articles*
WoS	2,570,556	23.7	26.0
SCI only	2,309,924	11.0	16.5
SSCI only	67,088	134.5	35.1
SCI & SSCI	192,749	137.1	32.1
A&HCI	2,810	74.4	35.9

As can be seen in Table 1, for both SSCI only and SCI & SSCI the incidences of interviews are over 12 times the incidence for SCI only and the incidence of qualitative interviews is 90% higher than the incidence for SCI only. These differences are likely to be partly due to the different sizes of the databases and partly due to differences in the proportion of articles that mention patients. The table also indicates that interviews are relatively prevalent in social science research relating to patients and rare in science research relating to patients. Because of the small number of A&HCI articles that contain 'patient*' in the topic, this paper does not further investigate this dataset.

In response to Question 1 (variation over time) the incidence of interviews for WoS rose by 175% between 1991 and 2013 (Figure 1, left). The incidence for SCI only undulated between 1998 and 2013, (10.2 in 1998, 11.1 in 2013), whereas, during the same period, the levels of SSCI only and SCI & SSCI rose steadily (the 2013 levels are respectively 48% and 36% higher than the 1998 levels). Thus, the use of interviews in patient-related research seems to have risen more rapidly in the social sciences than in science, despite the lower initial prevalence of interviews in science research. The use of qualitative methods in interviews appears to have risen substantially in all the areas investigated. However, the increase is more rapid in social sciences research than in science research (Figure 1, right).

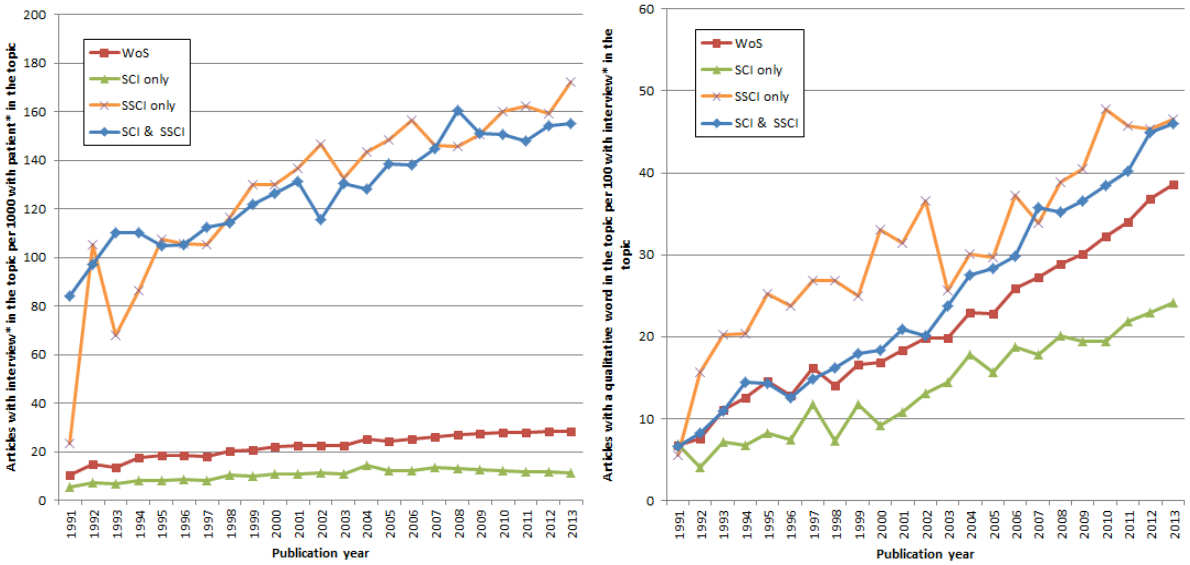


Figure 1. Annual incidence of interviews (left) and qualitative interviews (right).

In order to analyse disciplinary differences in more detail (Question 2), WoS categories were identified for each of the datasets SCI only, SSCI only and SCI & SSCI with at least 50 articles containing patient* and interview* in the topic. The ten categories identified were Clinical neurology, Health care sciences services, Health policy services, Nursing, Oncology, Pharmacology pharmacy, Psychiatry, Psychology, Public environmental occupational health and Rehabilitation. The incidence of interviews varies greatly between the ten categories, in addition to between science and social science research in the same category. The most extreme case is oncology, for which interviews are rare in science, but common in social science research (Table 2).

The incidence of qualitative interviews differs between science and social science in each individual category; qualitative interviews are more prevalent in social science research in 8 out of 10 categories (Table 2). For SCI only, the incidence of interviews is substantially lower for Clinical neurology, Oncology and Pharmacology pharmacy (average 12.0) than for the other seven categories (average 59.6). The incidence of qualitative interviews is also much lower for Clinical neurology, Oncology and Pharmacology pharmacy (average 14.0)

compared with the other seven categories (30.7). Hence, there are substantial disciplinary differences in the incidences of interviews and qualitative interviews within science.

Table 2: Incidence of interviews for ten WoS categories.

	Interviews			*Qualitative interviews*		
WoS category	*SCI*	*SSCI*	*Both*	*SCI*	*SSCI*	*Both*
Clinical neurology	16.5	65.1	107.3	11.9	25.0	17.6
Health care sciences services	92.2	99.9	157.5	41.4	30.3	46.5
Health policy services	76.0	182.7	125.4	31.6	47.6	39.0
Nursing	81.5	199.9	196.4	51.8	53.5	61.0
Oncology	7.2	226.3	195.2	15.0	47.7	45.7
Pharmacology pharmacy	12.3	199.2	67.6	15.2	58.0	17.8
Psychiatry	36.0	136.6	139.7	12.9	21.8	14.4
Psychology	46.0	102.2	115.5	25.9	17.7	19.4
Public environmental occupational health	53.3	219.8	170.6	20.0	44.3	37.0
Rehabilitation	32.5	86.7	137.9	31.0	34.5	52.4
Mean	**45.3**	**151.8**	**141.3**	**25.7**	**38.0**	**35.1**

For Clinical neurology, Oncology and Pharmacology, the percentage of articles in SCI only with patient* in the topic is particularly high: the percentage (in terms of articles in SCI or SSCI with patient* in the topic) for Clinical neurology is 89.3%, for Oncology is 96.4% and for Pharmacology pharmacy is 93.8%, whereas the average percentage for the other seven categories is 30.7%. There is a statistically significant Spearman correlation of -.81 between the interview incidence of SCI only and the percentage of articles with patient* in the topic that are in SCI only. This correlation reflects science categories having few interviews.

Limitations and conclusions

A limitation is that some studies with 'patient*' and 'interview*' in the topic do not interview patients (e.g., they interview physicians or carers of patients) and some studies with 'interview*' or indicators of qualitative in the topic do not conduct qualitative interviews (e.g., they combine quantitative interviews with qualitative analysis of patient records). But, as this research is comparative and the variations over time and between subjects are substantial, it seems likely that this limitation would not greatly affect the overall findings. Another limitation is that the results rely on the WoS journal subject classifications for journals. This may have a significant impact on the results for individual subject categories, as individual journals may have a substantial minority of the articles in a category. It would be useful to apply the techniques here to the full text of papers to help assess how often patient are involved in research but this is not discussed in the abstract of a paper.

After adjusting for the increase in the number of articles with 'patient*' in the topic, the number of WoS articles with 'interview*' in the topic increased by 175% from 1991 to 2013, suggesting that the use of patient interviews has increased substantially over the past 23 years. This may reflect a general trend towards involving patients more frequently in research, or an increase in the amount of research published, or indexed in WoS in research areas that typically involve patient interviews, such as nursing. In addition, after adjusting for the increase in the number of articles with 'patient*' and 'interview' in the topic, the number of articles that also had an indicator of qualitative in the topic increased by 511% from 1991 to 2013. This suggests that qualitative approaches are increasingly prevalent in health interviews, or that the qualitative nature of the research is more frequently specified. An

alternative explanation is that the amount of research published, or covered in WoS, has expanded in areas in which qualitative interviews are particularly common.

The incidences of interviews were particularly low amongst articles that were in SCI only; for 1991-2013 the incidence is less than one twelfth of the incidence for SSCI articles. When confining the study to categories present in both the SCI and the SSCI, there was a very marked difference between the datasets; however, the difference was substantially lower when excluding categories in which over 85% of the articles are in the SCI.

In the context of the NIHR aim of promoting a research-active population, the increased prevalence of patient interviews and qualitative interviews is encouraging, but categories with low percentages of interviews (e.g., Clinical neurology, Oncology and Pharmacology pharmacy) need to be further investigated to check whether individual subject areas are giving too little credence to patient interviews. Finally, this paper indicates that the technique of using simple co-word metrics based on the presence of words in the topic of WoS records can be applied usefully to informetric tasks. However, when investigating articles published prior to 1991, it is important to take into account that only a low percentage of WoS records for articles published in 1990 have abstracts.

References

An, X. Y., & Wu, Q. Q. (2011). Co-word analysis of the trends in stem cells field based on subject heading weighting. *Scientometrics, 88*(1), 133-144.

Charles, C., Gafni, A., & Whelan, T. (1997). Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). *Social Science & Medicine, 44*(5), 681-692.

Cotrell, V., & Schulz, R. (1993). The perspective of the patient with Alzheimer's disease: a neglected dimension of dementia research. *The Gerontologist, 33*(2), 205-211.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *IP&M, 41*(6), 1548-1572.

Greenfield, S., Kaplan, S., & Ware, J. E. (1985). Expanding patient involvement in care: Effects on patient outcomes. *Annals of Internal Medicine, 102*(4), 520-528.

Leydesdorff, L., & Zaal, R. (1988). Co-words and citations relations between document sets and environments. In: Rousseau, R., & Egghe, L. *Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval* (pp. 105-119).

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology, 62*(10), 1933-1953.

Morton, R. L., Tong, A., Howard, K., Snelling, P., & Webster, A. C. (2010). The views of patients and carers in treatment decision making for chronic kidney disease: systematic review and thematic synthesis of qualitative studies. *BMJ, 340*, c112. 10.1136/bmj.c112

NIHR. (2014). Promoting a research active nation, http://www.nihr.ac.uk/.

Nilsen, E. S., Myrhaug, H. T., Johansen, M., Oliver, S., & Oxman, A. D. (2006). Methods of consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material. *Cochrane Database Syst Rev, 3*.

Patsopoulos, N. A., Analatos, A. A., & Ioannidis, J. P. (2005). Relative citation impact of various study designs in the health sciences. *JAMA, 293*(19), 2362-2366.

Peters, H. P. F., & van Raan, A. F. (1993). Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling. *Research Policy, 22*(1), 23-45.

Serrano-Aguilar, P., Trujillo-Martin, M. M., Ramos-Goñi, J. M., Mahtani-Chugani, V., Perestelo-Pérez, L., & Posada-de la Paz, M. (2009). Patient involvement in health research: a contribution to a systematic review on the effectiveness of treatments for degenerative ataxias. *Social science & medicine, 69*(6), 920-925.

Thelwall, M. & Wilson, P. (in press). Does research with statistics have more impact? The citation rank advantage of structural equation modelling. *JASIST*.

van den Besselaar, P., & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: A method and an exploratory case study. *Scientometrics, 68*(3), 377-393.

Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords and co-word analysis. *Social Studies of Science, 19*(3), 473-496.

Normalized International Collaboration Score: A Novel Indicator for Measuring International Co-Authorship

Adam Finch[1], Kumara Henadeera[2] and Marcus Nicol[3]

[1] *adam.finch@csiro.au*
CSIRO, Waite Campus, Science Excellence Team, Waite Road, Urrbrae, SA 5064 (Australia)

[2] *kumara.henadeera@nhmrc.gov.au*
National Health & Medical Research Council, Strategic Policy Group, 16 Marcus Clarke Street, Canberra, ACT 2601 (Australia)

[3] *marcus.nicol@arc.gov.au*
Australian Research Council, Research Excellence Branch, 11 Lancaster Place, Canberra ACT 2609 (Australia)

Abstract
International collaboration on research publications is increasingly evaluated as part of a raft of performance measures. Levels of international co-authorship have increased substantially over the last few decades and vary substantially by research field and publication type; however, these variations are not typically accounted for by international collaboration indicators. In this research-in-progress paper, we introduce a novel metric, the Normalised International Collaboration Score, which adjusts the number of countries appearing on publication records using baselines relevant to the subject, age and type of the publication. A pilot analysis shows that these baselines vary substantially and that the application of this metric yields very different results to a more common measure of international collaboration. The limitations of the metric are discussed, along planned extensions for the full version of the study, as well as the relationship between normalised collaboration and citation.

Conference Topic
Indicators

Background and Purpose

Measuring international co-authorship

The availability of author address metadata on publication indices such as Web of Science and Scopus allows the analysis of patterns in co-authorship, including the collaboration by authors from different countries on research outputs. This approach has been used in many studies for decades (such as Glänzel & De Lange, 1997; Narin, Stevens, & Whitlow, 1991; Nederhof & Moed, 1993) and metrics describing international collaboration now appear regularly in bibliometric handbooks (Colledge, 2014; Rehn, Kronman, & Wadskog, 2007) and in reporting tools such as Thomson Reuters' InCites, Elsevier's SciVal and SCImago's Journal & Country Ranking. Such publications tend to receive higher levels of citation, an effect that is not due to the increased propensity for self citation arising from additional authors (Van Raan, 1998), but likely rather shared experience, knowledge and equipment.

Analysis of international co-authorship metadata has highlighted other important aspects of collaboration. Firstly, levels of international collaboration have increased substantially over the last quarter century (Leydesdorff & Wagner, 2008); and secondly, levels of international collaboration vary by field of research (Frame & Carpenter, 1979). A report on Thomson Reuters' InCites (retrieved 7 January 2015) indicates that 2013 articles, reviews and proceedings papers in Tropical Medicine involved international collaboration 46.7% of the time, while for History, this was only 4.3% of the time. Even within Medicine, Emergency Medicine saw only 9.9% foreign collaboration, far lower than Tropical Medicine. Variation is significant over time, with Astronomy & Astrophysics international collaboration rising from 19.4% in 1993 to 45.0% in 2013. To these two aspects, we must add publication type; 2013

Astronomy & Astrophysics articles saw 51.4% international collaboration, but its Proceedings Papers only 0.2%. Such variations exist across the full gamut of subject, years and publication types but most metrics used to evaluate collaboration do not take account of them.

Existing metrics

Frequently, analyses use either the number or proportion of collaborative publications (see for example Boekholt et al., 2009; Colledge, 2014; Luukkonen et al., 1993). Glänzel and De Lange (2002) use a Multilateral Collaboration Index to measure the number of collaborative links compared to the number of collaborative papers, establishing the intensity of collaboration.

Beaudet, Campbell, Côté, Haustein, Lefebvre and Roberge (2014) use a regression model based on power law relationships to establish the expected level of collaboration for a country and an Affinity Index to identify key partners. Degelsegger et al. (2013) propose thematic assessment, normalized either by relating it to the output of the country in the subject, or by comparing it to co-authored output in the same subject but with a different partner. Ding, Yang and Liu (2013) propose using network metrics to evaluate collaboration impact, which is a sound approach within a subject and time frame. Pohl, Warnan and Baas (2014) go the greatest distance to normalizing for the three aforementioned influences, by adjusting the proportion of publications with international collaboration by the number of collaborating countries in each subject. This study only considered a single year, however, did not adjust for publication type and was based on adjusting the share of research with a binary attribute (either internationally collaborative or non-internationally collaborative). The properties and results of this alternative will be compared to our metric in the full version of our study.

The Normalised International Collaboration Score (NICS)

The Normalised International Collaboration Score uses fundamentally the same calculation as the "new" Crown Indicator by which it was inspired (Waltmann, van Eck, van Leeuwen, Visser, & van Raan, 2011). For each publication, a global baseline is constructed, representing the average number of countries contributing to publications of the same type, from the same year and appearing in the same subject area(s). The number of countries contributing to the publication in question is then divided by the relevant baseline to yield a ratio. This ratio is then averaged for all publications in a set (for an institution, country, journal, etc). Our exploratory analysis uses both the mean (as in the Crown Indicator) and the statistically preferable median (Bornmann, Mutz, Neuhaus, & Daniel, 2008), for the purposes of comparison. While the present study only includes a selection of publication types, years and subjects, our full study will include all subjects and publication types back to 1996.

Methodology

The Advanced Search function on Web of Science was used to isolate publications of the Article, Review and Proceedings Paper types with issue cover dates in 1993, 2003 and 2013, and allocated to the subject categories Dance, Engineering (Manufacturing), Evolutionary Biology, Gastroenterology & Hepatology, Political Science, Psychology (Educational), Soil Science and Tropical Medicine. These publication types were selected as those most likely to contain address data; these years as spaced such to demonstrate evolution in collaboration trends and aspects of the data; and those subjects as representing a broad spectrum across science, social science, and the arts and humanities. The selection of a single discipline of period would not have illustrated any variation over time or theme. Record metadata were downloaded, tagged with the relevant subject name and recombined into a single dataset. Individual addresses were broken out, the non-country information in each field deleted and duplicate country entries deleted. A count of unique country contributions per publication was

made. The baselines were constructed by averaging all unique country contribution counts for each combination of year, publication type and subject, using the arithmetic mean and the median. These represented the denominator of the metric's article-level ratio.

The institutional data came from a database of Australian publication records from 2001 to 2014. A query extracted the unique identifier, selected subject areas, year, type and Crown Indicator of each publication, along with the author addresses. The addresses were subjected to a unique contributing country count, yielding the numerator of the metric's article-level ratio. The subject, publication type and publication year data were used to look up the mean and median baseline data (our ratio's denominator). Dividing the latter by the former yielded the article-level NICS, which was then averaged for each Australian institution – using the arithmetic mean and then the median, as appropriate for the baseline.

This gives the following notation for the mean form of NICS:

$$\frac{1}{p}\sum_{i=1}^{p}\frac{n_i}{g_i}$$

And the following notation for the median form of NICS:

$$\widetilde{\left(\frac{n_i}{m_i}\right)}$$

Where p denotes the number of publication produced by a unit of analysis, n_i denotes the number of countries contributing to the unit's publication i, g_i denotes the global mean number of countries contributing to publications of the same type, year and subject(s) as publication i and m_i denotes the global median number of countries contributing to publications of the same type, year and subject(s) as publication i. A third, "hybrid", version of was calculated, finding the median of article level ratios based on a mean:

$$\widetilde{\left(\frac{n_i}{g_i}\right)}$$

Results & Discussion

Table 1 shows the mean baselines for each year in each subject, combining publication types into a single entry. Several points are clear. Some subjects see a substantial increase in average country contributions over time – such as the increase from 1.15 to 1.71 for Evolutionary Biology – indicating a need to normalise for this change if fair comparisons are to be made among publication sets from different year ranges. There are also significant disparities between subjects, with the Engineering subject baseline 1.09 in 2013, compared to 1.78 for Tropical Medicine. It is also notable that, unlike citation counts, there does not seem to be a pattern of lower country contributions for social sciences as opposed to sciences, at least in this very limited dataset; Political Science has one of the higher baseline sets and Engineering, Manufacturing one of the lower. Lastly, some subjects, most likely those in the Arts & Humanities, may be difficult to assess using this metric, due to a paucity of address and a low publication count; the baselines would be based on too low a sample size and very prone to skew from outliers. It is also worth noting that, while country contributions are strongly positively skewed, the variance of the natural log of country contribution counts is lower than that of citation counts for publications of the same year, type and subject, in a all of a selection of the below instances that were considered.

Table 2 shows the number of publications missing address data in each of the three years for each subject. Coverage is a problem in Dance for all years and is more of a problem in the social science subjects than the sciences, but is an issue for all subjects in 1993. In the full analysis, work will be conducted to establish the point at which coverage is sufficient for robust analysis, but the institutional analysis in this pilot study exclude the 1993 publications.

Table 1. Mean Subject Country Contribution Baselines by Year.

	1993		*2003*		*2013*	
Table	*# Pubs*	*# Countries*	*# Pubs*	*# Countries*	*# Pubs*	*# Countries*
Dance	2	1.50	25	1.00	46	1.13
Engineering, Manuf.	1242	1.03	7935	1.06	14513	1.09
Evolutionary Biology	987	1.15	3900	1.38	5543	1.71
Gastroent. & Hepat.	3567	1.09	8595	1.15	11300	1.29
Political Science	987	1.15	3172	1.26	5549	1.76
Psychology, Educ.	483	1.05	1167	1.10	2253	1.20
Soil Science	1724	1.09	3890	1.23	4721	1.36
Tropical Medicine	798	1.45	1381	1.68	3128	1.78

Table 2. Instances of Publication Entries Missing Address Data by Year.

	1993			*2003*			*2013*		
Table	*No Address*	*Total Pubs*	*%*	*No Address*	*Total Pubs*	*%*	*No Address*	*Total Pubs*	*%*
Dance	245	247	99.2%	386	411	93.9%	184	230	80.0%
Eng., Manufact.	936	2178	43.0%	587	8522	6.9%	219	14732	1.5%
Evolutionary Biology	698	1685	41.4%	15	3915	0.4%	10	5553	0.2%
Gastro. & Hepat.	2080	5647	36.8%	158	8753	1.8%	68	11368	0.6%
Political Science	3421	4408	77.6%	965	4137	23.3%	636	6185	10.3%
Psych., Education.	573	1056	54.3%	20	1187	1.7%	47	2300	2.0%
Soil Science	1702	3426	49.7%	132	4022	3.3%	16	4737	0.3%
Tropical Medicine	475	1273	37.3%	11	1392	0.8%	24	3152	0.8%

Table 3 shows the mean baselines for each publication type in each subject, combining years into a single entry. It is clear that publication type is also a major factor for the baselines, with the Proceedings Papers consistently seeing fewer country contributions than other types. However, there is further variation; Political Science, for example, sees higher country counts for Articles than Reviews, while the reverse is true for Soil Science.

Table 3. Mean Subject Country Contribution Baselines by Publication Type.

	Articles		*Proceedings*		*Reviews*	
Table	*# Pubs*	*# Countries*	*# Pubs*	*# Countries*	*# Pubs*	*# Countries*
Dance	73	1.10	-	-	-	-
Engineering, Manuf.	9192	1.20	14398	1.00	100	1.23
Evolutionary Biology	9665	1.54	101	1.00	664	1.59
Gastroent. & Hepat.	20482	1.21	811	1.00	2169	1.24
Political Science	8650	1.57	790	1.18	268	1.28
Psychology, Educ.	3542	1.16	263	1.00	98	1.09
Soil Science	8547	1.31	1641	1.01	147	1.69
Tropical Medicine	5113	1.71	23	1.00	171	1.73

Table 4 shows a comparison of institutional collaboration analysis using the proportion of publications with international collaboration and each of the three variants of the NICS metric. The Median calculation appears the least useful; every baseline in each year, subject and document was 1, so this version essentially reports the median country contribution per article and cannot strongly differentiate among institutions. The version using mean baselines are more useful for ranking but, like the Crown Indicator, remains sensitive to outliers (as in the example of Flinders University, where performance was inflated by a single article with 35 contributing countries). Even though the full study will involve far larger sample sizes, which should be less susceptible to such outliers, it appears that the "hybrid" (median of ratios based on mean baselines) is the strongest option. This would preclude statistical analysis based on parametric data, but it is impossible to tell from the pilot study whether the article level results of the mean calculation would be normally distributed on a global scale either.

Table 4. Selected Australian Institution Ranking.

		% Collaboration		*NICS Mean*		*NICS Median*		*NICS 'Hybrid'*	
Table	*Pubs*	*Value*	*Rank*	*Value*	*Rank*	*Score*	*Rank*	*Score*	*Rank*
Queensland Inst Med Res	55	70.9%	1	1.61	3	2	1	1.19	3
James Cook Univ	98	65.3%	2	1.72	1	2	1	1.44	2
Charles Darwin Univ	40	62.5%	3	1.60	4	2	1	1.12	5
Univ Western Sydney	52	57.7%	4	1.55	6	2	1	1.45	1
Univ Western Australia	219	50.7%	6	1.26	15	2	1	1.12	5
Univ Melbourne	233	50.2%	7	1.48	7	2	1	1.12	5
Univ Adelaide	174	49.4%	9	1.24	22	1	9	0.86	19
Univ Sydney	286	48.6%	10	1.38	10	1	9	0.99	13
CSIRO	210	48.6%	11	1.24	21	1	9	0.99	12
Univ Queensland	271	48.3%	12	1.25	18	1	9	0.91	15
Queensland Univ Technol	58	48.3%	13	1.25	16	1	9	1.00	9
Murdoch Univ	45	46.7%	15	1.23	23	1	9	0.79	22
Univ Newcastle	48	45.8%	16	1.29	14	1	9	1.00	10
Australian Natl Univ	203	44.8%	17	1.08	27	1	9	0.79	22
Univ New S Wales	195	44.1%	18	1.24	20	1	9	0.88	16
Monash Univ	155	43.2%	19	1.35	11	1	9	0.88	16
Curtin Univ Technol	44	43.2%	20	1.20	24	1	9	1.00	11
Howard Florey Inst	48	35.4%	26	1.56	5	1	9	0.78	25
Flinders Univ S Australia	70	30.0%	27	1.65	2	1	9	0.78	25

Discussion

While only a few institutions see a large difference in ranking when applying NICS rather than proportion of international publications, the difference in results and the variations in baselines on which they are based suggest the metric has informational content. It is also worth noting that, at an article level, the Crown Indicator correlates positively and fairly strongly with NICS (Spearman's Rank r=0.384) and that at an institutional level, the two versions of NICS derived from mean baselines correlate more closely with NCI performance (r=0.289 and 0.148) than does share of publications with international collaboration (r=0.09).

There are clearly limitations to this approach. It does not account for collaboration intensity; eight co-authoring institutions in a specific foreign country count the same as one. The NICS

baselines could be rescaled to count not only contributing foreign countries but also the numbers of institutions in those countries, and even potentially types of institutions. As it would require a set of baselines for each country, this would be computationally intensive but will be explored in the full study. This approach would also normalise for the propensity of a country to collaborate, which many of the above-mentioned metrics are aimed at doing. Lower collaboration levels can arise from several causes, including a lower advantage yielded and having a large share of global output (therefore limiting the avenues available for external collaboration); normalising for national collaboration levels may obscure these differences and render accurate national comparisons challenging. In its present form, NICS serves best as a metric to compare the collaboration of countries and institutions, variations in which may then be considered in the context of national motivation and propensity to collaborate.

Other criticisms leveled at the Crown Indicator apply to NICS, most notably a limited representation of global output in some subjects and of some publication types, and the reliance on a subject taxonomy designed for information retrieval rather than bibliometric analysis. In the pilot study, moreover, many articles analysed here appeared in more than one subject area, and yet were normalised only with the baselines for one of those subject areas.

The full study will apply a wide range of statistical tests to the properties of the baselines, the country contribution counts and the resultant ratios; for now, however, and even with the aforementioned caveats, this metric shows potential for robust and meaningful analysis of institutional and national research collaboration abroad.

References

Beaudet, A., Campbell, D., Côté, G., Haustein, S., Lefebvre, C., & Roberge, G. (2014) *Bibliometric Study in Support of Norway's Strategy for International Research Collaboration.* Report to the Science Council of Norway, Science-Metrix. March 2014.

Boekholt, P., Edler, J., Cunningham, P., & Flanagan, K. (2009). *Drivers of international collaboration in research*. Report to EC, DG Research. Technopolis BV,Netherlands, April 2009.

Bornmann, L., Mutz, R., Neuhaus, C. & Daniel, H-D. (2008). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8*, 93-102.

Colledge, L. (2014). Snowball Metrics Recipe Book, Edition 2. Retrieved December 19, 2014 from: http://www.snowballmetrics.com/wp-content/uploads/snowball-recipe-book_HR.pdf

Degelsegger, A., Lampert, D., Büsel, K., Simon, J., Tschant, J., & Wagner, I. (2013) *Assessing international cooperation in S&T through bibliometric methods. Proc. ISSI*, 175-184.

Ding, J., Yang, L. & Liu, Q. (2013). Measuring the academic impact of researchers by combined citation and collaboration impact. *Proc. ISSI*, 1177-1187.

Frame, J.D., & Carpenter, M.P. (1979). International research collaboration. *Social Studies of Science,* 9, 481-497.

Glänzel, W., & De Lange, C. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part II. A Comparative study on the extent and change of international scientific collaboration links. *Scientometrics*, *40*, 605-626.

Glänzel, W., & De Lange, C. (2002). A distributional approach to multinationality measures of international scientific collaboration. *Scientometrics*, *54*, 75-89.

Leydesdorff, L. & Wagner, C.S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics, 2*, 317-325.

Luukkonen, T., Tijssen, R.J.W., Persson, O., & Sivertson, G. (1993). The measurement of international scientific collaboration. *Scientometrics, 28*, 15-36.

Narin, F., Stevens, K., & Whitlow, E.S. (1991). Scientific cooperation in europe and the citation of multinationally authored papers. *Scientometrics, 21*, 313-323.

Nederhof, A.J., & Moed, H.F. (1993). Modeling multinational publication - development of an online fractionation approach to measure national scientific output. *Scientometrics*, *27*, 39-52.

Pohl, H., Warnan, G., & Baas, J. (2014). Level the playing field in scientific international collaboration with the use of a new indicator: Field-Weighted Internationalization Score. *Research Trends, 9*, 3-8.

Rehn, C., Kronman, U., & Wadskog, D. (2007). *Bibliometric indicators – definitions and usage at Karolinska Institutet*. Retrieved December 19, 2014 from: http://kib.ki.se/sites/kib.ki.se/files/Bibliometric_indicators_definitions_1.0.pdf

van Raan, A.F.J. (1998). The influence of international collaboration on the impact of research results. *Scientometrics*, *43*, 423-428.

Waltmann, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., & van Raan, A.F.J. (2011). Towards a new crown indicator: an empirical analysis. *Scientometrics, 87*, 467-481.

Bibliometric Indicators of Interdisciplinarity Exploring New Class of Diversity Measures

Alexis-Michel Mugabushaka[1], Anthi Kyriakou & Theo Papazoglou

[1] *Alexis-Michel.MUGABUSHAKA@ec.europa.eu*
European Research Council Executive Agency[1], Brussels, (Belgium)

Abstract
In bibliometrics, interdsicsiplinatity is often measured in terms of the "diversity" of research areas in the references that an article cites. The standard indicators used are borrowed mostly from other research areas, notably from ecology (biodiversity measures) and economics (concentration measures). This paper discusses a new class of measures, which are used in the study of biodiversity and especially the Leinster-Cobbold diversity measure (Leinster Cobbold 2010). We present a case study based on previously published dataset of 12 journal articles from a group of five researchers from the bio-nano science described and published by Rafols and Meyer (2010). We replicate the findings of this study to show that the various interdisciplinarity measures are in fact special cases of the Cobbold-Leinster diversity measure. The paper discusses some interesting properties of the Cobbold-Leinster diversity measure, which makes it appealing in the study of disciplinary diversity than the standards diversity indicators used as proxy for interdisciplinarity.

Conference Topic
Indicators

Introduction
Considerable efforts have been made to operationalize and measure the concept of interdisciplinarity in bibliometrics (Porter et al., 2007; Rafols & Meyer, 2010). The most commonly used indicators of interdisciplinarity are mostly borrowed from other research areas, notably from ecology (biodiversity measures) and economics (concentration measures).
The purpose of this paper is to bring to discussion a relatively new class of diversity indicators which are used in ecology but so far not been used to investigate disciplinary diversity. Drawing from the literature in ecology, the paper highlights important properties of those measures and discusses how they can help the bibliometric study of interdisciplinarity.
The paper is divided in three parts. The next section briefly presents indicators of interdisciplinarity in bibliometrics. The second section discusses the development of new class of diversity measures used in ecology and presents the Leinster-Cobbold diversity measure, highlighting its properties and why they are relevant for bibliometric usage. The third section presents a case study to illustrate the potential of Leinster-Cobbold diversity indicators as a measure of disciplinary diversity.

Currently used Bibliometric indicators of interdisciplinarity
Bibliometric analyses of interdisciplinarity take as unit of analysis a scientific paper and assume that the extent to which it integrates elements of different disciplines is reflected in the references it cites. References in scientific papers are expected to reflect various aspects of interdisciplinary because researchers will credit what they are indebted to other disciplines: conceptually (concepts, ideas and approaches from other disciplines); analytically (methods for defining, collecting and analyze data) and technically (tools developed in other fields).

[1] The views expressed in this paper are the authors'. They do not necessarily reflect the views or official positions of the European Commission, the European Research Council Executive Agency or the ERC Scientific Council.

Porter et al. (2007) developed the integration score as measure of interdisciplinary which takes into account not only the distribution of the cited references in different subject categories but also how closely related those subject categories are (see also Porter et al., 2006; Porter et al., 2008). In line with Porter's conceptualization, Rafols and Meyer (2006, 2010) introduced a new set of bibliometric indicators to quantify the disciplinary diversity of references as a proxy measure of interdisciplinarity. They are mostly based on the general framework for analyzing diversity developed by Stirling (2007). The most commonly used indicators are summarized in table 1. We note that there are also efforts to use network based measures (Rafols & Meyer, 2010; Karlovčec & Mladenić, 2015) but here we focus on diversity measures.

Table 1. Most common indicators of interdisciplinarity in bibliometric studies .

Indicators	Definition/description	
Variety	The number of different disciplines that a given paper cites**	N
Shannon entropy	As measure of diversity the Shannon Entropy quantifies how diverse the subject categories in the references are.	$H_{SH} = -\sum_{i=1}^{S} p_i \log p_i$ Where p_i is the proportion of elements in a system and S the number of elements in the system.
Simpson diversity	It measures how references are distributed (or concentrated) in subject categories.	$H_{GS} = 1 - \sum_{i=1}^{S} p_i^2$ Where p_i is the proportion of elements in a system and S the number of elements in the system
Rao-Stirling index	Can be understood as the Simpson diversity which takes into account distance/similarity (between disciplines).	$= \sum_{i,j} d_{i,j}\, p_i p_j$ Where di,j is the distance between the ith and jth element in the distance matrix and pi is the proportion of element i

Source: Rafols & Meyer 2010, p. 267 **Its variants includes normalization by the total numbers of subject categories or the shares of references outside a given subject category

New classes of diversity measures in ecology

Effective numbers

The diversity measures listed in table are also among the commonly used indicators of biodiversity in ecology. However, they have recently faced strong criticisms (Jost, 2006; Chao & Lou, 2012).

The main criticism is that those measures fail to satisfy the most basic property that ecologist would expect from a meaningful measure of diversity, namely the replication principle. In simple term, the "replication principle" states that if you have two completely distinct communities (i.e. without any overlap in the species) with each community having a diversity measure X, one would expect that combining those two communities would result in a community with a diversity measure 2X.

One category of diversity measures, which satisfy this replication principle is the so called "Hill-numbers" (also called "effective numbers of species"). They can be interpreted as the

"number of equally abundant specifies that are needed to give the same value of the diversity measure (Chao & Lou, 2012, p. 204).

The Hill numbers have some properties that other measures of diversity based on entropy lack:

- They satisfy the replication principles. i.e. two communities with each 4 effective numbers of species will – if pooled together – result in a community whose effective number equal 8. They therefore give logically consistent answers.
- Their linear scale makes it easier to interpret the magnitude of their change.
- In addition to this this advantage of intuitive consistency, they have another interesting property that we call "unifying framework status". Jost (2006) has shown that practically all traditional measures of diversity can be easily converted to "Hill numbers/ "effective numbers" and vice-versa.

Leinster-Cobbold Diversity Measure

Leinster and Cobbold (2012) developed a measure, which extends the Hill numbers to include the similarities/differences between species. Their measure – called here the Leinster-Cobbold Diversity Measure - can be used with any similarity coefficient between each pair of the species. This extends the scope of its usage to other contexts such disciplinary diversity in bibliometrics. In the following, we first provide its formal definition and discuss its properties as well as its relation to other diversity measures. In the next section we provide a case study of its use in the study of disciplinary diversity.

Consider a system with S elements with relative frequencies translating in estimated probabilities p = (p1, …, pS) so that $\sum_{i=1}^{S} p_i = 1$

The similarity between the elements is encoded in an S x S Matrix Z.

Z = (Z_(i,j)), with Z_(i,j) measuring the similarity between the ith and jth elements.

Whereby 0 ≤Z_(i,j) ≤1, with 0 indicating total dissimilarity and 1 indicating identical elements.

The Leinster-Cobbold diversity measure is defined as

$$ {}^{q}D^{Z}(\boldsymbol{p}) = \begin{cases} \left(\sum_{i:p_i>0} p_i (Z\boldsymbol{p})_i^{q-1} \right)^{\frac{1}{1-q}} & q \neq 1, \\ \prod_{i:p_i>0} (Z\boldsymbol{p})_i^{-p_i} & q = 1, \\ \min_{i:p_i>0} \frac{1}{(Z\boldsymbol{p})_i} & q = \infty. \end{cases} $$

where

$$ (Z_p)i = \sum_{j=1}^{S} Z_{i,j} p_i $$

q is in number in range 0 ≤ q ≤ Infinity. It is called a sensitivity parameter and control the relative emphasize that the user wishes to place on common and rare species.

Case Study: Using the Leinster-Cobbold Diversity as a measure of disciplinary diversity

In our view, there are three main advantages in adopting the Leinster and Cobbold diversity measure in the study of disciplinary diversity as well:

- First, Leinster and Cobbold (2012) have discussed the relation between this measure and other diversity measures and showed that they can be seen as its special cases. The advantage here would be to have a single formula which would replace the Shannon entropy, the Simpson Diversity and the Rao-Stirling Index used in bibliometrics.
- Second, because the Leinster and Cobbold measure quantifies diversity on a spectrum which depends on how much emphasis should be given to relatively rare elements (sensitivity parameter q), it provides potentially more information than measures which consider only one value of this sensitivity parameter.
- The third advantage is the intuitive consistency of the Leinster and Cobbold measure. Because it directly produces "effective numbers" which obey the replication principle, the values can be easily interpreted and compared. Consider two publications: one with references from 2 (unrelated) categories and the other with reference from 4 (unrelated) categories. With the Leinster and Cobbold measure, they can be compared to say that the second has a twice as large diversity in references as the first one.

In the following, we present a case study to illustrate the potential of Leinster-Cobbold diversity profiles in quantifying disciplinary diversity.

Disciplinary diversity of selected papers in bio-nanoscience (Rafols & Meyer 2010)

The case study is based on a dataset of 12 journal articles from a group of five researchers from the bio-nano science described and published by Rafols and Meyer (2010). For those 12 papers, Rafols and Meyers published the distribution of their references in Web of Science Categories (Rafols & Meyers, 2010; p. 276, Table 3) as well as the scores on various indicators of diversity (ibid. p. 277, Table 4). The similarity/distance measures between the Web of Science subject categories are taken from the supplementary materials to the paper[2] by Chavarro et al. (2014).

Table 2. Diversity measures for the 12 papers in Rafols and Meyer (2010).

	not considering distance/similarity						considering distance/similarity					
sensitivity parameter q	**0**	**1**	**2**	**3**	**4**	**Inf**	**0**	**1**	**2**	**3**	**4**	**Inf**
Column no.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Papers												
Fun95	16	6,452	4,553	3,989	3,740	3,106	1,656	1,422	1,329	1,288	1,266	1,188
Koj97	17	5,526	4,232	3,848	3,652	2,880	1,479	1,284	1,225	1,203	1,192	1,143
Ish98	15	5,003	3,499	2,990	2,741	2,156	1,342	1,229	1,192	1,176	1,167	1,108
Noj97	16	4,532	3,120	2,665	2,447	1,967	1,280	1,172	1,141	1,128	1,122	1,077
Yas98	16	4,466	3,003	2,537	2,327	1,890	1,231	1,158	1,133	1,122	1,115	1,072
Oka99	16	4,857	3,814	3,557	3,439	3,062	1,253	1,190	1,165	1,154	1,148	1,108
Kik01	14	4,944	3,857	3,534	3,364	2,673	1,251	1,195	1,169	1,155	1,148	1,102
Sak99	14	5,103	4,040	3,764	3,641	3,184	1,245	1,181	1,159	1,149	1,143	1,098
Bur03	14	4,697	3,536	3,230	3,086	2,571	1,178	1,142	1,127	1,120	1,115	1,082
Tom00	15	4,841	3,846	3,625	3,530	3,028	1,227	1,165	1,145	1,136	1,132	1,095
Tom02	14	4,849	3,864	3,630	3,531	3,192	1,242	1,180	1,159	1,149	1,143	1,103

This case study illustrates that the various diversity measures are in fact special cases of the Leinster-Cobbold diversity profiles. We do this by replicating the diversity measures computed by Rafols and Meyer 2010 using the Leinster-Cobbold diversity profiles. We first compute the values of the Leinster Cobbold measure using different values for the sensitivity parameters (0, 1, 2, 3, 4 and infinity) and in two variants: without taking into account the

[2] http://www.interdisciplinaryscience.net/topics/interdisciplinarity-and-local-knowledge

distance/similarity between the subject categories (i.e. the matrix Z is an identity matrix) and by taking into account the distance/similarity between the subject categories (using the similarity data provided in supplementary materials of Chavarro et al. (2014). Using the conversion formulas in the first row of Table 3, we use those Leinster Cobbold values to derive the diversity measures provided in Rafols and Meyer 2010 (table 4 on page 277). The Table 3 below replicates the diversity values reported in Rafols and Meyer 2010. There are some differences, which are due to rounding but also to the fact that some indicators in Rafols and Meyer (2010) were given in normalized form.

Table 3. Deriving diversity measures commonly used in bibliometrics from the Leinster-Cobbold values.

	Variety	**Gini-Simpson**	**Shannon**	**Rao**
computation	**Col 1**	**1- (1/Col 3)**	**ln(Col 2)**	**1- (1/Col 9)**
Papers				
Fun95	16	0,78	1,86	0,25
Koj97	17	0,76	1,71	0,18
Ish98	15	0,71	1,61	0,16
Noj97	16	0,68	1,51	0,12
Yas98	16	0,67	1,5	0,12
Oka99	16	0,74	1,58	0,14
Kik01	14	0,74	1,6	0,14
Sak99	14	0,75	1,63	0,14
Bur03	14	0,72	1,55	0,11
Tom00	15	0,74	1,58	0,13
Tom02	14	0,74	1,58	0,14
Yil04	16	0,76	1,68	0,16

Concluding remarks

In bibliometrics, the interdisciplinarity is operationalized in terms of the diversity of the references in a scholarly article. The most commonly used indicators are derived from the fields of ecology (biodiversity measures) and from the fields of economics (concentration measures). We discuss a new class of biodiversity measures – the "effective numbers" - which not only generalize most of other diversity measures but also have some proprieties which make their interpretation intuitively consistent with the concept of diversity Jost (2006). They were further developed by Leinster-Cobbold (2012) to take into account the similarity/distance of elements (species) in a system (community). We provide an example on how the bibliometric indicators of interdisciplinarity are in fact special cases of this more general Leinster Cobbold indicator.

Future work should not only take a closer look at their statistical properties (distribution, parameters etc.) but also test their reliability and validity. In particular, it would be of interest to analyze how sensitive the indicators are to various degree of granularity of different classifications of research disciplines and to assess extent to which they depend on measures of distances used.

Acknowledgments

We thank Ismael Rafols for helpful comments on an earlier draft of the paper and Diego Chavarro for making the similarity matrix freely available.

References

Chao, A., & Jost, L. (2012). *Diversity measures. In Encyclopedia of Theoretical Ecology* (Eds. A. Hastings and L. Gross), pp. 203-207, Berkeley: University of California Press.

Chavarro, D., Tang, P., & Rafols, I. (2014). Interdisciplinarity and research on local issues: evidence from a developing country. *Research Evaluation*, *23*(3), 195-209.

Jost, L. (2006). Entropy and diversity. *Oikos*, *113*, 363-375.

Jost, L. (2007). Partitioning diversity into independent alpha and beta components. *Ecology*, *88*, 2427–2439.

Jost, L. (2009). Mismeasuring biological diversity: Response to Hoffmann and Hoffmann (2008). *Ecological Economics*, *68*,925–928.

Karlovčec, M., & Mladenić, D (2015) Interdisciplinarity of scientific fields and its evolution based on graph of project collaboration and co-authoring. *Scientometrics*, *102*(1), 433-454.

Leinster T., & Cobbold CA. (2012). Measuring diversity: the importance of species similarity. *Ecology*, *93*(3):477-489.

Porter et al., (2006). Interdisciplinary research: meaning, metrics and nurture. *Research Evaluation*, *15*(3), 187-195.

Porter, A.L. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics*, *81*(3), 719-745.

Porter, A.L., Cohen, A.S., Roessner, J.D., & Perreault, M. (2007), Measuring researcher interdisciplinarity, *Scientometrics*, *72*(1), 117-147.

Rafols, I. & Meyer, M. (2006). Diversity measures and network centralities as indicators of interdisciplinarity: case studies in bionanoscience. *SPRU working paper*.

Rafols, I. & Meyer, M. (2010). Diversity and Network Coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, *82*(2), 263-287.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707-719.

Modeling Time-dependent and -independent Indicators to Facilitate Identification of Breakthrough Research Papers

Holly N. Wolcott[1], Matthew J. Fouch[1], Elizabeth Hsu[2], Catherine Bernaciak[1], James Corrigan[2], and Duane Williams[1]

holly.wolcott@thomsonreuters.com
[1]Intellectual Property & Science, Thomson Reuters, Rockville, MD 20850 (USA)

corrigan@mail.nih.gov
[2]Office of Science Planning and Assessment, National Cancer Institute, Bethesda, MD 20892 (USA)

Abstract
Research funding organizations invest substantial resources to stay current with important research findings within their mission areas to identify and support promising new lines of inquiry. To that end, we continue to pursue the development of tools to identify research publications that have a strong likelihood of driving new avenues of research. This research-in- progress paper describes our work incorporating multiple time-dependent and -independent features of publications into a model that aims to identify candidate breakthrough papers as early as possible following publication. We used multiple Random Forest models to assess the ability of indicators to reliably distinguish a gold standard set of breakthrough publications as identified by subject matter experts from among a comparison group of similar *Thomson Reuters Web of Science*™ publications. These indicators will be selected for inclusion in a multi-variate model to test their predictive value. Prospective use of these indicators and models is planned to further establish their reliability.

Conference Topic
Indicators

Introduction
The National Cancer Institute (NCI) of the US National Institutes of Health (NIH) continues to show a commitment to encouraging transformative research, which the NIH recognizes on its Transformative Research Award website as "unconventional research projects that have the potential to create or overturn fundamental paradigms." Key requirements for identifying and nurturing these potential scientific breakthroughs are an enhanced understanding of the research landscape and awareness of novel approaches with great potential.

Defining Breakthrough Publications
The term "breakthroughs" has been used in prior work by Thomson Reuters (Ponomarev et al., 2014) and operationally, breakthrough publications have previously been defined as those that are highly cited and result in a change in research direction. The body of literature addressing breakthrough publications also uses the term "transformative research." Here, we define a breakthrough publication as an article that results from transformative research. In 2007, the National Science Board (NSB) defined transformative research as "research driven by ideas that have the potential to radically change our understanding of an important existing scientific or engineering concept or leading to the creation of a new paradigm or field of science or engineering. Such research also is characterized by its challenge to current understanding or its pathway to new frontiers" (NSB, 2007).

Prior Work Identifying Breakthrough Publications
Much of the research literature on breakthroughs focuses on retrospective identification of breakthroughs or pivotal points within a specific topic or field (Chen, 2006; Compañó & Hullmann, 2002; Fujita et al., 2012; Huang et al., 2013; Klavans et al., 2013; Ponomarev et

al., 2014). In addition, many of the current approaches require manual selection or curation of all data analysed (Chen, 2006; Klavans et al., 2012). Ponomarev et al. (2014) used variations of a single indicator, citation velocity, to predict highly cited papers while other groups made use of multiple indicators, full-text data and/or co-citation analysis to identify and characterize breakthrough publications in retrospective analyses (Chen, 2006, 2012; Klavans et al., 2012; Klavans et al., 2013). Other efforts focused on the development of analysis and visualization tools for quick visualization and assessment of potential turning points and breakthroughs (Boyack & Börner, 2003; Dunne et al., 2012).

Here, we aim to establish automated and semi-automated approaches to provide early indicators of published research with great potential. The goal is to provide program staff with a robust methodology that highlights pockets of breakthrough research, thereby enabling more informed program management. The methodology leverages an array of indicators to identify work that may contribute significantly to progress in its field. Here we describe work done to identify time-dependent and -independent publication indicators for differentiating breakthrough papers.

Data and Methods

Creating a Gold Standard Data Set

The first challenge in testing the importance of various publication features in predicting research breakthroughs is defining a core set of publications to be used as a gold standard. For our gold standard set of breakthroughs, we selected research articles from the following sources that highlight advances in cancer research:

1. The American Association of Cancer Research (AACR) publishes the AACR Cancer Progress Report annually (176 articles from the 2011-2014 reports).
2. The American Society of Clinical Oncology (ASCO) reports on key research in their annual Report, ASCO Clinical Cancer Advances. (58 articles from the 2009-2013 reports).
3. *Nature Medicine* 2011 special edition focused on advances in cancer research (74 articles spanning publication years 2008-2010).

Using these three sources we identified 287 distinct breakthrough publications that were indexed in the *Web of Science.* Table 1 shows the frequency by *Web of Science* Journal Subject Category. The inclusion of older publications (e.g., publication years of 2008 and 2009) enabled the curation of a dataset that included papers mature enough to have a range of breakthrough characteristics.

Table 1. Top 10 Web of Science Journal Subject Categories by Frequency for the Breakthrough Gold Standard Set (N=287).

Journal Subject Category	Count
Oncology	118
Medicine, General & Internal	109
Multidisciplinary Sciences	31
Cell Biology	17
Biochemistry & Molecular Biology	11
Public, Environmental & Occupational Health	7
Hematology	7
Genetics & Heredity	6
Immunology	6
Medicine, Research & Experimental	5

227 of the 287 breakthrough publications (81.7%) were published in journals in either the Oncology or Medicine, General & Internal *Web of Science* Journal Subject Categories.

Comparison Group Publication Set

We chose a comparison group of publications from a similar set of *Web of Science* Journal Subject Categories. We retrieved 647,879 publications from the 1) Oncology and 2) Medicine, General and Internal categories published between 2008 and 2014. We selected 2,500 publications at random from this dataset for use as the comparison group. We chose to select our control group by matching on the distribution of journal subject categories between the gold standard and comparison sets. However, we did not match the control group on publication year distribution due to the uneven publication year distribution resulting from the gold standard selection criteria.

Publication Indicators- bibliographic, citations, and altmetrics

We collected data from *Web of Science* to generate indicators for inclusion in our assessment. The majority of indicators were derived from the individual *Web of Science* citation records. These indicators were at the publication level (Table 2) and were collected in January 2015. While using a field-normalized Journal Impact Factor (JIF) would have been preferable, some publications in the gold standard set do not have JIFs determined for the publication journal, so we chose to use JIF best quartile as the best available alternative. Npayoffs reflects the inclusion of altmetrics gathered from *Web of Science* usage.

Table 2. Publication-level Indicators Considered For Inclusion in Random Forest Models.

Indicator level	Variable	Description
publication	TimesCitedTotal	total cites
	TimesNSCitedTotal	total cites (non-self)
	TimesCited2y	total cites in past 2 years
	TimeNSCited2y	total non-self cites in past 2 years
	NPages	total number of pages in an article
	NCitedRefs	number of references
	NAuthors	number of authors
	PubYear	publication year
	NCitedJSC	number of JSCs present in cited references
	NCountries	number of countries associated with publication authors
	NOrgs	number of institutions associated with publication authors
	CitVel6m	Citation velocity of specified time period (or maximum number of days since the article was published)
	CitVel1y	
	CitVel2y	
	CitVel5y	
	Bestquartile	Journal's best quartile from the 2013 Journal Citation Report
	DocumentTypeID	Describes publication type (article, review, etc.)
	Npayoffs	Total number of payoff events in Web of Science since January 2013 • A payoff event is when a WoS user downloaded the full-text article, added EndNote library, or saved for future use • Robot data filtered using multiple algorithms

Author-level indicators, person disambiguation

Some of the indicators in the study at the publication-level require a time lag after publication so we sought to increase the number of indicators that could identify potential breakthroughs immediately upon publication. Currently, these additional indicators are based on author publication history characteristics (Table 3). A critical aspect of author-based indicators is ensuring that each author's characteristics are correctly attributed. Therefore, we used a proprietary semi-automated algorithm to disambiguate authors and assign publications to each unique author.

Author-level indicators were assigned to each publication and computed in one of two ways: by averaging the indicator for all authors on a publication or by averaging the indicator for the top three authors on the paper as ranked by the indicator values.

Table 3. Author-level Indicators Considered for Inclusion in Random Forest Models.

Indicator level	Variable	Description
author	AvgNCoAuth	Number of distinct co-authors on all publications in the journal subject categories of oncology or general and internal medicine from 2008-2014
	AvgNCoAuth_Top3	
	AvgHindex	H-index based on all publications in the journal subject categories of oncology or general and internal medicine from 2008-2014
	AvgHindex_Top3	
	AvgPubHist	Total number of publications in the journal subject categories of oncology or general and internal medicine from 2008-2014 divided by six years
	AvgPubHist_Top3	
	NHighCitPubs	Highly cited publications defined by top 10% of publications in a particular year and journal subject category
	AvgNHighCitPubs	
	AvgNHighCitPubs_Top3	

Random Forest™ Model

We used the Random Forest™ machine learning algorithm (Brieman, 2001) as implemented by Liaw and Wiener (Liaw & Wiener, 2002) to assess the relative importance of each of the indicators listed above for differentiating breakthroughs from our comparison group. As Random Forest™ cannot handle null values; we were required to exclude all publications without citations and all publications where authors could not be disambiguated. This resulted in a final dataset of 223 breakthrough publications and 1,170 comparison publications.

The Random Forest™ algorithm is an example of a bagged decision tree algorithm (Breiman, 1996) that combines the classification results of some number N of individual decision trees. This set of N trees comprises the forest and is one of two input parameters that can be specified by the user. The other input parameter is an integer m which specifies the number of variables to consider when deciding how many variables to use for each node in the tree. Details on implementing this algorithm can be found in Liaw 2002 and references therein. As the random forest is built, a random subset of 2/3 of the data is used in the construction of each tree. The remaining 1/3 of the data is referred to as 'out-of-bag' (oob). For the analyses shown, the values $N = 500$ and $m = 4$ were found to minimize the out-of-bag error rate, which is a measure of the misclassification of the oob data by the random forest.

Results

We first examined the correlation among our publication indicators and removed the following indicators that were highly correlated: CitVel6m; CitVel2y; CitVel5y; TimesCitedTotal; TimesCited2y; AvgHindex_Top3; NHighCitedPubs_Top3. With the remaining set of indicators, we then ran the first Random Forest models using both the Mean

Decrease Accuracy (MDA) and Mean Decrease Gini (MDG) to determine the relative importance of the indicators, as shown in Figure 1. The indicators with the highest relative importance are time-dependent (left of the dotted line). However, in order to best inform program management, it would be preferable to predict breakthroughs soon after publication, requiring indicators that can be calculated at, or near, the time of publication.

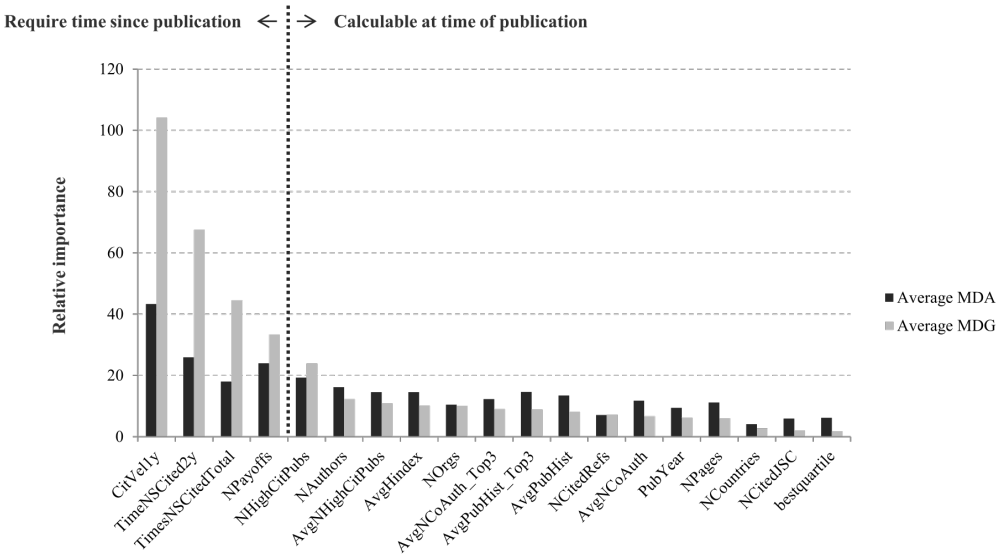


Figure 1. Relative Importance Ranking of Time-Independent and –Dependent Indicators based on Random Forest models (MDG and MDA). Out-of-bag error rate is 4.67%.

Because this work focuses on identification of publications with strong breakthrough potential near time of publication, we then considered only the time-independent indicators and produced new Random Forest models using these data. The relative importance ranking of the time-independent indicators are shown in Figure 2.

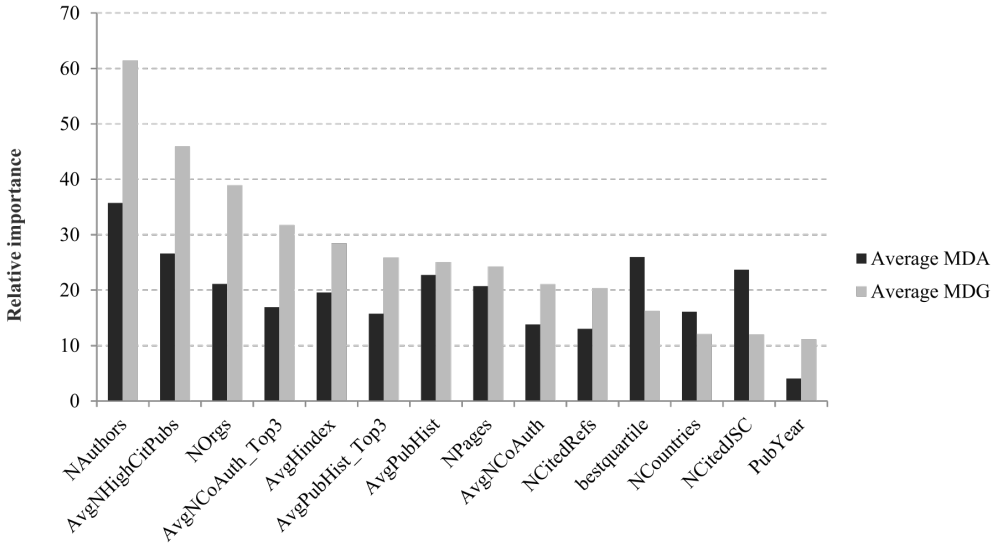


Figure 2. Relative Importance Ranking of Time-Independent Indicators based on Random Forest models (MDG and MDA). Out of bag error rate is 9.48%.

The highest ranked time-independent indicators, sorted by Average MDG, were: NAuthors, AvgNHighCitPubs, NOrgs, AvgNCoAuth_Top3, and AvgHindex. Sorting by Average MDA gives a slightly different set of top five variables: NAuthors, AvgNHighCitPubs, bestquartile,

NCited Journal Subject Category (JSC), and AvgPubHist. While the first two variables are the same for either type of ranking, it would be interesting to explore the divergence of the other variables between the two rankings. The relative importance of these time-independent indicators is consistent with breakthrough work being associated with teams and researchers with a history of strong performance.

Conclusions and Next Steps

We have identified and ranked a set of time-dependent and -independent indicators for their importance in differentiating a set of breakthrough publications from a comparison group. Our results are early steps in developing tools for potentially identify promising emerging research in a timely manner. Our next steps include using a subset of these indicators to establish a multivariate model where the outcome is the estimated probability of being a breakthrough paper based on the existing training set. Using this model, we will prospectively identify candidate breakthroughs and share the results with program officers within NCI to assess the practical value of the model. Future work could include efforts to determine which indicators gain or lose predictive value over time through iterative evaluation of the relative strength and importance of each indicator.

Acknowledgments

This study was improved by contributions from Danielle Daee (NCI); Di Cross, Leo DiJoseph and Joshua Schnell (Thomson Reuters); and extends work by Ilya Ponomarev (formerly Thomson Reuters) and Charles Hackett (National Institutes of Allergy and Infectious Diseases). This work was supported in part by NIH contract #HHS263201000058B.

References

Boyack, K.W., & Börner, K. (2003). Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers, *Journal of the American Society for Information Science and Technology*, *54*, 447-461.

Breiman, L. (1984). *Classification and regression trees.* Belmont, CA: Wadsworth International Group.

Breiman, L. (1996). Bagging Predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, *Journal of the American Society for information Science and Technology*, *57*, 359-377.

Compañó, R., & Hullmann, A. (2002). Forecasting the development of nanotechnology with the help of science and technology indicators, *Nanotechnology*, *13*, 243.

Dunne, C., Shneiderman, B., Gove, R., Klavans, J., & Dorr, B. (2012). Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization, *JASIST*, *63*, 2351-2369.

Fujita, K., Kajikawa, Y., Mori, J., & I. Sakata. (2012). Detecting Research Fronts Using Different Types of Combinational Citation, *Detecting Research Fronts Using Different Types of Combinational Citation*.

Huang, Y.H., Hsu, C.N., & Lerman, K. (2013). Identifying Transformative Scientific Research, *IEEE 13th International Conference on Data Mining* (ICDM), (pp. 291-300).

Klavans, R., Boyack, K.W., & Small, H. (2012). Indicators and precursors of "hot science", *17th International Conference on Science and Technology Indicators*, (pp. 475-487).

Klavans, R., Boyack, K.W., & Small, H. (2013). Identifying Emergent Opportunities in Science. Retrieved June 2, 2015 from: http://www.mapofscience.com/pdfs/EAGER_Final_v1.pdf

Liaw, A. & Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, 2/3, (pp. 18-22).

National Science Board. (2007). Enhancing Support of Transformative Research at the National Science Foundation, *National Science Foundation*, (p. 14).

ODNI, (2011). IARPA Launches New Program to Enable the Rapid Discovery of Emerging Technical Capabilities.

Ponomarev, I.V., Lawton, B.K., Williams, D.E., & Schnell, J.D. (2014). Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction?, *Scientometrics*, *100*, 755-765.

Reardon, S. (2014). Text-mining offers clues to success: *Nature*, *509*, 410.

Dimensions of the Author Citation Potential

Pablo Dorta-González[1], María-Isabel Dorta-González[2] and Rafael Suárez-Vega[3]

[1] *pablo.dorta@ulpgc.es,* [3] *rafael.suarez@ulpgc.es*
Universidad de Las Palmas de Gran Canaria (Spain)

[2] *isadorta@ull.es*
Universidad de La Laguna (Spain)

Introduction

It is well known that in some fields the average number of citations per publication is much higher than in others (Moed, 2005).

For decades, the number of publications and the number of citations have been the two accepted indicators in ranking authors. Recently, alternative indicators which consider both production and impact have been proposed (Dorta-González & Dorta-González, 2011; Egghe, 2013). However, these indicators based on the h-index do not solve the problem when comparing authors from different fields of science. Given the large differences in citation practices, the development of bibliometric indicators that allow for between-field comparisons is clearly a critical issue (Waltman & Van Eck, 2013).

Traditionally, normalization of field differences has usually been based on a field classification system. In said approach, each publication belongs to one or more categories and the citation impact of a publication is calculated relative to the other publications in the same field.

In our topic normalization we use the aggregate impact factor of three different sets of journals as a measure of the different dimensions in the citation potential of an author.

Dimensions of the author citation potential

Even within the same field, each researcher is working on one or several research lines that have specific characteristics, in most cases very distant from those of other researchers.

Generally, the citation potential in a field is determined within a predefined group of journals. This approach requires a classification scheme for assigning publications to fields. Given the fuzziness of disciplinary boundaries and the multidisciplinary character of many research topics, such a scheme will always involve some arbitrariness and will never be completely satisfactory. Therefore, we propose measuring the citation potential in the specific topic of each author and using this measure as an indicator of the probability of being cited in that topic.

The problem underlying the characterization of the author citation potential is as follows. Given a set of publications from an author in different journals and years, we will try to obtain a measure of the author topic defined by some dimensions of these publications so it can be compared with that of a different author (with publications in different journals and years).

Let us consider a 5-year time window Y. In this paper, we propose characterizing the topic of an author in period Y using three different dimensions (see Figure 1): the weighted average of the impacts in the journals containing the author's papers in Y (production dimension P), the weighted average of the impacts in the journals citing the author's papers in Y (impact dimension I), and the weighted average of the impacts in the journals included as references in the author's papers in Y (reference dimension R).

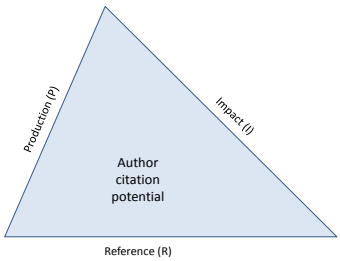


Figure 1. The three dimensions of the author citation potential.

In this characterization we propose the use of journal impact indicators instead of number of citations received by a particular paper. This is because it is necessary that several years pass after the publication of a document, so that the number of citations can be a consistent indicator in comparing similar documents of the same type published in the same year with that of other researchers in the same field. In some fields (e.g., Economics) more than 5 years are needed to obtain a consistent measure of impact (Dorta-González & Dorta-González, 2013). In many fields of the Humanities it is necessary to wait even longer (Dorta-González & Ramírez-Sánchez, 2014).

Materials and Methods

The bibliometric data was obtained from the online version of the Scopus database. Only journal papers in the period 2009-2013 were included, considering for each journal the Scimago Journal Ranking –SJR–. Four subject areas were considered: Chemistry, Computer Science, Medicine, and Physics & Astronomy. This was motivated in order to obtain authors with systematic differences in publication and citation behavior. We designed a random sample with a total of 120 authors (30 in each subject area). They were selected from the highly productive authors of the Consejo Superior de Investigaciones Científicas –CSIC– (Spain).

Results and discussion

The subject areas considered are very different in relation to the citation behavior. For this reason, in the sample there are important differences among the dimensions of the citation potential from one author to another. However, the proportion between production and impact dimensions is very close in all the subject areas considered (Figure 2).

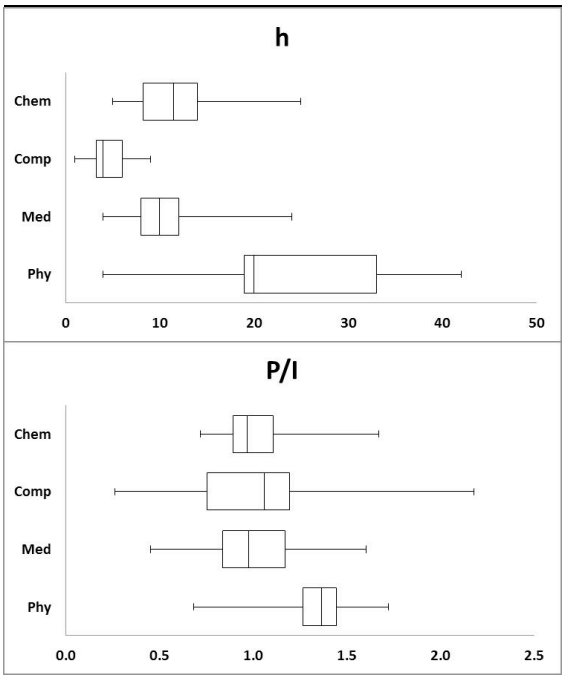


Figure 2. Box-plots comparing the subject areas.

Within- and between-group variability are both components of the total variability in the combined distributions. So: within variability + between variability = total variability.

Note in Table 1 that the proportion between production and impact dimensions produces the greatest percentage reduction of the variance. A more detailed analysis of the results can be found in Dorta-González et al. (2015).

Table 1. Central-tendency and variability.

	P	I	R	P/I
Median	1.521	1.526	2.564	1.065
Mean	1.719	1.546	2.759	1.093
Range of variation	3.692	3.776	7.527	1.915
Within-group variance	46.360	25.089	192.557	9.972
Between-group variance	39.434	17.325	54.463	2.358
Reduction in the variance	14.9%	30.9%	71.7%	76.3%

Conclusions

We have developed a measure of scientific performance whose distributional characteristics are invariant across scientific fields. Such a measure could be employed in the normalization of the impact at the author level in order to allow direct comparisons of scientists in different fields and permit a ranking of researchers that is not affected by differential publication and citation practices across fields.

References

Dorta-González, P., & Dorta-González, M. I. (2011). Central indexes to the citation distribution: A complement to the h-index. *Scientometrics*, *88*(3), 729-745.

Dorta-González, P., & Dorta-González, M. I. (2013). Comparing journals from different fields of science and social science through a JCR subject categories normalized impact factor. *Scientometrics*, *95*(2), 645-672.

Dorta-González, P., Dorta-González, M.I., & Suárez-Vega, R. (2015). An approach to the author citation potential: Measures of scientific performance which are invariant across scientific fields. *Scientometrics*, *102*(2), 1467-1496.

Dorta-González, P., & Ramírez-Sánchez, M. (2014). Producción e impacto de las instituciones españolas de investigación en Arts & Humanities Citation Index (2003-2012). *Arbor*, *190*(770), a191.

Egghe, L. (2013). Theoretical justification of the central area indices and the central interval indices. *Scientometrics*, *95*(1), 25-34.

Moed, H.F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Waltman, L., & Van Eck, N. J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics*, *96*(3), 699-716.

Scholarly Book Publishers in Spain: Relationship between Size, Price, Specialization and Prestige

Jorge Mañana-Rodríguez[1] and Elea Giménez Toledo[2]

[2] *jorge.mannana@cchs.csic.es*
Centre for Human and Social Sciences, Spanish National Research Council. C/ Albazanz, 26-28. Madrid. Spain.

Introduction

The prestige of book publishers is an important element for the assessment of SSH scholars in Spain. Until 2012, that 'prestige' remained based upon subjective, individual judgements from assessment committees' members. In order to provide a more objective reference for the prestige of book publishers, ÍLIA research group developed a ranking of book publishers (so called SPI) based on the opinion of almost three thousand experts from all SSH fields (Gimenez et al., 2013). Nevertheless, the factors underlying the perceived prestige are unknown. Some authors worked on the influence of marketing on the perception of books. Squires (2007) point out that 'we should not underestimate the value or efficiency that the association with a specific publisher provides to its contents'. It is hypothesized that three factors (among others) might be related to the perceived prestige: size of the book publisher (number of titles published), specialization (share of titles in each discipline) and price of the books. This research present the results of a correlational study on prestige, size, specialization and price of SSH book publishers in Spain.

The perception of 'prestige' strongly differs among different subjects to which the term can be applied. When the object is a product or a brand (with book publisher names as equivalent) the quantifiable variables related to the perception by different subjects of the different levels of prestige is relevant for explaining or defining the construct. The overall number of titles published by a book publisher could act as a reinforcement of the perception of prestige since the frequency with which the reader or consumer will be exposed to the brand is statistically more probable and this could lead to a perception of the publisher as able to publish more and better than others. In many goods, the perception of the prestige of competitors, in a similar way to how multi-branding strategies operate (Rahnamaee, A., & Berger, 2013). A brand prestige might also affected by the price (Yeoh & Paladino, 2013), and so the price of book might partially contribute, in a linear fashion, to the perceived prestige of book publishers.

Finally, specialization, as a factor, which might create a link between a specialized scholar with an specialized publisher, might contribute to influence the perception of the publisher as more prestigious in absolute terms. Since Scholarly Publishers Indicators (SPI) is being currently used as a source of information for assessment procedures in Spain (in some SSH fields), it is important to know whether the perceived prestige can be attributed to factors unrelated to the essential issues in research evaluation or if, by the opposite, the perceived prestige is not strongly (linearly) associated to these external factors.

Objectives

The objective of this research is to test the hypothesis stating that there is a linear relationship between prestige, size, specialization and price of books of book publishers in the case of Spain.

The information sources are the following:

-Prestige values: Scholarly Publishers Indicators (SPI, 2012).

-Size, price and specialization: DILVE (DILVE, 2013).

Variable definition:

-Prestige: ICEE (Prestige measure based on extensive survey to researchers and lecturers)

-Size: Raw number of different titles in DILVE for each discipline

-Mean price: the average price of all the titles published by the book publisher in the period analyzed.

-Max. Price: the maximum price of a single title in the whole set of titles published by each publisher.

-Specialization: Share of titles of publisher according to DILVE.

Methodology

For a total number of 119 book publishers (this number was fixed so that the number of lost cases is minimized), their ICEE was retrieved from SPI (2014, and the size, mean price and specialization degree obtained from the extensive database DILVE, for the years 2004 onwards up to 2012. The reason for including data from 2004 onwards is the fact that prestige, as other consumer perceptions, are developed over time so a smaller time span would not provide suitable. Data prior to 2004 is not fully consistent in DILVE database when compared with the publishers resulting from the questionnaire on publishers prestige due to the several changes (splits and merges) which took

place sin that date among book publishers, often involving the disappearance of book publishers names as they were and therefore requiring a much more complex codification of the previous names in order to keep the reliability of the data set. After a verification of the non-normality of the distribution of all the variables, using Kolmogorov-Smirnov nonparametric tests, Spearmans' Rho was selected as the appropriate technique contrasting the linear association hypothesis. The correlation matrix for all the variables was calculated using IBM SPSS (v. 19).

Results

Only significant results (p-value = .05) have been considered, since there is no reason for supposing any bias effect of *n* on the significance of the results (119, in all cases). The following table resumes these statistically significant correlations.

Table 1. Statistically significant correlations (Spearman's Rho).

ρ Publisher Prestige, Raw Size	.269; p < .05
ρ Publisher Prestige, Max Price	.217; p < .05
ρ Raw Size, Max Price	.198; p=.019
ρ Raw Size, Average price	-.232; p < .05
ρ Raw Size, Max Share	.433; p < .05
ρ Max Price, Average price	.593 p < .05

Conclusions

The main conclusion which can be drawn from the results is the seemingly (at least linear) independence of the construct 'prestige' from all the variables hypothesized as potentially influential in the values given to book publishers by the experts. The correlations of publishers' prestige with Raw Size (Number of Titles) and Max. Price, although statistically significant, are small enough as to suppose that the influence of these two variables in the perception of a publisher's prestige is not strong enough as to make necessary normalization measures. These results also suggest (at least from the perspective of a linear relationship) that the rankings in use are not biased by the possible influence of the great number of books, multiple branding and specialization or prices which sometimes can be displayed by some of the publishers belonging to big publishing houses which occupy the highest positions in the rankings.

Discussion

The fact that none of the variables analyzed is linearly related to the perceived prestige of book publishers is consistent with the multi-component structure generally involved in the composition of a concept such as 'prestige'. Also, since it is hardly possible to quantify the 'quality' (an also multi-faceted concept, particularly in the framework of research evaluation) of the contents of the books which, escalated to book publisher level of aggregation could contribute to the perceived prestige, the plausible influence of this factor remains unknown, although further research might offer new insight into this particular relationship. The existence of such relationship between the intrinsic quality of the contents and the prestige of a publisher is also plausible given that the use of books by those who have provided the prestige values presumably use the books as a source of information and as a form of scholarly communication where the quality of the contents might be the core of the perceived prestige, leaving behind other subjectively perceived variables. Also, given the relevance of peer review for assessment processes (Verleysen & Engels, 2013) as well as for the quality of the contents, the use of these filters might be related to the perceived prestige of book publishers.

References

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation*, 22(1), 64–77. doi:10.1093/reseval/rvs036

Rahnamaee, A., & Berger, P. D. (2013). Investigating consumers' online purchasing behavior: Single-brand e-retailers versus multi-brand e-retailers. *Journal of Marketing Analytic*s, 1(3), 138-148.

Squires, C. (2007). *Marketing Literature: The Making of Contemporary Literature.* Basingstoke: Palgrave Macmillan.

Verleysen, F. T. & Engels, T. C. E. (2013). A Label for Peer-Reviewed Books. *Journal of the American Society for Information Science and Technology*, 64, 428-430

Yeoh, M., & Paladino, A. (2013). Prestige and environmental behaviors: Does branding matter*? Journal of Brand Management*, *20*(4), 333-349.

Bootstrapping to Evaluate Accuracy of Citation-based Journal Indicators

Jens Peter Andersen[1] and Stefanie Haustein[2]

[1] *jepea@rn.dk*
Medical Library, Aalborg University Hospital, Sdr. Skovvej 15, 9000 Aalborg (Denmark)

[2] *stefanie.haustein@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal (Canada)

Introduction

Bibliometric indicators ranking aggregate units have a long tradition, including criticisms of methodology, interpretation and application. Despite the criticism, there is a demand for these indicators, and recent developments have led to improvements of methodology and interpretation. An essential element of these interpretations is to provide estimates of the accuracy, robustness, stability and confidence of bibliometric indicators, thereby providing the reader with data required to interpret results. This has, for example, been demonstrated for the set of indicators in the Leiden ranking (Waltman et al., 2012), the Journal Impact Factor (Chen, Jen, & Wu, 2014) and other journal indicators (Andersen, Christensen, & Schneider, 2012) as well as author metrics (Lehmann, Jackson, & Lautrup, 2008). The present study applies the same type of bootstrapping technique to estimate stability, as is used in the Leiden ranking (Waltman et al., 2012), on an array of citation-based journal indicators. The purpose of this analysis is to compare recent methodological advances, as well as traditional approaches. The study is based on clinical medicine journals in the Web of Science (WoS).

Methods

Data acquisition

The dataset contains all articles and reviews in the WoS, published in 2012 in journals classified as clinical medicine according to the National Science Foundation (NSF) classification system. This amounts to 362,556 papers and 2,699 journals from 34 different specialties within the discipline of clinical medicine. Each journal and paper is assigned to exactly one specialty. Citations are observed for a two-year window. In order to account for field differences in citation patterns, relative citations, $\hat{c}$, are computed by normalising observed against expected citations per specialty and year.

Journal indicators

The journal citation indicators selected for this study represent both traditional (means and medians of observed and relative) and novel (percentile) approaches. For a given journal *j*, we calculate the mean citations, μ_c, median citations, M_c, mean relative citations, $\mu_{\hat{c}}$, median relative citations, $M_{\hat{c}}$, top decile ratio of citations, N_{D10}, and relative citations. The top decile ratio for a journal is the percentage of papers present in the overall set of papers with citations in the highest decile range.

Indicator evaluation

Each indicator is evaluated for every journal by performing bootstrapping (Efron & Tibshirani, 1993). The technique involves resampling with replacement, i.e. for a given sample, all observed values are resampled so that a new sample of the same size is drawn randomly, but with the possibility that the same observation can be drawn multiple times. When repeating this resampling numerous times, we can calculate stability intervals to estimate how accurately the observed indicator value describes the underlying observations or whether it is influenced by outliers and thus less robust. To make our results comparable to those reported in the Leiden ranking, we have chosen to iterate each bootstrap 1,000 times and calculate 95% confidence intervals. In addition to this confidence interval we also calculate the standard deviation for each distribution. As the values of the different indicators are observed in very different ranges, we provide an additional mean-standardized version of every indicator. All calculations are performed using the *boot* package (Canty & Ripley, 2015) for *R* version 3.0.3 x64 (R Development Core Team, 2010).

Results and Discussion

We find that bootstrapping can identify outlying indicator scores within a specialty, by showing stability intervals (95% confidence intervals) for every indicator. As exemplified in Figure 1 for the subset of dentistry journals, the stability intervals demonstrate the robustness of rankings based on particular indicators. While, for example, the

stability intervals indicate that the citation impact of the 1st journal in Figure 1 is higher than that of the 5th, the first four journals cannot be clearly distinguished in terms of mean citation impact. Their mean citation rates are heavily influenced by a few highly cited papers.

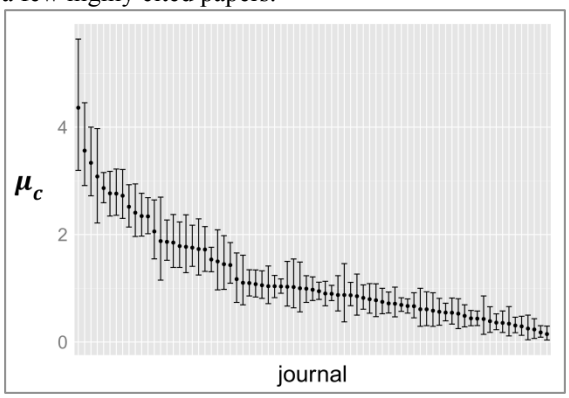


Figure 1. μ_c with stability intervals for all journals in the dentistry specialty.

The study also shows that the percentile-based indicators perform considerably better regarding stability than both mean- and median-based indicators (Figure 2 and Table 1). It is particularly interesting that the medians indicators do not seem to be more stable than the means.

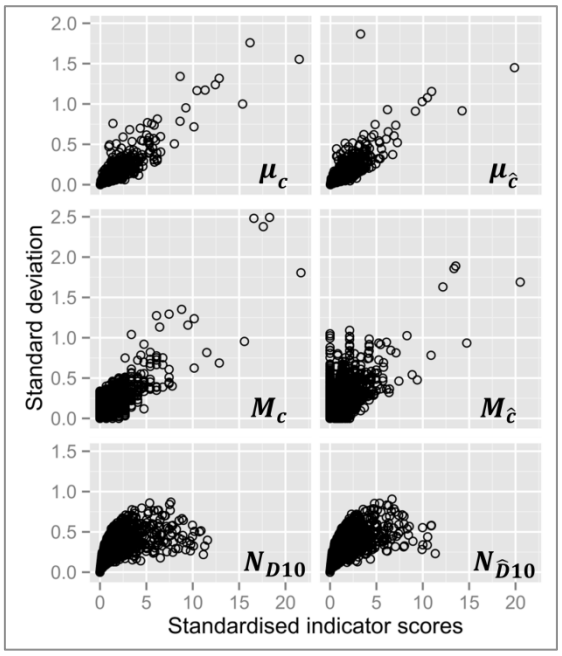


Figure 2. Standard deviation of bootstrapped scores as a function of standardised indicator scores, limited to journals with at least 50 papers.

Finally, we show that indicators are extremely sensitive to sample sizes. Journals with less than 50 papers published in the observation period show significantly larger variance than those publishing at least 50 papers (Table 1). Our results reiterate the importance of testing indicators and providing stability intervals to improve their interpretability. This would identify the limitations of rankings and avoid cases like the 24-fold increase of *Acta Crystallographica A*'s impact factor in 2009 (Haustein, 2012).

Table 1. Mean indicator values and standard deviations for all journals ("All") and journals publishing 50 or more papers ("≥50").

	All				≥50	
	Raw		Standardised			
Indicator	*mean*	*SD*	*mean*	*SD*	*mean*	*SD*
μ_c	2.321	3.897	1.000	1.679	1.052	1.261
M_c	1.477	2.278	1.000	1.543	1.079	1.471
$\mu_{\hat{c}}$	0.835	1.107	1.000	1.326	1.053	1.076
$M_{\hat{c}}$	0.520	0.717	1.000	1.381	1.075	1.297
N_{D10}	0.081	0.131	1.000	1.625	1.107	1.640
$N_{\hat{D}10}$	0.078	0.119	1.000	1.536	1.090	1.513

Further research will include in-depth analyses of multiple indicators and differences of stability intervals across specialties.

References

Andersen, J. P., Christensen, A. L., & Schneider, J. W. (2012). An approach for empirical validation of citation-based journal indicators. In E. Archambault, Y. Gingras, & V. Lariviére (Eds.), *Proc. of STI 2012* (pp. 71–81). Montréal, Canada: 17th International Conference on Science and Technology Indicators.

Canty, A., & Ripley, B. (2015). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-15.

Chen, K. M., Jen, T. H., & Wu, M. (2014). Estimating the accuracies of journal impact factor through bootstrap. *Journal of Informetrics*, *8*(1), 181–196.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the Bootstrap* (p. 456). New York: Chapman & Hall.

Haustein, S. (2012). *Multidimensional Journal Evaluation. Analyzing Scientific Periodicals beyond the Impact Factor*. Berlin / Boston: De Gruyter Saur.

Lehmann, S., Jackson, A. D., & Lautrup, B. E. (2008). A quantitative analysis of indicators of scientific performance. *Scientometrics*, *76*(2), 369–390.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., Wouters, P. F. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *JASIST*, *63*(12), 2419–2432.

The Lack of Stability of the Impact Factor of the Mathematical Journals

Antonia Ferrer-Sapena[1], Enrique A. Sánchez-Pérez[1], Fernanda Peset[1], Luis-Millán González[2] and Rafael Aleixandre-Benavent[3]

[1]*anfersa@upv.es, easancpe@mat.upv.es, mpesetm@upv.es*
Instituto de Diseño y Fabricación, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia (Spain)

[2]*luis.m.gonzalez@uv.es*
Departamento de Educación Física y Deporte, Universitat de València. Gascó Oliag, 3. 46010 Valencia (Spain)

[3]*aleixand@uv.es*
INGENIO (CSIC-Universitat Politècnica de València). UISYS-Universitat de València. Plaça Cisneros, 4. 46003-València (Spain)

Introduction

Although the 2-year Thomson-Reuters Impact Factor (IF) has become a usual tool for measuring the scientific productivity of all fields of the natural sciences (see Aleixandre-Benavent, Valderrama Zurián, & González Alcaide, 2007), its behavior in the particular case of the journals of pure mathematics (the area MATHEMATICS in the thematic directory of Thomson-Reuters) is far from being stable when its values in consecutive years are considered. If we consider the changes of the values of the IF of a given journal in the last decade, it can be easily seen that the variation of the values is surprisingly high if we compare with other disciplines. Mathematical journals seem to have the worst behavior regarding the time stability both of the IF and the position in the IF list.

A series analysis of a set of journals uniformly distributed in the IF list shows that the variations of the values of the IFs are very big when compared with other scientific disciplines, e.g., APPLIED PHYSICS and MICROBIOLOGY. The reader can see a representation of this behavior for three mathematical journals together with three journals of physics that have been chosen as representatives of these groups in the following graph (Fig. 1).

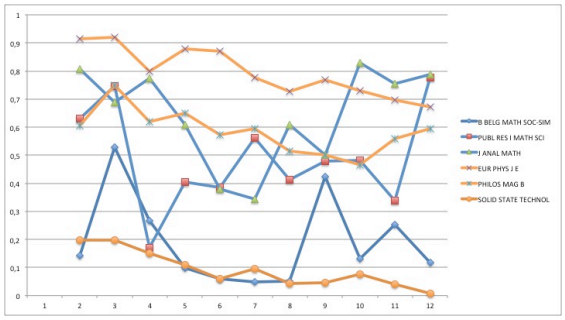


Figure 1. Variations of three journals of mathematics and three journals of physics.

In our study, we analyze the possible reasons for this fact, explaining some typical characteristics of the mathematical journals and of the research in mathematics, that make this science to have unusual properties from the point of view of the bibliometrics.

The research in pure mathematics

In general, mathematicians work in small groups of researchers from different parts of the world that are specialized in some topics, which have a long development period. For instance, it is usual that a group of mathematicians continue with some problems that appeared 50 years ago, or even before (see Behrens & Luksch, 2011). Although some of these topics were intensively studied some years ago, sometimes the research was left at that moment without having complete answers for some central questions, due to the fragility and the small size of the specialized group of researchers working on it. In this context, it is natural that after some years, a new group can recover the research and fruitfully continue with the investigation. The group of interested mathematicians is, almost in all cases, small. Even in new open topics, the size of the interested community of mathematicians is sparse and small. This of course changes when some particular theory becomes important due to the applications. But in these cases, the publication of the mathematical contents is redirected to more applied journals, or to journals of the fields where the theory finds applications.

This research dynamics is not usual at all, if we compare it with the pattern that can be observed in other fields. The main consequence is that the obsolescence of the scientific documents is faster in other sciences than in mathematics.

Mathematical journals

Classical journals that publish papers on pure mathematics follow also a different pattern that the

usual one in other scientific fields that are in some sense similar with respect to some descriptive parameters, as physics or other natural sciences. Although there are a lot of journals that are supported by big publishers—for example, Elsevier and Springer—, some of them preserve the editorial policy and the publication format that they used to have before. Another important group of journals is still published by national societies, universities and research institutes. Very often, these publications are small—in the sense that they publish a small number of papers per year—, but they are prestigious and serious papers are published in them.

This implies that the impact factor of these journals has a strong statistical variability, depending on the number of citations that a small number of papers can receive.

On the other hand, the publication of the papers is slow when compared with journals in other disciplines. Sometimes it takes more than two years for a paper from submission to publication. In general, this does not produce any problem for the dissemination and exchange of information, since the contents are often previously published by the authors in popular open access repositories as arXiv. Moreover, again the small size of the group of specialists interested in the topic reduces the pressure on the authors for a fast publication.

Conclusions: IF-based evaluation of the scientific productivity

The main direct consequence of the properties of the journals of mathematics together with the slow long-term activity in the research of the topics is the small rate of papers that are cited two years after their publication, when compared with other fields. This causes that the value of the IF of the journals is small even if they are prestigious and well-known in the field. For example, an IF of 0.5 is a reasonable impact factor for a journal, and enough to let it to be considered as a serious publication. This value is very small if we compare with other areas (see Bensman, Smolinsky & Pudovkin, 2010; Smolinsky & Lercher, 2012).

However, the 2-year IF is still the main tool in many countries—for example, Spain—to measure the production of a single mathematician or a research institute. This produces some fails in the evaluation systems, and lead the researchers to publish in journals that are considered by the community as less prestigious than others, as a consequence for example of the fact that these journals publish much more papers, and then have a better IF. Therefore, pure mathematics provides an example of a group of disciplines for which the IF-based evaluation clearly distorts the image of the scientific production.

Acknowledgments

This work has benefited from assistance by the National R+D+I of the Ministry of Economy and Competitiveness of the Spanish Government (CSO2012-39632-C02-01) and Prometeo Program for excellent research groups of Generalitat Valenciana (GVPROMETEO2013-041).

References

Aleixandre-Benavent, R., Valderrama Zurián, J. C., & González Alcaide, G. (2007). Scientific journals impact factor: limitations and alternative indicators. *El Profesional de la Informacion*, *16*(1): 4–11.

Behrens, H., & Luksch, P. (2011). Mathematics 1868–2008: a bibliometric analysis. *Scientometrics*, *86*, 179–194.

Bensman, S. J., Smolinsky, L. J., & Pudovkin, A. I. (2010). Mean citation rate per article in mathematics journals: Differences from the scientific model. *Journal of the American Society for Information Science and Technology*, *61*, 1440–1463.

Smolinsky, L, & Lercher, A. (2012). Citation rates in mathematics: a study of variation by subdiscipline. *Scientometrics*, *91*, 911–924.

Using Bibliometrics to Measure the Impact of Cancer Research on Health Service and Patient Care: Selecting and Testing Four Indicators

Frédérique Thonon[1,2], M. Saghatchian[1], R. Boulkedid[2] and C. Alberti[2]
[1]Gustave Roussy, European and International Affairs, Villejuif (France)
[2]Hôpital Robert Debré, unité d'épidémiologie clinique, Paris (France)

Introduction

Traditionally, biomedical research is measured by bibliometric indicators of scientific production and impact (such as number of publications and h-index) and indicators linked to clinical trial activities (Pozen & Kline, 2011). However, there has been an increasing demand in the last few years to measure the impact of medical research in terms of how it improves patients' well-being and public health (Wells & Whitworth, 2007; Ovseiko, Oancea, & Buchan, 2012). Measuring the final impact of research on patients' outcomes is difficult because of attribution problems and time lag between research and outcomes (Ovseiko, Oancea & Buchan, 2012). The aim of our research project is to select and test indicators measuring the impact of cancer research on health service and patient care

First step: indicators selection

See Figure 1 below for details of this process.

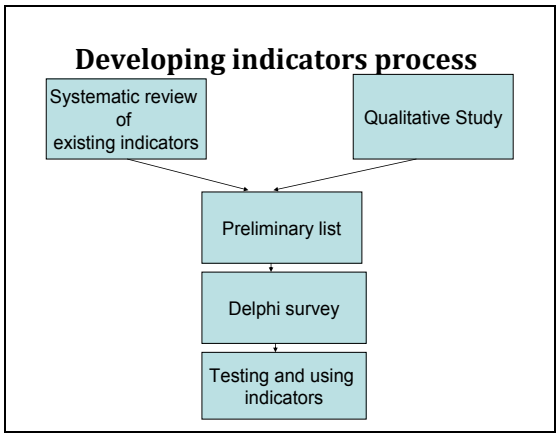


Figure 1: Indicators development process.

Systematic review of indicators

We firstly undertook a systematic review of existing indicators measuring the output and outcome of medical research in order to (1) enlist all the indicators that could potentially be used and (2) to describe their methodology, use, advantages and disadvantages. We took care of designing a study as comprehensive as possible, in order to include indicators ranging from those measuring research activity to those measuring the long-term impact of research. As a result we drew a detailed list of 57 indicators (Thonon et al., 2015).

Qualitative study of researchers

We wanted to develop indicators that would be accepted by those concerned by this evaluation system. Therefore, we undertook a qualitative study to explore the views of actors in translational research on the definitions, issues and evaluation modes of translational research. This study was done to complete the results of the systematic review with an input from the stakeholders directly involved. We interviewed 23 researchers, engineers, administrators and clinicians from diverse backgrounds and engaged in diverse fields of oncological translational research.

Delphi survey

Those two exploratory studies led us to the drawing of an initial list of 61 indicators. We submitted this list to all members of the platform for a modified Delphi survey (N=267). Participants were presented indicators, as well as their methodologies, advantages and disadvantages, and were asked to rate their feasibility and validity on a scale from 1 to 9, and to comment on them. Comments from participants were particularly useful to adjust the methodology of the indicators. In addition, a physical meeting was held where 26 participants discussed the inclusion and methodology of some indicators.

Results

As a result we were able to draw a list of 12 indicators, including 4 indicators that focused on measuring the impact of research on health service and patient care but not used in evaluation systems very often:

- Citation of research in clinical guidelines;
- Citation of research in public health guidelines;
- Number of clinical guidelines authored; and
- Number of validated biomarkers identified in publications.

Second step: indicators testing

We constructed the following methodology to measure those indicators: 17 European cancer centres have been selected in this study. We used

the Scopus database to extract all original articles published between 2000 and 2014 and analysed the data.

Citation of research in clinical guidelines

We selected clinical oncology guidelines published by the European Society of Medical Oncology (ESMO), the American Society for Clinical Oncology (ASCO), and the National Comprehensive Cancer Network. Those guidelines are published in, respectively, *Annals of Oncology, Journal of Clinical Oncology* and the *Journal of the National Comprehensive Cancer Network.* We analysed the number of publications cited in the 'clinical practice guidelines' issues of those journals. We searched the literature for data on the AGREE score of those guidelines to measure the validity of this indicator.

Authorship of clinical guidelines

We extracted and analysed data relative to the clinical oncology guidelines mentioned above.

Citation of research in public health guidelines

From the database of European publications (https://bookshop.europa.eu/en/home/) we searched for public health guidelines related to cancer. Then we extracted the references of the selected guidelines in Scopus and carried out a citation analysis.

Number of validated biomarkers identified in publications

We firstly performed a literature review to identify and list all validated biomarkers used in clinical practice for oncology patients. We then performed a search for all publications related to those biomarkers in the corpus of original articles.

Discussion

This study is still ongoing and the results will be available shortly. We believe those four indicators can provide an additional tool to measure the impact of cancer research on health service and patient care. Citation of research in clinical guidelines is the most investigated indicator (Lewison, 2003; Mostert et al., 2010). There is little literature on indicators linked to the citation of research in public health guidelines (Lewison, 2003) but none linked to indicators measuring the identification of biomarkers, despite the importance of their use for cancer patients' outcomes.

References

Lewison, G. (2003). Beyond outputs: New measures of biomedical research impact. *Aslib Proceedings, 55*, 32-42.

Mostert, S.P., Ellenbroek, S.P., Meijer, I., van Ark, G., & Klasen, E.C. (2010). Societal output and use of research performed by health research groups. *Health Research Policy and Systems, 8*, 30.

Ovseiko, P.V., Oancea, A., & Buchan, A.M. (2012). Assessing research impact in academic clinical medicine: a study using Research Excellence Framework pilot impact indicators. *BMC Health Services Research, 12*, 478.

Pozen, R., & Kline, H. (2011). Defining Success for Translational Research Organizations. *Science Translational Medicine*, *3*(94), 94cm20.

Thonon, F., Boulkedid, R., Delory, T., Rousseau, S., Saghatchian, M., van Harten, W., & Alberti, C. (2015, April 2). Measuring the outcome of biomedical research: a systematic literature review. *PLoS One*, 10(4):e0122239 doi: 10.1371/journal.pone.0122239.

Wells, R., & Whitworth, J.A. (2007). Assessing outcomes of health and medical research: do we measure what counts or count what we can measure? *Australia and New Zealand Health Policy*, *4*, 14.

A New Scale for Rating Scientific Publications

Răzvan Valentin Florian[1]

[1] *florian@epistemio.com*
Epistemio, str. Saturn nr. 26, 400504 Cluj-Napoca (Romania); 20-22 Wenlock Road, London N1 7GU (UK); and
Romanian Institute of Science and Technology, str. Cireşilor nr. 29, 400487 Cluj-Napoca (Romania)

Introduction

Citation-based bibliometric indicators are increasingly being used for evaluating research. This reflects the need of decision-makers to increase the efficiency of allocating resources to research institutions and scientists, while also keeping manageable and cost-effective the evaluation process that grounds the allocation of resources. There often is much room of improvement in how bibliometric indicators are being used in practice. But even state-of-the art bibliometric indicators suffer of a fundamental problem when used for evaluating research: the citations they are based upon are influenced by many factors beyond the quality of cited publications (Bornmann & Daniel, 2008) and these indicators need to be tested and validated against what it is that they purport to measure and predict, which is expert evaluation by peers (Harnad, 2008).

A solution to this problem is aggregating online ratings provided post-publication by the scientists who read the rated papers anyhow, for the purpose of their own research. Online-aggregated ratings are now a major factor in the decisions taken by consumers when choosing hotels, restaurants, movies and many other types of services or products. It is paradoxical that in science, a field for which peer review is a cornerstone, rating publications on dedicated online platforms is not yet a common behavior. For example, if each scientist would provide one rating weekly, it can be estimated that 52% of publications would get 10 ratings or more (Florian, 2012). This would be a significant enhancement for the evaluative information needed by decision makers that allocate resources to scientists and by other users of scientific publications.

For collecting this kind of ratings, a rating scale should be defined. Here I present the choices made during the development of the scale used at Epistemio, an online platform for aggregating ratings and reviews of scientific publications (www.epistemio.com).

Purpose

The expected usage of these ratings is: first, in steering of science by decision-makers, i.e. choosing to whom to allocate resources (typically contributed publicly), such as institutional funding, grants, jobs, positions, tenure, among the institutions, scientists, fields of science, etc. that compete for them; and second, in helping scientists to prioritize and filter the publications that they choose to read or use. For the first purpose, it is important to be possible to aggregate ratings across the set of publications of an individual, of a group of scientists or of an institution; and to be able to use the individual or aggregated ratings to rank the assessed entities. This implies that ratings should be unidimensional. While publications may be assessed across a number of characteristics, such as quality of research, quality of presentation, novelty, and interest, collecting individual ratings across all these dimensions reduces the response rates, and it is not clear how these multidimensional ratings may be aggregated into a scalar one. Therefore, it is desirable that an overall rating that reflects the overall properties of a publication is collected independently of ratings regarding individual characteristics of the publication. Collecting the latter may be left optional. This paper focuses on the overall rating.

What should be rated, exactly?

When experts are asked to rate a publication, the property that should be rated must be named. What is exactly this property? A proper discussion of this issue should analyze the foundations of scientific research, being outside the scope of the present paper. A different way of posing the problem is starting with the needs of expected users of the ratings, which were mentioned above. Typical desired properties of publications (and, therefore, of the results presented in these publications) that are mentioned in the context of steering of science is quality, importance, relevance, and impact. For usability purposes, the text of the question to raters should be kept brief; therefore, a choice must be made among the various wordings that may be used. Importance, long-term societal and scientific relevance, and long-term societal and scholarly impact seem to have similar semantics. Quality seems to be a complementary property: a publication may present potentially important results, but methodology and/or presentation may lack quality, therefore raising uncertainties about the real value of the publication; and a publication may be of high quality while the potential importance is low. We have thus chosen to use the wording “scientific quality and importance” for defining the variable that the ratings are supposed to estimate.

Scale type and range

Online ratings typically take the form of a five-star or ten-star discrete scale: this standard has been adopted by major players such as Amazon, Yelp, TripAdvisor and IMDb. However, these types of scales are likely not being able to measure well the quality and importance of scientific publications, because of the likely high skewness of the distribution of values of this target variable.

Let us consider the number of citations of scientific publications as a relevant proxy for the quality and importance of publications. About 44% of publications in Web of Science have zero citations, and the median number of citations is about 1, yet there is one paper having more than 305,000 citations and 148 papers having more than 10,000 citations (Van Noorden, Maher, & Nuzzo, 2014). In the case of patents, where the monetary value is defined by markets, the top 0.8% were valued at more than 1,000 times the median (Giuri et al., 2007). Let us assume that the main properties of these distributions generalize to the variable we want to measure, i.e. the maximum value can be of about 3 to 5 orders of magnitude larger than the median value. Therefore, a scale of 5, 10 or even 100 discrete categories cannot represent well this variability if the values that the scale represents vary linearly across categories. A logarithmic scale would be suitable, but it is psychologically difficult for most people to estimate values across so many orders of magnitude and to place them on a logarithmic scale.

A solution to this conundrum is asking experts to assess not the absolute value of the target variable, but its percentile rank. Then, the maximum value (100%) is represented by a number just 2 times larger than the median (50%), rather than several orders of magnitude larger. For usability and computational reasons, we limited the precision of the scale to 1%. Theoretically, this limits the capacity of indicating differences between top papers; in the case of the number of citations, in the top 1% the value varies from several hundreds to hundreds of thousands. In practice, test-retest reliability tends to decrease for scales with more than 10 response categories; users consider that a scale with 101 response categories allow them to best express their feelings adequately, but its ease and speed of use is slightly lower than of scales with 11 categories or less (Preston & Colman, 2000).

Because of the skewness of the distribution of absolute values, it is likely that experts are able to discriminate the percentile ranking of high quality papers better than the one of low quality papers. The confidence in rating papers also depends on how close the topic of the publication overlaps the expertise of the rater. For these reasons, raters should be able to express their uncertainty. Therefore, we allowed experts to give the rating as an interval of percentile rankings, rather than a single value. The rating is collected through a graphical interface representing the interval with sliding ends (Fig. 1). For ease of use on mobile devices, the interval can also be expressed using numerical selectors. A review may be associated to the rating, for explaining and supporting the rating.

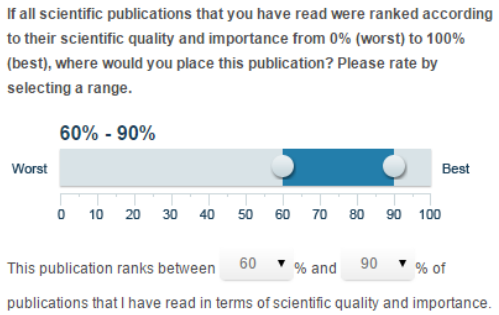


Figure 1. The Epistemio® rating scale for scientific publications.

Acknowledgments

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, project number PN-II-PT-PCCA-2011-3.2-0895.

References

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? *Journal of Documentation*, *64*(1), 45-80.

Florian, R. V. (2012). Aggregating post-publication peer reviews and ratings. *Frontiers in Computational Neuroscience*, 6(31).

Giuri, P., Mariani, M., Brusoni, S., Crespi, G., Francoz, D., Gambardella, A., et al. (2007). Inventors and invention processes in Europe: Results from the PatVal-EU survey. *Research Policy*, *36*(8), 1107–1127.

Harnad, S. (2008). Validating research performance metrics against peer rankings. *Ethics in Science and Environmental Politics*, *8*, 103–107.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*(2000), 1-15.

Van Noorden, R., Maher, B., & Nuzzo, R. (2014). The top 100 papers. *Nature*, *514*(7524), 550–553.

Analysis of the Factors Affecting Interdisciplinarity of Research in Library and Information Science

Chizuko Takei[1], Fuyuki Yoshikane[2] and Hiroshi Itsumura[3]

naoe.chizuko@ynu.ac.jp, fuyuki@slis.tsukuba.ac.jp, hits@slis.tsukuba.ac.jp

[1]University of Tsukuba, Graduate School of Library, Information and Media Studies, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

[2]University of Tsukuba, Faculty of Library, Information and Media Science, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

Introduction

In recent years, there has been a growing recognition of the necessity for interdisciplinary research that crosses disciplinary boundaries to deal with increasingly complex social issues (Rafols & Meyer, 2010). The relationship between the changes in interdisciplinarity of research over the years and researchers' attributions has rarely been investigated. Understanding the relationship between them will make it possible to gain useful information to foster interdisciplinary research, career-development of researchers, and development of research institutions. Thus, considering different periods, this study examines interdisciplinarity of research and the transdisciplinarity of researchers (targeted researchers themselves and their co-authors).

Methodology

This study targeted full-time faculty members of 2 iSchools, University of Pittsburgh (Pitt) and Syracuse University (SU), as of August 2014. The following data were employed: (1) information about targeted researchers and their co-authors, such as academic degrees or biographies, extracted from web pages; (2) bibliographic data of articles published by targeted researchers, which were extracted from Web of Science (WoS); (3) the title lists of WoS by subject categories acquired from the web site of Thomson Reuters; and (4) a matrix of the distance between categories of WoS, which was computed by Leydesdorff using Stirling's distance (http://www.leydesdorff.net/overlaytoolkit/stirling.htm). The procedure of this study was as follows: First, we examined transdisciplinarity of targeted researchers on the basis of the numbers of different disciplines where they had been engaged. We estimated their disciplines by several points of view such as belonging departments and academic degrees. As for their co-authors, though disciplines were estimated in the same way, we counted only disciplines that were different from those of the targeted researchers who had published the co-authored articles. Next, for each article of (2), by relating its reference list to (3) and (4), we computed indexes regarding interdisciplinarity that were used in later studies. This study applied the following indexes to the distribution of WoS categories assigned to the articles and their citing literature:

a. Total number of categories;
b. Simpson's Index (I);
c. Shannon's Index (entropy, H);
d. Distance between categories; and
e. The proportion of literature cited from different disciplines.

Indexes b and c evaluate the degree of diversity, taking into account both variety and equality in the frequency distribution. Index d indicates the distance between the categories of the articles and their citing literature. It ranges from −1 to 0, multiplying Stirling's distance by −1. As interdisciplinarity grows, their values become higher. Index e indicates the ratio of literature cited from different disciplines. Here, a different discipline is defined as a category with a distance over −0.7. Then, we performed a principal component analysis using these indexes and observed the correlation between the transdisciplinarity of targeted researchers or their co-authors and the interdisciplinarity of their articles along with its time-series variation. We discussed factors affecting the interdisciplinarity of research.

Results

Tendencies of indexes

Table 1 shows the basic statistics regarding transdisciplinarity of researchers and interdisciplinarity of their articles. We targeted 57 researchers, out of 73 faculty members, whose disciplines could be identified on the basis of information from university web sites and WoS.

The result of a principal component analysis for 5 indexes (C to G) revealed that the cumulative contribution rate of the first 2 principal components (PC1 and PC2) is 0.873. The characteristics of the 5 indexes can largely be explained by the first and second principal components. In Table 2, the principal component loading of PC1 suggests strong relationships between all 5 indexes. On the other hand, PC2 is characterized by large negative values of indexes F and G. Figure 1 is a plot of the first and second principal components and indicates that the 5 indexes can be divided into two groups (C, D, and E) and (F and G). It also implies that

highly interdisciplinary articles are remarkably diverse and rarely have common tendencies. In addition, we separated articles into two groups that were roughly equal in size (from 1981 to 2005 and from 2006 to 2014) to investigate the time-series variation related to the transdisciplinarity of researchers and the interdisciplinarity of research. The values of indexes concerning the interdisciplinarity of research (C to G) increased, while there were almost no changes in indexes concerning the transdisciplinarity of targeted researchers and their co-authors (A and B).

Table 1. Basic statistics regarding interdisciplinarity and transdisciplinarity.

		Pitt	SU	ALL
Targeted researchers/all faculties		23 / 30	34 / 43	57 / 73
Number of articles		267	259	526
Number of articles/targeted researchers	median	8	5	6
	range	1-33	1-31	1-33
A: Transdisciplinarity of targeted researchers	median	2	1	2
	range	1-2	1-3	1-3
B: Transdisciplinarity of co-authors	median	1	1	1
	range	0-6	0-4	0-6
C: Total number of categories	median	13	15	14
	range	1-79	1-59	1-79
D: Simpson's Index	median	0.781	0.767	0.777
	range	0-0.949	0-0.934	0-0.949
E: Shannon's Index	median	2.383	2.383	2.383
	range	0-4.385	0-4.061	0-4.385
F: Distance between categories	median	−0.438	−0.413	−0.424
	range	−1–−0.005	−1–−0.013	−1–−0.005
G: Proportion of literature cited from different disciplines	median	79%	79%	79%
	range	0%-100%	0%-100%	0%-100%

Table 2. Principal component loading for 5 indexes.

	PC1	PC2	PC3	PC4	PC5
C	−0.648	0.536	−0.540	0.002	−0.032
D	−0.876	0.301	0.345	0.037	−0.148
E	−0.898	0.350	0.202	−0.051	0.168
F	−0.717	−0.652	−0.089	−0.229	−0.031
G	−0.750	−0.610	−0.093	0.236	0.029

The relationship between transdisciplinarity of researchers and interdisciplinarity of their research

We computed Spearman's rank correlation coefficient for indexes A to G to survey the relationship between transdisciplinarity of researchers (A and B) and interdisciplinarity of their research (C to G) (Table 3). No strong correlation was found between them. However, comparing index A with B, we observed stronger and significant correlation between index B and the indexes concerning interdisciplinarity of research (C to G). In addition, we compared the articles before 2005 with those after 2006 to examine the time-series variation of correlation between indexes. Although there was no distinguished distinction between them, the degree of correlation tended to become stronger and the number of significant coefficients was increased for indexes A and B.

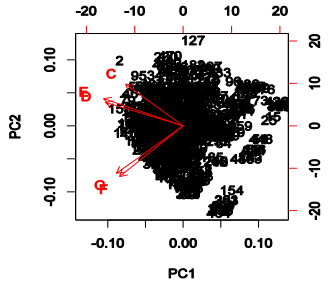


Figure 1. Plot of the first and second principal components.

Table 3. Rank correlation ρ among 7 indexes for all articles.

	A	B	C	D	E	F	G
A	1	0.23*	0.12*	0.17*	0.18*	0.05	0.06
B		1	0.21*	0.20*	0.21*	0.07	0.14*
C			1	0.69*	0.76*	0.17*	0.16*
D				1	0.99*	0.37*	0.30*
E					1	0.37*	0.30*
F						1	0.88*
G							1

*Significant ($p < 0.05$)

Discussion and Conclusions

This study computed indexes for interdisciplinarity of research in library and information science and performed principal component analysis to clarify the relationship among the indexes. The results indicate that the indexes considering the distance between subject categories of WoS have characteristics very different from the indexes considering only the number of categories and their frequency distributions. This suggests that we should consider a more multidimensional approach. Furthermore, we investigated changes over time in the indexes of interdisciplinarity, and observed the progress for interdisciplinarity of research in library and information science. As the results of the correlation analysis between interdisciplinarity of research and transdisciplinarity of researchers, stronger and significant correlations were seen with the transdisciplinarity of co-authors than with that of the targeted researchers themselves. This suggests that interdisciplinarity of research might be more affected by the transdisciplinarity of co-authors than by that of the researchers themselves. We will conduct further investigations with more samples.

Reference

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, *82*, 263-287.

An Analysis of Scientific Publications from Serbia: The Case of Computer Science

Miloš Pavković[1] and Jelica Protić[2]

[1]*milos_pakovic@yahoo.com* [2]*jelica.protic@etf.bg.ac.rs*
University of Belgrade, School of Electrical Engineering, Department of Computer Engineering and Informatics
Bulevar kralja Aleksandra 73, 11000 Belgrade (Serbia)

Introduction

In Serbia, like in other countries all over the world, career opportunities in computing are growing faster than most of the other professions. This trend should be in accordance with the growth of the number of study programs and consequently the number of teaching staff. The most important researchers' and university teaching staff's promotion criteria, according to the regulations in Serbia, are the papers published in journals from the JCR list, which is, for the area of computing, reduced to the SCIe list. The number of such papers is also relevant for projects financed by the Ministry of Education, Science and Technological Deve- lopment of the Republic of Serbia.

In this paper, we present an analysis of the references of Serbian researchers retrieved from the Web of Science. Using the bibliometric indicators from the Web of Science, we also examine the distribution of such references across WoS categories that belong to the broader area of computing. We show the distribution of such publications over the years, cities and universities and identify the relations with global trends in Serbian science.

Data Set

Data used in this paper were taken from Thompson Reuters Web of Science on 29 September 2014, selecting Science Citation Index Expanded (SCIe) journal articles. A basic search was conducted using the keyword ‘‘Serbia’’ in the field address and the retrieved results were limited to articles published during the period 2006–2013. All document information, including names of authors, titles, years of publications, source journals, contact addresses, and number of citations for each article, for every year, were downloaded into Microsoft Excel worksheets. The custom program in C# programming language was developed in order to perform data analysis.

The same data extraction was performed for WoS categories, that we considered the subcategories of the broader scientific area of Computer Science. The distribution of the number of papers from the year 2006 till 2013 (since results for 2014 were incomplete) is presented on Figure 1, and the number of papers over years and WoS categories is presented on Figure 2.

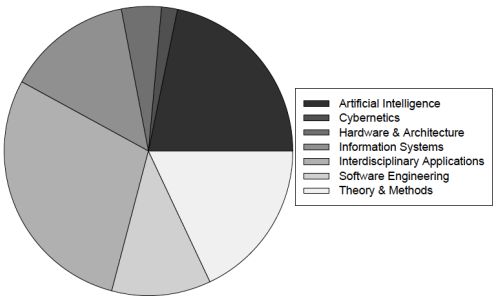


Figure 1. The number of papers in subcategories.

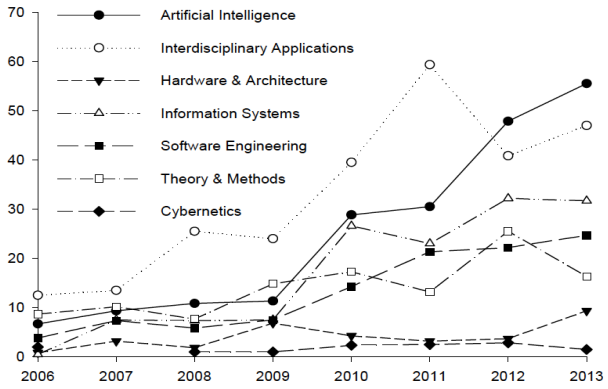


Figure 2. The number of papers in subcategories for each year.

To get numbers presented in Figure 2, disciplinary affiliation is computed fractionally, by assigning 1/N to each category, for a journal paper published in a journal indexed in N different categories.

The name of the country was not always correct for papers submitted before 2006, since our country changed its name to Serbia in 2006, and some papers had the former name Serbia and Montenegro, or even Yugoslavia in their affiliation. Therefore, the additional search was performed using only the names of significant Serbian cities and university centres. It was noticed that our dataset did not hold absolutely correct

information, because of unintentional mistakes in the authors' signatures or other elements of the affiliation. Incorrectly entered data propagate errors to later identification and grouping, as stated in Mitrovic (2014). This issue can be solved partially using text similarity matching algorithms. Our program uses Jaro-Winkler algorithm as proposed in Winkler (1995), also known as JWSF, "Jarod-Winkler similarity function" to overcome this problem.

Distribution of papers over major cities and institutions show interesting results. For the Serbian capital city of Belgrade, only 65.4% of all papers have affiliation of the state University of Belgrade, the biggest and oldest Serbian university, ranked between positions 300 and 400 on the ARWU list. For other university centres in Serbia, the share of publications of state universities is: 93.3% for Novi Sad, 87.4% for Niš and 97.9% for Kragujevac. We conclude that bigger cities have greater potential for scientific productivity outside the university, but this ratio also reflects some problems identified in the past, that institutes belonging to the University of Belgrade did not include the name of the University in affiliation before the initiative to do so, started during the procedure and efforts to qualify for ARWU ranking. The significant growth in the number of papers started in 2008, probably as the result of accreditation procedure regulated by national accreditation body CAQA (www.kapk.org).

Table 1. Journals with more than 20 papers published in the period from 2006 till 2013.

Journal Name	No.	5 years IF
MATCH-communications in mathematical and in computer chemistry	101	1.829
ComSIS - Computer science and information systems	67	0.575
Mathematical and computer modelling	47	2.020
Expert systems with applications	42	1.965
Advances in electrical and computer engineering	28	0.642
Fuzzy sets and systems	27	1.880
International journal of computers communications & control	23	0.694
Information sciences	21	3.893
Journal of multiple-valued logic and soft computing	21	0.667

The list of journals with more than 20 papers in the Table 1 shows that journals in multiple WoS categories are predominant. The journal MATCH publishes the mathematical results and applications in solving chemical problems, without significant content in computing research. The second journal on the list, ComSIS (Computer Science and Information Systems), is an international journal published in Serbia, dedicated to computing, that appeared for the first time on the SCIe list in 2010. In fractional counting, it has been shown that some other disciplines are represented in the comparable quantity to the basic computer science disciplines: Engineering, Electrical & Electronic (71.65), Mathematics, Applied (62.75) and Chemistry, Multidisciplinary (41.67) are in-between Computer Science, Theory & Methods (76.98) and Computer Science, Hardware and Architecture (22.83). Since the leading category is Computer Science, Interdisciplinary Applications (153.00), it is obvious that computer science in Serbia can be viewed predominantly as applied science, blended with electrical engineering, applied mathematics and multidisciplinary chemistry. The leading scientists are I. Gutman with 74 papers in total and 26 in fractional counting, and M. Ivanovic with 23 papers in total and 6.33 in fractional counting.

Conclusions

Considerable growth of publications from Serbia since 2006 was identified in Ivanovic (2014). Serbian national system that transfers data from WoS on weekly bases kobson.nb.rs shows that there were 1746 publications of Serbian authors during 2006 and the yearly production tripled in 2013. At the same time, the number of all publications in Computer Science categories in WoS core collection increased from 123 to 286, while articles only increased from 60 to 204, which was about 3.9% of total Serbian production and 0.47% of the world production in aforementioned categories in the year 2013. The ratio of total world production and total Serbian production is 0.39%, so the results of computer science disciplines are better than average, mostly due to the interdisciplinary approach.

Acknowledgments

We are grateful to Ms. Biljana Kosanović and Ms. Darija Dašić for their assistance in data retrieval, and for valuable advice.

References

Ivanovic, D., Ho, J.-S., (2014). Independent publications from Serbia in the Science Citation Index Expanded: a bibliometric analysis. *Scientometrics*, *101*(1), 603-622.

Winkler, W.E. (1995). Matching and record linkage. In Cox, B. (Ed.), *Business Survey Methods*, Wiley, London, pp. 355-84.

Mitrovic, I., Protic, J., (2014) Problems with affiliations, names and personal identity in the process of evaluating higher education institutions, *EDULEARN14 Proceedings,* Barcelona, 2524-2533.

SCIENCE POLICY AND RESEARCH ASSESSMENT

UNIVERSITY POLICY AND INSTITUTIONAL RANKINGS

SCIENTIFIC FRAUD AND DISHONESTY

A Computer System for Automatic Evaluation of Researchers' Performance

Ashkan Ebadi[1] and Andrea Schiffauerova[2]

[1] *a_ebad@encs.concordia.ca,* [2] *andrea@ciise.concordia.ca*
Concordia Institute for Information Systems Engineering (CIISE), Concordia University,
1515 Ste-Catherine Street West, Montreal, Quebec H3G 2W1 (Canada)

Abstract

The increasing number of researchers and the limited financial resources has caused a tight competition among scientists to secure research funding. On the other side, it has become even harder for funding allocation organizations to evaluate the performance of researchers and select the best candidates. However, it seems that the current evaluation methods are highly correlated with subjective criteria. In addition, the subjective nature of peer-review as one the most common methods in scientific evaluation calls itself for an accurate complementary quantitative method to help the decision makers. This paper proposes an automatic computer system, which is based on machine learning techniques for predicting the performance of researchers. The proposed system uses various features of different types as the input to a complex machine learning module to predict the performance of a researcher in a given year. The method provides the decision makers with fair comparative results regardless of any subjective criteria. Our results show the high accuracy of the proposed system in predicting the performance of researchers.

Conference Topic

Methods and techniques, Science policy and research assessment

Introduction

Research grants is known as one of the crucial drivers of scientific activities that can influence the size and efficiency of R&D sector and its productivity (Jacob & Lefgren, 2011). It can also affect the performance of researchers through providing them with a better access to the research resources (Lee & Bozeman, 2005). In the meantime, policies on R&D activities have evolved over the past fifty years (Elzinga & Jamison, 1995; Sanz-Menendez & Borras, 2000). Funding agencies put a lot of efforts on selecting the best candidates for allocating grants as well as on evaluating the performance of researchers in regards to the amount of funding that they have been receiving. On the other hand, the growing number of researchers worldwide has made the competition for securing the limited financial resources even harder. For example, according to Polster (2007) the contest for receiving research funding is on the rise in Canada especially among the academic researchers mainly due to the changes in federal funding policies, lack of university operating budgets, and increasing research costs. The researchers' demand for funding cannot be fully satisfied by the finite financial capacity of the funding agencies. However, the case could be even worse for the young researchers since the senior researchers are more known within their scientific community that might help them in getting money for research.

Peer review is the oldest measure that has been being used for evaluating researchers' performance and their proposals. Most of the funding agencies use a committee of independent researchers to review the researchers' proposals for funding and select the most appropriate researcher(s) through a competitive process. However, the peer review process has been widely criticized in the literature due to the potential biases since the accuracy of the procedure is highly dependent on the selected experts. For example, preferences of peers can affect the final decision or it can act as a gatekeeper for new research interests since peers may not come into an integrated conclusion (King, 1987). Despite the aforesaid drawbacks, the great advantage of peer review process is that the impact of the proposed research could

be assessed quite easily and accurately (Allen et al., 2009). For this important reason it has still remained as one of the most popular techniques in scientific evaluation. Though, the current trend is to combine the expert review with quantitative performance indicators (Butler, 2005; Hicks et al., 2004) in order to achieve a more balanced evaluation since it cannot be reliable enough as a single indicator. For this purpose, citation and publication counts based indicators are commonly used as the quantitative indicators of researchers' performance.

One of the reasons that scientists publish their work in the form of scientific papers is that in this way they can secure their priority in discoveries (De Bellis, 2009). According to the review of literature done by Tan (1986), performance evaluation of individual researchers and research departments are in most cases based on publication counts measures (at least partially). For the quality of publications, citation counts based indicators, first introduced by Gross and Gross in 1927, are commonly accepted as a proxy for the impact of a scientific publication (Gingras, 1996). In general, they count the number of citations received by an article after the date it is published; hence, papers with higher number of citations are assumed to have higher impact.

Invention of the Internet and availability of the digital data have made it feasible to extract and collect data in a very large scale. In addition, the rapid advancement in the field of computer science has made new ideas and algorithms available to the data scientists. Therefore, large scale digital data and complex algorithms provide researchers with novel opportunities to explore new directions of the information science as well as scientific evaluation. This paper presents an integrated highly accurate automatic productivity prediction system that can assist decision makers (and peers) to detect the most appropriate researchers for funding allocation. The remainder of the paper proceeds as follows: *Data and Methodology* section describes the data gathering procedure in detail while explaining the methods and methodologies that were used; the *Results* section presents the performance evaluation results and interpretations for the proposed system; the paper concludes in *Discussion* section; and limitations and future research directions are stated in the last section of the paper.

Data and Methodology

Data

We decided to focus on performance of the researchers who have been funded by the Natural Sciences and Engineering Research Council (NSERC)[1] of Canada. The main reasons for choosing NSERC was its role as the main federal funding organization in Canada, and the fact that almost all the Canadian researchers in natural sciences and engineering receive at least a basic research grant from NSERC (Godin, 2003). Therefore, as the first stage information about the funded researchers was collected from NSERC[2]. In the next phase, Elsevier's Scopus[3] was used to gather all the information about the funded researchers. The data spans from information about the authors themselves (*e.g.* Scopus ID, their affiliation, number of publications in a given year, *etc.*) to their articles (*e.g.* year of publication, authors of the paper, keywords, *etc.*).

The time interval of the research was set to the period of 1996 to 2010 since the data coverage of Scopus was better after 1996. Moreover, to have a proxy of the quality of the papers we

[1] For more information, see: http://www.nserc-crsng.gc.ca/index_eng.asp

[2] Students were excluded from the data as the goal of the paper is evaluating the performance of researchers.

[3] Scopus is a commercial database of scientific articles that has been launched by Elsevier in 2004. It is now one of the main competitors of Thomson Reuter's Web of Science.

used SCImago[4] to collect the impact factor information of the journals in which the articles were published. SCImago was chosen for two main reasons. Firstly, it provides annual data of the journal impact factors that enables us to perform a more accurate analysis since we are considering the impact factor of the journal in the year that an article was published not its impact in the current year. Secondly, SCImago is powered by Scopus that makes it more compatible with our publications database.

In the next phase of data preparation, we calculated several bibliometric features such as amount of funding received by a researcher in a given year, his/her career age, average number of co-authors, average number of publications, average number of citations, *etc.* In addition, using Pajek[5] software social network analysis techniques were employed to construct the collaboration networks of the researchers within the examined time interval. The created networks were used to calculate various network structure properties (*e.g.* betweenness centrality, eigenvector centrality, and clustering coefficient) of the researchers at the individual level. All the calculated features were integrated in a MySQL[6] dataset. The final database contains 117,942 records of researchers. In the next section, methodologies are discussed in more detail.

Methodology

Several features of various types and from different sources were selected for this study. Funding is acknowledged in the literature as one of the main drivers of scientific activities where a three-year (*e.g.* Payne & Siow, 2003) or a five-year (*e.g.* Jacob & Lefgren, 2007) time window is mostly considered for the funding to take effect. In this paper a three-year time window was considered for all the bibliometric variables, *e.g.* for assessing the productivity of a given researcher in year 1999 his/her amount of funding was summed up for the period of 1996 to 1998 (*sumFund3*). Intuitively, productive researchers are expected to at least maintain their performance level. Various past productivity features were hence included in the model reflecting the quality and quantity of the publications. As a proxy for the rate of publications, number of publications in a three-year time window (*noArt3*) was considered. Two indicators were used as proxies for the quality of publications, *i.e.* average number of citations in a three year time window (*avgCit3*) and the average impact factor of the journals in which the articles were published in a three year time interval (*avgIf3*). Both of the mentioned features can serve as a proxy for quality, but with a slightly different meaning. Impact factor indicates the respectability of the journal, *i.e.* the quality and the level of contribution perceived by the authors and the reviewers of the paper, whereas citation counts show the impact of the article on the scientific community and on the subsequent research.

A multi-level feature representing the scientific field of the researcher (*discip*) was also used in the model since publication and citation habits can be different in various scientific fields. For example, citing habits and the rate of citations may vary across different scientific fields in a way that in some scientific fields authors publish articles more frequently or the published papers contain more references (MacRoberts & MacRoberts, 1996; Phelan, 1999). It is argued in the literature that older researchers in general can be more productive (Merton, 1973; Kyvik & Olsen, 2008) due to several reasons (*e.g.* better access to the funding and expertise sources, more established collaboration network, better access to modern equipments). Hence, the career age of the researcher (*careerAge*) was included in the model representing the time difference between the date of his/her first article in the database and the given year. As a common indicator of the scientific collaboration, the average number of co-authors per paper was also included in the prediction model (*teamSize*). It is expected that

[4] For more information, see: http://www.scimagojr.com

[5] Social network analysis software, for more information see: http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[6] Open source relational database management system, for more information see: http://www.mysql.com/

researchers who have on average higher number of co-authors have more connections that might result in relatively higher number of projects or future publications, hence this feature was also considered as one of the influencing factors.

As discussed in the previous section, social network analysis was used to construct the collaboration networks and to measure the structural network properties of researchers. In particular, four network structure indicators were calculated namely betweenness centrality (*bc*), clustering coefficient (*cc*), eigenvector centrality (*ec*), and degree centrality (*dc*). Betweenness Centrality (*bc*) is an indicator of the important players (researchers) in a network who have a control over the flow of knowledge and resources. These players, who are also called as *gatekeepers*, are able to bridge different communities. Theoretically, betweenness centrality of the node *k* is measured based on the share of times that a node *i* reaches a node *j* via the shortest path passing from node *k* (Borgatti, 2005) and is calculated as follows (σ_{ij} is the total number of shortest paths from node *i* to j and $\sigma_{ij}(k)$ is the number of shortest paths from node *i* to node *j* that contains node *k*):

$$bc_k = \sum_{i \neq k \neq j} \frac{\sigma_{ij}(k)}{\sigma_{ij}} \qquad (1)$$

Clustering Coefficient (*cc*), also called *cliquishness*, indicates the tendency of researchers to cluster with other researchers in the network. Hence, researchers with high clustering coefficient may have a relatively high number of connections with the other team members who are collaborating in a tightly knit group. Therefore, this indicator was selected to represent the tight collaboration impact on the overall performance of the team. Theoretically, clustering coefficient of node *i* (cc_i) is defined based on the number of triangles (inter-connected sub-network of three nodes) that contains the node *i* (t_i) normalized by the maximum number of triangles in the given network (Watts & Strogatz, 1998). Let n_i denotes number of neighbors of the node *i*, hence:

$$cc_i = \frac{2t_i}{n_i(n_i - 1)} \qquad (2)$$

Degree Centrality (*dc*) that was also considered as one of the network variables is defined based on the number of ties that a node has (degree) in an undirected graph. Hence, researchers with high degree centrality should be more active since they have higher number of ties (links) to other researchers (Wasserman, 1994). Moreover, in co-authorship networks it can be regarded as the number of direct partners or team members of a given researcher. Hence, it is expected to have an influence on the scientific activities. Degree centrality for node *i* (dc_i) is thus defined based on the node's degree (deg_i) and then the values are normalized between 0 and 1 (dividing by the highest degree in the network) to be able to compare the centralities:

$$dc_i = \frac{deg_i}{deg_{high}} \qquad (3)$$

Eigenvector Centrality (*ec*) takes the importance of a node and its connections into the account. Hence, a researcher has high eigenvector centrality if he/she is connected with other important actors who are themselves occupying central positions in the network. These researchers can be identified as *leaders* in the scientific networks since they are connected

with too many other influential and highly central researchers, and it is hence expected that they shape the collaborations and play an important role in setting priorities in scientific projects that might affect the performance of researchers. A complete list of the selected features is shown in Table 1.

Table 1. List of attributes for the prediction models.[7]

No	*Attribute*
1	Scientific area in which the researcher is working (*discip*)
2	Total amount of funding received by each researcher in a 3 year time window (*sumFund3*)
3	Total number of publications of each researcher in a 3 year time window (*noArt3*)
4	Average number of citations received by researcher's articles in a 3 year time window (*avgCit3*)
5	Average impact factor of the journals in which researcher's articles were published in a 3 year time window (*avgIf3*)
6	Average betweenness centrality of each researcher in a 3 year time window (*btwn3*)
7	Average degree centrality for each researcher in a 3 year time window (*deg3*)
8	Average clustering coefficient of each researcher in a 3 year time window (*clust3*)
9	Average eigenvector centrality of each researcher in a 3 year time window (*eigen3*)
10	Average number of authors per paper for each researcher (*teamSize*)
11	Career age of the researcher (*careerAge*)

The mentioned features were used as an input to the prediction model. Figure 1 shows the whole process of the researchers' performance prediction. Number of publications was considered as the target variable for the performance prediction task. As it can be seen, data is first preprocessed and cleaned. For this purpose, several JAVA programs were coded to check the data for redundancy, out of range values, impossible combinations, errors, and missing values and then data was filtered based on the records that contained all the required data. The resulted data containing all the mentioned features was fed into the data preparation block where at first all the features were normalized to a value between 0 and 1. This was a crucial step since the features were of different units and scales. Local Outlier Factor (LOF) algorithm was then implemented to detect the outliers. LOF that was proposed by Breunig et al. (2000) is based on the local density concept in which the local deviation of a given data is measured with respect to its k nearest neighbors. A given data is outlier if it has a substantial different density from its k neighbors. The final step of the data preparation step was optimizing the attributes' weights. For this purpose we used an evolutionary attributes weights optimizer that employed genetic algorithm to calculate the weights of the attributes. The weighting procedure improved the accuracy of the system by giving more value to the most influential attributes. The resulted data was integrated into a single data repository named as the target data.

[7] The initial list of the selected features was prepared as a result of an intensive statistical analyses performed on the target data. The list was then refined and weighted within the proposed system.

After making the data ready for the analysis, a stratified 10-fold cross validation design was used for the model validation. Cross validation is an analytics tool that is used to design and develop fine tune models. In other words, the data is split into two disjoint sets where one part is used for training and fitting a model (training set) while the other part is employed for estimating the error of the model (test set) (Weiss & Kulikowski, 1991). We used a nested 10-fold cross validation in which the data is split into 10 disjoint subsets in a way that union of the 10 folds results the original data. The method runs 10 times and in each time one fold is considered as the test data while the rest are regarded as the training data.

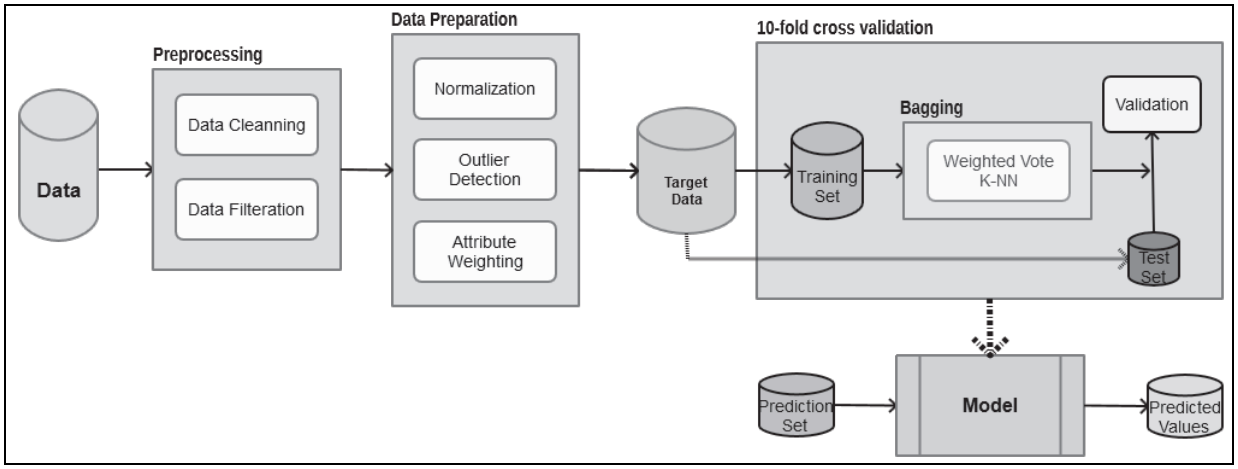


Figure 1. Proposed model for automatic evaluation of researchers' performance.

As mentioned earlier, number of publications was considered as the target variable. To further improve the accuracy of the prediction the ensemble meta-algorithm was employed. For this purpose, bootstrap aggregating (bagging) approach was used. Bagging is an ensemble method that makes random subsets of the data and trains them separately where the final result is obtained by averaging over the results of the separated models (Breiman, 1996). Bagging is a nested module in which we used weighted vote 10-Nearest Neighbor (10-NN) algorithm to train the data and to create the model. In weighted vote 10-NN the distance of the neighbors to the given data is considered as a weight in the prediction in a way that neighbors that are closer to the given data get higher weights. This particularly helped to increase the accuracy of the prediction. Data in the range of 1996 to 2009 was used to train and build the model while a separate disjoint data for 2010 (prediction set) was used for testing the accuracy of the prediction model. The final output of the proposed automatic computer system was the predicted number of publications for the researchers in the prediction set.

Results

In this section the results of the performance evaluation of the proposed automatic computer system (PACS) is presented. As discussed earlier, the model was trained on the data from 1996 to 2009 and a disjoint dataset for 2010 was used for the prediction and the accuracy tests. The accuracy of the proposed model was compared with several well-known machine learning algorithms, however, in this paper the results are presented and compared for the PACS model as well as two other algorithms that showed the highest accuracy in predicting the target variable.

Figure 2 shows the prediction errors of PACS, linear regression, and polynomial regression of degree three[8]. We considered three error measures for comparing the performance of the

[8] Other algorithms (*e.g.* decision trees) were also tested but these listed algorithms were the top two ones with the highest accuracy.

mentioned algorithms. Root mean squared error is one of the main measures for comparing the accuracy of the prediction models and is defined as the square root of the average of the squares of errors. According to Figure 2, PACS is predicating the number of publications of researchers with 1.451 average deviation between the predicted value and the real number of publications. Normalized absolute error is the absolute error (difference between the predicted value and the real value) divided by the error made if the average would have been predicted. The root relative squared error takes the average of the actual values as a simple predictor to calculate the total squared error. The result is then normalized by dividing it by the total squared error of the simple predictor and square root is taken to transform it to the same dimension as the predicted value. As it can be seen PACS is performing better in all the three measures where the degree 3 polynomial fit is the worst.

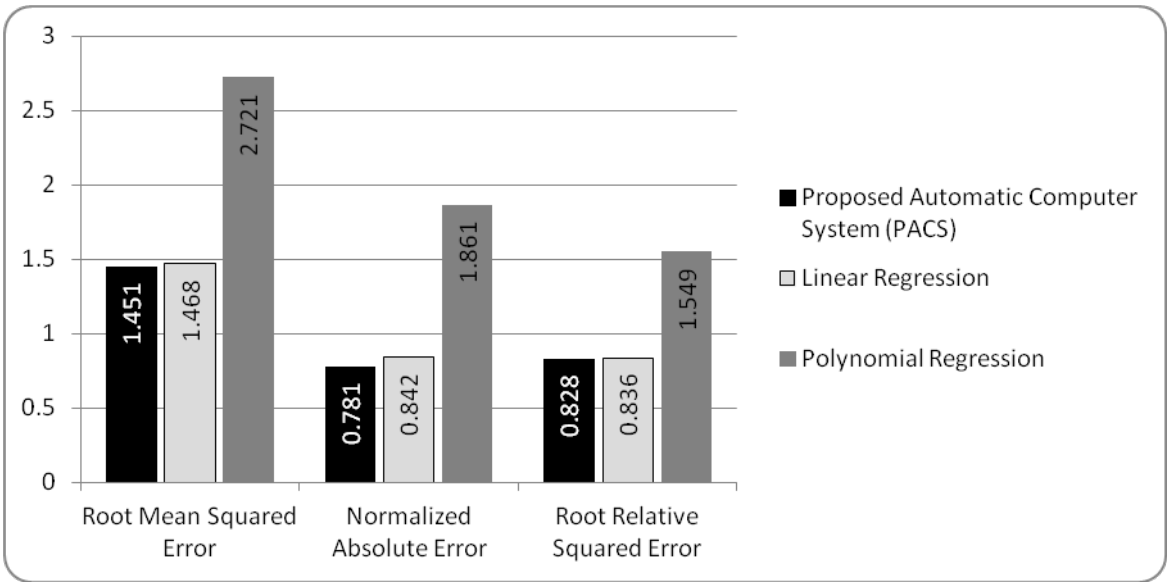


Figure 2. Accuracy test, PACS vs. other two top performing algorithms.

Table 2. Prediction results.

No	*Predicted no of articles*	*noArt*	*sum Fund3*	*avg If3*	*avg Cit3*	*teamSize*	*btwn3*	*clust3*	*deg3*	*eigen3*	*careerAge*	*discip*	*noArt3*
1	0.361	0	0.041	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.737	2	0
2	1.102	0	0.013	0.279	0.028	0.000	0.000	1.000	0.005	0.000	0.632	3	1
3	3.865	7	0.044	0.054	0.005	0.001	0.059	0.125	0.027	0.000	0.737	1	13
4	1.103	0	0.010	0.068	0.083	0.000	0.000	1.000	0.007	0.000	0.737	3	1
5	1.206	1	0.072	0.132	0.020	0.002	0.016	0.409	0.020	0.000	0.526	0	6
6	6.703	4	0.167	0.246	0.080	0.002	0.055	0.158	0.039	0.000	0.737	1	26
7	1.030	4	0.032	0.115	0.017	0.001	0.018	0.455	0.018	0.000	0.737	0	6
8	4.120	3	0.061	0.136	0.041	0.002	0.185	0.109	0.134	0.000	0.737	1	15
9	0.000	0	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.263	0	0
10	5.047	3	0.137	0.141	0.041	0.001	0.133	0.163	0.050	0.000	0.684	0	15
11	1.128	1	0.010	0.091	0.062	0.003	0.003	0.333	0.007	0.000	0.526	1	1
12	1.964	1	0.010	0.113	0.009	0.004	0.053	0.192	0.022	0.018	0.737	1	7
13	12.228	7	0.095	0.399	0.028	0.010	0.197	0.042	0.075	0.000	0.684	0	31
14	2.112	2	0.190	0.228	0.091	0.001	0.011	0.182	0.020	0.000	0.737	1	6
15	2.233	3	0.299	0.230	0.051	0.002	0.013	0.457	0.035	0.000	0.737	0	7
16	3.577	4	0.198	0.259	0.055	0.002	0.042	0.145	0.059	0.000	0.579	4	12
17	11.308	9	0.329	0.309	0.116	0.002	1.000	0.062	0.148	0.000	0.737	1	40
18	4.841	4	0.093	0.458	0.051	0.001	0.027	0.117	0.037	0.000	0.737	0	19
19	5.752	4	0.116	0.253	0.055	0.123	0.003	0.823	0.940	1.000	0.737	1	20
20	7.421	8	0.193	0.270	0.077	0.002	0.153	0.079	0.082	0.000	0.737	1	26

A randomly selected sample of the predictions is presented in Table 2. Each row represents a distinct researcher's profile in 2010 for whom several indicators have been calculated and used in the PACS model as the input features. The real number of articles is shown in *noArt* column that was not fed into the prediction model. Based on the other attributes the proposed system automatically predicted the number of publications of a researcher in 2010, *i.e.* column named *Predicted no of articles* in Table 2 and is highlighted in dark grey. As it can be seen using several features of different types and employing various techniques for data gathering (*e.g.* bibliometrics, social network analysis) and preparation provides the system with highly accurate high-dimensional input data that led to a low error rate and good predictions. Interestingly, it seems that the system successfully considered the differences between various scientific fields in performing scientific activities. According to the results, although the profile of the researchers numbered 1 and 9 in Table 2 are relatively similar, the predicted performance differs as they do not belong to the same scientific field. Hence, the results confirm the importance of the scientific disciplines in predicting the performance of researchers. In addition, comparison of the researchers numbered 6 and 7 highlights the importance of the past productivity as well as the quality of publications in predicting the number of publications.

Discussion

In this paper we used various bibliometric as well as network structural property features to build a model to predict the performance of researchers. Machine learning techniques and availability of the digital data has made it possible to use complex algorithms on high dimensional large scale data. This provides scientometrists with an opportunity to go beyond the current border of using common indicators or simple statistical analyses. Although some researchers recently worked on citation prediction using machine learning algorithms (*e.g.* Fu & Aliferis, 2010; Lokker et al., 2008) to our knowledge this is the first study that focused on the prediction of researchers' productivity using input features of different types and at the individual level of the researchers.

The attribute weighting method to rank features based on their importance that was implemented in the proposed model as well as the outlier detection module for data filtration increased the accuracy of the predictions significantly. Results of the attribute weighting module can also shed light on the most influential attributes in predicting the scientific activities of the target researchers. Another unique approach that was employed in designing the proposed system was using several features of similar nature in building the model that reinforced the prediction power of the system. For example, average number of citations and average impact factor of the journals were used to represent the quality of the paper. Another example is the degree centrality and scientific team size as the former represents the number of direct connections of a researcher while the latter indicates the average number of his/her co-authors. These attributes of similar nature surely empowered the accuracy of the model by providing it with more dimension and flexibility.

To conclude, as it was observed complex computer algorithms can be used to design automatic evaluation systems and prediction tools to evaluate different aspects of scientific activities of researchers. It is obvious that peer reviewing cannot be completely replaced by such tools. However, such systems can help decision makers in setting both long-run and short-term strategies in regard to the funding allocation and/or analyzing researchers' productivity. In addition, the availability of high-dimensional large scale data (in our case, a large dataset spanning from 1996 to 2010) that is intensively cleaned and preprocessed for learning the model will surely contribute to highly accurate predictions that are not based on a limited criteria or a limited feature set. Therefore, this can also help to establish a fairer funding allocation or scientific evaluation system.

Limitations and Future Work

We were exposed to some limitations in this paper. Firstly, Scopus was selected for gathering information about the funded researchers' articles. Since Scopus and other similar databases are English biased, hence, non-English articles are underrepresented (Okubo, 1997). Secondly, due to the better coverage of Scopus before 1996, the time interval of 1996 to 2010 was selected for the analysis. Although Scopus is confirmed in the literature to have a good coverage of articles, as a future work it would be recommended to focus on other similar databases to compare the results.

Furthermore, we were exposed to some limitations in measuring scientific collaboration among the researchers where we used the network structure properties. In particular, we were unable to capture other links that might exist among the researchers like informal relationships since these types of connections are never recorded and thus cannot be quantified. In addition, there are also some drawbacks in using co-authorship as an indicator of scientific collaboration since collaboration does not necessarily result in a joint article (Tijssen, 2004). An example could be the case when two scientists cooperate together on a research project and then decide to publish their results separately (Katz & Martin, 1997). For assessing the quality of the papers based on citation counts we did not account for self citations, negative citations, or special inter-citation patterns among a number of researchers. Although we also used another proxy (average impact factor of journals) to overcome this limitation, it can be addressed in the future works.

References

Allen, L., Jones, C., Dolby, K., Lynn, D., & Walport, M. (2009). Looking for landmarks: The role of expert review and bibliometric analysis in evaluating scientific publication outputs. *PLoS One*, *4*(6), e5910.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55-71.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM Sigmod Record*, *29*(2), pp. 93-104.

Butler, L. (2005). What happens when funding is linked to publication counts? Handbook of quantitative science and technology research (pp. 389-405), Springer.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Scarecrow Press.

Elzinga, A. & Jamison, A. (1995). *Changing policy agendas in science and technology. Handbook of Science and Technology Studies* Ed.by Sheila Jasanoff et al. London: Sage.

Fu, L. D. & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics*, *85*(1), 257-270.

Gingras, Y. (1996). Bibliometric analysis of funded research. A feasibility study. *Report to the Program Evaluation Committee of NSERC.*

Godin, B. (2003). The impact of research grants on the productivity and quality of scientific research. No. 2003. *INRS Working Paper.*

Gross, P., & Gross, E. (1927). College libraries and chemical education *Science, 66*(1713), 385-389.

Hicks, D., Tomizawa, H., Saitoh, Y., & Kobayashi, S. (2004). Bibliometric techniques in the evaluation of federally funded research in the United States. *Research Evaluation*, *13*(2), 76-86.

Jacob, B. A. & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics*, *95*(9), 1168-1177.

Katz, J. S. & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1-18.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science*, *13*(5), 261-276.

Kyvik, S. & Olsen, T. B. (2008). Does the aging of tenured academic staff affect the research performance of universities? *Scientometrics*, *76*(3), 439-455.

Lee, S. & Bozeman, B. (2005). The impact of research collaboration on scientific productivity. *Social Studies of Science*, *35*(5), 673-702.

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *BMJ* (Clinical Research Ed.), *336*(7645), 655-657.

MacRoberts, M. H. & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435-444.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. University of Chicago press.

Okubo, Y. (1997). Bibliometric indicators and analysis of research systems: Methods and examples. No. 1997/1. *OECD Publishing*.

Payne, A. A. & Siow, A. (2003). Does federal research funding increase university research output? *Advances in Economic Analysis & Policy*, *3*(1).

Phelan, T. (1999). A compendium of issues for citation analysis. *Scientometrics*, *45*(1), 117-136.

Polster, C. (2007). The nature and implications of the growing importance of research grants to Canadian universities and academics. *Higher Education*, *53*(5), 599-622.

Sanz Menéndez, L., & Borrás, S. (2000). Explaining changes and continuity in EU technology policy: The politics of ideas.

Tan, D. L. (1986). The assessment of quality in higher education: A critical review of the literature and research. *Research in Higher Education*, *24*(3), 223-265.

Tijssen, R. J. (2004). Is the commercialisation of scientific research affecting the production of public knowledge? Global trends in the output of corporate research articles. *Research Policy*, *33*(5), 709-733.

Wasserman, S. (1994). Social network analysis: Methods and applications. *Cambridge university press*.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440-442.

Weiss, S., & Kulikowski, C. (1991). *Computer Systems that Learn*: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems. San Francisco: Morgan Kaufmann.

Grading Countries/Territories Using DEA Frontiers

Guo-liang Yang[1], Per Ahlgren[2], Li-ying Yang[3], Ronald Rousseau[4], Jie-lan Ding[3]

[1]*glyang@casipm.ac.cn*
Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190 (China)

[2]*perahl@kth.se*
School of Education and Communication in Engineering Sciences (ECE), KTH Royal Institute of Technology, 100 44 Stockholm (Sweden)

[3]*yangly@mail.las.ac.cn, dingjielan@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing 100190 (China)

[4]*Ronald.Rousseau@kuleuven.be*
Institute for Education and Information Sciences, IBW, University of Antwerp (UA), Antwerp B-2000 (Belgium)
KU Leuven, Department of Mathematics, Leuven B-3000 (Belgium)

Abstract

Several approaches exist related to categorizing academic journals/institutions/countries into different levels. Most existing grading methods use either a weighted sum of quantitative indicators (including the case of one properly defined quantitative indicator) or quantified peer review results. An important issue of concern for science and technology management is the efficiency of resource utilization. In this paper we deal with this issue and use multi-level frontiers of data envelopment analysis (DEA) models to grade countries/territories. Research funding and numbers of researchers as used as inputs, while papers and citations are output variables. The research results show that using DEA frontiers we can grade countries/territories on six levels. These levels reflect the corresponding countries' level of efficiency in S&T resource utilization. Furthermore, we use papers and citations as single outputs (with research funding and researchers as inputs) to show changes in country/territory level.

Conference Topic

Science Policy and Research Assessment

Introduction

The efficiency of science and technology (S&T) resource utilization is one of the important issues for S&T management (Yang et al., 2013a; Yang et al., 2014a). Johnes and Johnes (1992) evaluated the efficiency of S&T organizations using data envelopment analysis (DEA) as a performance analysis tool. Rousseau and Rousseau (1997, 1998) assessed the efficiency of countries using gross domestic product, active population and research and development (R&D) expenditure as inputs, and publications and patents as outputs. They showed that DEA can be used in scientometrics as a tool to measure the efficiency of decision making units (DMUs, e.g., countries) by gauging closeness to the efficiency frontier. Similar techniques have been used by other researchers (Kao & Lin, 1999; Roy & Nagpaul, 2001; Shim & Kantor, 1998). Yang and Chang (2009) used DEA under constant and variable returns to scale (RTS) to measure firms' efficiency. Worthington (2001) conducted an empirical survey of frontier efficiency measurement techniques in education. Other researchers have analyzed the efficiency or productivity in the education sector, (e.g., Abbott & Doucouliagos, 2003, Avkiran, 2001, Carrington et al., 2005, Worthington & Lee, 2008, Flegg et al., 2004, Johnes & Johnes, 1995, Johnes, 2006a,b, Kempkes & Pohl, 2010, Wolszczak-Derlacz & Parteka, 2011, and Aristovnik, 2012). When studying the standard university model, Brandt and Schubert (2013) observed that universities are large agglomerations of many (often loosely

affiliated) small research groups. They explained this observation by typical features of the scientific production process. In particular, they argued that there are decreasing RTS on the level of the individual research groups. RTS is a concept with strong relation to scale efficiency. Somewhat similar observations (decreasing RTS) were published earlier by Bonaccorsi and Daraio (2005). Schubert (2014) used non-parametric techniques of multidimensional efficiency measurement, such as DEA, to analyse the RTS in scientific production based on survey data for German research groups from three scientific fields. Based on DEA models, Yang et al. (2013a, 2014a) analyzed the directional RTS of a couple of biological institutes in the Chinese Academy of Sciences (CAS).

Some fairly recent studies have examined the efficiency of countries or regions in utilizing R&D expenditures or other resources. Lee and Park (2005) evaluated R&D efficiency across nations using patents, technology balance of receipts and journal articles as outputs. Wang and Huang (2007) analyzed R&D efficiency of nations by considering patents and papers as outputs. Lee et al. (2009) used DEA to measure and compare the performance of national R&D programs in South Korea. Sharma and Thomas (2008) investigated the R&D efficiency of developing countries in relation to developed countries, taking into account time lags. Other, and similar, studies include Chen et al. (2011), Sueyoshi and Goto (2013), and Zhong et al. (2011).

The literature referred to hitherto focuses on the quantitative measurement of efficiency of resource utilization. In this context, DEA is one of the most popular mathematical tools for estimating the relative efficiency of DMUs. However, Banker (1993) pointed out that DEA efficiency scores usually overestimate efficiency and are biased. Smith (1997) argued that the extent of the overestimation is highly dependent on sample size and the complexity of the production process (as indicated by the numbers of inputs and outputs). However, in many cases we only need to know the general level (grade) of DMUs in terms of efficiency instead of their exact scores or complete ranking.

Several efforts have been made regarding categorization of academic journals, institutions and countries into different levels of standing or quality. Since 2007, the Association of Business School (ABS) has issued the Academic Journal Quality Guide, which classifies journals in business and management into four categories (grade 1 to 4) recognizing the quality of those journals based on a survey of hundreds of experts in the field (Harvey et al., 2007a,b; 2008). From 2010, a new category, termed 4*, was added to the four existing categories to recognize the quality of the top journals (Harvey et al., 2010). Bandyopadhyay (2013) categorized business and management journals into four categories (Excellent, Very Good, Standard, Satisfactory) based on multiple inputs, including Thomson Reuters' Social Science Citation lists of ranked journals and WoS impact factor analyses. In 2005, CAS evaluated its dependent institutes and classified them into three grades (Excellent, Good, and Satisfactory) (CAS, 2006). Glänzel (2011) used characteristic scores and scales as parameter-free tools to identify top journals. Yang et al. (2013b) analyzed the overall development and the balance of the disciplinary structure of China's science based on papers covered by Science Citation Index and with the use of bibliometric methods. These authors further categorized selected countries to reflect their developmental status.

The grading methods in the research reported above use either a weighted sum of quantitative indicators (including the case of one properly defined quantitative indicator) or quantified peer review results. In general, the weighted sum approach normally needs indicator weights and corresponding threshold values as a priori information, while the peer review process usually costs a lot of time and expenditures (Smith, 1996). In the light of these downsides, this paper presents an alternative approach, involving multiple DEA frontiers, to divide various countries/territories into different levels with respect to the efficiency of their S&T resource utilization.

The rest of the paper is organized as follows. The next section introduces the input and output indicators, and the corresponding dataset used in the analysis. The used methods are described in the third section, in which we treat multi-level efficient frontiers and show how to divide the countries/territories into grades using these frontiers. In the fourth section, the results of the study are given, whereas conclusions appear in the final section.

Indicators and data

In this work, research funding and researchers are used as input indicators. Research funding here means Gross Domestic Expenditure on R&D (million current PPP$). The total number of researchers (full time equivalents, FTEs) in one country is used as indicator for researchers. For the output indicators, we used the number of papers covered by the Science Citation Index (SCI) and Social Science Citation Index (SSCI) from the Web of Science (WoS), and the number of citations to these papers in the year 2011. We use OECD statistics and Thomson Reuters' research evaluation tool InCites as sources for input and output data, respectively. All 34 OECD member countries and seven non-OECD member countries/territories were selected for the study. The other non-OECD member countries, covered by OECD statistics, were excluded due to lack of input data. This also holds for the two OECD members Australia and Switzerland (the Gross Domestic Expenditure in 2011 on R&D of these two countries is missing), and thereby the number of OECD member countries included in the study is 32. See Table 1 for details.

Methods

DEA models and their frontiers

DEA is an approach based on linear programming for analyzing performance of organizations and operational processes. This approach was first proposed by Charnes et al. (1978). All DEA models use input and output data to evaluate the relative efficiency of DMUs without prior knowledge of input/output functions and the weights for indicators. Nowadays, numerous theoretical and empirical works on this method have been published, extending the original approach in different ways, and applying them to many areas, including the private and the public sector (e.g., Cooper et al., 2007).

Let $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_s)$ be input and output vectors of n DMUs, respectively of m and s dimensions. Then the Production Possibility Set (*PPS*) is defined by

$$PPS = \{(X, Y): X \; can \; produce \; Y\} \quad (1)$$

There can be different forms of PPS based on different assumptions. Banker (1984) defined the PPS under the assumption of variable RTS to obtain the BCC-DEA model:

$$PPS(X,Y) = \left\{(X,Y) | X \geq \sum_{j=1}^{n} \lambda_j X_j \, , Y \leq \sum_{j=1}^{n} \lambda_j Y_j \, , \sum_{j=1}^{n} \lambda_j = 1 \, , \lambda_j \geq 0, j = 1, \dots, n\right\} \quad (2)$$

where λ_j is a coefficient.

The *PPS* implied in the CCR-DEA model, which was proposed by Charnes et al. (1978) under the assumption of constant RTS, is defined as follows:

$$PPS(X,Y) = \left\{(X,Y) | X \geq \sum_{j=1}^{n} \lambda_j X_j \, , Y \leq \sum_{j=1}^{n} \lambda_j Y_j \, , \lambda_j \geq 0, j = 1, \dots, n\right\} \quad (3)$$

The boundary of the *PPS* is referred to as the production technology or production frontier.

Table 1. Values of input and output indicators across 39 countries/territories.

No.	*Countries/Territories*	*Output*		*Input*	
		Papers	*Citations*	*Research Funding (PPP)*	*Researcher (FTE)*
1	Argentina	8136	40201	4592.313295	50340
2	Austria	12843	100412	9971.246479	37113.8
3	Belgium	18876	152731	9739.425206	42685.77
4	Canada	59025	427079	24756.76203	157360
5	Chile	5795	31737	1172.833167	6082.9
6	China	162794	846720	247808.3033	1318086
7	Czech Republic	9866	55662	4659.446488	30681.59
8	Denmark	13608	124330	6934.707773	37944.1
9	Estonia	1509	10731	733.5776566	4511
10	Finland	10761	82802	7897.729287	40002.61
11	France	67407	480151	53310.69922	249086.3
12	Germany	95935	738284	96971.46462	338608
13	Greece	10819	62818	2006.921474	24674.25
14	Hungary	5934	36137	2721.690282	23019
15	Iceland	815	9013	317.6389104	2258.3
16	Ireland	7438	57682	3169.659323	15172
17	Israel	12478	88753	9306.312467	49797
18	Italy	55338	385416	25780.80141	106151.3
19	Japan	77453	429710	148389.2294	656651
20	Luxembourg	678	4480	660.3865084	3031
21	Mexico	10490	46668	8058.470588	46124.96
22	Netherlands	33845	302477	14597.91748	58447.26
23	New Zealand	8181	50974	1766.588573	16300
24	Norway	10825	78889	5064.393225	27228
25	Poland	21057	91097	6409.165974	64132.8
26	Portugal	10789	66489	4152.692178	50061.2
27	Romania	6927	24373	1725.931612	16080
28	Russia	29072	85915	35192.07719	447579
29	Singapore	9950	82648	6922.39777	33718.5
30	Slovakia	3083	13861	921.2876157	15325.9
31	Slovenia	3776	17682	1429.743722	8774
32	South Africa	9477	48450	4652.174133	20115.06
33	South Korea	45588	222201	58379.65416	288901
34	Spain	50677	332172	20106.98571	130234.9
35	Sweden	21568	172220	13366.28061	48589
36	Taiwan	27283	129286	26184.28683	134047.7
37	Turkey	23920	72981	11301.84442	72108.6
38	UK	100895	784071	39217.4483	251357.6
39	USA	364548	2774572	429143	1252948

Data sources: Input: OECD Statistics. http://www.oecd-ilibrary.org/statistics; Output: InCites. http://incites.isiknowledge.com/Home.action.

Definition 1: The efficient frontier of *PPS* is defined as follows:

$$EF = \{(X,Y) \in PPS | there\ is\ no\ (\bar{X},\bar{Y}) \in PPS\ such\ that\ (-\bar{X},\bar{Y}) > (-X,Y)\} \quad (4)$$

Note: This unobservable production frontier is called the true efficient frontier hereinafter. When there is only a single output, the production frontier is known in the economic literature as the production function. DMUs, which are technically efficient, operate on the frontier, while technically inefficient DMUs operate at points in the interior of the *PPS*. Thus it is rational to rank DMUs according to their distance to the true frontier.

The core idea of classic DEA is to identify first the production frontier. DMUs on the frontier are regarded as efficient. DMUs not situated on the frontier are compared with their peers or projections on the frontier to measure their relative efficiency. All DMUs on the frontier are considered to represent the best practices and have the same level of performance.

Let $\{(x_j, y_j)|j = 1, \dots, n\}$ be a group of observed input and output data. Based on such observations, DEA models construct a piecewise linear production frontier, a non-parametric estimate of the unobservable true frontier. Then DEA models measure the efficiency of a DMU via its distance to the estimated frontier. Using radial measurement and input orientation, we have the following input-based CCR-DEA model (Charnes et al., 1978):

$$\theta_c^* = min\,\theta$$
$$s.t.\begin{cases} \sum_{j=1}^{n} x_{ij}\lambda_j \le \theta x_{i0}, i = 1, \dots, m, \\ \sum_{j=1}^{n} y_{rj}\lambda_j \ge y_{r0}, r = 1, \dots, s, \\ \lambda_j \ge 0, j = 1, \dots, n. \end{cases} \quad (5)$$

where $\lambda_j \ge 0$ are the multipliers of inputs and outputs. Here θ_c^* measures the degree of efficiency by radial measurement under the assumption of constant RTS.

If we assume that the production technology satisfies the variable returns to scale assumption, we have the following input-based BCC-DEA model (Banker et al., 1984):

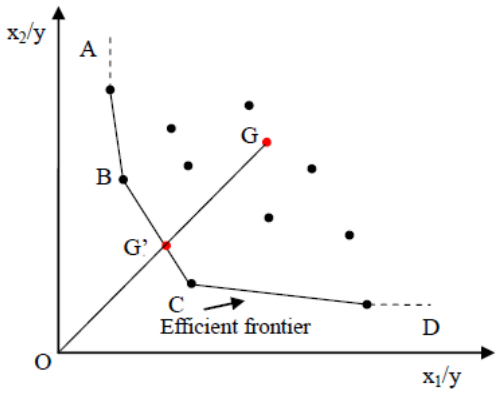
$$\theta_b^* = min\,\theta$$
$$s.t.\begin{cases} \sum_{j=1}^{n} x_{ij}\lambda_j \le \theta x_{i0}, i = 1, \dots, m, \\ \sum_{j=1}^{n} y_{rj}\lambda_j \ge y_{r0}, r = 1, \dots, s, \\ \sum_{j=1}^{n} \lambda_j = 1, \\ \lambda_j \ge 0, j = 1, \dots, n. \end{cases} \quad (6)$$

where θ_b^* measures the degree of efficiency by radial measurement under the assumption of variable returns to scale. It should be noted that Model (6) differs from Model (5) only regarding the constraint $\sum_{j=1}^{n} \lambda_j = 1$, which yields that the variable RTS assumption is satisfied.

Obviously, if $\theta_c^* = 1$ in model (5) or $\theta_b^* = 1$ in Model (6), then the DMU is situated on the efficient frontier in CCR-DEA or BCC-DEA, respectively.

We visualize the frontier of a DEA model in Figure 1, using two inputs (x_1 and x_2) and one output (y). The piecewise linear line ABCD defines the efficient frontier of the existing observations. For example, for point G, representing a DMU, its efficiency score can be calculated as the ratio of distance OG' to distance OG.

We now give an example to illustrate the detection of the efficient frontier and the evaluation of DMUs using a DEA model. We suppose there are six DMUs with two inputs and a single output. In Table 2, hypothetical data is given.



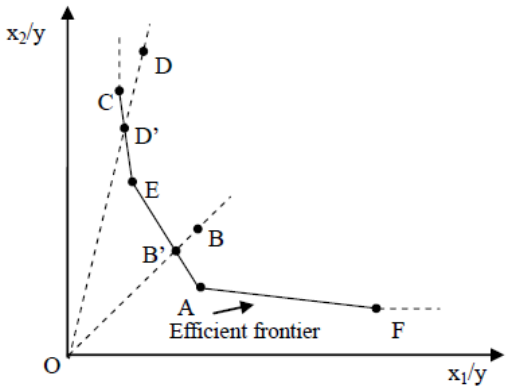


Figure 1. Efficient Frontier of a DEA model. Figure 2. Efficient Frontier and DMUs.

First, for comparison, we expand the inputs and output of each DMU proportionally and let the output of each DMU be 120 (Table 3).

Table 2. 6 DMUs with 2 inputs and a single output.

DMUs	DMU_1	DMU_2	DMU_3	DMU_4	DMU_5	DMU_6
Output (y)	120	8	24	40	120	24
Input 1 (x_1)	19	1	1	2	10	8
Input 2 (x_2)	10	1	6	15	17	1

We show these six DMUs in Figure 2 (which gives projections in input space) using points A-F to denote DMU_1-DMU_6.

Table 3. Expanded DMUs with 2 inputs and single output.

DMUs	DMU_1	DMU_2	DMU_3	DMU_4	DMU_5	DMU_6
Output(y)	120	120	120	120	120	120
Input 1(x_1)	19	15	5	6	10	40
Input 2(x_2)	10	15	30	45	17	5

We use a piecewise linear curve to link points C, E, A, F and merge it with the horizontal and vertical lines from point F and C, respectively, to obtain the piecewise linear convex hull, which is the efficient frontier produced from this DEA model. Points C, E, A, F are on the efficient frontier and their efficiencies are all unity. On the contrary, points B and D are inside the convex hull, so these two DMUs are inefficient compared with their peers or projections (points B' and D') on the efficient frontier. Taking point B as example, the DEA model uses the ratio of distance OB' to the distance OB to measure point Bs relative efficiency.

Decomposition of countries/territories based on multi-level frontiers in DEA

In the preceding section, we showed how the effective frontier can be detected. If we remove the efficient DMUs on the frontier, we can use the DEA model again to obtain a new frontier. We do this repeatedly in order to decompose DMUs into different levels. This process is illustrated in Figure 3. In this figure, the first tier of the efficient frontier is the piecewise line ABCD (Efficient frontier – tier1), on which the DMUs with the best level of efficiency are located. After we remove the DMUs on the Efficient frontier – tier1, we rerun the DEA model, obtaining the DMUs on the efficient frontier – tier2 as the second group, and so on.

This process is iterated until there is no DMU left, and the grading of the DMUs ends. The efficient frontier in Figure 1 is the same as the efficient frontier– tier1 in Figure 3.

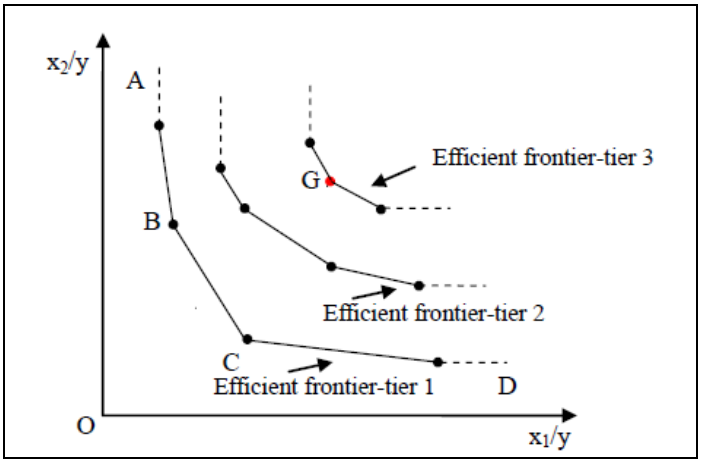


Figure 3. Multi-level efficient frontiers of a DEA model.

In earlier works, DEA frontiers have been used either to measure the relative efficiency of the DMUs (e.g., Charnes et al., 1978; Cook and Seiford, 2009) by comparing them with their peers or projections on the frontier, or to estimate the RTS by the frontier's shape (Banker et al., 2004). To the best of our knowledge, no research similar to the research reported in this paper has used multi-level frontiers in DEA models to decompose DMUs into different grades to reflect different levels of performance.

In the process of decomposing the DMUs into different grades, we need to ensure that a given DMU can only be assigned to one level to avoid conflicts. An efficient frontier is a convex hull. This implies that if a point belongs to F_k it cannot belong to any other F_{k+l} (if it exists, where l is a positive integer). Indeed a point on the frontier is a convex linear combination of efficient points on the frontier. If point P would belong to F_k and F_{k+l} this would mean that P is a convex linear combination of points that do not belong to F_k, which is not possible. Thus, one country/territory can only be assigned to one level.

Results

The BCC-DEA model was applied to produce multi-level efficient frontiers, and these were used to decompose the countries/territories of the study into different grades. Table 4 reports the levels of the countries/territories for the three experiments: two inputs & two outputs, two inputs & the first output (papers), and two inputs & the second output (citations).

We first consider the case of two inputs and two outputs. The results show that Chile, Greece, Iceland, Italy, Netherlands, UK and USA are the first level countries in the sense of efficiency of S&T resource utilization (Table 4). Mexico is the least efficient unit among the 39 countries/territories and belongs to the last level (Tier 6).

We reused the multi-level efficient frontiers in the BCC-DEA model on the 39 countries/territories with two inputs and the first output (papers) to decompose the countries/territories into different grades. We can see that now Chile, Greece, Iceland, Italy, Netherlands, UK and USA are the most efficient countries/territories (Table 4). Mexico, Finland, Israel and Singapore have with the lowest efficiencies.

We also used the multi-level efficient frontiers in the BCC-DEA model on the 39 countries/territories with two inputs and the second output (Citations), which is shown in table 4. Also in this case Chile, Greece, Iceland, Netherlands, UK and USA are first level countries, and Italy has moved into Tier 2. The latter means that Italy performs better for papers than for citations. Mexico and Turkey are in the last tier, Tier 7. It is interesting to see

that Turkey is in Tier 3 in the case of two inputs and two outputs while in Tier 7 in the case of two inputs and the second output, which means that the citation performance of Turkey is considerably worse than its performance for papers.

Table 4. Levels of the countries/territories.

No.	*Countries /Territories*	*two inputs & two outputs*	*two inputs & first output(paper)*	*two inputs & second output(citation)*
1	Chile	1	1	1
2	Greece	1	1	1
3	Iceland	1	1	1
4	Netherlands	1	1	1
5	UK	1	1	1
6	USA	1	1	1
7	Italy	1	1	2
8	Canada	2	2	2
9	China	2	2	2
10	Estonia	2	2	2
11	Germany	2	2	2
12	Luxembourg	2	2	2
13	New Zealand	2	2	2
14	Spain	2	2	2
15	Belgium	2	2	3
16	Slovakia	2	2	3
17	Sweden	2	2	3
18	Poland	2	2	4
19	Ireland	2	3	2
20	Denmark	2	4	3
21	France	3	3	3
22	Slovenia	3	3	3
23	Japan	3	3	4
24	Romania	3	3	4
25	South Africa	3	3	4
26	Turkey	3	3	7
27	Norway	3	4	4
28	Portugal	3	4	4
29	Austria	3	5	4
30	South Korea	4	4	4
31	Hungary	4	4	5
32	Taiwan	4	4	5
33	Czech Republic	4	5	6
34	Israel	4	6	5
35	Singapore	4	6	5
36	Argentina	5	5	6
37	Russia	5	5	6
38	Finland	5	7	6
39	Mexico	6	8	7

Figure 4 corresponds to Table 4 and visualizes the levels of the countries/territories when using two inputs and two outputs, two inputs and the first output (paper), and two inputs and the second output (citation). From this figure, it is clear that some countries/territories (e.g.,

Argentina, Belgium, Czech Republic, Turkey) belong to a lower level in the case of two inputs & the second output (citations) compared to the case of two inputs & the first output (papers), which indicates that these countries perform more efficient for papers than for citations. Inversely, some countries (e.g., Austria, Denmark, Finland) perform more efficient for citations than for papers.

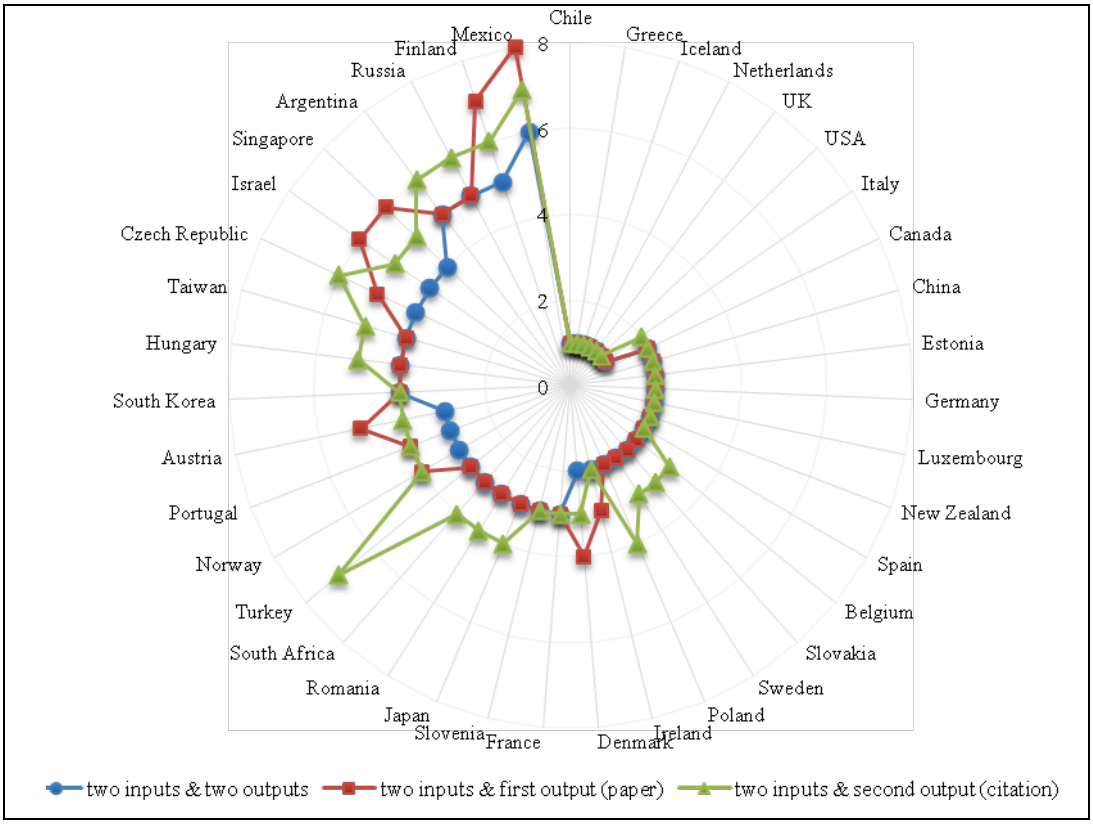


Figure 4. Visualisation of the levels of the countries/territories.

It is surprising that Greece and Chile are rated first level countries together with S&T-developed countries like USA and UK. For papers as output, we can verify this result using the ratios Papers to Researcher and Papers to Research Funding. From Table 5, we can see that Greece and Chile perform very well for these two ratios. On the contrary, we can see China, Japan and South Korea have low performance compared to other countries. We believe that a reason for this is that researchers from these countries publish relatively frequently in domestic journals that are not covered by WoS. We do not tabulate the values of the corresponding two ratios for citations, but it turned out that Chile and Greece perform well also with respect to these ratios.

Discussion and conclusions

In this paper we have shown that multi-level frontiers of DEA can be used to decompose countries/territories into different levels, reflecting the efficiency of S&T resource utilization of the countries/territories. The approach put forward is not restricted to the grading of countries/territories. It can also be used to grade, for instance, journals and research institutions based on properly selected indicators. In case of no explicit inputs, e.g., when journals should be graded, we can assume that there is single constant input, which is equal to unity for all observations (e.g., Yang et al. 2014b).

There are two main advantages of the grading approach proposed in this paper. First, it is a nonparametric and recursive approach, which needs no a priori information such as indicator

weights and threshold values for different grading levels. Second, the observations within the same level are indifferent in the sense of efficiency of resource utilization. The main disadvantage of the approach is that in some cases there are too few indicators (single input and single output). Under such circumstances, it might be the case that each level includes exactly one observation (in our case, exactly one DMU). Thus, the approach is more suitable for grading observations with multiple input and output indicators.

For future research, we would like to investigate the multiple DEA frontiers regarding weight restrictions in DEA models. There are at least four types of restrictions on the weights of input and output variables (e.g., Allen et al., 1997), and the efficient frontiers will vary accordingly and show different properties. Furthermore, this grading approach can be easily extended to the classification of scientific journals, research institutions, etc.

Table 5. Ratios of Papers to Researcher and Research Funding.

No.	*Countries/Territories*	*Papers/Researcher*	*Papers/Research Funding*	*No.*	*Countries/Territories*	*Papers/Researcher*	*Papers/Research Funding*
1	Argentina	0.1616	1.7717	21	Mexico	0.2274	1.3017
2	Austria	0.3460	1.2880	22	Netherlands	0.5791	2.3185
3	Belgium	0.4422	1.9381	23	New Zealand	0.5019	4.6310
4	Canada	0.3751	2.3842	24	Norway	0.3976	2.1375
5	Chile	0.9527	4.9410	25	Poland	0.3283	3.2855
6	China	0.1235	0.6569	26	Portugal	0.2155	2.5981
7	Czech Republic	0.3216	2.1174	27	Romania	0.4308	4.0135
8	Denmark	0.3586	1.9623	28	Russia	0.0650	0.8261
9	Estonia	0.3345	2.0570	29	Singapore	0.2951	1.4374
10	Finland	0.2690	1.3625	30	Slovakia	0.2012	3.3464
11	France	0.2706	1.2644	31	Slovenia	0.4304	2.6410
12	Germany	0.2833	0.9893	32	South Africa	0.4711	2.0371
13	Greece	0.4385	5.3908	33	South Korea	0.1578	0.7809
14	Hungary	0.2578	2.1803	34	Spain	0.3891	2.5204
15	Iceland	0.3609	2.5658	35	Sweden	0.4439	1.6136
16	Ireland	0.4902	2.3466	36	Taiwan	0.2035	1.0420
17	Israel	0.2506	1.3408	37	Turkey	0.3317	2.1165
18	Italy	0.5213	2.1465	38	UK	0.4014	2.5727
19	Japan	0.1180	0.5220	39	USA	0.2910	0.8495
20	Luxembourg	0.2237	1.0267				

Acknowledgements. We would like to acknowledge the support of the National Natural Science Foundation of China (NSFC, No.71201158).

References

Abbott, M. & Doucouliagos, C. (2003). The efficiency of Australian universities: a data envelopment analysis. *Economic of Education Review*, *22*(1), 89-97.

Allen, R., Athanassopoulos, A., Dyson, R.G., & Thanassoulis, E. (1997). Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. *Annals of Operations Research*, *73*, 13-34.

Aristovnik, A. (2012). The relative efficiency of education and R&D expenditures in the new EU member states. *Journal of Business Economics and Management*, *13*(5), 832-848.

Avkiran, N.K. (2001). Investigating technical and scale efficiencies of Australian universities through data envelopment analysis. *Socio-Economic Planning Sciences*, *35*(1), 57-80.

Bandyopadhyay, A. (2013). *Ranking of Business School Journals: A Rating Guide for Researchers.* Retrieved August 23, 2014 from: http://mpra.ub.uni-muenchen.de/49608/1/MPRA_paper_49608.pdf.

Banker, R.D. (1993) Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Management Science*, *39*(10), 1265–1273.

Banker, R.D., Charnes, A., & Cooper, W.W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*(9), 1078-1092.

Banker, R.D., Cooper, W.W., Seiford, L.M., Thrall, R.M., & Zhu, J. (2004). Returns to scale in different DEA models. *European Journal of Operational Research*, *154*, 345-362.

Bonaccorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics*, *63*(1), 87-120.

Brandt, T., & Schubert, T. (2013). Is the university model an organizational necessity? Scale and agglomeration effects in science. *Scientometrics*, *94*(2), 541-565.

Carrington, R., Coelli, T., & Rao, P.D.S. (2005). The performance of Australian universities: Conceptual issues and preliminary results. *Economic Papers-Economic Society of Australia*, *24*(2), 145-163.

CAS (2006). *Report on Comprehensive Quality Evaluation of Institutes in Chinese Academy of Sciences (CAS)*. Retrieved August 23, 2014 from http://www.bps.cas.cn/kjpj/xgzl/200905/t20090527_232560.html.

Charnes, A., Cooper, W.W., & Rhodes, E.L. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429-444.

Chen, C.P., Hu, J.L., & Yang, C.H. (2011). R&D efficiency of multiple innovative outputs: The role of the national innovation system. *Innovation: Management, policy & practice*, *13*, 341-360.

Cook, W.D., & Seiford, L.M. (2009). Data Envelopment Analysis (DEA)-Thirty years on. *European Journal of Operational Research*, *192*, 1-17.

Cooper, W.W., Seiford, L.M., & Tone, K. (2007). *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software (Second Edition)*. New York: Springer.

Flegg, A.T., Allen, D.O., Field, K., & Thurlow, T.W. (2004). Measuring the Efficiency of British Universities: A Multi-Period Data Envelopment Analysis. *Education Economics*, *12*(3), 231-249.

Glänzel, W. (2011). The application of characteristic scores and scales to the evaluation and ranking of scientific journals. *Journal of Information Science*, *37*(1), 40-48.

Harvey, C., Kelly, A., Morris, H., & Rowlinson, M. (2010). *Academic Journal Quality Guide—Version 4*. London: Association of Business Schools.

Harvey, C., Morris, H., & Kelly, A. (2007a). *Academic Journal Quality Guide: Context, Purpose and Methodology*.London: Association of Business Schools.

Harvey, C., Morris, H., & Kelly, A. (2007b). *Academic Journal Quality Guide*. London: Association of Business Schools.

Harvey, C., Morris, H., & Kelly, A. (2008). *Academic Journal Quality Guide Version 2: Context, Purpose and Methodology*. London: Association of Business Schools.

Johnes, G., & Johnes, J. (1992). Apples and oranges: The aggregation problem in publications analysis. *Scientometrics*, *25*(2), 353-365.

Kao, C., & Lin, Y.C. (1999). Comparing university libraries of different university size. *Libri*, *49*(3), 150-158.

Kempkes, G., & Loikkanen, H.A. (1998). The efficiency of German universities: Some evidence from nonparametric and parametric methods. *Applied Economics*, *42*, 2063-2079.

Lee, H.Y., & Park, Y.T. (2005). An international comparison of R&D efficiency: DEA approach. *Asian Journal of Technology Innovation*, 13(2), 207-222.

Lee, H.Y., Park, Y.T., & Choi, H. (2009). Comparative evaluation of performance of national R&D programs with heterogeneous objectives: A DEA approach. *European Journal of Operational Research*, *196*(3), 847-855.

REIST-2 (1997). *European Commission, Second European Report on S&T Indicators.* (EUR 17639). Brussels: Luxembourg.

Rousseau, S. & Rousseau, R. (1997). Data envelopment analysis as a tool for constructing scientometric indicators. *Scientometrics, 40*(1), 45-56.

Rousseau, S., & Rousseau, R. (1998). The scientific wealth of European nations: taking effectiveness into account. *Scientometrics, 42*(1), 75-87.

Roy, S., & Nagpaul, P.S. (2001). A quantitative evaluation of relative efficiencies of research and development laboratories: A data envelopment analysis approach. In: (M. Davis & C.S. Wilson. Eds.) *Proceedings of the 8th International Conference on Scientometrics & Informetrics* (pp. 629-638). Sydney (Australia): BIRG.

Sharma, S., & Thomas, V.J. (2008). Inter-country R&D efficiency analysis: an application of data envelopment analysis. *Scientometrics*, *76*(3), 483-501.

Shim, W., & Kantor, P.B. (1998). A novel economic approach to the evaluation of academic research libraries. *Proceedings of the ASIS Annual Meeting*, *35*, 400-410.

Smith, R. (1996). Peer review: A flawed process in the heart of science and journals. *Journal of the Royal Society of Medicine*, 99(4), 178-182.

Smith, P. (1997). Model misspecification in data envelopment analysis. *Annals of Operations Research*, *73*, 233-252.

Schubert, T. (2014). Are there scale economies in scientific production? On the topic of locally increasing returns to scale. *Scientometrics*, (to appear), DOI 10.1007/s11192-013-1207-1.

Sueyoshi, T., & Goto, M. (2013). A use of DEA-DA to measure importance of R&D expenditure in Japanese information technology industry. *Decision Support Systems*, *54*, 941-952.

Wang, E.C., & Huang, W.C. (2007). Relative efficiency of R&D activities: A cross-country study accounting for environmental factors in the DEA approach. *Research Policy*, *36*(2), 260-273.

Wolszczak-Derlacz, J., & Parteka, A. (2011). Efficiency of European public higher education institutions: a two-stage multicountry approach. *Scientometrics*, *89*, 887-917.

Worthington, A.C. (2001). An empirical survey of frontier efficiency measurement techniques in education. *Education Economics*, *9*(3), 245-268.

Worthington, A.C., & Lee, B.L. (2008). Efficiency, technology and productivity change in Australian universities, 1998-2003. *Economics of Education Review*, *27*(3), 285-298.

Yang, G.L., Yang, L.Y., Liu, W.B., Li, X.X., & Fan, C.L. (2013a). Directional returns to scale of biological institutes in Chinese Academy of Sciences. In: (J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, H. Moed, Eds.). *Proceedings of ISSI 2013* (pp. 551-566).Vienna: Austrian Institute of Technology (AIT).

Yang, G.L., Rousseau, R., Yang, L.Y., & Liu, W.B. (2014a). A study on directional returns to scale. *Journal of Informetrics*, *8*(3), 628-641.

Yang, G.L., Shen, W.F., Zhang, D.Q., & Liu, W.B. (2014b). Extended utility and DEA models without explicit input. *Journal of the Operational Research Society*, *65*, 1212–1220.

Yang, L.Y., Zhou, Q.J., & Yue, T. (2013b). China's Science: The Overall Development and the Balance of Disciplinary Structure—Statistics and Analysis of SCI-indexed Papers in 2012. *Science Focus*, 8(1), 23-50. (In Chinese).

Yang, H.-H., & Chang, C.-Y. (2009). Using DEA window analysis to measure efficiencies of Taiwan's integrated telecommunication firms. *Telecommunications Policy*, *33*(1-2), 98-108.

Zhong, W., Yuan, W., Li, S.X., & Huang, Z.M. (2011). The performance evaluation of regional R&D investments in China: An application of DEA based on the first official China economic census data. *Omega-the International Journal of Management Science*, *39*(4), 447-455.

Continuous, Dynamic and Comprehensive Article-Level Evaluation of Scientific Literature

Xianwen Wang[1], Zhichao Fang[1] and Yang Yang[1]

[1] xianwenwang@dlut.edu.cn, fzc0225@dlut.edu.cn, yangyang0477@mail.dlut.edu.cn
WISE Lab, Faculty of Humanities and Social Sciences, Dalian University of Technology, Dalian 116085 (China)

Abstract

Current research assessment is built on the basis of core-journals-selection system. Journal evaluation is not equal to article evaluation, evaluating scientists, institutions and countries based on article-level evaluation is more reasonable than the current journal-based evaluation. Different from the current research evaluation tools and databases, e.g., ESI and Nature Index, in this study, we propose the idea of continuous, dynamic and comprehensive article-level-evaluation based on article-level-metrics data. Different kinds and sources of metrics are integrated into a comprehensive indicator, to quantify both the long-term academic and short term societal impact of the article. At different phases after the publication, the weights of different metrics are dynamically adjusted to mediate the long term and short-term impact of the paper. Using the sample data, we collect the metrics data over two years for each sample article, and make empirical study of the article-level-evaluation method. The original data and interactive visualization of this research is available at http://xianwenwang.com/research/ale/.

Conference Topic

Altmetrics; Indicators; Science policy and research assessment

Introduction

For decades, citation has been regarded as the sole indicator to evaluate the impact of a paper, a paper that is cited more frequently means the research results gained more recognition. However, citations need a long time (often over two years) to accumulate. In many situations, e.g., funding decisions, hiring tenure and promotion, people need to make evaluations for newly published papers. Alternatively, some people begin to use journal based metrics, e.g., Journal Impact Factor, as an alternative way to quantify the qualities of individual research articles (Alberts, 2013). There are many debates about the abuse of Impact Factor (Bordons, Fernández, & Gomez, 2002; Garfield, 2006; Opthof, 1997; PLoS_Medicine_Editors, 2006; Seglen, 1997), applying Journal Impact Factor to assess the research excellence is not the most appropriate way. In addition, only tracking citation metrics could not tell the whole story about the influence of a paper. Besides citation, the impact of scientific papers could be reflected with article usage (browser views and pdf downloads), captures (bookmarks and readership), online mentions (blog posts, social media discussions and news reports) (Priem, Taraborelli, Groth, & Neylon, 2010). Therein, the idea of altmetrics comes into being. Different from citation, which puts particular emphasis on describing the academic impact of articles, altmetrics is based on data gathered from social media platforms and focuses on the societal impact (Kwok, 2013; Sud & Thelwall, 2014; Zahedi, Costas, & Wouters, 2014). Compared with the long time for papers to reach their citation peaks, it takes a short period for newly published articles to peak for altmetric scores. In summary, citation is an indicator to measure the long-term academic impact, when the indicator of altmetrics reflects short term societal impact. Neither citations nor altmetrics individually could fully indicate the complete impact of a paper, we cannot accurately conjecture the results of one metric by the results of another.

It is necessary to find a way to quantify both the academic and societal impact together, and mediate the long term and short-term impact of the paper. Some publishers have already listed

the different types of metrics for an individual article, e.g., PLOS, when some altmetrics tools and services are also available, e.g., Impact Story, Altmetric.com, Plum Analytics, etc. Although altmetric score from altmetric.com is a weighted count that integrates different online mentions of the paper. If we go further on this way, taking all available metrics (e.g., citation, usage, online attention, etc.) into consideration to design a comprehensive metric, which could be used to evaluate the complete impacts of articles.

Based on the calculated total impacts, the comprehensive metric makes it possible to rank articles on a unified dimension, which solo academic or societal impact indicator could not.

The absence of evaluating data source

According to the official statement of Web of Science, it is designed for researchers to "find high-impact article". Nowadays, with the absence of specialized evaluating data source, Web of Science has been adopted by many scientometrics researchers and institutions as the primary data source of article evaluation. In some countries, e.g., China, articles indexed in Science Citation Index/Social Science Citation Index or not is a very important criterion to judge the quality of the research.

However, applying Web of Science to assess the research performance and research excellence is not a good choice. Web of Science is designed and created on the basis of journal selection, it collectively index journals cover-to-cover. However, articles published in the same journal, the same issue, have totally different impacts. Even for those high impact factor journals, there are many articles have few citations.

We check the articles published in 2000 and indexed in Science Citation Index Expanded, as Table 1 shows. For example, 2901 of the total 13660 articles in Chemical Engineering have never been cited. For the area of Condensed Matter Physics, the zero-citation percentage is 10.91%, for the area of Biochemistry, Molecular Biology, the zero-citation percentage is 3.23%.

Table 1. Number of Zero-citation articles in 2000 indexed in Science Citation Index Expanded.

	Total	*Zero-citation*	*Percentage*
Engineering, Chemical	13660	2901	21.24
Physics, Condensed Matter	21974	2397	10.91
Biochemistry, Molecular Biology	42710	1380	3.23

There are also some publishers regard Web of Science as a profit-making tool. For example, *Academic Journals* charges a US$550-$750 manuscript handling fee from the author for each accepted article (http://www.harzing.com/esi_highcite.htm). Among which, several ISI-listed journals publish more than 1,000 articles per year, e.g., in 2007, *African Journal of Business Management* only published 28 articles, in 2010, it published 446, when in 2011, as many as 1350 articles were published by this single journal. Thomson Reuters has the mechanism to review the exiting journal coverage constantly, some journals that have become less useful would be deleted. However, this kind of mechanism does not apply to the articles, even some journals are deleted from the coverage, numerous low-quality papers published by these journals are still indexed in Web of Science.

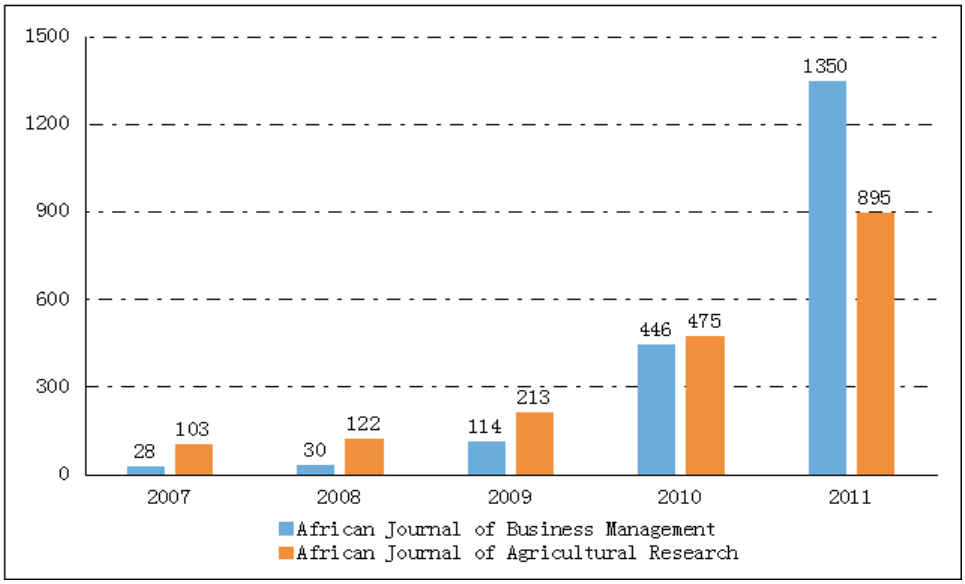


Figure 1. Rapid growth of yearly indexed articles of two journals.

With the same idea of Web of Science, Nature Publishing Group (NPG) introduced the Nature Index in November 2014, which is "a database of author affiliation information collated from research articles published in an independently selected group of 68 high-quality science journals" (Nature, 2014). The 68 journals are selected by a group of professors and validated by 2,800 responses to a large-scale survey, when these 68 journals account for approximate 30% of total citations to natural science journals (http://www.nature.com/press_releases/nature-index.html).

Based on journal article publication counts and citation data from Thomson Scientific databases (mainly from Web of Science), ISI/Thomson (now Thomson Reuters) proposed Essential Science Indicators (ESI), which is an in-depth analytical tool and also a database where citations are analyzed, so that scientists, journals, institutions, and countries can be ranked and compared, for example, most cited scientists rankings, institutions rankings and countries rankings. Ranking in ESI is made by the citations, it has nothing to do with the Impact Factors of journals, which means that whichever journal the paper is published in, citations is the only factor to be taken into account. Although ESI set a relatively low selection criterion for newly published papers (http://www.in-cites.com/thresholds-highly-cited.html), using cited times to evaluate is not a good choice.

Compared to 8670 journals covered by Science Citation Index Expanded, the journals selected by Nature Index is so much less, which makes Nature Index become an elite database. The aim of Nature Index is "intended to be one of a number of metrics to assess research excellence and institutional performance" (http://www.natureindex.com/faq). However, we think journal-based database is not appropriate for research evaluation, including research excellence and institutional performance, which should be on the basis of article-level metrics. Because of the great influence of Nature Publishing Group, the Nature Index will definitely make great changes to the academia and research evaluation system.

It is necessary to make changes to the current evaluating way of scientific literature. In this research, our purpose is to design a new method, through which the continuous, dynamic and comprehensive evaluation of scientific literature could be made. This new method will be valuable to the research community. With this evaluating method and system, we could make a better evaluation of articles, scientists, journals, institutions, and even countries.

Design a new evaluation way

Considering both academic and societal impact of a paper

As mentioned above, the impact of a paper could be measured by citation, article usage and online mentions, etc., as Table 2 shows.

Table 2. Types and metrics of the impact of a paper.

Type	*Metric*
Article usage	browser views (abstract, full-text), pdf downloads
Captures	bookmarks (CiteUlike), readers (Mendeley)
Online mentions	blog posts, news reports, likes (Facebook), shares (Facebook), Tweets, +1 (Google plus)
Citations	citations

The Issue 6, Volume 8 of PLOS Computational Biology is selected as our research object. It was published in June 2012, and includes 46 research articles.

In November 2012, PLOS began to provide a regular report covering a wide range of article-level-metrics covering all of its journals via the platform http://article-level-metrics.plos.org/. In this research, the cumulative article-level-metrics data for the entire PLOS corpus are harvested from the PLOS ALM platform. From October 2012 to October 2014, PLOS has provided the ALM reports for 8 times, when the provided date are Oct. 10, 2012, Dec. 12, 2012, Jan. 8, 2013, Apr. 11, 2013, May. 20, 2013, Aug. 27, 2013, Mar. 10, 2014 and Oct. 1, 2014. Factor analysis is employed to study the metrics data of the 46 articles, Table 3 shows the results of the data extracted from the ALM report of Oct. 2014.

Table 3. Rotated Component Matrix.

	Factor 1: Academic impact	*Factor 2: Societal impact*
CiteUlike	0.775	
Mendeley	0.856	
HTML views	0.692	0.672
PDF downloads	0.917	
Scopus	0.751	
Facebook		0.745
Twitter		0.709

Note. Factor loadings < .5 are suppressed

7 metrics data of Oct. 10, 2012 are factor analyzed by using principal component analysis with Varimax (orthogonal) rotation. The analysis yields two factors explaining a total of 73.709% of the variance for the entire set of variables. Factor 1 is labeled academic impact to the high loadings by the following items: CiteUlike bookmarks, Mendeley readership, PDF downloads and Scopus citations. This first factor explained 48.691% of the variance. The second factor derived is labeled societal impact. This factor is labeled as such due to the high loadings by the two indicators of Facebook and Twitter. The variance explained by this factor is 25.018%. For the indicator of HTML views, the both factor loadings are greater than 0.65, which means that browser HTML views has both academic and societal impact.

The Altmetric score is a quantitative measure of the attention that a scholarly article has received. It is a weighted count of the different online platform sources (newspaper stories, tweets, blog posts, comments) that mention the paper. Downloads, citations and reader counts

from Mendeley or CiteULike are not used in the score calculation. So, Altmetric score could be regarded as a comprehensive indicator that measures the societal impact of paper partially.

Dual function of societal impact

The value of societal metrics is not only reflected by the social effects of the diffusing of the knowledge embodied in the literature, but also reflected by the possible additional academic impact caused by social online attention.

Social media make the research achievements and scientific discoveries spread to the general public, which is just the goal of scientific researches. From the other hand, wide spreading of scientific literature could lead to more scholarly citations. The mechanism from online attention to citation is very complicated, but social attention do have the potentiality to contribute some extra citations to a paper (Wang, Liu, Fang, & Mao, 2014; Wang, Mao, Zhang, & Liu, 2013).

Dynamic patterns of article-level metrics

For the 46 selected articles published in June 2012, we sum the metrics data at the 8 time periods separately, as Figure 2 shows. Different metrics show different dynamic evolution patterns. In October 2012, when the articles had been published for about 4 months, there is few citations. The curve of citations begins a sharp rise at the phase of May 2013, one year after the publication. However, for the Facebook and Twitter data, the two curves have almost reached their summits at the very first phase. During the next periods, there is little increase for the Facebook and Twitter data. And for the views data, which is placed on the secondary Y axis in Figure 2, the situation is somehow between the citations and Facebook/Twitter. At the first phase, there is considerable data. During the following 7 periods, there is a steady growth trend for the curve of views.

Dynamic patterns for the different metrics are distinct. Social attention comes to go, citation takes a long time to know, when article view also comes fast but keeps a steady growth.

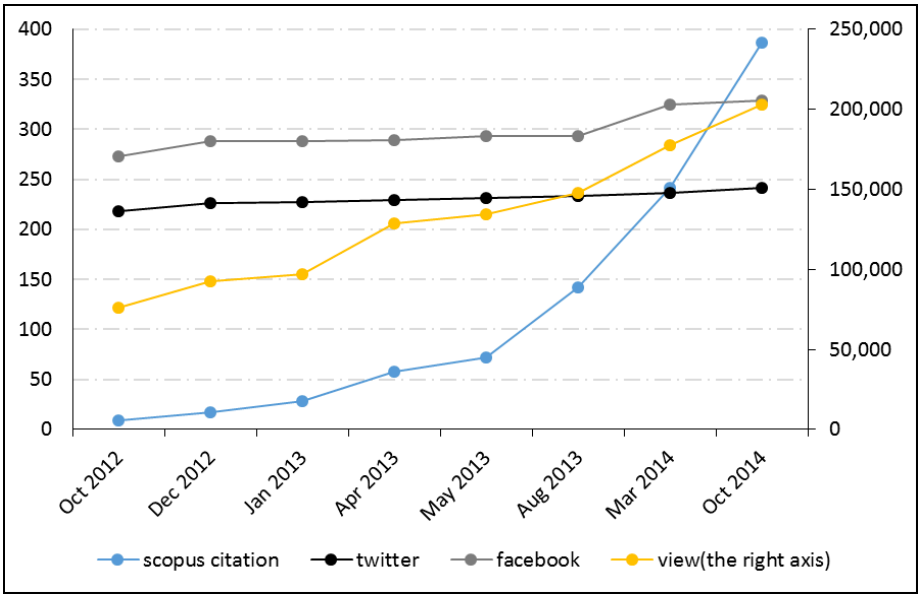


Figure 2. Temporal trend of different metrics of 46 articles published in June 2012.

Article-level evaluation based on Article-level-metrics

In the era of print, the article could not be separated from the whole issue. For example, libraries could provide the borrowing statistical data, however, it's difficult to know which

single article or articles readers are interested in. In the digital era, the situation has been changed greatly. Metrics data for each article are easy to know, including the views, downloads, altmetric score and citations. Of course, some data are easy for publishers to know but not released to public. As early in March 2009, PLOS inaugurated a program to provide "article-level metrics" on an article across all PLOS journals. The metrics data include five main categories, which are Viewed, Cited, Saved, Discussed and Recommended. Following PLOS, more and more publishers began to provide detailed article-level metrics data for readers and researchers. For example, in October 2012, Nature began to provide a real-time online count of article-level metrics for its published research papers, including citation data, news mentions, blog posts and details of sharing through social networks, such as Facebook and Twitter (http://www.nature.com/news/nature-metrics-1.11681). In 2014, the article-level metrics data are also available for PNAS and Science. The growing article-level metrics dataset provides us with the possibility to design a new evaluating way to make article-level evaluation.

Problems need to be solved

The first problem is there are too many indicators need to be considered. Citation has been regarded as the single indicator for the past tens of years, nowadays there are much more indicators which are worth being considered, including article views, bookmarks and readership, online discussion, news reports and citations, etc. So many indicators mean a lot of dimensions of the impact, different papers may have different values for the indicators, for example, paper A has been downloaded many times but retweeted few times, when paper B may has opposite situation, so it is very difficult to compare the impact of these two articles, especially when these articles are newly published.

Could these so many indicators be synthesized to one single comprehensive indicator, which could reflect the most of information of the original data and make the papers in diverse situations comparable?

The second problem is the dynamic adjustment of the results. At different phases after publication, the same indicator may have different effects on the impact of the paper. For the newly published articles, because the citations are generally low, it is difficult to judge the qualities and compare the new articles. At the early phase, it is a better choice to use article usage data, online mention data to make evaluation of the newly published articles. As time goes by, the evaluation is gradually dominated by citation metrics, which means that citation would play the most important role in the evaluation when the article has been published for a relatively long time. To solve these two problems, we propose the idea of designing a comprehensive indicator to reflect all the impacts of an article. The weights of the indicators at different phases should be adjusted dynamically due to the change of relative importance of metrics, just like Table 4 shows.

To integrate different metrics into a comprehensive indicator, the first problem needs to be solved is weighting. Here we use Analytic hierarchy process (AHP) to calculate the weights of different metrics. The AHP methodology was developed by Thomas L. Saaty in the 1970s (Saaty, 1980). It allows users to assess the relative weight of multiple criteria in an intuitive manner, so it has both advantages of quantitative criteria and qualitative judgment provided by the users. Using pairwise comparisons (X is more important than Y), the relative importance (priority) of one criterion over another can be expressed. To calculate the weights for the different criteria, a pairwise comparison matrix needs to be created. The matrix is a matrix A, where m is the number of evaluation criteria considered, denotes the entry in the *i*th row and the *j*th column of matrix. Each entry of the matrix represents the importance of the *i*th criterion relative to the *j*th criterion. If the cell value in the entry is greater than 1, then the *i*th criterion is more important than the *j*th criterion, and vice versa. If two criteria have the

same importance, then the cell value in the entry is 1. The relative importance between two criteria is measured according to a numerical scale from 1 to 9 or 1/9 to 1.

Table 4. Relative importance of metrics at different phases.

Phase	*Relative importance*	*Selection standard*
1 (0-6 months)	PDF downloads > HTML views > Twitter > Facebook > Mendeley > CiteUlike > Citation	Top 80% of all articles of same month and subject
2 (6 months-2 years)	PDF downloads > HTML views > Mendeley > CiteUlike > Citation > Twitter > Facebook	Top 70% of all articles of same month and subject
3 (2 -5 years)	Citation > Mendeley > CiteUlike > PDF downloads > HTML views > Twitter > Facebook	Top 50% of all articles of same year and subject
4 (5 years-)	Citation > Mendeley > CiteUlike > PDF downloads > HTML views > Twitter > Facebook	Top 30% of all articles of same year and subject

According to the definition of relative importance of different metrics, we need to construct different pairwise comparison matrixes at different phases. The pairwise comparison matrix at phase 1 is shown in Table 5. The higher the weight is, the more important the corresponding criterion becomes, which is represented by the cell value in the matrix. For example, the values in the cells where the row of CiteUlike, the column of HTML views and PDF downloads intersect are less than 1, moreover, the ratio of CiteUlike and PDF downloads is less than the ratio of CiteUlike and HTML views, it means that at phase 1, CiteUlike is less important than HTML views, and much less important than PDF downloads.

Table 5. Pairwise Comparison Matrix at phase 1.

	CiteUlike	*Mendeley*	*HTML views*	*PDF downloads*	*Citation*	*Facebook*	*Twitter*
CiteUlike	1	1	1/4	1/6	4	1/4	1/6
Mendeley		1	1/4	1/6	4	1/4	1/6
HTML views			1	1/4	6	3	2
PDF downloads				1	9	4	3
Citation					1	1/4	1/7
Facebook						1	1/2
Twitter							1

At phase 4, there is much change in the relative importance of the metrics, as Table 6 shows. CiteUlike and Mendeley become more important than HTML views, so the cell values get greater than 1. At this phase, citation is the most important criterion.

In this study, the weights and CI values of AHP models are calculated by a CGI system (http://www.isc.senshu-u.ac.jp/~thc0456/EAHP/AHPweb.html). The results are shown in Table 7.

In Figure 3, we show the change of the weights of metrics. At Phase 1 and 2, the metric of PDF downloads has the greatest weight. From Phase 1 to 4, the curve of PDF downloads shows a downward trend, when the weight of citation is upward.

Empirical Study

The weights in Table 7 are applied to calculate the comprehensive scores of the metrics data of the 46 articles. Metrics data of Oct. 10, 2012 is calculated with the weights of phase 1,

Table 6. Pairwise Comparison Matrix at phase 4.

	CiteUlike	*Mendeley*	*HTML views*	*PDF downloads*	*Citation*	*Facebook*	*Twitter*
CiteUlike	1	1	3	2	1/7	3	2
Mendeley		1	3	2	1/7	3	2
HTML views			1	1/4	1/9	1	1
PDF downloads				1	1/6	1	1
Citation					1	4	3
Facebook						1	1/2
Twitter							1

Table 7. Weights of AHP models at different phases.

	CiteUlike	*Mendeley*	*HTML views*	*PDF downloads*	*Citation*	*Facebook*	*Twitter*
Phase 1	0.0477	0.0477	0.1996	0.3901	0.0234	0.1109	0.1806
Phase 2	0.1723	0.1723	0.1182	0.2108	0.1321	0.0828	0.1116
Phase 3	0.1514	0.1514	0.0481	0.0921	0.3979	0.0644	0.0947
Phase 4	0.1269	0.1269	0.0455	0.0809	0.4819	0.0570	0.0810

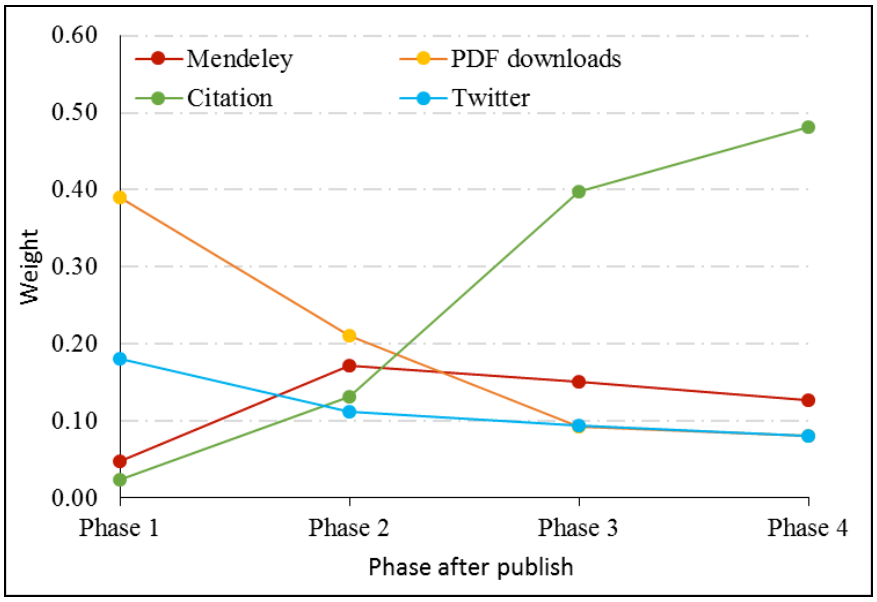


Figure 3. The change of the weights of different metrics.

when weights of phase 2 and 3 are used for metrics data of Aug. 27, 2013 and Oct. 1, 2014 separately. All the original metrics data are normalized to the range of 0-1. The normalized value of e_i for variable E in the ith row is calculated as:

$$Normalized\ (e_i) = \frac{e_i - E_{min}}{E_{max} - E_{min}}$$

Where $E_{\min}$ and $E_{\max}$ are the minimum and maximum value for variable E correspondingly.

In Table 8, the values of 7 metrics are original data, when the scores are calculated with the normalized data instead of the original metrics data.

Table 8. Top 25% articles with greatest score at 3 phases.

phase	*rank*	*doi*	*citeulike*	*mendeley*	*html*	*pdf*	*citation*	*facebook*	*twitter*	*score*
	1	1002358	16	81	5060	1733	3	8	12	0.7906
	2	1002543	14	0	4041	871	0	2	31	0.5653
	3	1002590	0	18	4302	469	0	73	11	0.4413
	4	1002561	3	37	3579	721	0	0	9	0.3671
	5	1002519	3	17	2516	648	0	0	13	0.3146
1	6	1002538	3	6	1777	394	0	22	15	0.2603
	7	1002541	13	24	1794	354	0	3	12	0.2456
	8	1002527	3	12	1818	373	0	6	14	0.2305
	9	1002572	6	18	2045	489	0	0	6	0.2248
	10	1002588	0	13	1809	454	1	0	7	0.1989
	11	1002531	4	20	1519	522	1	2	1	0.1865
	1	1002358	16	170	11720	3236	30	7	14	0.8579
	2	1002543	16	72	5389	1103	1	2	34	0.4739
	3	1002561	3	79	9669	1242	5	2	11	0.3408
	4	1002541	15	57	3609	665	3	4	13	0.3395
	5	1002590	1	36	6024	627	1	91	13	0.2622
2	6	1002531	8	39	3389	912	11	3	1	0.2552
	7	1002519	3	39	5515	1262	1	0	13	0.2419
	8	1002572	6	44	3273	754	2	0	6	0.2006
	9	1002538	3	14	3155	668	4	22	15	0.1889
	10	1002577	2	25	5063	1141	2	0	5	0.1816
	11	1002527	3	21	3266	638	1	6	14	0.1641
	1	1002358	18	324	19909	4651	73	23	14	0.8942
	2	1002543	16	95	6071	1241	1	2	36	0.3113
	3	1002541	16	91	4896	824	11	4	13	0.2931
	4	1002531	9	77	5670	1229	26	3	1	0.2874
	5	1002561	4	121	11231	1577	21	2	11	0.2866
3	6	1002588	0	56	6112	1314	19	3	8	0.1849
	7	1002572	9	62	3803	910	6	0	6	0.1707
	8	1002519	3	69	8233	1653	6	0	13	0.1692
	9	1002590	1	42	7101	904	3	90	13	0.1690
	10	1002555	3	31	5048	701	13	22	4	0.1531
	11	1002562	7	58	2840	529	10	0	0	0.1476

Note: (1) Because of the limited layout space, the first half of the doi is omitted. For example, for the doi 10.1371/journal.pcbi.1002358, we only keep 1002358 in Table 8.

(2) Detailed information of Table 8 is available at http://xianwenwang.com/research/ale

Table 8 lists the top 11 (top 25% of 46) articles of each phase. At phase 1, when the 46 articles had been published for 4 months, article 10.1371/journal.pcbi.1002358 has 16

CiteUlike bookmarks, 81 Mendeley readers, 5060 HTML views, 1733 PDF downloads and 3 Scopus citations, etc., when the comprehensive score of this article is 0.7906, ranks top 1. At phase 2, the values of the metrics of Mendeley, HTML views, PDF downloads and Scopus citations have risen sharply, but not for the metrics of Facebook and Twitter, when the score is 0.8579 and still ranks top 1. From phase 1 to 2 and 3, there is much change for the top 11 articles. The ranks of some articles rise, when others may fall. For example, article 10.1371/journal.pcbi.1002538 ranks 6th at phase 1, downs to 9 at phase 3, and is disappeared from the top 11 at phase 3; article 10.1371/journal.pcbi.1002531 ranks 11 at phase 1, and rises to top 4 at phase 3.

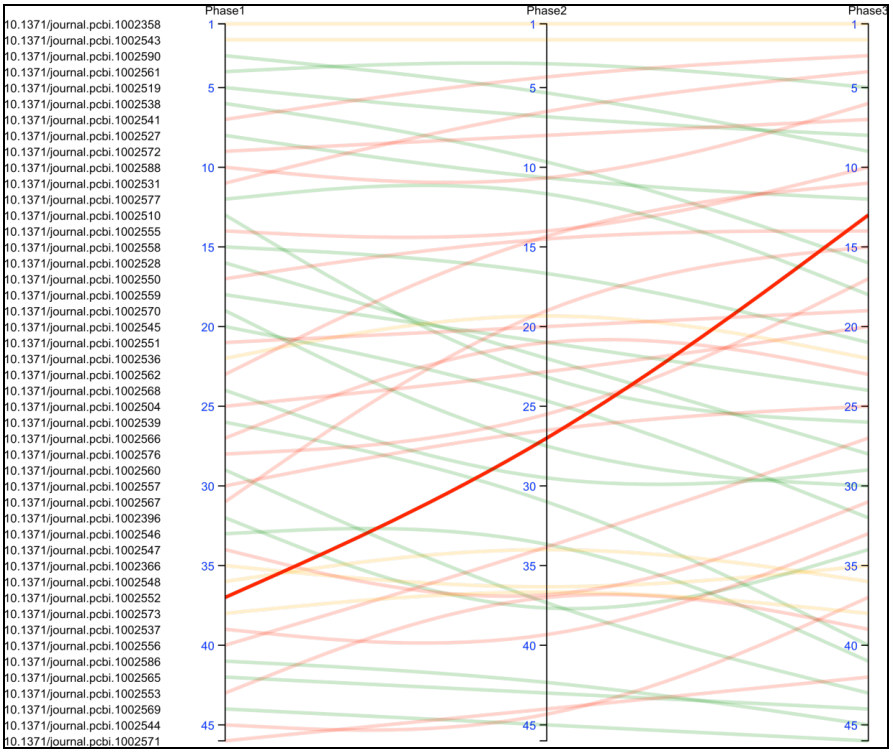


Figure 4. Dynamic changes according to the ranking at different phases.[1]

The dynamic changes of the scores and rankings of the 46 articles from phase 1 to 3 are shown in Figure 4. The DOIs of 46 articles are listed on the leftmost column, and ranked according to the scores at phase 1. The position of article at the certain phase is decided by the ranking of score at that phase. 46 articles could be only compared at the same phase. Articles at different phases, and even the same article at different phases are not comparable. As shown in Figure 4, if the rank of an article from phase 1 to 3 shows an upward trend, it is displayed with a red curve, there are 20 papers with red curves. We use green curve to represent the downward trend, there are also 20 papers with green curves. Otherwise, if the rank of the article has not changed, the color of the curve is yellow, there are 6 yellow curves. In Figure 4, one red curve with dramatic upward trend is highlighted, indicating that the performance of this paper is rising. The doi of this article is 10.1371/journal.pcbi.1002552, it only ranks 37 at phase 1, rises to 28 at phase 2 and continue to rise to 13 at phase 3.

According to the rankings calculated by the comprehensive metric, articles with the highest impact are selected into the database. There are different selection standards at different phases, as Table 4 shows. As time goes on, the data of the original indicators become

[1] An interactive version of Figure 4 is available at http://xianwenwang.com/research/ale/dynamic.html

increasingly sufficient, the accuracy of the results becomes higher. Due to the dynamic changes of the rankings of articles, the database is also dynamic, it ensures the articles included are always has the highest impact at each phase. It would be much easier for researchers to index the high quality articles through the dynamic database.

Discussion

In the 1950s, people read papers from printed journals. A group of articles are bundled together to form an issue of journal, it is difficult to separate single article from the whole issue, which is the carrier of articles. For example, if we want to know which paper the readers are interested in when they borrow the journal from the library, which seems to be an extremely difficult task. At that time, journal evaluation is the most important and basic issue. SCI is designed on the basis of core journals selection, specialized indicators and tools are proposed to evaluate journals, e.g., Impact Factor and Journal Citation Reports.

Compared to fifty years ago, scholarly communicating ways have changed a lot. With the advent and fast development of computers, internet and digital libraries, the transformation from print to electronic publishing is accelerating, just as the digital music revolution set music free from the carriers of cassette tape and CD, the concept of printed journals or even journals in the conventional sense is not important any more. Actually, for some new journals, articles are not organized and published by issues and volumes, e.g., PLOS ONE, Scientific Reports, eLIFE and Peer J, etc.

It is necessary to make changes to the current research evaluation way rooted in the journal selection system. We should be aware of that journal evaluation is not equal to article evaluation, evaluating scientists, institutions and countries based on article-level evaluation is more reasonable than the current journal-based evaluation. It would "be better to measure the performance of countries and institutions on the basis of individual papers, rather than on the journals in which they are published" (Haunschild & Bornmann, 2015). In order to make better assessment of research performance and research excellence, we propose the idea of article level evaluation system and database. Using metrics data at different time periods of 46 articles in one issue, we make empirical test of the article level evaluation method.

Firstly, the basic function of this evaluation system is to assess the qualities of articles. Based on article level evaluation, it is also available to assess the research excellence of scientists, journals, institutions and countries. For example, how many articles tracked in phase 3 and 4 are published by one specific institution? What are the top institutions in one specific field? Secondly, both scholarly and societal impact of articles are taken into account. Thirdly, using the article usage data and online mention data, we can make evaluation of newly published papers. At different phases after publication, the comprehensive score of the paper is calculated with different weights of metrics, so the score and rank of a paper in different phases change.

To accomplish this, the biggest problem needs to be solved is the availability of metrics data. The citation data could be obtained from Web of Science, Scopus, Google Scholar, etc. The online attention data, e.g., social media, news reports, Mendeley readership is also available from various but certain data sources. However, for the article usage data, only part of academic publishers and journals provide usage data to public, including Nature Publishing Group, Science, PLOS, Taylor & Francis, ACM Digital Library, IEEE Xplore Digital Library, etc. (Wang, Mao, Xu, & Zhang, 2013). For many others, e.g., Elsevier, Sage and Wiley, they may provide the metrics data of each article to some specific users and subscribers, but not free to public. If we want to evaluate all the papers whatever the publishers are, metrics data from publishers is indispensable.

With the movement from print to electronic publishing and the diversification of article-level-metrics, it is time to make change to the current research evaluation system. To better assess

scientists' research and satisfy the evaluation needs in many situations, ranging from funding decisions to hiring tenure and promotion, we need to build an article-level-evaluation system.

Limitation

In this study, we interpret the idea of building such a kind of system and make empirical study using a relative small size dataset, and we only track the metrics data of the sample articles in the last two years. To build the article-level-evaluation system is not an easy job, of course there are lots of problems need to be solved, including a bigger dataset, longer time period, more detailed metrics and maybe more scientific weighting methods, but we think it is the right way to make assessment of research, we are moving on the right direction.

Acknowledgments

The work was supported by the project of ''National Natural Science Foundation of China'' (61301227), the project of "Growth Plan of Distinguished Young Scholar in Liaoning Province" (WJQ2014009), and the project of "the Fundamental Research Funds for the Central Universities" (DUT15YQ111).

References

Alberts, B. (2013). Impact factor distortions. *Science*, *340*(6134), 787-787.

Bordons, M., Fernández, M. T., & Gomez, I. (2002). Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics*, *53*(2), 195-206.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, 295(1), 90-93.

Haunschild, R. & Bornmann, L. (2015). Publishing: Criteria for Nature Index questioned. *Nature*, *517*(7532), 21-21.

Kwok, R. (2013). Research impact: Altmetrics make their mark. *Nature*, 500(7463), 491-493.

Nature (2014). Launch of the Nature Index. Retrieved December 15, 2014 from: http://www.nature.com/news/launch-of-the-nature-index-1.16310

Opthof, T. (1997). Sense and nonsense about the impact factor. *Cardiovascular Research*, *33*(1), 1-7.

PLoS_Medicine_Editors. (2006). The impact factor game. *PLoS Medicine*, 3(6), e291.

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. In (Vol. 2014). Retrieved June 14, 2015 from: http://altmetrics.org/manifesto.

Saaty, T. L. (1980). *The analytic hierarchy process: planning, priority setting, resources allocation*. New York: McGraw.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, *314*(7079), 497.

Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics*, *98*(2), 1131-1143.

Wang, X., Liu, C., Fang, Z., & Mao, W. (2014). From attention to citation, what and how does altmetrics work? *arXiv preprint arXiv:1409.4269*.

Wang, X., Mao, W., Xu, S., & Zhang, C. (2013). Usage history of scientific literature: Nature metrics and metrics of nature publications. *Scientometrics*, *98*(3), 1923-1933.

Wang, X., Mao, W., Zhang, C., & Liu, Z. (2013). The diffusion of scientific literature in web. In S*TI 2013 – 18th International Conference on Science and Technology Indicators* (pp. 415-426). Berlin.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, *101*(2), 1-23.

Interdisciplinarity and Impact: Distinct Effects of Variety, Balance, and Disparity

Jian Wang[1], Bart Thijs[1] and Wolfgang Glänzel[1]

{Jian.Wang, Bart.Thijs, Wolfgang.Glanzel}@kuleuven.be
[1] KU Leuven (Belgium)

Abstract

Interdisciplinary research is increasingly recognized as the solution to today's challenging scientific and societal problems, but the relationship between interdisciplinary research and scientific impact is still unclear. This paper studies the relationship between interdisciplinarity and citations at the paper level. Different from previous literature compositing various aspects of interdisciplinarity into a single indicator, this paper uses factor analysis to uncover distinct aspects of interdisciplinarity and investigates their independent dynamics with scientific impact. Three uncovered factors correspond to variety, balance, and disparity. Subsequently, we estimate Poisson models with journal fixed effects and robust standard errors to investigate the relationship between these three factor and citations. We find that the number of citations (1) increase at an increasing rate with variety, (2) decrease with balance, and (3) increase at a decreasing rate with disparity. These findings have important implications for interdisciplinarity research and science policy.

Conference Topic

Science policy and research assessment

Introduction

Interdisciplinary research has been increasingly viewed as the remedy for the challenging contemporary scientific and societal problems. As important ideas often transcend the scope of a single discipline, interdisciplinary research is the key to accelerate scientific discoveries and solve societal problems. Given the normative interest in and the policy push for interdisciplinary research, it's important to empirically investigate the consequences of interdisciplinary research. Bibliometric studies have explored the relationship between interdisciplinary research and citation impact, but findings are mixed. For example, Steele and Stier (2000) found a positive effect of interdisciplinarity on citation impact for environmental sciences papers, where interdisciplinarity was measured as the disciplinary diversity of the cited references. Rinia, van Leeuwen, van Vuren, and van Raan (2001) studied physics programs in the Netherlands and operationalized interdisciplinarity as the ratio of non-physics publications. They found significantly negative correlations between interdisciplinarity and non-normalized citation-based metrics, but correlations became insignificant when field-normalization took place. Levitt and Thelwall (2008) found that interdisciplinary papers received fewer citations in life and physical sciences but not in social sciences, and interdisciplinary papers were defined as papers published in journals assigned to multiple subject categories. Larivière and Gingras (2010) analyzed all Web of Science (WoS) articles published in 2000, measured interdisciplinarity as the percentage of its cited references to other disciplines, and found an inverted *U*-shaped relationship between interdisciplinarity and citations.

One possible explanation for these conflicting results pertains to their different choices of the interdisciplinarity measure. On the one hand, a number of interdisciplinarity indicators have been proposed, at various levels (e.g., paper, journal, institution, and fields) and using various bilometric information (e.g., disciplinary memberships of authors, published journals, or cited references). On the other hand, the concept of interdisciplinarity remains an abstract and complex one (Wagner et al., 2011). One useful conceptualization is to view interdisciplinarity

as the diversity of disciplines invoked in the research (Porter & Rafols, 2009; Stirling, 1998, 2007). Furthermore, diversity has three distinct components (Stirling, 2007, p. 709):

> *Variety is the number of categories into which system elements are apportioned. It is the answer to the question: 'how many types of thing do we have?'*
>
> *Balance is a function of the pattern of apportionment of elements across categories. It is the answer to the question: 'how much of each type of thing do we have?'*
>
> *Disparity refers to the manner and degree in which the elements may be distinguished. It is the answer to the question: 'how different from each other are the types of thing that we have?'*

Many studies have devoted to compositing all aspects of interdisciplinarity into one single indicator. However, this paper adopts an opposite approach: we decompose different aspects of interdisciplinarity and explore their unique relationships with citation impact, at the individual paper level. Given that interdisciplinarity is an abstract and multidimensional concept, there might not be a straightforward answer to the question of whether interdisciplinary research draws higher impact. Instead, we should ask the question in another way: what kinds of interdisciplinarity have positive/negative relationships with citation impact? In addition, nuanced understanding of the divergent dynamics underlying different aspects of interdisciplinarity is also important for informing interdisciplinary research and science policy.

Data and methods

We analyzed all the journal articles published in 2001 indexed in the Thomson Reuters Web of Science Core Collection (WoS). Only articles were analyzed, while all other document types such as reviews and letters were excluded. The year 2001 was chosen so that studied papers could have a sufficiently long period to accumulate their citations (Wang, 2013).

Interdisciplinarity measures

Following previous literature, we constructed interdisciplinarity measures for each individual articles based on the disciplinary profile of its cited references, since referencing to prior literature in various disciplines indicates drawing and integrating knowledge pieces from these disciplines. Specifically, we constructed interdisciplinarity measures based on the WoS subject categories (SCs) referenced by each article. Interdisciplinarity measures constructed in this paper are listed in Table 1, which have been commonly used in the literature (Leydesdorff & Rafols, 2011; Rafols et al., 2012; Stirling, 2007). Because the last two interdisciplinarity measures cannot be constructed if the focal article references fewer than two subject categories, we excluded these articles from the analysis. Nevertheless, regressions using the whole dataset for the other measures yielded consistent results. In total, our data have 646,669 papers.

Factor analysis

We used factor analysis to uncover components underlying these interdisciplinarity measures. The first step was to determine the number of factors to retain. A classic approach is Kaiser's eigenvalue greater than one rule (Kaiser, 1960). The idea is that the retained factor should explain more variance than the original standardized variables. Horn's parallel analysis

Table 1. Interdisciplinarity measures.

Measure	*Description*
Ratio of references to other subject categories	
Number of referenced subject categories	n
1 – Gini	$1 - \frac{\sum(2i - n - 1)x_i}{n \sum x_i}$ where i is the index, x_i is the number of references to the i-th subject category, and subject categories are sorted by x_i in non-decreasing order.
Simpson index	$1 - \sum {p_i}^2$ where $p_i = x_i/X$, and $X = \sum x_i$
Shannon entropy	$-\sum p_i log\ (p_i)$
Average dissimilarity between referenced subject categories	$\frac{1}{n(n-1)} \sum_{i \neq j} d_{ij}$ where d_{ij} is the dissimilarity between subject category i and j. Specifically, $d_{ij} = 1 - s_{ij}$, where s_{ij} is the cosine similarity between subject category i and j based on their co-citation matrix.
Rao-Stirling diversity	$\sum_{i \neq j} p_i p_j d_{ij}$

modified Kaiser's rule, where the criterion for each eigenvalue is different and also superior to one, and these criteria are obtained from a Monte-Carlo simulation (Horn, 1965). Cattell's scree test provided a graphical strategy: plotting the eigenvalues against the component numbers and searching for the elbow point (Cattell, 1966). However it does not yield a definitive number of factors to retain, which still relies on subjective judgments of the researcher. Recently, Raiche, Walls, Magis, Riopel, and Blais (2013) developed numerical solutions for Cattell's scree test: (1) the optimal coordinate solution for the location of the scree and (2) the acceleration factor solution for the location of the elbow. We implemented all these methods to determine the number of factors. After determining the number of factors to retain, we extracted these factors using the varimax rotated principal components method. In addition, the number of referenced subject categories is highly skewed, so its nature logarithm was used in the factor analysis.

Regression analysis

To study the relationship between interdisciplinarity and citation impact at the article level, we ran regressions, using the number of long-term citations (in a 13-year time window from 2001 to the end of 2013) as the dependent variable and the interdisciplinarity measures and extracted factors as explanatory variables.

For all our regressions, we incorporated journal fixed effects to control for (1) unobserved topic/subfield heterogeneities at a very refined level and (2) journal reputation effects (Judge et al., 2007). Therefore, we estimated the within-journal effects, in other words, we were evaluating the association between interdisciplinarity and citations among papers published in the same journal. In addition, the following variables were incorporated as controls: the number of authors, the number of countries, the number of pages, and the number of references. The numbers of authors, pages, and references are skewed so that their natural

logarithms were used in regression analyses. The number of countries is still highly skewed after logarithm transformation, so we created a dummy variable, international: 1 if the paper has authors from more than one country, and 0 otherwise. In our sample, about 19% of the papers are internationally coauthored.

Because citation counts are over-dispersed count variables, we used Poisson regression with robust standard errors, following previous literature (Hall & Ziedonis, 2001; Hottenrott & Lopes-Bento, In Press; Somaya, Williamson, & Zhang, 2007). An alternative is the negative binomial model. However, because the Poisson model is in the linear exponential class, Gourieroux, Monfort, and Trognon (1984) have shown that the Poisson estimator and the robust standard errors are consistent so long as the mean is correctly specified even under misspecification of the distribution, but the negative binomial estimator is inconsistent if the true underlying distribution is not negative binomial. Therefore, we adopted the Poisson model with robust standard errors in our empirical analysis. Furthermore, we incorporated journal fixed effects. Such fixed effects Poisson models can be fitted by conditioning out the individual fixed effects (Hausman, Hall, & Griliches, 1984).

Results

Decomposing interdisciplinarity

We used the following variables in the factor analysis: log number of referenced subject categories, ratio of references to other subject categories, 1 – Gini, Simpson index, Shannon entropy, average dissimilarity between referenced subject categories, and Rao-Stirling diversity. The first three eigenvalues are greater than 1, so 3 factors should be retained according to Kaiser's rule. Horn's parallel analysis also suggests 3 factors. Raiche's nongraphic solutions for Cattell's scree test lead to conflicting conclusions: the optimal coordinate approach suggests 3 factors, while the acceleration factor approach suggests 1 factor to retain. Considering (1) the consensus between the classic Kaiser's rule and Horn's parallel analysis, (2) the divergence in this recent nongraphic solution for Cattell's scree test, and (3) that the optimal coordinate solution actually agrees with the more conventional approaches. We decided to retain 3 factors. Subsequently, we extracted 3 factors using the varimax rotated principal components method, and the cumulative proportion variance explained is 0.89. Factor loadings are reported in Table 2. Simpson index and Shannon entropy have the highest loading on the first factor, which reflects the variety aspect of disciplinary diversity. 1 – Gini has the highest loading on the second factor, which reflects balance, and the average dissimilarity between referenced subject categories has the highest loading on the third factor, which reflects disparity. The results are also in line with Harrison and Klein (2007) that Simpson index and Shannon entropy reflect more on variety, while Gini reflects more on unbalance.

Table 2. Factor loading.

	Factor 1	*Factor 2*	*Factor 3*
ln(referenced SCs)	0.78	-0.59	0.15
Ratio oth-disc refs	0.67	0.35	-0.17
1 – Gini	-0.07	0.94	0.05
Simpson	0.93	-0.11	0.18
Shannon	0.91	-0.32	0.18
Avg dissimilarity	0.09	0.00	0.95
Rao-Stirling	0.77	0.04	0.59

Data sourced from Thomson Reuters Web of Science Core Collection.

Interdisciplinarity and impact

We first estimated the fixed effects Poisson models using the citation counts as the dependent variable and original interdisciplinarity measures as the independent variables (Fig. 1A, regression table not reported). The divergent results suggest that the low consensus in previous literature regarding the relationship between interdisciplinarity and citation impact may be partially explained by their different choice of the interdisciplinarity measures.

Table 3 reports fixed effects Poisson models using the extracted interdisciplinarity factors as independent variables. Variety, balance, and disparity are the three extracted factors, and they follow the standard normal distribution with mean equals to 0 and standard deviation equals to 1. Holding that the papers are published in the same journal, with the same number of authors, pages and references, and have the same status in terms of whether being internationally coauthored, the expected number of citations increases by 1.48% as variety increases by 1 standard deviation (column 1), decreases by 2.45% as balance increases by 1 standard deviation (column 3), and increases by 5.77% as disparity increases by 1 standard deviation. Squared terms are subsequently added to test the non-linearity in these relationships. On the one hand, the square terms of variety and disparity are significant, suggesting nonlinear relationships. On the other hand, the squared term of balance is insignificant, suggesting a simply linear relationship. Fig. 1B plots the estimated number of citations with variety, balance, and disparity, based on column 2, 4, and 6 in Table 3, respectively. Again, for these estimations, we fix journal fixed effect at 0, international at 0, and all other variables at their mean.

We observe that long-term citations increase at an increasing rate with variety, which is in line with the information processing perspective that cognitive variety is very important for creative and innovative work (Lee, Walsh, & Wang, In Press; Page, 2007; Simonton, 2003). For interdisciplinary research, integrating knowledge from more disciplines contributes to potentially more broadly useful outcomes.

We also observe a negative relationship between balance and citation impact, which is also in line with Uzzi, Mukherjee, Stringer, and Jones (2013) that a paper with both higher novelty and conventionality are more likely to be a top cited paper. In other words, a paper is more likely to be top cited if it is embedded at the core of a discipline (drawing most of its prior knowledge/references from one discipline) while at the same time borrows some knowledge from some remote disciplines. However, the reason for this negative association between long-term citations and balance is still unclear. On the one hand, it could be that interdisciplinary research driving evenly by different disciplinary logics is more likely to fail in integrating these logics into something useful. Therefore, having one disciplinary core and simultaneously borrowing knowledge from other disciplines is a more effective research strategy, compared with drawing knowledge evenly from multiple disciplines. On the other hand, it could be that the current science system is biases against balanced interdisciplinary research. There are anecdotes that balanced interdisciplinary research which truly transcend disciplinary boundaries is difficult to evaluate and more likely to be unnoticed, simply because most scientists are trained within a discipline and unable to realize its value, although such balanced interdisciplinary research is very novel and broadly useful.

In addition, we observe that the number of citations increases with disparity but at a decreasing rate. This is in line with the combinatorial novelty literature that combining more remote disciplines is more novel than combining neighboring disciplines (Lee et al., In Press; Uzzi et al., 2013). Furthermore, there is a rather complex dynamics between novelty and impact. On the one hand, novelty is important for generating impact. On the other hand, a highly novel paper might not be useful or helpful for other scientists to further build on it, and therefore would fail to generate high impact (Latour & Woolgar, 1986; Merton, 1973; Whitley, 2000). We do observe that that the marginal return from disparity is decreasing. It's

possible that the effect of disparity on long-term citations might turn into a negative one after certain point, but this threshold is about six standard deviations above the mean, which is beyond the maximum disparity value in our data.

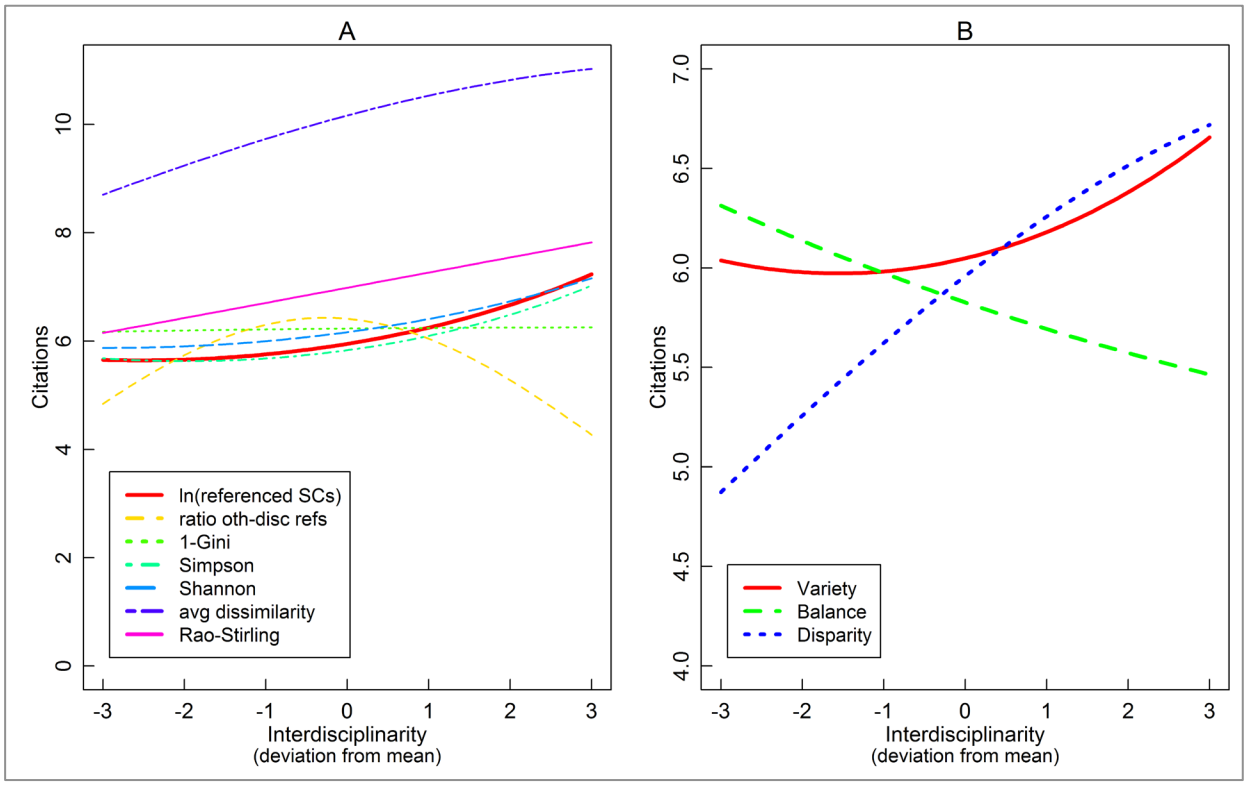


Figure 1. Interdisciplinarity and citations. Data sourced from Thomson Reuters Web of Science Core Collection.

Conclusions

This paper studies three different aspects of interdisciplinarity and investigates their distinct relationships with citation impact. The factor analysis extracts three main factors underlying various interdisciplinarity measures, and these three factors correspond to variety, balance, and disparity. Regression analysis further uncovers their different relationships with long-term citation impact: citations (1) increase at an increasing rate with variety, (2) decrease with balance, and (3) increase at a decreasing rate with disparity.

This paper contributes to future interdisciplinarity research and science policy. First, we advocate the idea of using different interdisciplinarity measures in different contexts. This paper demonstrates that various interdisciplinarity measures bear non-identical relationships with citation impact. Interdisciplinarity is an abstract and multidimensional concept, and different aspects of interdisciplinarity may (1) respond to certain individual, team, or institutional factors in completely different ways, and (2) have unique consequences in terms of usefulness or impact. Furthermore, various theories which might shed light on interdisciplinarity research have their own unique focuses. For example, the information processing perspective focuses on cognitive variety, while the combinatorial novelty literature emphasizes disparity. Therefore, it's important to choose a suitable interdisciplinarity measure consistent with the invoked theory and focal research question.

Table 3. Fixed effects Poisson models: interdisciplinarity and long-term impact (N = 646223).

	Citations							
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	*(7)*	*(8)*
ln(authors)	0.1588***	0.1586***	0.1600***	0.1600***	0.1590***	0.1586***	0.1578***	0.1575***
	(0.0105)	(0.0105)	(0.0106)	(0.0106)	(0.0110)	(0.0110)	(0.0107)	(0.0107)
International	-0.0009	-0.0008	-0.0013	-0.0013	-0.0025	-0.0025	-0.0023	-0.0022
	(0.0130)	(0.0130)	(0.0130)	(0.0130)	(0.0135)	(0.0135)	(0.0133)	(0.0133)
ln(pages)	0.4054***	0.4055***	0.4022***	0.4019***	0.3958***	0.3963***	0.3965***	0.3965***
	(0.0295)	(0.0295)	(0.0295)	(0.0294)	(0.0301)	(0.0302)	(0.0300)	(0.0300)
ln(refs)	0.3021***	0.3013***	0.2868***	0.2871***	0.3056***	0.3045***	0.2855***	0.2836***
	(0.0078)	(0.0077)	(0.0105)	(0.0105)	(0.0082)	(0.0083)	(0.0118)	(0.0119)
Variety	0.0148*	0.0162*					0.0137^{+}	0.0154^{+}
	(0.0061)	(0.0064)					(0.0078)	(0.0083)
Variety^2		0.0052*						0.0044^{+}
		(0.0026)						(0.0026)
Balance			-0.0245**	-0.0241**			-0.0194^{+}	-0.0194^{+}
			(0.0074)	(0.0073)			(0.0106)	(0.0108)
Balance^2				0.0009				0.0021
				(0.0033)				(0.0030)
Disparity					0.0577***	0.0535***	0.0528***	0.0488***
					(0.0075)	(0.0074)	(0.0088)	(0.0087)
Disparity^2						-0.0045^{+}		-0.0036
						(0.0025)		(0.0025)
Journal fixed effects	YES	YES	YES	YES	YES	YES	YES	YES
Log pseudolikelihood	-8642990	-8642683	-8642595	-8642588	-8629711	-8629503	-8628738	-8628365
χ^2	2946***	2957***	2967***	2961***	4450***	4438***	4552***	4807***

Cluster-robust standard errors in parentheses.
*** p<.001, ** p<.01, * p<.05, $^{+}$ p<.10.
Data sourced from Thomson Reuters Web of Science Core Collection.

Second, this paper suggests a more refined policy agenda for encouraging interdisciplinary research. This paper pushes forward the research on the relationship between interdisciplinarity and scientific impact: from a dichotomous question of whether interdisciplinary research draws higher impact towards a more complicated question about differentiated dynamics underlying different aspects of interdisciplinarity. Answers to this more complicated question is also important for more effective science policies. As science increasingly deals with boundary-spanning problems, various policy and funding initiatives have been developed to encourage interdisciplinary research, such as the US National Science Foundation (NSF) solicited interdisciplinary programs, the US National Institutes of Health (NIH) common fund's interdisciplinary research program, European Research Council (ERC) synergy grants, and UK Research Councils' cross-council funding agreement. However, interdisciplinarity is an abstract and multidimensional concept, and nuanced understanding of these different dimensions and their consequences are important for effective policies. Specifically, the positive relationship between variety and citation impact demonstrates the benefits of cognitive variety for creative work. Therefore, policy and funding initiatives can encourage research across more disciplinary boundaries and integrating knowledge from more disciplines. Furthermore, the positive relationship between disparity and citation impact also suggests potential improvements from encouraging interdisciplinary research across more remotely connected disciplines. However, since the positive marginal effect is decreasing, the policy might not want to push too far. It's possible that disparity effect on citations might turn into a negative one when the disparity is too high, that is, integrating disciplines too far apart may fail to find a common ground to produce something useful. In addition, the negative relationship between balance and citation impact may suggest that the most effective interdisciplinary research strategy in terms of generating impact is to have one disciplinary core and simultaneously borrow knowledge from some other disciplines, instead of drawing knowledge evenly from multiple disciplines without a disciplinary core. It's possible that research driving evenly by different disciplinary logics fails to integrate these logics into something useful. On the other hand, this might also suggest that balanced interdisciplinary research is biased against in the current discipline-based science system, in which scientists are mostly trained within a single discipline and therefore fail to realize the value of balanced interdisciplinary work which truly transcends interdisciplinary bounties. However, further research is required to better understand this problem. Specifically, to claim the bias against balanced interdisciplinary research, we need to estimate the unbiased should-be scientific impact first and then compare it with the observed citations. To recommend policies encouraging unbalanced instead of balanced interdisciplinary research, we would also need to test the usefulness or value of the papers directly, instead of only examining citation counts.

Acknowledgments

The authors thank You-Na Lee, Diana Hicks, Paula Stephan, and Reinhilde Veugelers for their helpful comments and suggestions.

References

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica, 52*(3), 701-720.

Hall, B.H. & Ziedonis, R. H. (2001). The patent paradox revisited: an empirical study of patenting in the US semiconductor industry, 1979-1995. *Rand Journal of Economics, 32*(1), 101-128.

Harrison, D.A. & Klein, K.J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review, 32*(4), 1199-1228.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents R&D relationship. *Econometrica, 52*(4), 909-938.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Hottenrott, H. & Lopes-Bento, C. (In Press). Quantity or quality? Knowledge alliances and their effects on patenting. *Industrial and Corporate Change*.

Judge, T. A., Cable, D.M., Colbert, A.E., & Rynes, S.L. (2007). What causes a management article to be cited: Article, author, or journal? *Academy of Management Journal, 50*(3), 491-506.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141-151.

Larivière, V. & Gingras, Y. (2010). On the relationship between interdisciplinarity and scientific impact. *Journal of the American Society for Information Science and Technology, 61*(1), 126-131.

Latour, B. & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Lee, Y.-N., Walsh, J.P., & Wang, J. (In Press). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*.

Levitt, J. M. & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science and Technology, 59*(12), 1973-1984.

Leydesdorff, L. & Rafols, I. (2011). Indicators of the interdisciplinarity of journals: Diversity, centrality, and citations. *Journal of Informetrics, 5*(1), 87-100.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago, IL: University of Chicago Press.

Page, S. E. (2007). *The difference: how the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.

Porter, A. & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics, 81*(3), 719-745.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management. *Research Policy, 41*(7), 1262-1282.

Raiche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 9*(1), 23-29.

Rinia, E., van Leeuwen, T.N., van Vuren, H.G., & van Raan, A.F.J. (2001). Influence of interdisciplinarity on peer-review and bibliometric evaluations in physics research. *Research Policy, 30*(3), 357-361.

Simonton, D.K. (2003). Scientific creativity as constrained stochastic behavior: the integration of product, person, and process perspectives. *Psychological Bulletin, 129*(4), 475-494.

Somaya, D., Williamson, I.O., & Zhang, X.M. (2007). Combining patent law expertise with R&D for patenting performance. *Organization Science, 18*(6), 922-937.

Steele, T.W., & Stier, J. C. (2000). The impact of interdisciplinary research in the environmental sciences: a forestry case study. *Journal of the American Society for Information Science, 51*(5), 476-484.

Stirling, A. (1998). On the economics and analysis of diversity. *SPRU Electronic Working Papers, 28*.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707-719.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*(6157), 468-472.

Wagner, C.S., Roessner, J.D., Bobb, K., Klein, J.T., Boyack, K.W., Keyton, J., . . . Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics, 5*(1), 14-26.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics, 94*(3), 851-872.

Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences* (2nd ed.). Oxford, UK; New York, NY: Oxford University Press.

The Evaluation of Scholarly Books as a Research Output. Current Developments in Europe

Elea Giménez-Toledo[1], Jorge Mañana-Rodríguez[1], Tim Engels[2], Peter Ingwersen[3], Janne Pölönen[4], Gunnar Sivertsen[5], Frederik Verleysen[6] and Alesia Zuccala[7]

{elea.gimenez, jorge.mannana}@cchs.csic.es
[1]Centro de Ciencias Humanas y Sociales, ÍLIA Research Group, CSIC, Albasanz Street, 28037, Madrid (Spain)

[2]*tim.engels@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp (Belgium); Antwerp Maritime Academy, Noordkasteel Oost 6, B-2030 Antwerp (Belgium)

[3]clb798@iva.ku.dk, [7]*spl465@iva.ku.dk*
[3, 7]Royal School of Library and Information Science, University of Copenhagen (Denmark)

[4]*janne.polonen@tsv.fi*
Publication Forum, Federation of Finnish Learned Societies, Snellmaninkatu 13, 00170 Helsinki (Finland)

[5]*gunnar.sivertsen@nifu.no*
NIFU Nordic Institute for Studies in Innovation, Research and Education, PO Box 5183 Majorstuen, 0302 Oslo (Norway)

[6]*frederik.verleysen@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp (Belgium)

Abstract

The relevance and value of books in scholarly communication from both sides, the scholars who chose this format as a communication channel and the instances assessing the scholarly and scientific output is undisputed. Nevertheless, the absence of worldwide comprehensive databases covering the items and information needed for the assessment of this type of publication has urged several European countries to develop custom-built information systems for the registration of books, weighting procedures and funding allocation practices enabling a proper assessment of books and book-type publications. For the first time, these systems make the assessment of books as a research output feasible. This paper resumes the main features of the assessment systems developed in five European countries / regions (Spain, Denmark, Flanders, Finland and Norway), focusing on the processes involved in the collection and processing of data on books, weighting, as well as their application in the context of research funding assessment.

Conference Topic

Science policy and research assessment and/or University policy and institutional rankings

Introduction

Scholarly books are key for the communication of research outputs in Social Sciences and Humanities (Hicks, D., 2004; Thompson, 2002; Engels, Ossenbklok & Spruyt, 2012). At the same time, performance-based assessment and funding allocation systems, as well as evaluation exercises at an individual level are widespread throughout Europe, affecting all instances of universities and research institutions (Hicks, D., 2012; Frølich, N., 2011). Despite developments such as Book Citation Index (Adams & Testa, 2011) there still exist a clear need for comprehensive databases collecting 'quality' indicators for books and book publishers. Quality in books is a multi-faceted concept and translating it into indicators is a

difficult task, in many occasions closely oriented to the specific research and assessment policies of each country. This diversity at the policy level is matched by an intrinsic heterogeneity of scholarly books themselves (e.g. disciplines, languages, formats, peer review and other editorial standards, etc.). In the past, the vast variety of books has made their reliable and comprehensive registration notoriously difficult and, consequently, their inclusion in research assessments unrewarding. By introducing the information systems presented in this paper, five European countries/regions have sought to redress the balance.

Objectives

The aim of this paper is to compare different approaches for assessing books across Europe. To do so, the context of each assessment exercise -where books evaluation occurs- is presented. The existence of valid peer review processes, the prestige of book publishers or the division in tiers according to the quality of the communication channel and the specific features of each discipline are some of the elements on which Spain, Denmark, Flanders, Finland and Norway have developed assessment systems for books. These developments are the result of applied research and also the object of a research-in-progress. This paper summarizes the main features of the current registration and assessment systems developed in the five countries in their present state. After a detailed discussion of each system, preliminary conclusions are presented, as well as a perspective on possible future developments.

Results

Scholarly Book's evaluation practices at the micro level

Spain

Scholarly books are taken into account in various assessment processes on the research outputs of scholars. As an example, both ANECA and CNEAI (Spanish assessment agencies) include various aspects of books and book publishers among their assessment criteria at the individual level. One of them is the prestige of the publisher (the latest, being CNEAI Resolution of November 26, 2014, but included as quality criteria various years backwards). Given the lack of specific data on the prestige of book publishers, the Research Group on Scholarly Books (ÍLIA) at CSIC developed Scholarly Publishers Indicators (SPI) on the grounds of the research conducted in previous years (Giménez-Toledo & Román Román, 2009). SPI ranks the perceived prestige of book publishers in the social sciences and humanities (SSH), both Spanish and non-Spanish, according to the scores resulting from an extensive survey to Spanish lecturers, researchers and scholars specializing in all fields of SSH. The system is based on more than 3,000 usable responses in 2012 and almost 3,000 in 2013. The responses are given to the question of which are the first, second or third (and from first to tenth in the 2013 edition) most prestigious book publishers in the responder's field; only specialists with positive assessment of their research are susceptible of being included among the respondents. Once collected, the responses are summarized using a simple weighting algorithm based on the share of scores in each position (1^{st}, 2^{nd}, etc.). The results are summarized in an indicator: ICEE. This indicator serves as a ranking item, both at the general level and specifically for each discipline, since the assigned weights are related to each discipline's distribution of scores (Giménez-Toledo, Tejada-Artigas & Mañana Rodríguez, 2012). The weighting procedure involves no arbitrary intervention from its designers and permits certain normalization per discipline. The ranking is publicly available at (http://ilia.cchs.csic.es/SPI/) and the users can access both discipline-level and general rankings for Spanish and non-Spanish publishers.

The main advantage of this system is the wide population on which it is based (more than 11,000 experts), while the main disadvantage lies in the difficulty to control for possible bias in the surveying process. The ranking was first used for assessment purposes in 2013 and is increasingly being included in the current evaluation framework as a reference for the assessment of SSH books and book chapters, together with other criteria. It is important to note that SPI is a reference tool for assessment exercises. It is meant to inform, not to perform, the research evaluation.
SPI also includes interactive charts as well as a 'specialization profile' of publishers obtained from the DILVE database (collecting the editorial production of Spanish publishers). Specialization is a point where evaluation agencies may focus their attention. In progress is the research into the use of different peer review systems with the use of surveys to book publishers as well as information about the transparency of their websites. These are qualitative indicators which aim is to serve as supporting information in the assessment processes.

Book's evaluation practices at meso or macro-level

Denmark

The performance indicator model (BFI/BRI, the Bibliometric Research Indicator) was started up in 2009. For each year 68 groups of academics selected by the Danish Research Agency from the Danish universities list all available knowledge resources and assign points to peer reviewed journals, publishers and conferences that publish scientific material authored by Danish academics from the previous year. Each of the 68 groups represents an academic field or specialty. The bibliometric research indicator takes into account published peer reviewed research and review articles, monographs as well as anthology and proceedings papers published by the Danish research institutions, which provide the input metadata for the system. In the period 2008–2012 proceedings (and anthology) papers were assigned .75 points. Journal articles received 1.0 point in Level 1 journals and 3.0 points in Level 2 journals, i.e. the leading journals of a field as judged by the relevant researcher group and covering maximum 20% of the field journal output. From 2013 proceedings papers and articles receive similar points as journal articles, depending on the level of the conference or publisher, as assessed by the relevant academic group. Monographs are assessed according to two publisher levels, Level 1 (5 points) and Level 2 (8 points). Anthology papers and chapters receive 0.5 and 2 points depending on publisher level. For each document the points are fractionalized (min 0.1) according to number of collaborating universities, including non-Danish universities. The model encourages collaboration by multiplying the institutional fraction by 1.25. The previous year's cumulated points per university is used to distribute a substantial portion (in 2013 it was 25%) of public basic research funding among the universities the following year. Only the cumulated results are publicly available per university and major academic area, such as the Humanities, Social Sciences, Natural Sciences or Medicine/Health sciences via the Danish Research Agency's web page (https://bfi.fi.dk/). The intermediate or more detailed publication point distributions and document lists per unit and department will be publicly accessible from 2015. This is in difference to Norway where no multiplication of fractions takes place and all the documents and their point assignments are transparent as well as publicly accessible through an open access database. In the Finnish system and in Belgium the Flemish BOF-key applies whole counting at the institutional level (Debackere & Glänzel, 2004; Engels, Ossenblok & Spruyt, 2012). The output of the Danish BRI system can, as a spin-off, be used for assessment purposes. See also Ingwersen & Larsen (2014).

Flanders (Belgium)

The Flemish Academic Bibliographic Database for the Social Sciences and Humanities ('Vlaams Academisch Bestand voor de Sociale en Humane Wetenschappen', or VABB-SHW) has been developed to allow for the inclusion of the peer reviewed academic publication output in the Social Sciences and Humanities (SSH) in the regional performance-based research funding model. As such, in 2015 the VABB-SHW accounts for 6.62% of the University Research Fund (or BOF), distributing over 150 million euro per year over the five universities. As the BOF-key is also re-used for the distribution of other research funding, the actual impact of the VABB-SHW is even greater. In a secondary role, the VABB-SHW supports research assessments at various levels. As all information in the VABB-SHW is available to both the universities and the Flemish national science foundation (FWO), data is harvested and integrated into each institution's repository. In a third role, the VABB-SHW's comprehensive publication coverage (peer reviewed or otherwise) allows for in-depth research on publication practices in the SSH (Engels, Ossenblok, & Spruyt, 2012; Verleysen, Ghesquière, & Engels, 2014). The database covers the comprehensive publication output of academic research in 16 SSH disciplines and 3 general categories. Three types of book publications are included: 1° monographs, 2° edited books, 3° book chapters, weighted 4, 1 and 1 for the funding model, respectively. Journal articles also receive a weight of 1 and proceedings papers a weight of 0.5. No prestige levels are distinguished. For funding calculation, a ten-year timeframe is used. For research purposes, coverage extends back to the year 2000. For books, four aggregation levels are in use: 1° publisher names (as collections of ISBN-roots), 2° book series, 3° books published in Flanders and labeled as Guaranteed Peer Reviewed Content (GPRC-label (Verleysen & Engels, 2013), and 4° individual books identified as peer reviewed by the Authoritative Panel ('Gezaghebbende Panel' or GP, a committee of full professors installed by the government and responsible for decisions regarding the content of the VABB-SHW). The information system is fed through a yearly upload (May 1st) of all SSH publications from the two preceding years newly registered in the five universities' academic bibliographies. Data is managed at the Flemish Centre for R&D monitoring (ECOOM), University of Antwerp, through its custom-built Brocade library services (http://en.wikipedia.org/wiki/Brocade_Library_Services). Each individual publication receives a unique identifier, contributing to maximum granularity and reliability of the data both for funding calculation as well as for retrieval and research. Consolidation processes making use of algorithmic identification allow a systematic de-duplication of records that are submitted more than once. Publications are identified algorithmically at the publisher, series or journal level by their ISBN-prefix or ISSN. Each year all new publishers, series, books and journals are classified by the Authoritative Panel as peer reviewed and presenting new content (or not). At the public interface www.ecoom.be/en/vabb, online access is provided to the database itself, lists of publishers, journals and series, explanation of procedures, FAQ's, and background information.

Finland

In Finland, the use of publications in the performance based funding model is based on two components: the publication metadata consisting of the entire output of universities, and a quality index of outlets. Universities have their own registries of publications, including peer-reviewed and non-peer-reviewed articles in journals, conferences and anthologies, as well as monographs. Universities report their publication data, with full bibliographic details, once a year to the ministry of education and culture (Puuska 2014). The publication data is processed (including deduplication) at CSC - IT Centre for Science, which is a company owned by the ministry. The bibliographic details of publications are matched against the list of serials, conferences and book publishers classified in three quality levels by 23 expert panels

coordinated by the Federation of Finnish Learned Societies (FFLS). This quality index of outlets is called Julkaisufoorumi (JUFO) -luokitus (Publication Forum Classification). The universities' publication metadata collected by the ministry is known as OKM-julkaisuaineisto (MinEdu publication data).

In the Publication Forum classification, published for the first time in 2012, the level 2 comprises 20 % of the leading serials and conferences and 10% of the leading book publishers (Auranen & Pölönen, 2012). Most peer-reviewed outlets belong to the level 1, and those that fail to meet the criteria of scientific publication channel are listed as the level 0. For serials there is also a level 3, in which are classified 25% of the level 2 titles, but in the funding model it is not differentiated from the level 2. Updated classifications have been published in the beginning of 2015. In the new classification, as in Denmark, the level 2 serials and conferences comprise at most 20% share of the world production of articles in each panel's field. The level 3 was added also for book publishers. The new classifications will be applied on articles and books published in 2015. The classification of book publishers is used specifically to determine the level of monographs and articles in anthologies when the publication does not come out in a book series or the series has not been classified. The main rule is that the Finnish book series are classified, while those of foreign book publishers are not classified separately.

In the current funding model for 2015 and 2016, which still uses the 2012 Publication Forum classifications, 13% of all budget-funding is allocated on basis of publications (Ministry of Education and Culture, 2014). The peer-reviewed articles in journals, conferences and anthologies published in the level 0 channels will have the weighting coefficient 1, those of the level 1 have the coefficient 1.5, and for the level 2 and 3 channels the coefficient is 3. The weighting coefficient of non-peer-reviewed (scholarly, professional and general public) articles is 0.1 regardless of outlet. Weighting coefficient of peer-reviewed monographs is four times higher than that of articles: 4 in the level 0, 6 in the level 1, and 12 in the level 2. For non-peer-reviewed monographs, as well as all edited volumes, the weight is 0.4. There is no fractionalization of co-publications at the institutional or author level. The Ministry has instituted a working-group to determine the weights and calculation method of publications used in the funding model from 2017 onwards.

The MinEdu publication data, which covers Finnish universities output since 2010, is openly available through Vipunen-portal (www.vipunen.fi) for statistics, as well as Juuli-portal (www.juuli.fi) for browsing the publication information. The quality index of outlets is openly available on the Publication Forum website (www.tsv.fi/julkaisufoorumi).

Norway

The Norwegian model (Sivertsen, 2010; Sivertsen & Larsen, 2012) consists of three main elements: 1) A national database containing comprehensive and unified bibliographic metadata for the peer reviewed literature in all areas of research; 2) a publication indicator making field-specific publishing traditions comparable in the same measurement; and 3) a performance based funding model.

The national database is called CRISTIN (Current Research Information System in Norway). It is shared by all research organizations in the public sector: universities, university colleges, university hospitals, and independent research institutes. The institutions provide quality-assured and complete bibliographic about articles in journals and series (ISSN), articles in books (ISBN), and books (ISBN) that can be included according to a definition of peer-reviewed scholarly literature.

The indicator is based on a division of publication channels (journals, series, book publishers) in two levels: level 1 and level 2. Level 2 contains the most selective international journals, series and book publishers and may not contain more than 20 per cent of the publications

worldwide in each field of research. Articles in journals and series are given 1 point on level 1 and 3 points on level 2. Articles in books (with ISBN only) are given 0.7 1 points on level 1 and 1 point on level 2. Monographs are given 5 points in level 1 and 8 points on level 2. The points are fractionalized in the level of institutions according to the institution's share of contributing authors.

Although less than two per cent of the total expenses reallocated by the use of the indicator in Norway, it has attracted a lot of attention among researchers and resulted in increased productivity (Aagaard et al., 2014).

Conclusions

One of the first conclusions which stand out is the lack of use of citation metrics in any of the five systems. This might be the result of a lack of fit, lack of acceptance or the irrelevance as a quality indicator for books of the traditional measures for journals. Another element is the incomprehensiveness for many scholarly fields of the current citation indexes. Equally remarkable is the clear convergence as regards criteria and procedures among the Nordic countries and Flanders, not only in the registration of books, but also in the funding and/or assessment policies making use of book data. For assessments, in Northern Europe data is used mainly at the institutional level, despite its collection and registration being nationally coordinated in the context of a performance-based research funding system. This is clearly not the case for Spain, where data is used for assessments at the individual level, while university budgets are not calculated in a performance-based, centralized system. Also, the different policies show great divergences regarding the much higher weight given to scholarly books in the Nordic systems, while in Spain the tendency is just the opposite (more weight is given to papers than is to books). It is also remarkable that the most frequently used aggregation level is that of book publishers, although in the case of Flanders the Guaranteed Peer Reviewed Content-label allows for the inclusion of individual books in the regional system as well, while Finland currently counts with a Peer Review Mark similar to the already mentioned, making feasible that possibility. This involves that the expected coherence in the practices underlying to the concept of quality is sufficient at the level of book publishers, since the congruent use of this level of aggregation (from which the positioning in tiers of each individual contribution is derived) is common to all systems analyzed. Nevertheless, future developments may well see a stronger interest in the registration of book data at lower aggregation levels as well (e.g. that of the book series), as this evidently implies a more fine-grained approach to the comprehensive registration and the validation in assessments of books. In Spain, that specific level of aggregation (book series) is the object of a current initiative by UNE (University Presses Union) in collaboration with three research teams.

Finally, it will be interesting to see whether the on-going internationalization of research and the growing collaboration between scholars worldwide will contribute to a greater harmonization at the European level of the assessment systems for books and book publishers. Such developments could indeed provide scholars with new opportunities to assert the (often under-rated) value of their books, although some hypotheses regarding the role of the book in the scholarly communication shall be addressed in the close future.

Acknowledgements

This research is partially the result of the project 'Evaluación de editoriales científicas (españolas y extranjeras) de libros en Ciencias Humanas y Sociales a través de la opinión de los expertos y del análisis de los procesos HAR2011-30383-C02-01 (Ministerio de Economía y Competitividad. Plan Nacional de I+D+I).

Table 1. Comparison of the main features of the information systems for the assessment of books.

ITEM	SPI	BFI/BRI*	VABB-SHW	MinEdu Data/JUFO	CRISTIN
Country	Spain	Denmark	Flanders	Finland	Norway
Reasons for its development	Assessment at the individual level and research evaluation (unknown uses at institutional level)	Research funds allocation among universities and measures of research activities at institutional levels.	Inclusion of the peer reviewed scholarly publication output in the regional performance-based research funding model.	Funding allocation, research information and quality promotion.	Research information and fund allocation in the public sector. National statistics.
Object of study/ aggregation level	Book publishers / specialization from book-level information.	Book publishers, books and book parts (anthologies); journal articles and proceeding papers.	Book publishers, book series, GPRC**-labeled books published in Flanders and individual books assessed by the Authoritative Panel.	Book publishers and monographic series / peer reviewed monographs and articles in books at university level.	Bibliographic references to all scholarly publications in books, book articles and journal papers.
Stage	Already published and applied in assessment.	Already published and applied in assessment and funding since 2009.	Applied for funding allocation and institution-level assessment since 2010.	Published in 2012 and applied in funding since 2015.	Applied in assessment and funding since 2005.
Coverage	All Spanish and non-Spanish book publishers mentioned by experts in each field.	All scholarly publishers worldwide with publications from Danish scholars since 2009.	The comprehensive peer reviewed publication output of academic research in the Social Sciences and Humanities since 2000.	National and international scholarly book publishers and Finnish book series	All scholarly publishers worldwide with publications from Norwegian scholars since 2004.
Information feeding the system	Survey to experts and book publishers / database analysis.	Metadata for scholarly publications from all Danish universities.	Yearly upload from the academic bibliographies of the five Flemish universities, of all newly registered publications of the previous two years.	Metadata for universities' scholarly publications and new additions suggested by researchers	Metadata for scholarly publications from all Norwegian institutions in (CRISTIN).
Information processing	Votes from respondents are summarized in the ICEE indicator. DILVE database is statistically analyzed. Surveys to book publishers are summarized. Done by ILIA research group (CSIC).	Quality level assessments of publishers and journals by 67 topical peer groups plus a central coordination council, providing authoritative lists from which each publication is assigned a score by the system.	Data input from the universities processed by ECOOM / University of Antwerp Scientific steering and assessment of publication channels by a central Authoritative Panel.	In order to assign weight to universities' publications in the funding model, the metadata of publications is collected and matched against the list of serials, conferences and book publishers classified in quality levels by 23 panels.	Input from the institutions of metadata for individual publications is connected to a centrally monitored dynamic register of approved scholarly publication channels (journals, series, and book publishers)
Operative results	Ranking of book publisher's prestige / specialization charts / peer review info.	Annual number of publications and number of publication points per university and per larger academic topic.	A growing database of 125,000 scholarly peer reviewed and other publications. Publicly available lists of assessed book publishers, book series, journals and conference proceedings.	List of quality-classified outlets and database of universities' all publications from 2011 that can be analyzed by type, field and outlet.	A database of so far 70,000 scholarly publications that can be analyzed by type, field, language, institution, and publication channel
Use for research assessment and aggregation level	Used at the individual level by ANECA and CNEAI, two Spanish assessment agencies.	Funding allocation in the following year; Institutional level; also used as promotion or 'extras' factor (local incentive). Individual level in the future.	Funding allocation to five universities; support of internal assessments at individual universities, and assessments by the Flemish national science foundation (FWO)	Funding allocation to universities; internal assessment and planning at universities (also funding allocation); use for assessment at individual level is discouraged.	Funding allocation, stats for field and/or institution research evaluation, administrative information at institutions and annual reports.
Public availability	Yes (from 2012)	Yes (from 2015)	Yes	Yes	Yes (from 2004)
Book / paper weighting	Approx. 1 to 3 (as defined by assessment agencies, but not by SPI)	From 5 to 8 and from 0.5 to 2 (anthology items) and from 1 to 3.	From 4 to 1 and from 1 to 0.5	From 0.4 to 12 and from 0.1 to 3.	From 8 to 3 and from 3 to 1.

* BFI/BRI = Bibliometric Forskningsindokator / Bibliometric Research Indicator, **GPRC = Guaranteed Peer Reviewed Content

References

Aagaard, K., C.W. Bloch, & J.W. Schneider. (2014). Impacts of Performance-based Research Funding Systems: the case of the Norwegian Publication Indicator. *Research Evaluation,* (forthcoming).

Adams, J. & Testa, J. (2011). Thomson Reuters Book Citation Index. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *The 13th Conference of the International Society for Scientometrics and Informetrics* (pp. 13–18). Durban, South Africa: ISSI, Leiden University and University of Zululand.

Auranen, O., & Pölönen, J. (2012). *Classification of scientific publication channels: Final report of the Publication Forum project (2010–2012)*. Federation of Finnish Learned Societies: http://www.tsv.fi/files/yleinen/publication_forum_project_final_report.pdf.

Debackere, K., & Glänzel, W. (2004). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics*, *59*(2), 253-276.

Engels, T.C.E., Ossenblok, T.L.B., & Spruyt, E.H.J. (2012). Changing publication patterns in the Social Sciences and Humanities, 2000–2009. *Scientometrics*, *93*, 373–390.

Frølich, N. (2011). Multi-layered accountability. Performance-based funding of universities. *Public Administration*, *89*(3), 840-859.

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation*, *22*(1), 64–77. doi:10.1093/reseval/rvs036.

Giménez-Toledo, E. & Román-Román, A. (2009). Assessment of humanities and social sciences monographs through their publishers: a review and a study towards a model of evaluation. *Research Evaluation*, 18(3), 201-213.

Hicks, D. (2004). The four literatures of social science. In H.F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 473–496). Dordrecht, The Netherlands: Kluwer Academic.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41(2), 251-261.

Ingwersen, P. & Larsen, B. (2014). Influence of a performance indicator on Danish research production and citation impact 2000–12. *Scientometrics*, DOI 10.1007/s11192-014-1291-x.

Ministry of Education and Culture (2014). *Greater incentives for strengthening quality in education and research: A proposal for revising the funding model for universities as of 2015*: http://www.minedu.fi/export/sites/default/OPM/Julkaisut/2014/liitteet/tr07.pdf?lang=fi.

Puuska, H.-M. (2014). *Scholarly Publishing Patterns in Finland: A comparison of disciplinary groups*. University of Tampere.

Schneider, J.W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway, *European Political Science*, *8*(3), 364-378.

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. *ISSI Newsletter*, *6*(1), 22–28.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics, 91*(2), 567- 575.

Thompson, J.W. (2002). 'The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship'. *Libri*, *52*, 121–36.

UNE. (2014). España crea un sello de calidad para reconocer la excelencia científica del proceso editorial de las colecciones publicadas por las universidades. http://www.une.es/Ent/Items/ItemDetail.aspx?ID=9610

Verleysen, F. T., Ghesquière, P., & Engels, T. C. E. (2014). The objectives, design and selection process of the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (VABB-SHW). In W.Blockmans, et al., (Eds.). *The use and abuse of bibliometrics* (pp. 115-125). Academiae Europaea; Portland Press.

Verleysen, F.T. & Engels, T.C.E. (2013). A label for peer-reviewed books. *Journal of the American Society for Information Science and Technology*, *64*, 428-430.

Zuccala, A., Costas, R., & van Leeuwen, T.N. (2010). Evaluating research departments using individual level bibliometrics. In: *Eleventh International Conference on Science and Technology Indicators* (p. 314).

Publications or Citations – Does it Matter? Beneficiaries in Two Different Versions of a National Bibliometric Performance Model, an Existing Publication-based and a Suggested Citation-based Model

Jesper W. Schneider

jws@ps.au.dk
The Danish Centre for Studies in Research and Research Policy, Department of Political Science, Aarhus University (Denmark)

Abstract

The paper discusses the adoption of the Norwegian Publication Model in a Danish context and examines arguments for supplementing or substitution the current mechanism where reward is based on publication activity with one based on citations. Based on national publication data from 2009 from the Danish model, belonging to the science and technology research area, and corresponding citation data, we examine the Danish universities' relative input when it comes to publications and subsequently examine the relative output from these publications, i.e., the "returns on investment" from the model, either the current publication points, or the alternative, citations. Findings support the claims that high-performing units would benefit more from a citation-based approach, but at the same time also show, contrary to what was conjectured, that in the present case the same university also benefits the most from the current publication model. Based on the findings, we discuss the publication versus citation-based models, or hybrids between them, and argue that citation-based models in performance-based funding context are harder to influence and most likely will support already existing cumulative advantages.

Conference Topic

Science policy and research assessment

Introduction

In recent decades several countries have introduced performance-based research funding among their universities (Hicks, 2012). The performance-based research funding systems (PRFS) vary considerably between countries, from panel-based peer review evaluations, to systems based on citation or publication metrics, or various hybrids of these three basic forms (see Hicks, 2012). Generally, peer review systems are considered superior to systems based on bibliometric indicators (see Gläser & Laudel, 2007). Nevertheless, large-scale panel evaluations are very expensive, and several *post hoc* comparisons between panel results and citation metrics, for example from the UK Research Assessment Exercises, suggest that the latter could be an effective, and cost-effective, supplement or even substitute to peer reviews (e.g., Oppenheim, 1996; Moed, 2008). Among PRFS based on bibliometric indicators, citation-based systems are considered by some to be superior due to the assumption that citation indicators to some extent are able to measure aspects of research quality by focusing on impact (Gläser & Laudel, 2007). But citation indicators also have obvious deficiencies especially when implemented in PRFS which in principle are supposed to cover all fields of research (Schneider, 2009). It is well-known that citation indicators are not equally valid across all fields of research and even where relevant, coverage in the citation databases is also restricted (Moed, 2005). Consequently, PRFS based on citation indicators severely restricts the measurable outcome of research basically to journal articles indexed in one of the two major citation databases. But there are other issues with citation indicators which can be considered inadequate when used in PRFS, especially when such systems are supposed to (re)distribute funding on a regular basis, most often annually, and at the same time also give

universities (and their researchers) incentives to improve performance (e.g., Gläser & Laudel, 2007; Schneider, 2009). Citation indicators reflect research done in the past often a considerable number years prior to the actual funding year. It is also very difficult to directly influence citations when conceived of as an incentive system, in fact the well-known cumulative advantages could be detrimental to such an incentive system if it is supposed to be fair for all involved (Merton, 1988). Such features are seen by some as undesirable if PRFS as supposed to cover all research fields with their different publication traditions, and be able to reflect recent research performance in a dynamic model, as well as give transparent behavioural incentives to change performance (Schneider, 2009; Hicks, 2012).

PRFS based on publication activity have been introduced as an alternative to citation-based systems (Butler, 2002; Schneider, 2009). There are some apparent "benefits" with publication-based systems compared to citation-based systems. They can reflect short-term research activity making them more up-to-date when it comes to redistributing funding. In principle they can encompass all desired publication types and they can provide straightforward behavioural incentives. But it is important to emphasise that the two approaches measure different constructs. It would be naïve to suppose that incentives directed at publication behaviour, i.e., quantity and/or supposed status of the publication outlet, encompass the same aspects of perceived "quality" that citation impact is thought to reflect (Schneider, 2009). Experiences from Australia testify to this. In a succession of papers, Linda Butler demonstrated how researchers in Australia responded when funding, at least partially, was linked to publication counts undifferentiated by any measure of supposed "quality" in the early 1990s (e.g., Butler, 2003a; Butler, 2003b). Australian publication output increased considerably with the highest percentage increase in lower impact journals. For a consecutive number of years, this lead to a general drop in overall citation impact for Australia. Since Butler's documentation of the adverse effects, the experience from Australia has stood as a "warning" for what would most likely happen if funding was linked to publication activity. Nonetheless, in the early 2000s a so-called "quality reform" of the higher education sector in Norway introduced a PRFS where publication activity again was linked to funding. The main political intention with the model was in fact to encourage more research activity and thereby also more publication activity, and preferably more international publication activity, in the university sector[1].

The so-called *Norwegian Publication Model* (NPM) is interesting in in relation to PRFS. Obviously, the designers of the NPM were well-aware of the adverse behavioural effects documented in the Australian case. As a consequence, a slightly more sophisticated model was developed (Schneider, 2009; Sivertsen, 2010). A primacy of the model was to reflect the encouragement to publish in international outlets (i.e., international journals and academic book publishers) and at the same time to counter so-called adverse publication effects like the Australian case, where researchers seek to publish more but with less effort. Hence, a differentiated publication model was constructed where publication channels were classified on two levels. Level one comprises in principle all scholarly eligible publication channels, where eligibility criteria are some basic norms such as a standard external peer review process. Level two, is an exclusive number of publication channels, which are deemed to be leading in a field and preferably with an international audience. Level two is exclusive in as much as the number of publication channels designated at any given time to this level should produce roughly one-fifth of the publications produced in a field "world-wide". Correspondingly, three different types of scholarly publications are included in the model: journal publications (articles and reviews), articles in books (contributions to anthologies and

[1] http://www.uhr.no/documents/Rapport_fra_UHR_prosjektet_4_11_engCJS_endelig_versjon_av_hele_oversettelsen.pdf.

conference papers) and books. A two dimensional point system was implemented where the different publication types yield different points within the same level and between the two levels depending on the outlet status. Hence, the basic idea behind this two-tiered classification system is that publications on level two receive more publication points than publications on level one. Finally, publication points are fractioned $1/n$ so that an institution eventually receives $1/n$ points depending on their number of contributing authors.

Eventually the annual sum of publication points for an institution is exchanged for funds, where the exchange rate is determined by the amount of money available for redistribution and the total number of publication points in the system in a given year. A noticeable assumption in the NPM is that publication behaviour, publication activity and publication types across all fields can be treated identically. Consequently, all research fields' eligible research publications are included in the model, which for example means that a level one journal article with one author is worth the same in physics and literature studies. It is assumed that the differentiated point system together with fractionalized counting will level out the major differences in publication behaviour between the fields and also to some extent will discourage researchers to speculate in "easy publications" resulting in a levelling out effect at the aggregate level. Consequently, in the Norwegian PRFS funding is competitive not only between institutions but also across all fields. Hence, the subject composition within and between the research institutions is interesting as performance improvement in one major area, in principle can lead to improved funding at the expense of another major area due to the basic zero-sum situation.

The NPM has recently been "adopted" in several European countries, for example in Denmark, Finland and Flanders (Hicks, 2012; Verleysen, Ghesquière & Engels, 2014). In the present paper we look at the "adoption" of the indicator in Denmark and examine the overall distributional consequences of focusing on publication activity and not impact.

It is important to accentuate that in Norway the publication model was to a large extent developed to support overall political goals, i.e., more international research activity. As it were, Norway's internationalization in research and general citation impact, were considerably lower, than for example Denmark, at the time of the introduction of the model. Since then Norway's international publication output has risen considerably, albeit rise in citation impact has been meagre (e.g., Aagaard, Bloch & Schneider, 2015). Nonetheless, the NPM was developed and implemented with a legitimate goal which to some extent seems to have been achieved seen from the national policy perspective.

During a reform of the Danish research funding system in the mid-2000s it was decided to implement a PRFS officially in order to enlarge competition among universities for funding, although the board of university rectors probably more saw it as management tool that should legitimize their overall research activity to the public (Schneider & Aagaard, 2012). The political process leading to the "adoption" of the NPM in Denmark is complex and documented in Aagaard (2011). It is not totally clear why the choice fell upon the NPM, although its coverage of all areas, transparency and clear incentive system were no doubt deemed viable, yet some actors actually indicated that it would probably be "the one that would cause the least damage" (Aagaard, 2011). Most interesting, contrary to Norway, there were no immediate strategies or goals for research and publication behaviour behind the "adoption" of the NPM in Denmark.

Denmark was the first country to adopt the NPM at a time when the model was still in its infancy in Norway and little empirical evidence of its potential effects was available. The NPM was adopted with very few moderations, as if the model was a one-size fit all package suitable for all contexts. Most notably, the simple two-tiered classification system was kept and considerations about expanding or adapting the classification to a Danish context were not done. Nevertheless, some seemingly minor moderations turned out to be imperative,

including a maximum fractionalization of contributions at 1/10th; but perhaps most important, performance-based publication activity was locked between the major research areas: science and technology, health sciences, social sciences and humanities. Consequently, in the Danish adoption of the NPM, funding is not competitive across areas only within areas. Further, politically it was decided to more or less keep the old annual allocation model between the areas which effectively meant that a publication point, contrary the Norwegian PRRS, have different monetary values across the four main research areas. This is an extremely important deviation from NPM and it gives rise to some questions about the Danish adoption of the NPM, popularly known by the acronym BFI (bibliometric research indicator).

One can argue that the model is transparent, seemingly coherent and all-inclusive when it comes to research areas. All areas are measured with same indicator. But since competition is restricted to within areas and as a consequence publication points have different values across areas, one could also ask why the model still assumes equality of publication practices across areas? And to go further, with the locking of the competition to within areas, there is basically no reason why fields where citation analysis could be a reasonable and indeed preferred indicator could implement such devices either in combination with a publication model or alone. Of course the latter would muddle the overall model, although it would probably satisfy many of the critics of the publication-based model, arguing for more emphasis on impact.

Indeed, the Technical University of Denmark (DTU) has been an ardent critic of the adoption of the NPM in Denmark. A common argument goes: Why implement an incentive model that reward publication activity in international outlets when "we" already do that and do it well? More generally the critics stated that the behavioural goals with the model in Norway were irrelevant in a Danish context, because Denmark, contrary to Norway, has 1) for decades consistently been among the top five highest performing countries when it comes to impact; 2) has consistently four of its eight universities in the top 200 of the Leiden Ranking[2]; and 3) the Danish research system has had a long trajectory of internationalization (e.g., Karlsson & Persson, 2012). According to DTU, what should be procured and rewarded is impact and not publication activity. While the argument is relevant, it is also self-serving. DTU happens to be the highest performing Danish university when it comes to impact and is ranked in the top 50 of the Leiden Ranking. DTU has a very strong focus upon science and technology and close to no medical, social or humanistic research activities. Also, DTU has the lowest student to researcher ratio in Denmark. Obviously, DTU would fit very-well to a model based on citations. DTU has continued the criticism over the years claiming that they are the actually "losers" in the current Danish PRFS. According to DTU, universities are reward for quantity and not "quality" which should always be the focus in research. Why risk the current impact status by increasing output for some marginal gains? This cannot be a national interest.

So goes the argument - what we examine in this paper is to what extent the argument holds. Who benefits from the current Danish publication-based model and is DTU the current "losers"? What would be the differences if a citation-based approach was applied instead?

The aim of the analysis is to examine the universities' "return on investment". We take a simple approach where we examine the relative input of the universities when it comes to publications and subsequently examine the relative output from these publications, i.e., the rewards in the model, either the current publication points, or the alternative, citations. We keep the analysis simple using basically a zero-sum approach, like the current model, where gains somewhere mean losses elsewhere.

[2] www.leidenranking.com

The next section briefly presents the data and main methods and indicators used for the analyses. The subsequent section presents main results, and the final section contains a brief discussion of the findings.

Data and methods

The paper examines the first full publication year (2009) used for redistributing funds in the Danish model. We are able to measure the citation impact of the Danish journal publications from 2009 and make comparisons between the Danish universities and examine their potential gains and/or losses by using either differentiated publication counts or citations. We compare publication counts and points derived from the BFI model between Danish universities, and we likewise compare the impact between these universities for the 2009 journal publications indexed in Web of Science (WoS). As argued in the introduction section, locking the main research areas in principle means that the current publication-based model could be adapted to specific behaviours and wishes, or even supplemented or exchanged with a citation approach, in the individual areas, although citations would only be relevant in the areas: science and technology and medical and health sciences. In this paper we focus the analysis on the main research area of science and technology. We do this because the issue concerning citation impact versus publication activity raised by DTU is directly linked to this area due to DTUs research profile. We have done a corresponding analysis for the medical and health sciences but due to limited space we will not address them in this paper.

The publication activity in 2009 in the main research area of science and technology is around 8700 publications of all types eligible in the BFI model, books constituted 2%, articles in books 19% and journal articles 79%. It is reasonable to argue that (international) journal publication is the primary publication activity in this area, which means that citation analysis of eligible articles is a sensible endeavour. However, as the area includes some fields known to have their main publication activity in conference proceedings (i.e., articles in books), we do scrutinize the influence of proceedings papers on the total number of BFI points acquired for the individual universities and discuss that in relation to the citation analysis where proceedings papers are excluded. Notice, we do not include conference papers in the citation analysis due to the meagre quality of the current proceedings citation indices.

All journal publications published in 2009 reported by the universities to the BFI-indicator were extracted from the BFI database. Subsequently, paper titles were extracted, and so were first author names and journal names. These parameters were used to match the publications with Danish WoS journal publications from 2009 using CWTS's in house version of WoS. Eligible publication types are research articles and reviews. The match rate is 77% of the initial journal articles. Among the non-matched publications were non-English language articles, as well as false positive articles, articles not eligible for the BFI model, but still succeeded in accruing points.

As indicated in the introduction section, the BFI model applies a fractional counting method at the institutional level where articles are fractioned up to $1/10^{th}$ among the participating institutions. We do not apply the exact same counting formula for the WoS publications going into the citation analysis. Here we simply do a straightforward fractional counting on the institutional level. As will be clear from the results section, this small deviance had no practical relevance on relative publication shares.

We use standard CWTS citation indicators from the Leiden Ranking (www.leidenranking: P_{frac} (fractionalized publications), TNCS (total number of normalized citations), MNCS (mean normalized citation score) and PPtop10% (proportion of papers for a unit among the 10 percent most cited in the database) (Waltman et al., 2012).

Eight universities are included in the Danish PRFS. The universities differ considerable in both subject/faculty composition and size. We have two "old" universities basically covering

all four main research areas included in the BFI model: Copenhagen University (KU) and Aarhus University (AU). These universities are also the largest universities in Denmark with long research traditions and strong science faculties. University of Southern Denmark (SDU) is a younger university, but its subject/faculty composition is basically a reflection of KU and AU, although the size is considerably lower. Roskilde University (RU) and Aalborg University (AAU) are even younger, from the mid-1970s. These universities have regional obligations with a substantial emphasis on teaching. Nevertheless, both universities have developed unique research profiles, both universities have focused on interdisciplinary research, where RU has a strong focus on the social sciences and AAU has focused strongly on engineering. Both universities have science and technology faculties, albeit at RU the size is only comparable to a large department. The Information-Technology University is the youngest and smallest university in Denmark. Their focus is mainly outside the science and technology areas but we include them here for numbers to add up. Likewise, Copenhagen Business School (CBS) is also included for matters of completeness in the analyses, their publication activity in the science and technology area are scanty. Finally, as discussed in the introduction, the Technical University of Denmark (DTU) is basically a "mono-faculty" university, albeit its activities are spread between science and technology. It is important to emphasise that while the university is known for primarily educating engineers, it has a considerable research activity in what would be considered basic natural science fields as well. In fact DTU can be dated back to the early nineteenth century where it was part of Copenhagen University, making it the second oldest university in Denmark. We recapitulate, DTU has been particularly dissatisfied with the Danish PRFS arguing that - for them at least - citations would be a more appropriate and valid performance-based indicator. In the next section we examine the consequences of this claim.

We calculate basic statistics based on individual articles both for the publication-based model and the simple citation approach we apply. As stated in the introduction, we take a simple approach where we examine the relative input of the universities when it comes to publication shares and subsequently examine the relative "rewards" the universities archives from these publications, i.e., the output in the model, either shares of the total publication points, or the alternative, shares of the total number of citations. Also, we keep the analysis simple using basically a zero-sum approach, like the current PRFS, where gains somewhere mean losses elsewhere.

Results

Table 1 below shows the eight universities' total number of matched fractionalized WoS publications belonging to the science and technology area, as well as their accumulated number of normalized citations after four years. Notice, these are fractionalized WoS publications, the absolute number of publications is 6,117.

Table 1 also shows relative citation performance for the eight universities using the MNCS and PPtop10% field normalized indicators.

The three main actors measured by volume is not surprisingly KU (32.9%), DTU (28.7%) and AU (21.4%), the volumes for AAU and SDU are considerably lower, both universities have a share of 7.2% of the total volume. DTU has the largest number of normalized citations among the eight universities. It is noticeable that DTU's share of citations (34.8%) is markedly higher than their share of publications (28.7%). Obviously, this is also reflected in the relative citation indicators. The MNCS at 1.66 is considerably higher than the average of the database and a score that would rank DTU among the top 30 in the Leiden Ranking if we only focused on science and technology, and among the top 50 for all fields combined.

Table 1. Science and technology: Number of fractionalized publications in WoS, total number of citations and relative citation indicators.

	WoS pubs (P_{frac})	TNCS	MNCS	Share of total P_{frac}	Share of total no. of NCS	PPtop10%
AAU	225.3	284.4	1.26	7.2%	6.6%	12.3%
AU	673.0	874.5	1.30	21.4%	20.3%	14.6%
CBS	13.2	12.2	0.93	0.4%	0.3%	
DTU	904.9	1498.8	1.66	28.7%	34.8%	17.0%
ITU	11.5	9.1	0.79	0.4%	0.2%	
KU	1035.9	1281.0	1.24	32.9%	29.7%	13.4%
RU	56.9	61.3	1.08	1.8%	1.4%	10.7%
SDU	227.7	284.8	1.25	7.2%	6.6%	15.8%
Total	3148.2	4306.3		100%	100%	100%

Interestingly, we also see that the minor universities, SDU and AAU, have relative citation indicator scores comparable to the larger universities KU and AU. In fact, SDU has more of their 2009 publications among the 10% most cited in the database compared to KU and AU. Overall, these results confirm what we suspect and are essentially the basis for the argument about including citations in the BFI model advanced by DTU.

In order to examine "return on investment", i.e., the institutions' reward for their publication input, we have calculated their share of BFI publications and BFI points for 2009 for the science and technology area, as well as the shares of fractionalized WoS publications and the total number of field normalized (fractionalized) citations. We thereby assume that shares of BFI points and shares of normalized citations can be treated equally. In the final discussion section we reflect upon this. We do, however, think that the straightforward approach taken can give a rudimentary indication of potential differences in "returns" for the individual institutions if one was to apply a citation based approach instead of or as a supplement to the current differentiated publication-based indicator in the science and technology area.

Table 2 below shows the shares of BFI publications and BFI points, where all publication types used in the science and technology fields are included (e.g., also conference proceedings), as well as shares of fractionalized WoS journal articles and normalized citations.

Table 2. Science and technology: Distribution and shares of BFI-points, BFI-publications, plus fractionalized publications from WoS and total number of normalized citations; notice <u>all</u> BFI-publication types are included.

	BFI-points	BFI-publications (P)	Share of BFI-points	Share of total BFI P	Share of P_{frac} (WoS)	Share of total no. of TNCS
AU	1814.9	1766	19.1%	20.4%	21.4%	20.3%
CBS	6.9	6	0.1%	0.1%	0.4%	0.3%
DTU	2854.1	2378	30.1%	27.5%	28.7%	34.8%
ITU	117.4	107	1.2%	1.2%	0.4%	0.2%
KU	2730.9	2457	28.8%	28.4%	32.9%	29.7%
RUC	185.9	157	2.0%	1.8%	1.8%	1.4%
SDU	571.0	572	6.0%	6.6%	7.2%	6.6%
AAU	1203.6	1219	12.7%	14.1%	7.2%	6.6%
	9484.8	8662	100%	100%	100%	100%

Table 3 below shows the same variables as Table 2, but in this case we *only* use the BFI publication type journal articles and the points derived from these articles. Table 3 is included for comparison because the citation analysis in reality only deals with journal articles. Notice, the BFI journal articles include non-WoS indexed articles, which give points in the indicator, however, the numbers are very low, the coverage of the science area in WoS is very high.

Table 3. Science and technology: Science and Technology: Distribution and shares of BFI-points, BFI-publications, plus fractionalized publications from WoS and total number of normalized citations; notice only the BFI-publication type journal article is included.

	BFI-points (journals only)	BFI-publications (P) (journals only)	Share of BFI-points (journals only)	Share of total BFI P (journals only)	Share of P_{frac} (WoS)	Share of total no. of NCS
AU	1526.2	1515	21.9%	23.5%	21.4%	20.3%
CBS	6.9	6	0.1%	0.1%	0.4%	0.3%
DTU	2007.4	1663	28.8%	25.8%	28.7%	34.8%
ITU	53.3	39	0.8%	0.6%	0.4%	0.2%
KU	2166.8	2047	31.1%	31.8%	32.9%	29.7%
RUC	139.5	126	2.0%	2.0%	1.8%	1.4%
SDU	420.1	442	6.0%	6.9%	7.2%	6.6%
AAU	657.5	596	9.4%	9.3%	7.2%	6.6%
Total	6977.7	6434	100%	100%	100%	100%

For analytical and illustrative reasons we plot the results from Table 2 and 3 in Figures 1 and 2 below. Figure 1 shows the results based on all BFI publication types, whereas Figure 2 shows the results where only BFI journal articles are included.

The figures are simple plots were the shares of the total number of publications (i.e., both BFI publications and fractionalized publications from WoS) for the eight universities constitute the x-axis, this is the "input", i.e. what the individual institutions "invested" in the Danish performance-based model for science and technology in 2009. The y-axis shows the shares of BFI points and citations, this is the "output", i.e. the institutions' "return on their investment" in the Danish performance-based model for science and technology in 2009. The axes are symmetrical and the diagonal shows the point where the institution has the same relative share of input (publications) and output (BFI points or citations). The distance from the university to the diagonal suggests whether input is larger than the return (output), which means that the institution will be below the diagonal, or the return (output) is larger, in which case the university is placed above the diagonal. Further, each university is plotted two times, one for the BFI data and one for the WoS citation data. Significant changes between these two representations for a university up and down the diagonal, suggest that the university receives a substantial number of BFI points from publication types other than journal articles. Notice in order to avoid confusion when examining the figures, shares of BFI publications on the x-axis should be compared to shares of BFI points on the y-axis, and likewise shares of WoS publications on the x-axis should be compared with shares of citations on the y-axis.

It is clear from Figure 1 that RU, CBS and ITU are not interesting for the current analysis as their numbers and shares are too low. We are interested in the other five universities, which all have a faculty of some size within science and technology. Interestingly, from Figure 1, where *all* BFI publication types are included, we can see that DTU actually has a larger output than input with a ratio of 1.09. This is somewhat unexpected and contrary to the conjecture that DTU is not gaining much from the current model. If we then turn to the

citation analysis, then we can see an even larger distance from the diagonal to DTU, compared to the BFI data, but also all other universities. The ratio is 1.25, so in line with the previous findings, DTUs WoS publications receive considerably more citations than the other Danish universities in 2009 but also the average paper in the WoS database. If a citation-based indicator of some sort were constructed where points were given based on citations, as implied in the arguments from DTU, then it seems that DTU would benefit from such a model, obviously conditioned on how it was designed. However, the most interesting finding here is perhaps that DTU within the science and technology area also seems to be the largest beneficiary when it comes to BFI points earned per input publication. Notice, like the current PRFS, we also treat it as a zero-sum game. If all universities improve then we have status quo. As it is in Figure 1, only DTU seems to really benefit from the citation approach. While KU seems to be in balance with the BFI data, they experience a smaller drop in returns on their input in the citation approach. Perhaps the most remarkable result from Figure 1 is the dramatic drop on the diagonal between BFI data and WoS citation data for AAU. We return to this below.

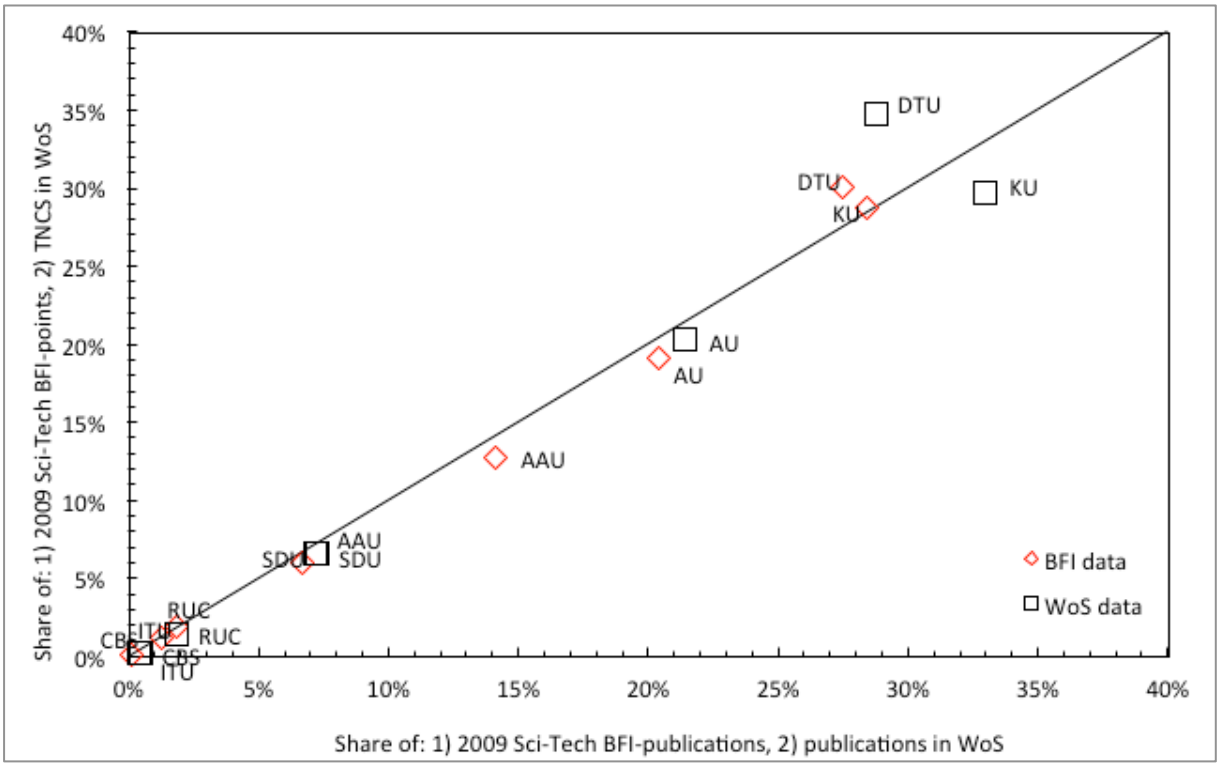


Figure 1. Science and technology: 1) Shares of 2009 Sci-Tech BFI publications as a function of shares of 2009 BFI-points; and 2) Shares of 2009 Sci-Tech WoS publications as a function of shares of total number of citations to these publications; notice BFI data includes all publication types.

Figure 2 depicts the same analysis but this time we have reduced the BFI data to include only journal publications in order to compare like with like, i.e., BFI journal data with WoS journal data. Obviously, the WoS data are identical to Figure 1, what is changing is the relative shares of BFI data (i.e., shares of publications and shares of points). There are some minor repositions, but the two major differences are the large drop on the diagonal for AAU and the corresponding smaller drop above the diagonal for DTU. Notice, the input-output is in balance for AAU, whereas DTU still has a substantial "return on investments" when it comes BFI journal data.

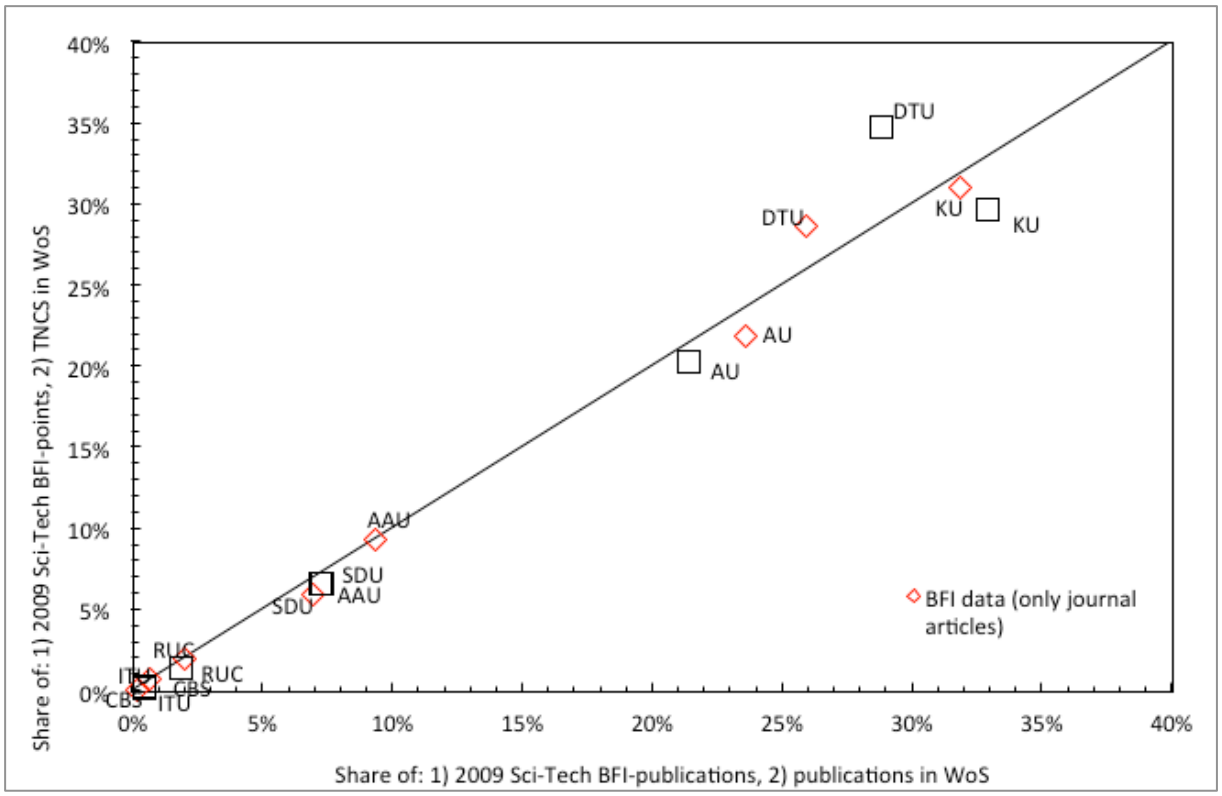


Figure 2. Science and technology: 1) Shares of 2009 Sci-Tech BFI publications as a function of shares of 2009 BFI-points; and 2) Shares of 2009 Sci-Tech WoS publications as a function of shares of total number of citations to these publications; notice BFI data only includes the publication type journal articles.

The drop of AAU along the diagonal was foretold in the WoS data in Figure 1. Here we saw a considerable distance between the BFI data when they included all publication types and the restricted WoS journal data needed for the citation analysis. For obvious reasons, this gap has been shortened considerably in Figure 2 since both data sets are restricted to journal articles.

The discrepancy in Figure 1 and the drop in Figure 2 are caused by the deviant publication profile for AAU compared to the other four universities with substantial publication activity in the science and technology area. Interestingly, 41% of the BFI publication activity in 2009 for AAU is in the category "articles in books", which in this case essentially means conference papers, and 49% is journal articles. For a comparison, 21% of DTUs activity is in "articles in books" and 70% in journal articles. These are both universities with strong focus on the technical sciences where publication in conference proceedings is very important. To contrast these profiles, the three other universities, KU, AU and SDU, all have more traditional science faculties and their relative publication activity in "articles in books" is 9%, 9% and 14% respectively. For these universities, due to their strong focus on science and less focus on technology, journal publication is the main activity 83% for KU, 86% for AU and 77% for SDU. However, we can also see that DTU does indeed have a strong science focus judged from their strong journal publication profile.

Considering the impetus for DTU to argue for a citation model, it is interesting to notice that while DTU clearly has the highest citation performance among the eight universities based on the 2009 journal publications, as we expected, they also have the highest performance when it comes to BFI publication points. Indeed, it seems that DTU would benefit even more in the science and technology area if they were to be rewarded for their relative share of the total number of citations, but contrary to the expected and suggested, DTU also benefit the most when it comes shares of BFI publication points compared to their relative input in the science

and technology area. DTU seems not only to be the most efficient when it comes to citations, this is also the case when it comes to BFI publication points. For example, the size of KUs activity in the science and technology area is larger than DTUs, but DTUs average point per publication is 1.20 for both of the above-mentioned analyses, considerably higher than KUs at 1.11.

Discussion

The main immediate findings in the present case study is that DTU will most probably benefit from a citation model, but perhaps more important, that they also seem to be the relatively most efficient university when it comes to BFI publication points. What are the more general implications of these findings seen in relation the current spread of the NPM to a number of European countries? The Danish case is special because competition is locked within the main areas this opens up for adapted models across areas including citation models where relevant. In Sweden a citation model is currently in use encompassing all fields. This is undesirable for several reasons; one of them is clearly demonstrated in this analysis, the desire to embrace all major publication behaviours, one of the rationales for the original NPM. A citation model alone restricts data to journal articles indexed in one of the two major citation databases. It was clear from Figure 1, that a university with an emphasis on technical sciences, like AAU, will be reduced in relative size when it comes to sharing the output.

The NPM is a differentiated publication indicator where points are graded for where you publish. Incentives to improve performance are clear and straightforward. Citation indicators reflect short term impact upon the scientific communication system. Citation indicators are retrospective and quite stable. It is very difficult to directly try to improve performance when it comes to impact. While one can argue that a publication-based model support the publish and perish culture with the ever increasing publication pressure, one could also argue that a citation model at the university level, due to its stability or conservative nature, and the fact that preferential attachment is at play for some universities, most likely would give cumulative advantages to those "who already have plenty", and potential changes brought about by incentives, are certainly not a short term phenomena.

There have been suggestions in Denmark to meet some of the requirements from DTU to focus more on citation impact. In order to keep the existing differentiated publication model intact, suggestions have been presented to bring in a third level especially in relation to journal outlets. This should be a category for the few hyped journals and publishing in these should be rewarded more lavishly. There may be good reasons for extending the levels in the model, but it is a flawed argument to claim to compensate wishes for more focus on impact by rewarding publication activity in "high impact" outlets. As it is well-known, article citation rates and journal citation impact have meagre correlations and the latter is a rather poor predictor of the former (Seglen, 1997).

A citation-based indicator or a hybrid indicator based on both publications and citations can be conceived in many ways, the question is whether the former or the latter is desirable. As discussed in the introduction, publication activity and citation impact are two different phenomena with substantially different prospects when it comes to incentives and behavioural adjustments. In the present analysis we could of course have experimented with more sophisticated citation-based approaches, for instance by constructing a mirror of the current publication-based model, where an arbitrary system allocates points according to which percentile group in the citation distribution they belonged to. We actually did that with a three-tiered point system, both the results were in line with the ones presented here.

As it is, based on the 2009 data, the BFI model in Denmark seems to work. Claims of more focus on citation impact seem only to speed up the cumulative advantage for "those who

already have" and at the same downgrade the influence of certain publication behaviours and muddling the transparent incentive structure.

References

Aagaard, K. (2011). Kampen om basismidlerne. Historisk institutionel analyse af basisbevillingsmodellens udvikling på universitetsområdet i danmark. PhD, Aarhus University, Aarhus.

Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: The case of the Norwegian publication indicator. Research Evaluation, 24(2), 106-117.

Butler, L. (2002). A list of published papers is no measure of value - the present system rewards quantity, not quality - but hasty changes could be as bad. Nature, 419(6910), 877-877.

Butler, L. (2003a). Explaining Australia's increased share of ISI publications - the effects of a funding formula based on publication counts. Research Policy, 32(1), 143-155.

Butler, L. (2003b). Modifying publication practices in response to funding formulas. Research Evaluation, 12(1), 39-46.

Gläser, J., & Laudel, G. (2007). The social construction of bibliometric evaluations. In R. Whitley & J. Gläser (Eds.), The changing governance of the sciences (Vol. 26, pp. 101-123): Springer Netherlands.

Hicks, D. (2012). Performance-based university research funding systems. Research Policy, 41(2), 251-261.

Karlsson, S. & Persson, O. (2012). The swedish production of highly cited papers Vetenskabsrådets lilla rapportserie. Stockholm, SWE.

Merton, R. K. (1988). The Matthew effect in science, ii: Cumulative advantage and the symbolism of intellectual property. Isis, 79(4), 606-623.

Moed, H. F. (2005). Citation analysis in research evaluation. Dordrecht, NL: Springer.

Moed, H. F. (2008). UK research assessment exercises: Informed judgments on research quality or quantity? Scientometrics, 74(1), 153-161.

Oppenheim, C. (1996). Do citations count? Citation indexing and the research assessment exercise (rae). Serials: The Journal for the Serials Community, 9(2), 155-161.

Schneider, J. W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in norway. European Political Science, 8(3), 364-378.

Schneider, J. W., & Aagaard, K. (2012). "Stor ståhej for ingenting" - den danske bibliometriske indikator. In K. Aagaard & N. Mejlgaard (Eds.), Dansk forskningspolitik efter årtusindskiftet (pp. 229-260). Aarhus: Aarhus Universitetsforlag.

Seglen, P. O. (1997). Citations and journal impact factors: Questionable indicators of research quality. Allergy, 52(11), 1050-1056.

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. ISSI Newsletter (International Society for Scientometrics and Informetrics), 6(1), 22-28.

Verleysen, F. T., Ghesquière, P., & Engels, T. (2014). The objectives, design and selection process of the Flemish academic bibliographic database for the social sciences and humanities (vabb-shw). In W. Blockmans, L. Engwall & D. Weaire (Eds.), Bibliometrics: Use and abuse in the review of research performance (pp. 117-127). London: Portland Press Ltd.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., . . . Wouters, P. (2012). The leiden ranking 2011/2012: Data collection, indicators, and interpretation. Journal of the American Society for Information Science and Technology, 63(12), 2419-2432.

The Effect of Having a Research Chair on Scientists' Productivity

Seyed Reza Mirnezami[1] and Catherine Beaudry[2]

[1]*seyed-reza.mirnezami@polymtl.ca* [2]*catherine.beaudry@polymtl.ca*
Polytechnique Montreal, P.O. Box. 6079, Montreal, QC, H3C 3A7 (Canada)

Abstract

Having combined data on Quebec scientists' funding and journal publication, this paper tests the effect of having a research chair on the scientists' performance. The novelty of this paper is to use matching technique to understand whether having a research chair is a real cause for better scientific performance. This method compares two different sets of regressions, which are conducted on different data sets: the one with all records and another with records of matched scientists only. Two chair and non-chair scientists are called matched with each other when they have closest propensity score in terms of age, number of articles, and amount of funding. The result shows that research chair is a significant determinant in complete data set but it is insignificant when only matched scientists are kept in data set. In other words, in the case of two scientists with similarity in terms of three mentioned factors, having a chair cannot significantly affect the scientific performance.

Conference Topic

Science policy and research assessment

Introduction

The scientists' academic performance has been extensively discussed and many of its determinants are currently known as potential motives for publishing papers in peer reviewed journals. Among others, age, gender, private and public funding, institutional setting, field and context are the most important determinants. The funding definitely plays the major role in knowledge production and shaping scientific productivity. Its positive effect has been extensively investigated in literature (Crespi & Geuna, 2008; Pavitt, 2000, 2001; Salter & Martin, 2001).

However, having a great academic performance does not depend solely on funding. The networking capability of scientist can also explain the number of journal papers. Most of the studies on the effects of network rely on co-authorship as a proxy of scientific collaboration (Katz & Martin, 1997; Melin & Persson, 1996). In addition to direct collaboration, there are also some other networking measures, which are known in the literature as determinant of publication. For instance, it is possible to show how a researcher links two other researchers by making separate collaborations with them. Newman (2001a, 2001b) finds that in physics, biomedical research, and computer science, most of the authors are connected with each other via one or two of their collaborators, a concept generally referred to as betweenness centrality. Beaudry and Allaoui (2012) also show a positive effect of betweenness centrality on the scientific productivity of Quebec's scientists.

In addition to the above measures of networking effect, the networking capacity of scientists partially depends on prestige of their academic affiliation. Turner and Mairesse (2005) show it for the outstanding performance of 'Grandes Ecoles' in France. Beside the name and brand of academic institutions, centers with specific research orientations such as 'centers for excellence' are also effective. According to Niosi (2002), the government of Canada launched 7 centers for biotechnology sectors in 1988, which financially supports the collaboration of university research, the specialized biotechnology firms, and the governmental laboratories. In addition to the funding support, however, this program comes up with improving intellectual property regulations, and developing human resources.

There are some other desirable factors similar to 'centers for excellence', which increase an individual's research motivation and influence the willingness or ability of scientists for conducting original research. In this paper, we focus on the effect having a 'research chair' as a possible determinant of scientific publication. On the one hand, it helps the holder of this position to absorb more money or to construct more effective network, which results in propelling future knowledge production. On the other hand, it may be the effect of past super performance of scientist, implying the intrinsic ability of scientists in conducting research or referring to the chair-holder extensive networking capacity.

By analysing data in an econometric model, it is possible to test the significant effect of 'being a chair holder' on the scientific productivity. The rest of paper is followings: Section 1 reviews theoretical framework and literature review. Section 2 explains how data is gathered and what the variables represent. In addition, it raises the related hypotheses and explains which econometric models can test these hypotheses. Section 3 presents the results of econometric model and the result of testing hypotheses. A conclusion will summarize the results of the paper.

Section 1 - Theoretical framework

The literature relevant to this article brushes on the importance of having a prestigious academic position or affiliation. Focusing on the role of university prestige in academic performance, Long, Allison, and McGinnis (1979) found a positive and significant correlation between the prestige of the scientist alma matter and prestige of subsequent employment affiliation. The authors also indicated that graduating from a prestigious university has a positive effect on citations (but not on publication counts). The paper also provides a justification for the effect of prestige arguing that the best students are admitted to the most prestigious universities and subsequently the graduates of the prestigious universities are generally recruited by other similar institutions. Furthermore, such scientists who studied in and have been recruited by prestigious universities are better able to interact with new gifted students (Long et al., 1979). This paper tries to argue that academic prestige can push forward research and its quality. More recently, Zhou, Lü, and Li (2012) show that papers cited by prestigious scientists, regardless of the number of citations, are of a higher quality than papers which are cited by 'ordinary' scientists.

The prestige can be seen from the reverse direction of causality. West, Smith, Feng, and Lawthom (1998) investigate the relationship between departmental climate, such as degree of formalization, support for career development and support for innovation on the one hand, and official rated effectiveness of universities on the other hand. They conclude that the causality direction is from former to latter, showing that prestige of universities is an effect and not a cause for appropriate departmental climate and necessary institutional setting for conducting research.

Nevertheless, measuring academic prestige itself is another story. Frey and Rost (2010) compare three types of university ranking based on the number of articles, number of citations, and membership of editorial board or of academic associations. The paper indicates that these rankings are not compatible with each other and suggests the use of multiple measurements. Van Raan (2005) criticize the applicability of university rankings such as the Shanghai ranking for evaluating academic excellence by noting that the 'affiliation', as an important factor reflecting research atmosphere, is not well addressed in those ranking. In addition to the university ranking, it is important to assess individual research productivity to have a better sense of prestige. Henrekson and Waldenström (2007) introduce three types of indicators, measuring research performance: (1) measures based on weighted journal publications, (2) measures based on citations to most cited works, and (3) measures based on the number of publications.

To measure prestige with more robust measure, it is possible to consider the honor as the measure of prestige, which is awarded based on a deliberate assessment in specialized and independent committees. Different types of research chair are example of awards. In Canada, there are three types of research chair: (1) the research chairs which are awarded by industry and called industrial chair; (2) the research chairs which are awarded by Canadian funding agencies such as NSERC, SSHRC, and CIHR; and (3) the 'Canada research chairs', whose holders are assumed to already achieve research excellence in one main fields of research: engineering and the natural sciences, health sciences, humanities, and social sciences. The purpose of this program is to "improve our depth of knowledge and quality of life, strengthen Canada's international competitiveness, and help train the next generation of highly skilled people through student supervision, teaching, and the coordination of other researchers' work".[1] Considering this specific measure of prestige, it is possible to find out the effect of being a 'chair-holder' on scientific productivity. Therefore, our first hypothesis reads as:

Hypothesis 1

Being chair-holder increases the scientist's number of publications.

The hypothesis 1 just tests the performance of chair-holders compared to other scientists and it does not seek for the cause and effect. Considering the fact that the chair-holders are the well-funded scientists too, this hypothesis cannot detach the funding effect of chair from its other effects (mainly from prestige and networking effect). In other words, there are evidences in literature about the benefits and goals of research chair program other than funding, but hypothesis 1 is not able to test them.

Some articles try to highlight the functions and characteristics of research chair. Cantu, Bustani, Molina, and Moreira (2009) show the research chair program would be a good strategy for implementing knowledge-based development. In study on German universities, Schimank (2005) argues that chair-holders are small businessmen with high job security and no bankruptcy in addition to the good level of freedom of teaching and research, indicating that research chair has characteristics of job security and sovereignty.

According to some official documents, affecting scientific productivity is not the direct goal of research chair. In the tenth-year evaluation report for Canada research chair (CRC),[2] the authors conclude that CRC program is an effective way for Canadian universities to "attract and retain leading researchers" from other countries. The report does not say that having a research chair is determinant and cause of chair's scientific production: "the extent to which this success can be related directly to the CRC is difficult to quantify". It is also possible to bring some evidence that having a research chair is not a cause for other factors such as salary. Courty and Sim (2012) show that although having Canada Research Chair (CRC) initially increases the professors' salary, such increase erodes quickly over the time. This means that getting a research chair does not necessarily result in long term salary jump.

Regarding the mentioned points, it is possible to look at the research chair as the effect of scientists' characteristics (including age, number of articles, and number of citations), while it aims to expand academic network and absorb highly skilled talents. To control for the effect of scientist's past performance on having a research chair and to detach the funding advantage of chair, we propose our second hypothesis:

Hypothesis 2

Keeping the main scientists' characteristics (age, number of articles, and amount of grant) constant, having a research chair does not have significant positive effect on scientists' productivity.

[1] http://www.chairs-chaires.gc.ca/about_us-a_notre_sujet/index-eng.aspx

[2] http://www.chairs-chaires.gc.ca/about_us-a_notre_sujet/publications/ten_year_evaluation_e.pdf

This hypothesis can be tested by matching technique, which will be explained in the methodology section. The important note here is that 'being a research chair' cannot be the only determinant in right-hand-side of regression equations. We should look for some control variables, which are mentioned in literature as determinants of scientific production. Among others, age, gender, funding, field, and university characteristics are the most important determinants of scientific production which should be controlled when the effect of research chair on scientific productivity are being tested.

In terms of age, there are two groups of evidences in literature about its effect on scientific productivity. First, some articles assess the life cycle trend in economic activity, referring to the non-linearity of human productivity during life (Becker, 1962). The second group of articles generally find that scientists' academic performance (number of articles and number of citations) decreases as they age (Bonaccorsi & Daraio, 2003; Diamond, 1986; Levin & Stephan, 1991). Some articles like Gonzalez-Brambila and Veloso (2007) also indicate that age does not have any effect on the number of articles but it positively affect the number of citations. Gender effect is known as a significant determinant of scientific productivity in literature. Long (1990) explains that women's opportunities for collaboration are significantly less than those of men's because women have young children. However, in another study, Long (1992) shows that women are less productive in the first decade of their career but are more productive afterwards. Research funding is another important determinant of scientific productivity. Pavitt (2001) also refers to the importance of public support for scientific infrastructure development and highlights its role in the effectiveness of public grants. In another study, Pavitt (2000) argues that fudging for infrastructure of expertise, equipment and networks is necessary for development and implementation of research. A body of literature investigates the effect of university characteristics on the scientific productivity. There are also some papers about the effect of faculty size. Buchmueller, Dominitz, and Lee Hansen (1999) indicate that graduate school faculty size is a significant determinant of the research proficiency of graduates. Jordan, Meador, and Walters (1988, 1989) indicate that research productivity is positively associated with department size but that effect becomes weaker as the size increases. In an opposite direction, Kyvik (1995) rejects both hypotheses that large departments are more productive and that faculty members of large departments better assess the research environment.

There also some evidences about differences between fields and context. Blackburn, Behymer, and Hall (1978) show that the fields of humanities and sciences have different pattern of scientific production. To justify the differences between disciplines, Baird (1986) shows that for instance large research laboratory in chemistry, scholarly apprenticeship approach in history, and research over practice in psychology are important factors in scientists' productivity, which are field-dependent factors. In another comprehensive study, Baird (1991) refers to the productivity and citation pattern differences among disciplines and argues that size, internal university support and federal support can explain such differences. All of the mentioned evidence in literature shows that scientific productivity may have different determinants including academic prestige and other control variables such as funding, gender, age, and university-specific characteristics.

Section 2 - Data and methodology

Data and variables

In order to validate these two hypotheses, we built a data set based on the integration of data on funding and journal publications for Quebec scientists. For publications, Elsevier's Scopus provides information on scientific articles (date of publication, journal name, authors and their affiliations). In terms of funding, there is a database for researchers in Quebec universities

(*Système d'information sur la recherche universitaire* or SIRU) gathered and combined by the Ministry of Education, Leisure and Sports. The SIRU database lists the grants and contracts information, including yearly amount, source, and type during the period of 2000-2010 for all Quebec university scientists. The appendix 1 reviews the names and description of variables in data set.

Methodology and econometrics model

To measure the effect of 'being a research chair' on the scientist's performance, a regression equation is fitted to the available data using a panel regression. In such regression, the left-hand-side (LHS) variable of regression is the number of articles [*ln(nbArticle)*] as a measure of scientific productivity. In terms of right-hand-side (RHS) variable, the main independent variables are the dummy variables of research chair [*dChair1*, *dChair2*, *dChair3*, *dChair4*, *dChair5*]. However, the dependent variable of regression in LHS should be also controlled for the other determinants of articles count. Among others, age [*Age*], gender [*dFemale*], and funding are the important ones. We also control for the fixed effect of university, year, and research field in order to account for any impact that our explanatory variables do not cover.

It is important to note that two variables of [*ln(PublicfundingO)*] and [*ln(nbArticle)*] are determined by each other and co-evolved during time, which is the source of endogeneity. Thus it means that simple ordinary least square or panel models are biased. The main reason for this potential endogeneity is that scientists are assessed for public funding based on their CV and past performance while at the same time, publication and research quality significantly depends on the funding capability of researchers. Using instrumental variables (IV) instead of endogenous variable is a common suggested method in literature to address endogeneity problem. If there is more than one instrument for an endogenous variable, it is necessary to perform a two-stage regression, in which the first stage estimates the endogenous variable (named here as instrumented variable) based on a list of instrumental variables. In the first stage of our model, the amount of public funding [*ln(PublicfundingO)*] is estimated by the rank of scientist in the field in terms of three-year average of funding (for the purpose of operational costs and direct expenditure of research) [*PubORank*], the rank of scientist in the field in terms of three-year average of articles count [*PublRank*], and natural logarithm of three-year average of aggregate public sector funding in the field [*ln(totFund)*]. These three variables play the role of instruments for public funding. It should also be noted that public funding is not determined by the instruments in the same year. Hence the one-year lags of instruments are being used in the first-stage regression. The second stage is similar to the previous model in which there is no endogeneity.

1st stage: $ln(PublicfundingO)_{it} = f(\ PubORank_{it-1}, PublRank_{it-1}, ln(totFund)_{it-1}\ )$

2nd stage: $ln(nbCitation)_{it} = f(\ ln(PublicfundingO)_{it}\ ,\ ln(PrivatefundingO)_{it}\ ,\ ln(NFPfundingO)_{it}\ ,\ (dChair1|dChair2|dChair3|dChair4|dChair5)_{it}$, *dFemale, Age,* Age^2*, research field dummies, year dummies, university dummies)*

The main purpose of this research is to show how much having a research chair as an external support is important and significant in promoting scientific publication. To test the first hypothesis, it is sufficient to run the two-stage panel regression on the whole data set whether 'having a research chair' is a significant RHS variable, either as a real cause or a channel for other variables/causes. According to the chair characteristics, the networking and prestige effect of 'having a research chair' may be mixed with the effect of funding. To address this issue, we use matching technique and compare two chair and non-chair scientists who have close funding to each other (and have some other similar characteristics). Like what Bérubé and Mohnen (2009) did, it is possible to find pairs of chair and non-chair by using the psmatch2 command in Stata and delete the unmatched records. The selection is made by

generating propensity score and choosing the pairs of scientists with closest scores to each other. The new data set consists of twin scientists who are similar to each other in terms of funding, gender, and division of studies.[3]

By controlling the mentioned criteria and keeping matched scientists only, 'having a research chair' becomes a better and more informative signal for the prestige and networking of scientists. In this case, the effect of 'being chair' on scientific productivity does not include funding effect or it is not related to the division or gender of scientist. To test the second hypothesis, only matched pairs of scientists are being used in regression analysis to identify whether having a research chair is a significant cause for scientific productivity.

One of the important stages in matching technique is to check the quality of matching. It means there should be no difference between the averages of mentioned criteria (gender, funding, and division of studies) when the comparison is made between chair and non-chair scientists among the matched pairs. However, there can be a difference when the comparison is made in original database and before any entry deletion. Table 1 summarizes such comparisons to show that the matching is done with an acceptable quality for *dChair3*, *dChair4*, and *dChair5*.

Table 1. Make a comparison between mean to show the quality of matching.

	Comparison over whole database				Comparison over matched scientists "**After Matching**"			
	Gender	Funding	Research field[4]	number of scientist	Gender	Funding	Research field	number of scientist
dChair3=0	0.2959	86217	0.4284	7359	0.1023	403051	0.2286	293
dChair3=1	0.2013	464106	0.3447	293	0.2013	464106	0.3447	293
Is difference significant at 5% level?	Yes	Yes	Yes		Yes	No	No	
dChair4=0	0.2954	95871	0.4318	7508	0.1111	369080	0.0416	144
dChair4=1	0.1319	351785	0.0833	144	0.1319	351785	0.0833	144
Is difference significant at 5% level?	Yes	Yes	Yes		No	No	No	
dChair5=0	0.2987	82183	0.4344	7234	0.1483	367494	0.1698	418
dChair5=1	0.1818	420920	0.2655	418	0.1818	420920	0.2655	418
Is difference significant at 5% level?	Yes	Yes	Yes		No	No	No	

Section 3 - Result and discussion

Based on the models presented in methodology section, we need to first run the regressions on the whole dataset (Table 2) which show that all types of chair have positive and significant effect on scientific productivity. However after keeping only matched scientists in dataset, who are similar to each other in terms of gender, funding, and research field, the regression equations indicate significant and positive result only for Canada research chair (Table 3) Industrial chairs and chairs appointed by Canada research council (NSER, SSHRC, and CIHR) do not have an independent positive effect on scientific productivity. Considering the hypotheses in previous section, it possible to validate the first hypothesis and partially validate the second hypothesis. One may question whether research chairs in general are independent cause for research productivity or they are proxy for other known factors in literature. Considering literature and mentioned mission of research chairs in their mandate,

[3] We have three divisions: 'engineering and the natural sciences', 'health sciences', and' humanities, and social sciences'

[4] Test whether dummy variable of Social Science and Humanities is equal to 1.

Table 2. Regression results over all samples for *dChair3* and *dChair4* (the second stage of 2SLS).[1]

$ln(nbArticle)_{it}$	*IV1*	*IV2*	*IV3*	*IV4*	*IV5*	*IV6*	*IV7*	*IV8* `	*IV9*	*IV10*	*IV11*
$ln(PublicfundingO)_{it}$	0.0433 ***	0.0417 ***	0.0417 ***	0.0416 ***	0.0417 ***	0.0416 ***	0.0415 ***	0.0415 ***	0.0417 ***	0.0416 ***	0.0416 ***
	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011	0.0011
$ln(PrivatefundingO)_{it}$	0.0112 ***	0.0109 ***	0.0105 ***	0.0109 ***	0.0105 ***	0.0113 ***	0.0108 ***	0.0111 ***	0.0110 ***	0.0109 ***	0.0110 ***
	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007	0.0007
$ln(NFPfundingO)_{it}$	0.0076 ***	0.0074 ***	0.0074 ***	0.0075 ***	0.0075 ***	0.0074 ***	0.0092 ***	0.0092 ***	0.0074 ***	0.0075 ***	0.0074 ***
	0.0006	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	0.0006	0.0006	0.0006
Age_{it}	0.0021	0.0038	0.0038	0.0038	0.0038	0.0037	0.0036	0.0036	0.0038	0.0038	0.0038
	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025
sq_Age_{it}	-0.0001 ***	-0.0001 ***	-0.0001 ***	-0.0001 ***	-0.0001 ***	-0.0001 ***	-0.0001 ***	-0.0001 ***	-0.0001 ***	-0.0001 ***	-0.0001 ***
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$dFemale_i$	-0.0911 ***	-0.0847 ***	-0.0848 ***	-0.0847 ***	-0.0848 ***	-0.0815 ***	-0.0700 ***	-0.0686 ***	-0.0832 ***	-0.0841 ***	-0.0827 ***
	0.0109	0.0108	0.0108	0.0108	0.0108	0.0110	0.0112	0.0113	0.0109	0.0109	0.0109
$dFemale_i*ln(PrivatefundingO)_{it}$						-0.0023		-0.0013			
						0.0016		0.0016			
$dFemale_i*ln(NFPfundingO)_{it}$							-0.0065 ***	-0.0064 ***			
							0.0013	0.0013			
$dChair3_{it}$		0.3331 ***	0.3105 ***	0.3444 ***	0.3233 ***	0.3332 ***	0.3323 ***	0.3324 ***	0.3330 ***	0.3413 ***	0.3404 ***
		0.0249	0.0268	0.0271	0.0284	0.0249	0.0249	0.0249	0.0251	0.0252	0.0254
$dChair4_{it}$		0.1025 ***	0.1006 **	0.0891 **	0.0894 **	0.1020 ***	0.0998 ***	0.0996 ***	0.1195 ***	0.0942 ***	0.1114 ***
		0.0352	0.0432	0.0387	0.0451	0.0352	0.0352	0.0352	0.0360	0.0356	0.0362
$dChair3_{it}*ln(PrivatefundingO)_{it}$			0.0060 **		0.0064 **						
			0.0026		0.0027						
$dChair4_{it}*ln(PrivatefundingO)_{it}$			0.0000		-0.0003						
			0.0033		0.0034						
$dChair3_{it}*ln(NFPfundingO)_{it}$				-0.0026	-0.0033						
				0.0024	0.0024						
$dChair4_{it}*ln(NFPfundingO)_{it}$				0.0026	0.0026						
				0.0031	0.0031						
$dFemale_i*ln(PrivatefundingO)_{it}*dChair3_{it}$									0.0005		0.0024
									0.0063		0.0064
$dFemale_i*ln(PrivatefundingO)_{it}*dChair4_{it}$									-0.0177 **		-0.0212 **
									0.0077		0.0079
$dFemale_i*ln(NFPfundingO)_{it}*dChair3_{it}$										-0.0102 **	-0.0104 **
										0.0050	0.0050
$dFemale_i*ln(NFPfundingO)_{it}*dChair4_{it}$										0.0125	0.0175 **

[1] *, **, and *** show the significance level at 0.05, 0.02, and 0.01 respectively - Year dummies, field dummies, and university dummies are significant. The minimum year activity, average year activity, and maximum year activity are 1, 10.6, and 12 respectively.

$ln(nbArticle)_{it}$	*IV1*	*IV2*	*IV3*	*IV4*	*IV5*	*IV6*	*IV7*	*IV8* `	*IV9*	*IV10*	*IV11*
										0.0081	0.0083
Constant term	0.4681 ***	0.4210 ***	0.4218 ***	0.4200 ***	0.4204 ***	0.4222 ***	0.4218 ***	0.4223 ***	0.4202 ***	0.4210 ***	0.4205 ***
	0.0683	0.0680	0.0680	0.0680	0.0680	0.0680	0.0680	0.0680	0.0680	0.0680	0.0680
Number of observations	80772	80772	80772	80772	80772	80772	80772	80772	80772	80772	80772
Number of scientists	7652	7652	7652	7652	7652	7652	7652	7652	7652	7652	7652
$\chi 2$	13859.3	14234.6	14251.6	14244.3	14258.4	14236.5	14277.1	14277.9	14239.7	14241.7	14246.4
sigma	0.5689	0.5664	0.5661	0.5662	0.5660	0.5664	0.5662	0.5662	0.5664	0.5664	0.5664
rho	0.4235	0.4183	0.4178	0.4180	0.4176	0.4184	0.4181	0.4182	0.4183	0.4183	0.4184
R^2 ***within groups***	0.0617	0.0630	0.0629	0.0631	0.0630	0.0631	0.0633	0.0634	0.0631	0.0631	0.0632
R^2 ***overall***	0.3367	0.3456	0.3460	0.3455	0.3458	0.3457	0.3464	0.3464	0.3457	0.3456	0.3457
R^2 ***between groups***	0.5044	0.5148	0.5154	0.5145	0.5151	0.5148	0.5156	0.5156	0.5148	0.5147	0.5148

Table 3. Regression results over only matched pairs of scientists for *dChair3* and *dChair4* (the second stage of 2SLS).[2]

$ln(nbArticle)_{it}$	*IV23*	*IV24*	*IV25*	*IV26*	*IV27*	*IV28*	*IV29*	*IV30*	*IV31*	*IV32*	*IV33*
$ln(PublicfundingO)_{it}$	0.0702 ***	0.0680 ***	0.0692 ***	0.0680 ***	0.0691 ***	0.0680 ***	0.0679 ***	0.0679 ***	0.0682 ***	0.0678 ***	0.0679 ***
	0.0059	0.0059	0.0060	0.0059	0.0060	0.0059	0.0059	0.0059	0.0059	0.0059	0.0059
$ln(PrivatefundingO)_{it}$	0.0053 ***	0.0059 ***	0.0076 ***	0.0059 ***	0.0072 ***	0.0062 ***	0.0059 ***	0.0062 ***	0.0066 ***	0.0060 ***	0.0067 ***
	0.0019	0.0019	0.0025	0.0019	0.0026	0.0021	0.0019	0.0021	0.0020	0.0019	0.0020
$ln(NFPfundingO)_{it}$	0.0038 **	0.0042 **	0.0041 **	0.0077 **	0.0074 **	0.0041 **	0.0045 **	0.0044 **	0.0041 **	0.0045 **	0.0043 **
	0.0018	0.0018	0.0018	0.0025	0.0026	0.0018	0.0019	0.0019	0.0018	0.0019	0.0019
Age_{it}	0.0217 **	0.0244 **	0.0260 **	0.0249 **	0.0265 **	0.0244 **	0.0244 **	0.0244 **	0.0246 **	0.0242 **	0.0243 **
	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104	0.0104
sq_Age_{it}	-0.0003 ***	-0.0003 ***	-0.0003 ***	-0.0003 ***	-0.0003 ***	-0.0003 ***	-0.0003 ***	-0.0003 ***	-0.0003 ***	-0.0003 ***	-0.0003 ***
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
$dFemale_i$	-0.1217 **	-0.1230 **	-0.1215 **	-0.1224 **	-0.1210 **	-0.1160 **	-0.1138 **	-0.1081 **	-0.0889 **	-0.1173 **	-0.0848 **
	0.0545	0.0533	0.0533	0.0532	0.0533	0.0562	0.0572	0.0595	0.0555	0.0549	0.0567
$dFemale_i*ln(PrivatefundingO)_{it}$						-0.0020		-0.0018			
						0.0051		0.0051			
$dFemale_i*ln(NFPfundingO)_{it}$							-0.0022	-0.0020			
							0.0049	0.0049			
$dChair3_{it}$		0.1696 ***	0.1625 ***	0.2062 ***	0.1954 ***	0.1697 ***	0.1696 ***	0.1698 ***	0.1689 ***	0.1766 ***	0.1756 ***
		0.0451	0.0483	0.0483	0.0506	0.0451	0.0451	0.0452	0.0453	0.0454	0.0456

[2] *, **, and *** show the significance level at 0.05, 0.02, and 0.01 respectively - Year dummies, field dummies, and university dummies are significant. The minimum year activity, average year activity, and maximum year activity are 1, 10.9, and 12 respectively.

$ln(nbArticle)_{it}$	*IV23*	*IV24*	*IV25*	*IV26*	*IV27*	*IV28*	*IV29*	*IV30*	*IV31*	*IV32*	*IV33*
$dChair4_{it}$		-0.0401	0.0475	-0.0267	0.0524	-0.0398	-0.0400	-0.0397	-0.0157	-0.0479	-0.0240
		0.0553	0.0650	0.0595	0.0677	0.0553	0.0553	0.0553	0.0560	0.0556	0.0562
$dChair3_{it}*ln(PrivatefundingO)_{it}$			0.0015		0.0026						
			0.0040		0.0040						
$dChair4_{it}*ln(PrivatefundingO)_{it}$			-0.0122 **		-0.0118 **						
			0.0048		0.0048						
$dChair3_{it}*ln(NFPfundingO)_{it}$				-0.0078 **	-0.0080 **						
				0.0037	0.0037						
$dChair4_{it}*ln(NFPfundingO)_{it}$				-0.0031	-0.0019						
				0.0044	0.0044						
$dFemale_i*ln(PrivatefundingO)_{it}*dChair3_{it}$									-0.0012		0.0001
									0.0081		0.0082
$dFemale_i*ln(PrivatefundingO)_{it}*dChair4_{it}$									-0.0280 ***		-0.0311 ***
									0.0102		0.0103
$dFemale_i*ln(NFPfundingO)_{it}*dChair3_{it}$										-0.0087	-0.0091
										0.0065	0.0065
$dFemale_i*ln(NFPfundingO)_{it}*dChair4_{it}$										0.0120	0.0174 *
										0.0103	0.0104
Constant term	-0.0326	-0.1656	-0.2236	-0.2009	-0.2565	-0.1650	-0.1656	-0.1649	-0.1795	-0.1607	-0.1719
	0.2714	0.2711	0.2719	0.2715	0.2723	0.2712	0.2712	0.2712	0.2711	0.2712	0.2712
Number of observations	9097	9097	9097	9097	9097	9097	9097	9097	9097	9097	9097
Number of scientists	836	836	836	836	836	836	836	836	836	836	836
χ2	2185.96	2231.62	2230.58	2237.66	2237.25	2231.39	2230.92	2230.54	2235.76	2234.8	2239.7
sigma	0.6921	0.6842	0.6844	0.6835	0.6836	0.6843	0.6844	0.6845	0.6840	0.6842	0.6843
rho	0.4798	0.4675	0.4677	0.4672	0.4672	0.4676	0.4678	0.4680	0.4675	0.4678	0.4682
R2 within groups	0.1385	0.1393	0.1392	0.1398	0.1398	0.1394	0.1394	0.1394	0.1399	0.1399	0.1406
R2 overall	0.3300	0.3409	0.3406	0.3411	0.3407	0.3411	0.3410	0.3411	0.3409	0.3408	0.3413
R2 between groups	0.4584	0.4730	0.4729	0.4728	0.4726	0.4731	0.4729	0.4731	0.4724	0.4722	0.4726

it is possible to argue that having a chair improve networking capability or funding amount of scientists.

In the second hypothesis we try to make a distinction between the effect of funding and having a research chair. By running regression model only on matched pairs of scientists, having a chair cannot be a proxy for criteria of matching (age, gender, and research field) anymore. We can verify the hypothesis 2 for industrial chair and research chairs appointed by research council but this hypothesis cannot be validated for 'Canada research chair' because its effect is still positive and significant even after matching. Some justification can be provided for this finding. The first is that Canada research chair intends to be prestige sign of research in Canada. Based on its mandate, the Canada research chair program aims to attract and retain some of most accomplished and promising minds in the world. It is more prestigious than other research chairs and other scientists may also have more willingness to conduct collaborative research with the Canada research chair holders. As the second justification, it should be noted that industrial chairs are appointed by firms to promote research, probably with major benefits for firms. In other words, this type of chair is not necessarily and originally designed for the sake of scientific publication. The chairs appointed by research councils may have quite similar characteristic. Looking at these chairs' description, most of chair holders are appointed as industrial chair. There are some evidence in literature indicating that industrial funding forces researchers to shift to more applied research, neglecting their normative responsibilities for knowledge development (Geuna & Nesta, 2003; Partha & David, 1994).

In addition to the effect of chair on scientific productivity, there are also some interesting results for other control variables in econometric model. Funding from different sources is always a significant determinant of scientific productivity, which has positive sign. Funding from private sector and funding from not-for-profit sector are directly put in regression equation while funding from public sector is first estimated by instrumental variables and then inserted to regression model.

The age of scientists seems to affect scientific productivity with an inverted-U shape pattern. However, considering its peak, which is 10 years old and less than the normal age for scientific activity, it is possible to argue that scientific productivity of scientists decreases in age. The gender of scientist, as another individual attribute, shows a significant impact. It indicates that women are less likely to publish journal paper compared with men. Both of these findings have some similar evidence in literature as discussed in previous section for age (Bonaccorsi & Daraio, 2003; Diamond, 1986; Levin & Stephan, 1991) and gender (Long, 1990).

The results verify the fixed effect of university and research discipline in addition to the year-specific effect on scientific production. Our regression analysis also tests the interactive effects of RHS variables. The first interactive effect is the interaction between gender and funding. From technical point of view, it is not possible to estimate the interactive effect with an endogenous variable in 2SLS models because its amount is estimated in the first stage and we are not using the raw value reported in dataset. However, we can estimate the effect of interaction with private funding and not-for-profit funding, which both are not significant. The only exception is in table 2 where the regression is run on whole dataset and interaction of gender and not-for-profit is negative and significant, which means that women may benefit from not-for-profit funding less efficiently compared with men.

The variables measuring interaction between having a chair and amount of funding are the next possible interaction in regression analysis, most of which are not significant. However, if there is significance, it is positive before matching and negative after matching. It refers to the more impact of funding for the chair people in general (complete data set) but when the chairs are compared to scientists, who are similar to them in terms of funding, gender, and research

field, they benefit from the funding less efficiently compared to non-chairs. The last group of interactive variables are the combination of two previous groups: interaction between funding, chair, and gender. There are some negative and significant effects for this type of interaction, showing the combined results of previous interactive variables.

Conclusion

In this article we show that having a research chair is a significant determinant of scientific publication when the regression is run over whole data set. As previously explained, a distinction should be made to clarify different attributes of research chair and their effect on scientific productivity. For instance, it is a fact that research chairs receive more grants due to their chair so the question here is to check if positive effect of research chair on scientific productivity remains significant after controlling for the funding amount of chair. To investigate the causality of this relationship, the matching technique is applied to control for some common characteristics of chair and non-chair scientists and to highlight the channel through which this positive effect has happened.

To conduct this matching technique, we only keep pairs of chair and non-chair scientists, matched together based on funding, gender, and research field, and delete the rest of scientists from data set. This methodology is effective to understand other attributes of research chair (except funding) that have significant and positive effect on scientific productivity. After such matching, the results show that the effect of Canada research chair on scientific productivity remains significant and positive while the effect of industrial chairs and the chairs appointed by Canada research council (NSER, SSHRC, and CIHR) become insignificant. This finding indicates that there are some special attributes in Cana research chair, which do not exist in other chairs. Those attributes may significantly push scientific productivity. Among other attributes, Canada research chairs may have better prestige to absorb talents or they are well designed to conduct scientific research for publication.

References

Baird, L.L. (1986). What characterizes a productive research department? *Research in Higher Education, 25*(3), 211-225.

Baird, L.L. (1991). Publication productivity in doctoral research departments: Interdisciplinary and intradisciplinary factors. *Research in Higher Education, 32*(3), 303-318.

Beaudry, C. & Allaoui, S. (2012). Impact of public and private research funding on scientific production: The case of nanotechnology. [Working paper].

Becker, G.S. (1962). Investment in human capital: a theoretical analysis. *The journal of political economy, 70*(5), 9-49.

Bérubé, C. & Mohnen, P. (2009). Are firms that receive R&D subsidies more innovative? *Canadian Journal of Economics/Revue canadienne d'économique, 42*(1), 206-225.

Blackburn, R.T., Behymer, C.E., & Hall, D.E. (1978). Research note: Correlates of faculty publications. *Sociology of Education*, *51*(2), 132-141.

Bonaccorsi, A., & Daraio, C. (2003). Age effects in scientific productivity. *Scientometrics, 58*(1), 49-90.

Buchmueller, T.C., Dominitz, J., & Lee Hansen, W. (1999). Graduate training and the early career productivity of Ph.D. economists. *Economics of Education Review,* 18(1), 65-77. doi: 10.1016/s0272-7757(98)00019-3

Cantu, F.J., Bustani, A., Molina, A., & Moreira, H. (2009). A knowledge-based development model: the research chair strategy. *Journal of Knowledge Management, 13*(1), 154-170.

Courty, P., & Sim, J. (2012). What is the cost of retaining and attracting exceptional talents? Evidence from the Canada Research Chair program: Queen's Economics Department Working Paper.

Crespi, G.A., & Geuna, A. (2008). An empirical study of scientific production: A cross country analysis, 1981–2002. *Research Policy, 37*(4), 565-579.

Diamond, A.M. (1986). The life-cycle research productivity of mathematicians and scientists. *Journal of Gerontology, 41*(4), 520.

Frey, B.S., & Rost, K. (2010). Do rankings reflect research quality? *Journal of Applied Economics, 13*(1), 1-38.

Geuna, A. & Nesta, L. (2003). University patenting and its effects on academic research. *SEWPS Paper* (99).

Gonzalez-Brambila, C., & Veloso, F.M. (2007). The determinants of research output and impact: A study of Mexican researchers. *Research Policy, 36*(7), 1035-1051.

Henrekson, M., & Waldenström, D. (2007). How should research performance be measured: IFN Working Paper.

Jordan, J.M., Meador, M., & Walters, S.J.K. (1988). Effects of department size and organization on the research productivity of academic economists. *Economics of Education Review, 7*(2), 251-255.

Jordan, J.M., Meador, M., & Walters, S.J.K. (1989). Academic research productivity, department size and organization: Further results. *Economics of Education Review, 8*(4), 345-352.

Katz, J. S., & Martin, B.R. (1997). What is research collaboration? *Research Policy, 26*(1), 1-18.

Kyvik, S. (1995). Are big university departments better than small ones? *Higher Education, 30*(3), 295-304.

Levin, S. G. & Stephan, P. E. (1991). Research productivity over the life cycle: evidence for academic scientists. *The American Economic Review*, 114-132.

Long, J.S. (1990). The origins of sex differences in science. *Social Forces, 68*(4), 1297-1316.

Long, J.S. (1992). Measures of sex differences in scientific productivity. *Social Forces, 71*(1), 159-178.

Long, J.S., Allison, P.D., & McGinnis, R. (1979). Entrance into the academic career. *American Sociological Review*, *44*(5), 816-830.

Melin, G. & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics, 36*(3), 363-377.

Newman, M.E.J. (2001a). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 025102.

Newman, M.E.J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review-Series E-, 64*(1; Part 2), 16132-16132.

Niosi, J. (2002). Regional systems of innovation: Market pull and government push. *Holbrook, J.-A and D. Wolfe, (eds.) Knowledge, Clusters and Regional Innovation. Montréal & Kingston, McGill-Queen's University Press*, 39-55.

Partha, D. & David, P.A. (1994). Toward a new economics of science. *Research Policy, 23*(5), 487-521.

Pavitt, K. (2000). Why European Union funding of academic research should be increased: a radical proposal. *Science and Public Policy, 27*(6), 455-460.

Pavitt, K. (2001). Public policies to support basic research: What can the rest of the world learn from US theory and practice? (And what they should not learn). *Industrial and corporate change, 10*(3), 761-779.

Salter, A.J. & Martin, B.R. (2001). The economic benefits of publicly funded basic research: a critical review. *Research Policy, 30*(3), 509-532.

Schimank, U. (2005). 'New Public Management' and the academic profession: Reflections on the German situation. *Minerva, 43*(4), 361-376.

Turner, L. & Mairesse, J. (2005). Individual Productivity Differences in Public Research: How important are non-individual determinants? An Econometric Study of French Physicists' publications and citations (1986-1997). *Centre National de la Recherche Scientifique*.

Van Raan, A.F.J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133-143.

West, M.A., Smith, H., Feng, W.L., & Lawthom, R. (1998). Research excellence and departmental climate in British universities. *Journal of Occupational and Organizational Psychology, 71*(3), 261-281.

Zhou, Y.B., Lü, L., & Li, M. (2012). Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. *New Journal of Physics, 14*, 033033.

Appendix 1 – Variable description.

Variable name	*Variable description*
dChair1	Dummy variables taking the value 1 if a scientist has a research chair awarded by industry (industrial chair)
dChair2	Dummy variables taking the value 1 if a scientist has a research chair awarded by Canadian funding agencies (NSERC, SSHRC, and CIHR)
dChair3	Dummy variables taking the value 1 if a scientist has a Canada research chair
dChair4	Dummy variables taking the value 1 if *dChair1* or *dChair2* are equal to 1
dChair5	Dummy variables taking the value 1 if any of *dChair1*, *dChair2*, or *dChair3* is equal to 1
ln(PublicfundingO)	Natural logarithm of the three-year average of public sector funding for the purpose of operational costs and direct expenditure of research
ln(PrivatefundingO)	Natural logarithm of the three-year average of private sector funding for the purpose of operational costs and direct expenditure of research
ln(NFPfundingO)	Natural logarithm of three-year average of funding from not-for-profit institutions (NFP) for the purpose of operational costs and direct expenditure of research
ln(nbArticle)	Natural logarithm of the yearly number of articles
PubORank	Normalized rank of scientist in the field in terms of three-year average of funding for the purpose of operational costs and direct expenditure of research
PublRank	Normalized rank of scientist in the field in terms of three-year average of articles count
ln(totFund)	Natural logarithm of three-year average of aggregate public sector funding in the field
Age	Age of a scientist
dFemale	Dummy variable taking the value 1 if the scientist is a woman and 0 otherwise
dULaval, dUMcGill, …, dUdeM	Dummy variables indicating the university affiliation of researcher
dMedical, dHumanities, …, dScience	Dummy variables indicating the field of researcher
d2000, d2001, …., d2012	Dummy variables indicating the year

Drivers of Higher Education Institutions' Visibility: A Study of UK HEIs Social Media Use vs. Organizational Characteristics

Julie M. Birkholz[1], Marco Seeber[1] and Kim Holmberg[2]

{Julie.Birkholz, Marco.Seeber}@UGent.be
[1]Centre for Higher Education Governance Ghent, Ghent University, Korte Meer 5, 9000, Gent (Belgium)

Kim.J.Holmberg@utu.fi
[2]Research Unit for the Sociology of Education, University of Turku, 20014 Turku (Finland)

Abstract

Social media is increasingly used in higher education settings by researchers, students and institutions. Whether it is researchers conversing with other researchers, or universities seeking to communicate to a wider audience, social media platforms serve as a tool for users to communicate and increase visibility. Scholarly communication in social media and investigations about social media metrics is of increasing interest for scientometric researchers, and to the emergence of altmetrics. Less understood is the role of organizational characteristics in garnering social media visibility, through for instance liking and following mechanisms. In this study we aim to contribute to the understanding of the effect of specific social media use by investigating higher education institutions' presence on Twitter. We investigate the possible connections between followers on Twitter and the use of Twitter and the organizational characteristics of the HEIs. We find that HEIs' social media visibility on Twitter are only partly explained by social media use and that organizational characteristics also play a role in garnering these followers. Although, there is an advantage in garnering followers for those first adopters of Twitter. These findings emphasize the importance of considering a range of factors to understand impact online for organizations and HEIs in particular.

Conference Topic

Science policy and research assessment, Country-level studies, Webometrics, Altmetrics

Introduction

The use of social media increases visibility of users (Constantinides & Zinck, 2011). This online visibility garners success and performance (Schindler & Bickar, 2005; Dellarocas, 2003; Duan et al., 2008). Less understood is the role of offline effects in garnering this visibility. For example, how do organizational characteristics influence an organization's visibility on social media? The understanding of the potential dual role of organizational characteristics and social media use in explaining visibility allows us to delineate how traditional characteristics such as status or reputation of organization play a role in generating attention on social media and how best to measure this impact.

We explore this through the lens of higher education. Social media is increasingly used in scholarly communication. Higher education institutions (HEIs), in particular, are increasingly using social media platforms as tools to communicate to prospective and current students, alumni and society at large (Gibbs, 2002; Helgesen 2008; Hemsley-Brown & Oplatka, 2006). Thus, the case of higher education and institutions' social media use in particular provides a valuable case to explore the possible dual role of organizational characteristics and the use of social media by these institutions in explaining garnered visibility.

In this paper we review literature on visibility of organizations and identify the potential role of social media use and organizational characteristics in explaining this visibility. We propose a number of hypotheses in which social media visibility is dependent on the two. To test these effects, we investigate 137 UK higher education institutions, collecting data of their Twitter activities and characteristics to explain social media visibility. Findings suggest that organizational characteristics of HEIs play a large role in their social media visibilities on

Twitter, compared to social media use alone. This emphasizes the importance of considering a range of factors in understand impact online for both organizations and HEIs in particular. This topic is of interest for scientometric researchers, as it is an additional avenue from bibliometrics to evaluate potential impact of a HEIs. In particular this work contributes to recent research on altmetrics. Altmetrics seeks to investigate the potential use of social media metrics for research evaluation and mapping of scholarly communication (Priem et al., 2010). The delineation of this mechanism advances our understanding of metrics validity and sheds light on the practical questions of how organizations can garner visibility online.

Social media and organizations

Organizational visibility is generated by the organization itself, and the users that engage with organizations. Organizational visibility is partly generated through word-of-mouth (WOM). WOM is the practice of communication where information is spread between individuals about a product or a service of a given organization (Richins, 1983). This mechanism allows individuals to share information and opinions to others about specific products, brands and services (Hawkins, Best, Coney, 2004; Westbrook, 1987) and to attach sentiment to messages. Positive WOM influences the awareness, image, decisions, evaluation and interest of potential consumers and stakeholders (Ozcan & Ramaswamy, 2004; Price, Feick & Guskey, 1995).

Organizations in particular are keen to attempt to achieve or maintain this positive WOM through different strategies of communication about the product or service they offer. With nearly half of all US internet users engaging on social networking sites (Smith, 2011), and with the numbers increasing worldwide, it is not a surprise that organizations are also getting involved in communicating via social media. The use of social media by organizations has largely been seen as marketing strategy to increase visibility (Constantinides & Zinck, 2011).

Social media in particular serve as platforms for electronic WOM where entities spread and share information, but also as a medium where identification of organizational interests is transparent through online liking or following mechanisms (Dellarocas, 2003). Social media platforms serve as sites of social interaction, communication and marketing. This is achieved through socializing and networking online through text, images and videos. These platforms are largely made of user-generated content and facilitated through peer-to-peer communication and participation (Nambisan & Nambisan, 2008; Shankar & Malthouse, 2009).

A number of positive outcomes have been attributed to the use of social media by organizations. The use of social media platforms and thus consequent eWOM around a product or service of an organization influences attitudes, intentions and buying decisions (Schindler & Bickart, 2005; Goldsmith & Horowitz, 2006; Yao, Dresner & Palmer, 2009). The use of social media has also been attributed to increased economic impact (Chevalier & Mayzlin, 2006; Dellarocas, 2003). Recent work has questioned the impact of social media use on outcomes, suggesting that online content is solely a predictor of economic success, and not a factor that influences buying decisions (Chen et al., 2011; Duan et al., 2008). Follow-up studies suggest that user consult the Web for a confirmation of a decision they have made about a product, service or organization (Schindler & Bickart, 2005). Thus, this questions the explanatory power of social media use in garnering different outcomes, suggesting that other information about an organization or its product or service may play a role in understanding this garnered visibility online.

External to social media, the organization has a reputation, status and perceived legitimacy of an organization (Baum & Oliver, 1991; DiMaggio & Powell, 1983). Qualities such as status are said to determine a part of users'/consumers' expectations of future qualities of organizations (Podolny, 1993), which aid in defining the visibility and positions of an

organization in a field (Wry et al., 2009). Consequently, the degree of visibility of higher education institutions is not only dependent on the institution's use of social media for exposure, but also on certain organizational characteristics. Thus, we question: in addition to the use of social media platforms, how do organizational characteristics influence online visibility?

Higher education institutions and social media

In this paper we investigate organizations in the system of higher education. With higher education we mean the organizations that organize education and research, such as universities. Higher education is an industry in which consumers are often under informed in the sense that they cannot objectively evaluate the quality of the service before they actually "purchase it" (Jongbloed, 2003). Thus visibility about the organization is highly dependent on word-of-mouth practices to foster interest of potential students, research funding, and public support.

There is a rise of social media use by higher education institutions as tools in communicating information about the organization to prospective and current students, alumni and society at large (Gibbs, 2002; Helgesen, 2008; Hemsley-Brown & Oplatka, 2006). Social media fill a gap in the information that these groups cannot find in other forms of communication (Hemsley-Brown & Oplatka, 2006) such as alternative contact points for education and campus life (Yu et al., 2010; Mason & Rennie, 2007). Research shows that social media serves to fill a gap in the information that those interested in a university cannot find on the websites (Hemsley-Brown & Oplatka, 2006). Studies have found a significant relationship between those who logged onto the social media platform and the likelihood of them applying to the university (Hayes, Ruschman, & Walker, 2009). Thus, social media by higher education institutions serves said to play a positive role in garnering visibility through different methods.

On the other hand, recent studies in webometric studies of scholarly communication Web indicators or altmetrics have frequently been compared against more traditional indicators of research productivity (such as number of publications) and research impact (citations). Studies on the individual level found significant correlations between traditional bibliometric metrics, for instance research productivity and online visibility (Bar-Ilan, 2004; Thelwall & Tang, 2003). This relationship has been attributed to highly cited scholars producing more content on the web, which then attracted more attention (Thelwall & Harries, 2003). This has also been found in recent studies on HEIs, questioning how social media platforms play a lesser role than other forms of communication in attracting students in particular (Constantinides & Zinck Stagnothe, 2011), as well as the role of geographical proximity in the likelihood of universities in particular to link with other universities (Heimeriks & Van den Besselaar, 2006).

This is not necessarily striking given that HEIs have reputations external to the messages disseminated on social media platforms. Organizations are expected to capitalize on a baseline visibility as scholars have shown that organizations with a central position in the system, related to the organizational size, status and reputation, receive more attention from audiences and stakeholders (Wry et al., 2011, Podolny, 1993). Recent works in webometrics have also demonstrated that core organizational attributes matter in explaining online communication; where status, reputation and size are important predictors of hyperlink connections and centrality (Seeber et al., 2012, Lepori et al., 2013). Thus, using a social media platform does not alone garner visibility or interest from others. Given this we propose:

Hypothesis 1: Social media visibility can be explained by the social media use of the organization.

Hypothesis 2: The social media visibility of the organization can be explained by a HEIs organizational characteristics related to organizational size, status and reputation.
Hypothesis 3: The social media visibility of the organization can be explained by both the HEI's social media activity and organizational characteristics related to organizational size, status and reputation.

Methodology

We explore in this study UK universities, investigating both their Twitter activity and organizational characteristics. In selecting a social media platform where HEIs are active we have selected Twitter. Twitter is especially efficient for word-of-mouth marketing, given the ability to *re-tweet* – forward messages from users (Jansen et al., 2009) In addition tweets often contain expressions of sentiments (ibid), which makes it a valid source for identifying practices driven by potential eWOM. Following the theoretical framework, we assume that followers are a function of the organizational attributes and the social media use of the university.

Sample

Alike most European universities, UK universities are public institutions and the State and related funding bodies represent the most important funding sources.[1] On the other hand, UK universities are autonomous institutions, provided with strong decision making hierarchies and operating in a competitive system, they are expected to be able and in need of developing strategies to actively improve their position in the system (de Boer & Jongbloed, 2012; Seeber, et al., forthcoming). In turn, the UK Higher Education is a suitable case to explore what determines social media visibility in a quasi-market public system. Our sample includes 137 UK HEIs included in the European Micro Data dataset (Eumida) - a database containing the structural characteristics of 2,457 Higher Education institutions in twenty-eight European countries (Bonaccorsi et al., 2010; Eumida, 2009).[2]

Measures

We retrieved data from the HEIs' Twitter accounts manually. This data was collected on 24 November 2014 to measure the dependent variable of visibility and the independent variable - social media use. We also collected data on the organizational characteristics of the institutions, the second independent variable, for measuring a number of characteristics of the HEIs.

Visibility

We focus in this paper on social media visibility. This is a count variable that identifies the number of followers of each UK HEIs.

[1] HESA statistics on finance of UK universities available at: https://www.hesa.ac.uk/
[2] EUMIDA data refer to year 2007. Originally it included 148 universities, although four institutions have merged in the meanwhile, leading to a sample of 144. The Institute of Cancer Research and the London School of Hygiene and Tropical Medicine were excluded, as they are research institutes rather than HEIs; as well as the University of Southampton as it missed a value on coreness, one of the major predicting variables. Four outliers cases in terms of the number of followers were also excluded, leading to a sample of 137 UK HEIs; the University of Oxford, with 175,000 followers, The University of Cambridge 151,000, the Open university 100,000 and the London Business School 69,800, compared to a mean of 20,217 and standard deviation of 21,466.

Social media use

Scholarly communication in social media has been measured in a number of ways. Following literature suggesting a combination of activities we seek to identify attributes of the ways that HEIs use social media. Aguillo (2009) suggested using Web data as indicators related to 1) activity, 2) impact, and 3) usage. Indicators related to activity include measurements of the efforts made to actively create and establish a Web presence, while impact is the mentions on and linking from other websites. Usage is a proxy for the number of downloads or how users engage with the organization on the web. Given these metrics we sought to collect any query-able data on Twitter use. We collected data on the total number of tweets sent, the number of users that the HEIs themselves are following as a measure of their activity. Data was collected the date of HEI's first tweets obtained from the Twitter website[3]. In addition we collected data on the HEIs using Twitter to disseminate and share news and events or targeting students, as indicated by the HEIs in their profiles.

Organizational characteristics

We selected organizational characteristics that are deemed to be particularly relevant for the visibility of universities. We sought to identify on a number of measure of the universities' size, age, resources and status. The organizational features were constructed by using information from Eumida (Bonaccorsi, et al., 2010; Eumida, 2009). We considered, in particular; a) the size of the university, in terms of the number of staff units and undergraduate students; b) the university reputation in the core activities of research, measured through the scientific productivity and the research intensity, and teaching, measured through the teaching burden c) the university status, which is measured through the relational centrality of the university in the system (Owen-Smith & Powell, 2008). As control variables we considered; a) the discipline profile, as some disciplines may attract more attention than others because of the societal salience of the topics addressed, and b) the geographical context, in terms of the urban centrality of the city where the university is located, which may indirectly benefit the university's visibility. Table 1 describes the characteristics of each variable.

Results

Descriptive Statistics

Tables 2 and 3 present respectively the descriptive statistics and the Pearson correlation of the considered variables. The distribution of followers is moderately right skewed, as well as the number of Tweets, whereas the number of following is strongly right skewed. The days on Twitter is left skewed, as most universities started using twitter in early days and a small number of universities are late adopters (Table 2). Pearson correlations show that the number of followers is significantly correlated to most of the considered variables, and in particular to the status-coreness of the university (0.693), size measured by units of staff (0.642) and students (0.477), and scientific productivity (0.452). These organizational characteristics are strongly correlated with each other, so that high status universities are also large, and have a good scientific reputation. Variables of social media use are weakly correlated among each other and the organizational characteristics, with the highest correlations existing between the number of tweets and the size in terms of number of undergraduate students (0.264) (Table 3). The descriptive statistics show that the number of Twitter followers are characterized by over dispersion (i.e., the variance increases faster than the mean).

[3] https://discover.twitter.com/first-tweet#username

Table 1. Description of organizational characteristics.

Size	The *number of total staff* (Full Time Equivalents measured in thousands), including academic as well as administrative and technical staff. The *number of undergraduate students*. (Eumida)
Reputation in research	Universities reputation in research activity is strongly related both to the *scientific productivity*, e.g. the quantity and quality of scientific publications. The indicator results from the product between the total number of publications multiplied by their field-normalized impact factor and divided by the number of academic staff. Data for two-thirds of the universities could be derived from the SCIMAGO institutional rankings for the year 2011 (http://www.scimagoir.com/), which is based on publications from the period 2005-2009; One-third of the universities are not covered since they had less than 100 publications in Scopus in the considered period. For these universities the indicator was set to zero. In fact, the scaling properties of research output (van Raan, 2007) maintain that the individual productivity tend to correlate with the organizational output, so that the indicator approaches zero when the level of output approaches the threshold of 100 publications. A second indicator of reputation in research considers the *research intensity*, as measured by the ratio between the number of PhD students over undergraduate students (Bonaccorsi, et al., 2007). (Eumida)
Reputation in teaching intensity	Teaching quality can be expected to be inversely related to the *teaching burden*, as measured by the ratio between the number of undergraduate students per unit of academic staff. (Eumida)
Status	University status is measured through the relational centrality or *coreness* in the system, estimated by considering web links connections between universities. Weblinks are receiving increasing attention in the study of inter-organizational relationships (Bar-Ilan 2009). European national higher education systems have been shown to conform to a core-periphery structure, where a status hierarchy is in place, core actors holding higher status and the *coreness* measuring the proximity to the network center (Borgatti & Everett, 1999; Lepori, et al., 2013; Owen-Smith & Powell, 2008).
Control: discipline profile	The disciplinary profile is defined by the share of academic staff employed in each of six subject domains considered in Science classification statistics (Eumida, 2009; Uoe, 2006). A Factor Analysis identifies three factors; separately employed as predicting variable. (Eumida)
Control: geographical context	The *Urban centrality* of the city where the university is located is measured through the Globalization and World Cities Network (GARC) scale of cities 2010 (Taylor, 2004) http://www.lboro.ac.uk/gawc/world2010.html). Accordingly, we ranked the universities with a numeric score from 9 (alpha++ cities) to 1 (gamma- cities), setting to zero the cities that are not in the list[1]

Table 2. Variables' descriptive statistics.

		Mean	*Median*	*Maximum*	*Minimum*	*Standard Deviation*
1	size - units of staff	2.001	1.665	9.498	68	1.675
2	size - undergraduate students	13.826	13.356	33.640	351	8.462
3	reputation - scientific productivity	274,66	72,50	1.828,00	0,00	389,03
4	reputation - research intensity	0,04	0,02	0,27	0,00	0,05
5	reputation - teaching burden	8,14	7,89	28,03	1,78	3,80
6	status - coreness	68	66	173	0	45
7	urban centrality	2,2	0,0	9,0	0,0	3,5
8	number of followers	17.189	15.900	46.200	1.233	10.085
9	number of tweets	6.792	5.598	19.000	300	4.220
10	days on twitter	1.918	2.019	2.644	305	342
11	number of following	1.312	832	12.700	107	1.506

Results of models

The dependent variable is represented by the number of Twitter followers, and assume that the number of followers is a function of the organizational attributes and the social media use of the university. Hence, we rely on techniques used for modelling count data for series of non-negative integers. If individual events are independent and their number is sufficiently large, the resulting probability distribution for the counts follows a Poisson distribution. Unlike linear regressions, the Poisson regression model does not assume that observations are normally distributed around the conditional mean, see Table 3. The descriptive statistics show that the number of Twitter followers are characterized by over dispersion (i.e., the variance increases faster than the mean). We then employ a negative binomial regression, which includes a parameter to model over dispersion. Table 4 presents the results of models: i) the empty model; ii) the model including the significant organizational characteristics; iii) the model employing the variables of social media use; iv) the full model including significant organizational characteristics and social media use variables.

Findings show that the social media are significant predictors of the number of followers, with the exception of the *number of following* (Hypothesis 1). In particular the *number of tweets* and *days on Twitter* have a positive effect; the orientation towards *news and events* has a positive and highly significant effect when compared to a *general* orientation. Findings also show that the organizational characteristics are predictive of the number of followers (Hypothesis 2). The *size in terms of undergraduate students* and the *research intensity* have a positive and strongly significant effect. Despite the lower correlation with the number of followers, these two measures are better predictors, respectively than the *size as number of staff units* and the *scientific reputation*. The variable on *status – coreness* is also strongly significant and positive. The *teaching burden,* the *discipline profile* as well as the *urban centrality* of the university' location are not significant predictors.

Comparatively speaking, the organizational characteristics model perform considerably better than the model on social media use.[4] However, the final model (Hypothesis 3) displays that the better fit includes both organizational characteristics and social media use variables as regards the *number of tweets* and the *days on Twitter*.[5] All variables have a positive and strongly significant effect. In order to assess the predictive capability of the full model we cannot rely on usual fit measures, like the R^2, which assume a normal distribution. The model provides expected count values of followers, so that the fit can be judged by: a) computing a pseudo R^2 based on the formula: 1 – (Total Sum Squared/Residual Sum Squared); b) computing the percentage of observed counts correctly predicted. The Pseudo R^2 is 0.66.[6] Further, we consider the capability of the full model to correctly predict values below and above the median of 15,900 followers. The model correctly identifies 92% of the values below the median (*sensitivity*) and, when it predicts a value below the median, it is correct in 79% of the cases (*positive predictive value*). The performance is also good in terms of detecting the values above the median (67%, *specificity*); when the model predicts a value above the median, it is correct in 80% of the cases (*negative predictive value*). In sum, the overall predicting capability of the full model is fairly good. Figure 1. below displays a graphical depiction of these results, related to Twitter followers and organizational characteristics.

Binomial regression coefficients are exponential and multiplicative: if the coefficient for an antecedent is β, then the percentage change in the expected number of counts for unit a

[4] Akaike Information Criterion - AIC (Akaike, 1998) of the null model is 2898.6, social media model AIC 2822.1 vs. organizational characteristics model AIC 2871.3, where lower values indicate a better fit.

[5] Test for multicollinearity, VIF variance inflation factor, all variables well below the threshold of 10, the highest value observed for coreness at 2.62.

[6] Pearson correlation between predicted and actual values is 0.826.

change in the antecedent is e^{β}. For instance, if the university "A" have 8,462 students more than university "B" (one standard deviation), it is predicted that A will have 1.16 times the number of followers of "B" (+16%).[7] The observed coefficients confirm that both organizational characteristics and the specific use of social media have an important impact on the number of followers (Table 5).

Outliers

As a final test, we explore the capability of the full model to predict the four outlier cases that were excluded from the sample in a first stance. Whereas the number of followers of the Open University is reasonably well predicted (129,825 vs. 100,000 followers), the University of Oxford (60,180 vs. 175,000), the University of Cambridge (85,692 vs. 151,000), and the London Business School (12,624 vs. 69,800), attract a much larger number of followers than predicted by the model.

Table 3. Pearson correlation between the selected variables.

Table 3 - Pearson correlation between the selected variables

		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	size - units of staff	1	,683**	,575**	,427**	-,291**	,804**	-,006	,513**	-,098	-,182*	,642**	,112	-,159	,183*
2	size - undergraduate students	,683**	1	,187*	-,065	,176*	,564**	-,152	,459**	,057	-,208*	,477**	,264**	,046	,106
3	reputation - scientific productivity	,575**	,187*	1	,495**	-,370**	,596**	,065	,465**	-,175*	-,100	,452**	-,035	-,107	,188*
4	reputation - research intensity	,427**	-,065	,495**	1	-,411**	,444**	,238**	,246**	-,038	-,019	,347**	-,185*	-,147	,029
5	reputation - teaching burden	-,291**	,176*	-,370**	-,411**	1	-,298**	-,107	-,173*	,095	-,056	-,230**	,090	,091	-,092
6	status - coreness	,804**	,564**	,596**	,444**	-,298**	1	-,046	,566**	,132	-,219*	,693**	,159	-,052	,145
7	urban centrality	-,006	-,152	,065	,238**	-,107	-,046	1	-,147	-,162	,044	-,052	-,290**	-,142	,017
8	discipline profile - factor 1	,513**	,459**	,465**	,246**	-,173*	,566**	-,147	1	,000	,000	,336**	,107	-,076	,085
9	discipline profile - factor 2	-,098	,057	-,175*	-,038	,095	,132	-,162	,000	1	,000	,066	,089	,060	-,069
10	discipline profile - factor 3	-,182*	-,208*	-,100	-,019	-,056	-,219*	,044	,000	,000	1	-,252**	-,121	-,114	-,058
11	number of followers	,642**	,477**	,452**	,347**	-,230**	,693**	-,052	,336**	,066	-,252**	1	,323**	,294**	,326**
12	number of tweets	,112	,264**	-,035	-,185*	,090	,159	-,290**	,107	,089	-,121	,323**	1	,120	,158
13	days on twitter	-,159	,046	-,107	-,147	,091	-,052	-,142	-,076	,060	-,114	,294**	,120	1	,033
14	number of following	,183*	,106	,188*	,029	-,092	,145	,017	,085	-,069	-,058	,326**	,158	,033	1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 4. Negative Binomial regression models.

Table 4 - Negative Binomial regressions models

	Empty Model			Organizational variables Model			Social media use Model			Full Model		
	Estimate	*S.E.*	*Pr(>|z|)*	*Estimate*	*S.E.*	*Pr(>|z|)*	*Estimate*	*S.E.*	*Pr(>|z|)*	*Estimate*	*S.E.*	*Pr(>|z|)*
Intercept	9,752	0,054	<2e-16 ***	8,862	0,089	<2e-16 ***	8,371	289,300	<2e-16 ***	7,671	0,230	<2e-16 ***
size - undergraduate students				0,000023	0,000007	0,0007***				0,000018	0,000006	0,0035**
research intensity				2,774	1,088	0,01*				3,416	1,013	0,0007***
coreness				0,005	0,001	0,0002***				0,005	0,001	0,0004***
Tweets							0,00004	0,00001	0,0003***	0,00003	0,00001	0,0004***
days twitter							0,001	0,000	0,0003***	0,00053	0,00011	0,000002***
orientation: news and events							0,296	0,113	0,009**			
orientation:students							-0,312	0,183	0,09 .			
Null deviance	145,96	on 136 df		252,25	on 136 df		183,82	on 136 df		304,75	on 136 df	
Residual	142,49	on 136 df		142,25	on 133 df		144,15	on 132 df		141,37	on 131 df	
AIC:	2898,6			2822,1			2871,3			2798,4		
log-likelihood:	-2894,6			-2812,1			-2859,3			-2784,4		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[7] Changes in different antecedents have a multiplicative impact on expected number of followers. Hence, for instance, a university that is a standard deviation larger and research intensive than a university B will have 37% more followers (1.16*1.18 = 1.37).

Table 5. Negative binomial regression model: comparing the impact of the variables.

		delta: standard deviation	proportion in number of followers
1	size - undergraduate students	8'462	1.16
2	research intensity	0.049	1.18
3	Status - coreness	45	1.24
4	Tweets	4'220	1.15
5	days twitter	342	1.20

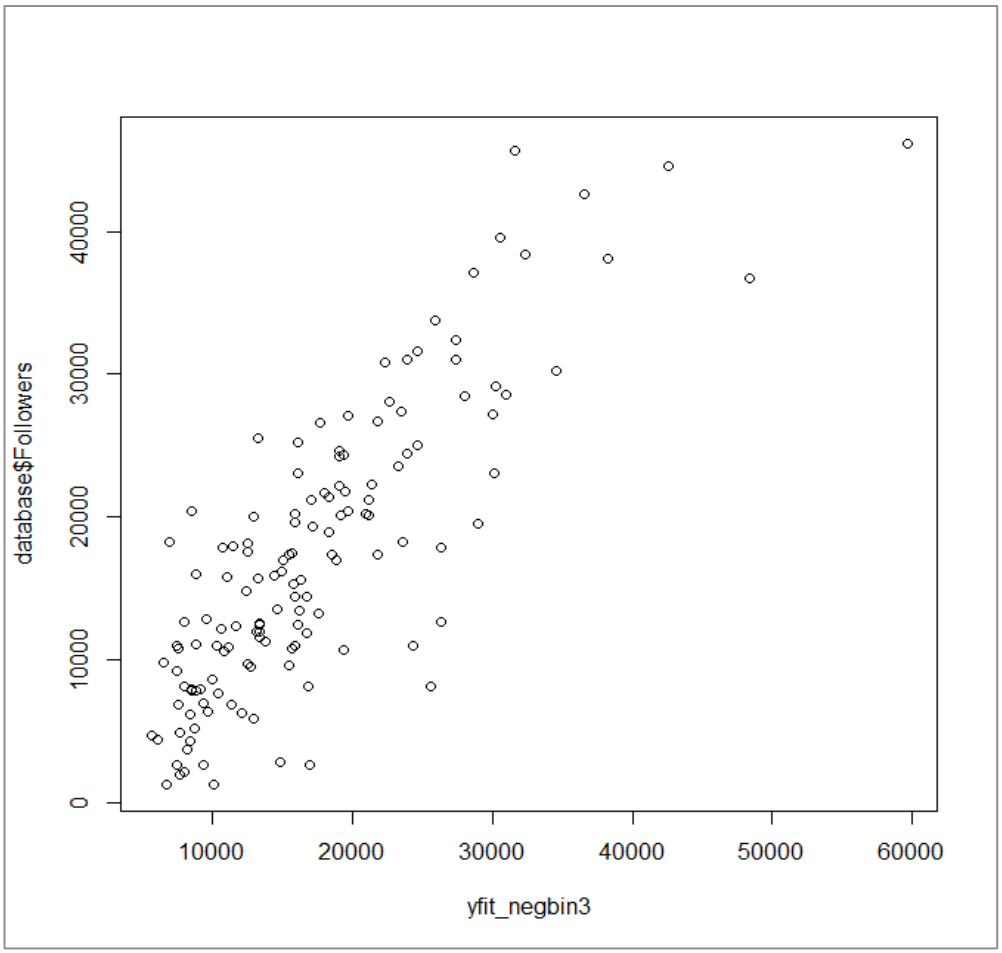


Figure 1. Results from full model.

Discussion and conclusions

Findings indicate that both social media use and organizational characteristics explain the social media visibility of HEIs. Thus, organizations may be successful in garnering followers through their Twitter activity, but these high number of followers is also attributed to the organizational characteristics of size, status, and reputation. Notable is that these characteristics were better predictors of followers than the use of Twitter, suggesting that visibility is highly influenced by offline activities and traditional WOM, compared to eWOM. Although in regards to altmetrics – these online metrics do provide valid proxies for understanding dynamics, the addition of organization characteristics allows us to question how they serve as proxies, as the correlations suggest followers and following seem to be related to organizations size, and reputation, although the organizations own activities of tweeting and experience on Twitter are not related. That does not discard the power of social

media platforms as a tool for garnering visibility, although emphasizes that it is not a replacement for building reputation external to online domains.

Findings show that the social media are significant predictive of the number of followers, with the exception of the *number of following* (Hypothesis 1). In particular the *number of tweets* and *days on twitter* have a positive effect. Findings also show that the combined organizational characteristics are predictive of the number of followers (Hypothesis 2). The *size in terms of undergraduate students* and the *research intensity* have a positive and strongly significant effect. Despite the lower correlation with the number of followers, these two measures are better predictors, respectively than the *size as number of staff units* and the *scientific reputation*. The variable on *status – coreness* is also strongly significant and positive. The *teaching burden,* the *discipline profile* as well as the *urban centrality* of the university's location are not significant predictors.

In addition to the specific a number of notable findings emerged with regards to the specific variables. First, the importance of length of time on Twitter suggests a "first mover advantage", where first adopters have yielded higher numbers of followers. HEIs Twitter accounts that had an orientation towards news and events play a more significant role in garnering online visibility through followers. Secondly, in regards to the organizational characteristics *size in terms of undergraduate students* and *research intensity* played the most significant role in explaining online followers. These two measures reflect the two core tasks of HEIs – research and education. That is HEIs that are able to attract a high number of students as well as sustain a higher number of PhD candidate to conduct research, which again garners increased social media visibility.

This study provides clear support for a causal mechanism that stipulates that both organizational characteristics and social media use explain social media visibility as measure by followers. This provides additional evidence to scientometricians of the importance of considering a combination of metrics in explaining impact and scholar impact in particular. Although, in this research we have analyzed basic descriptors. There is margin for improving explanation of social media use. Future research should investigate, for instance, the content of tweets, as well as the strategies for managing eWOM (Bao & Chang 2014). In addition, the existence of a few outliers suggests that few actors attract a disproportionally high attention from the public. Future research may investigate why this occurs. Given the state of literature we did not have evidence at the onset of our model to suggest an interaction effect, although given that the explanatory power of an organizations social media visibility is explained by both organizational characteristics and social media use, an interaction effect is a natural next step. For example, to investigate the effect of social media use by HEI on (social media) visibility is enhanced in HEIs with a large size, high status and high reputation.

References

Aguillo, I. (2009). Measuring the institutions' footprints in the web. *Library High Tech*, *27*(4), 540-556.

Akaike H. (1998). Information theory and an extension of the maximum likelihood principle, in (eds) *Selected Papers of Hirotugu Akaike*, 199-213. New York: Springer.

Bao, T., & Chang, T. L. S. (2014). Finding disseminators via electronic word of mouth message for effective marketing communications. *Decision Support Systems*, *67*, 21-29.

Baum, J. A. C., & Oliver, C. (1991). Institutional linkages and organizational mortality. *AdministrativeScience Quarterly*, *36*, 187–218.

Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country – the case of Israel. *Scientometrics*, *59*(3), 391-403.

Bar-Ilan J. (2009). Infometrics at the beginning of the 21st century - A review. *Journal of Infometrics*, *2*(1), 1-52.

Bonaccorsi A., Daraio, C., Lepori, B. & Slipersaeter, S. (2007). Indicators on individual higher education institutions: addressing data problems and comparability issues. *Research Evaluation*, *16*(2), 66-78.

Bonaccorsi A., Lepori, B., Brandt, T., De Filippo, D., Niederl, A., Schmoch, U., Schubert, T. & Slipersaeter, S. (2010). Mapping the European higher education landscape. New insights from the EUMIDA project. Science and Technology Indicators Conference, Leiden, the Netherlands, 9-11 September.

Borgatti S. P. & Everett, M. G. (1999). Models of core/periphery structures. *Social Networks*, *21*, 375-395.

Chevalier, J. A. & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345-354.

de Boer H. & Jongbloed, B. (2012). A Cross-National Comparison of Higher Education Markets in Western Europe, in A. Curaj, P. Scott, L. Vlasceanu and L. Wilson (eds) *European Higher Education at the Crossroads: Between the Bologna Process and National Reforms*, 553-571. Dordrecht: Springer Netherlands.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, *48*, 147-160.

Chen, Y., Fay, S., & Wang, Q. (2011). The role of marketing in social media: How online consumer reviews evolve. *Journal of Interactive Marketing*, *25*(2), 85-94.

Constantinides, E., & Zinck Stagno, M. C. (2011). Potential of the social media as instruments of higher education marketing: a segmentation study. *Journal of Marketing for Higher Education*, *21*(1), 7-24.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*, 49(10), 1407-1424.

Duan, W., Gu, B., & Whinston, A. B. (2008). The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. *Journal of Retailing*, 84(2), 233-242.

Eumida (2009). EUMIDA Handbook - Collection of institutional-level data on tertiary educational institutions at the European level. Eumida.

Gibbs, G. (2002). Institutional strategies for linking research and teaching. *Exchange*, *3*, 8–11.

Goldsmith, R. E., & Horowitz, D. (2006). Measuring motivations for online opinion seeking. *Journal of Interactive Advertising*, *6*(2), 2-14.

Hawkins, D.I., Best, R. & Coney, K.A. (2004). *Consumer Behavior: Building Marketing Strategy*. 9th ed. Boston: McGraw Hill.

Hayes, T. J., Ruschman, D., & Walker, M. M. (2009). Social networking as an admission tool: A case study in success. *Journal of Marketing for Higher Education*, *19*(2), 109-124.

Heimeriks, G. & Van den Besselaar, P. (2006). Analyzing hyperlinks networks: the meaning of hyperlink based indicators of knowledge production. *Cybermetrics*, *10*(1), paper 1.

Helgesen, Ø. (2008). Marketing for higher education: A relationship marketing approach. *Journal of Marketing for Higher Education*, *18*(1), 50-78.

Hemsley-Brown, J., & Oplatka, I. (2006). Universities in a competitive global marketplace: systematic review of the literature on higher education marketing. *International Journal of Public Sector Management*, *19*(4), 316-338.

Jansen, B.J., Zhang, M., Sobel, K. & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, *60*(11), 2169-2188.

Jongbloed, B. (2003). Marketisation in higher education, Clark's triangle and the essential ingredients of markets. *Higher Education Quarterly*, *57*(2), 110-135.

Lepori B., Barberio, V., Seeber, M. & Aguillo, I. (2013). Core-periphery structures in national highereducation systems. A cross-country analysis using interlinking data, *Journal of Infometrics*, *7*(3), 622-34.

Mason, R. & Rennie, F. (2007). Using web 2.0 for learning in the community. *Internet and Higher Education*, *10*, 196–203.

Nambisan, S., & Nambisan, P. (2008). How to Profit From a Better Virtual Customer Environment. *MIT Sloan Management Review*, *49*(3), 53.

Owen-Smith J. & Powell, W. W. (2008) 'Networks and Institutions', in R. Greenwood, C. Oliver, K. Shalin & R. Suddaby (eds) *The Sage handbook of organizational institutionalism*, 594-621. London.

Ozcan, K. & Ramaswamy, V. (2004) Word-of-mouth as dialogic discourse: A critical review, synthesis, new perspective, and research agenda. Working Paper. Retrieved June 15, 2015 from: http://kerimcanozcan.com/portal/downloads/Word-ofMouth%20as%20Dialogic%20 Discourse.pdf.

Price, L. L., Feick, L. F., & Guskey, A. (1995). Everyday market helping behavior. *Journal of Public Policy & Marketing*, *14*(2), 255-266.

Priem J., Taraborelli, D., Groth, P., & Neylon, C. (2010). alt-metrics: a manifesto. Retrieved June 15, 2015 from: http://altmetrics.org/manifesto/.

Podolny, J. M. (1993). A status-based model of market competition. *American Journal of Sociology*, *98*(4), 829-872.

Richins, M. L. (1983). Negative word-of-mouth by dissatisfied consumers: a pilot study. *The Journal of Marketing*, *47*(1), 68-78.

Schindler, R. M., & Bickart, B. (2005). Published word of mouth: Referable, consumer generated information on the Internet. In *Online consumer psychology: Understanding and influencing consumer behavior in the virtual world*, 35-61.

Seeber M., Lepori, B., Montauti, M., Enders, J., De Boer, H., Weyer, E., Bleiklie, I., Hope, K., Michelsen, S., Nyhagen, G., Frohlich, N., Scordato, L., Stensaker, B., Waagene, E., Dragsic, Z.,Kretek, P., Krücken, G., Magalhanes, A., Ribeiro, F., Sousa, S., Veiga, A., Santiago, P., Marini, G. & Reale, E. (2015). European Universities as Complete Organizations? Understanding Identity, Hierarchy and Rationality in Public Organizations. *Public Management Review*.

Seeber, M., Lepori, B., Lomi, A., Aguillo, I., & Barberio, V. (2012). Factors affecting web links between European higher education institutions. *Journal of Informetrics*, *6*(3), 435-447.

Shankar, V., & Batra, R. (2009). The growing influence of online marketing communications. *Journal of Interactive Marketing*, *23*(4), 285-287.

Smith, A. (2011). Why Americans use social media. Pew Research Internet Project. Retrieved June 15, 2015 from: http://www.pewinternet.org/2011/11/15/why-americans-use-social-media/.

Uoe (2006). *UOE data collection on education systems. Manual: Concepts, definitions, classifications*, Montreal, Paris, Luxembourg: UNESCO, OECD, Eurostat.

Taylor P. J. (2004). World City Network: a Global Urban Analysis. London: Routledge.

Thelwall, M. & Harries, G. (2003). Do the Web sites of higher rated scholars have significantly more online impact? *Journal of the American Society for Information Science and Technology*, *55*(2), 149-159.

Thelwall, M. & Tang, R. (2003). Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics*, *58*(1), 155-181.

van Raan A.F.J. (2007). Bibliometric statistical properties of the 100 largest European universities: prevalent scaling rules in the science system. Available: *arXiv:0704.0889*.

Westbrook, R. A. (1987). Product/consumption-based affective responses and post purchase processes. *Journal of Marketing Research*, *24*(3), 258-270.

Wry, T., Lounsbury, M., & Glynn, M.A. Legitimating new categories of organizations: Stories as distributed cultural entrepreneurship. *Organization Science*, *22*, 449-463.

Yu, A., Tian, S., Vogel, D. & Kwok, R (2010). Embedded social learning in online social networking, in ICIS 2010 Proceedings 2010, Retrieved June 15, 2015 from: http://aisel.aisnet.org/icis2010_submissions/100.

Yao, Y., Dresner, M., & Palmer, J. W. (2009). Impact of Boundary-Spanning Information Technology and position in Chain on Firm Performance. *Journal of Supply Chain Management*, *45*(4), 3-16.

A Computing Environment to Support Repeatable Scientific Big Data Experimentation of World-Wide Scientific Literature

Bob G. Schlicher[1], James J. Kulesz[2], Robert K. Abercrombie[3], and Kara L. Kruse[4]

[1] *schlicherbg@ornl.gov,* [2] *jim.kulesz@gmail.com,* [3] *abercrombier@ornl.gov,* [4] *krusekl@ornl.gov*
Oak Ridge National Laboratory, Computational Sciences and Engineering Division, 1 Bethel Valley Road, Oak Ridge, TN 37830-6085 (USA)

Abstract

A principal tenet of the scientific method is that experiments must be repeatable. This tenet relies on *ceteris paribus* (i.e., all other things being equal). As a scientific community, involved in data sciences, we must investigate ways to establish an environment where experiments can be repeated. We can no longer allude to where the data comes from, we must add rigor to the data collection and management process from which our analysis is conducted. This paper describes a computing environment to support repeatable scientific big data experimentation of world-wide scientific literature, and recommends a system that is housed at the Oak Ridge National Laboratory in order to provide value to investigators from government agencies, academic institutions, and industry entities. The described computing environment also adheres to the recently instituted digital data management plan, which involves all stages of the digital data life cycle including capture, analysis, sharing, and preservation, as mandated by multiple United States government agencies. It particularly focuses on the sharing and preservation of digital research data. The details of this computing environment are explained within the context of cloud services by the three layer classification of "Software as a Service", "Platform as a Service", and "Infrastructure as a Service".

Conference Topic

Science policy and research assessment, Methods and techniques

Introduction[1]

The scientific policy and research assessment community is investigating methods and techniques to establish an environment where experiments can be repeated through the use of data management. This approach attempts to ensure the integrity of scientific findings and the processes from which scientific literature analysis is conducted.

Data Science is the study of the generalizable extraction of knowledge from data (Dhar, 2013). From this definition, scientific development thus becomes the piecemeal process by which these items have been added, singly and in combination, to the ever growing stockpile that constitutes scientific technique and knowledge (Kuhn, 1970). Scientific literature analysis, or Scientometrics, is the study of measuring and analysing science, technology and innovation. Organizations, such as Thomson Reuters, have long used these analyses to identify the most influential papers or researchers in a field. Recently, Foresight and Understanding from Scientific Exposition (Murdick, 2011) takes this further by mining millions of papers and patents in both English and Chinese, two of the most commonly used languages in scientific literature (Readron, 2014).

[1] This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the United States Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

Scientometrics and its related research activities in today's world make extensive use of digital research data. The data management of this digital research data is, in essence, the quintessential requirement for repeatable scientific experimentation. This term, digital research data, encompasses a wide variety of information stored in digital form including: experimental, observational, and simulation data, codes, software and algorithms, text, numeric information, images, video, audio, and associated metadata. It also encompasses information in a variety of different forms including raw, processed, and analysed data, and published and archived data ("Statement on Digital Data Management," 2014). More specifically, research data are defined in regulation ("Intangible property - Code of Federal Regulations 2 CFR 200.315," 2014), continuing the definition in further statues and United States Government Directives ("2 CFR 215 - Uniform Administration Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (OMB Circular A-110) ", 2012) as follows:

- "Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This 'recorded' material excludes physical objects (e.g., laboratory samples). Research data also do not include:
    - Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and
    - Personnel and medical information and similar information, which the disclosure would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study."

Purpose of the Study

When addressing the reality of allocating the scarce resources of the current research budget constraints, the current institutions of science today operate, essentially the same, as from the time period just after the Second World War (Azoulay, 2012). Azoulay further argues it would be a fortuitous coincidence if the systems that served us so well in the twentieth century were equally adapted to twenty-first-century needs. Such is not the case. To leverage these finite resources and to adhere to the principle of the scientific method that all experiments must be repeatable, we, as a scientific community must investigate ways to establish environments where experiments can be repeated. We can no longer allude to from where the data come, we must add rigor to the data collection and data management process from which our analysis is conducted.

Data management involves all stages of the digital data life cycle including capture, analysis, sharing, and preservation. The focus of this statement is the sharing and preservation of digital research data. The following principles apply to the effective management of digital research data ("Statement on Digital Data Management," 2014):

- Effective data management has the potential to increase the pace of scientific discovery and promote more efficient and effective use of government funding and resources. Data management planning should be an integral part of research planning.
- Sharing and preserving data are central to protecting the integrity of science by facilitating validation of results and to advancing science by broadening the value of research data to disciplines other than the originating one and to society at large. To the greatest extent and with the fewest constraints possible, and consistent with the requirements and other principles of this statement, data sharing should make digital

research data available to and useful for the scientific community, industry, and the public.

- Not all data need to be shared or preserved. The costs and benefits of doing so should be considered in data management planning.

Procedure for a Computing Environment to Support Repeatable Scientific Big Data Experimentation

A data management plan is a formal document that outlines how a research institution and program will handle data both during research and after the project is completed ("Data management plan," 2014). The goal of a data management plan is to consider the many aspects of data management, metadata generation, data preservation, and analysis before the project begins. This ensures that data are well-managed in the present and prepared for preservation in the future. Multiple United States government agencies now require proposals submitted to include a supplementary document labelled “Data Management Plan” (Collins, 2014; "Dissemination and Sharing of Research Results," 2010). These supplementary documents describe how the proposal will conform to scientific policy on the dissemination and sharing of research results.

FUSEnet is a data analytics cloud specializing in managing both data and computational processes for assessing technical knowledge for identifying emergent technologies and capabilities. Under a multi-year United States Government research effort sponsored by Intelligence Advanced Research Projects Activity (IARPA), the overall goal of the FUSE program is to produce a new capability to accelerate the process of identifying and prioritizing emerging technologies across the globe (Murdick, 2011). The FUSE Program was established to develop automated methods that aid in the systematic, continuous, and comprehensive assessment of technical emergence using information found in published scientific, technical, and patent literature. A concise description is as follows (Murdick, 2011):

> A fundamental hypothesis of the FUSE Program is that real-world processes of technical emergence leave discernible traces in the public scientific, technical, and patent literature. FUSE envisions a system that can (1) process the massive, multi-discipline, growing, noisy, and multilingual body of full-text scientific, technical, and patent literature from around the world; (2) automatically generate and prioritize technical terms within emerging technical areas, nominate those that exhibit technical emergence, and provide compelling evidence for the emergence; and (3) provide this capability for literature in English and at least two non-English languages. Technology developed from the FUSE Program would automatically nominate both known and novel technical areas based on quantified indicators of technical emergence with sufficient supporting evidence and arguments for that nomination. The FUSE Program also addresses the vital challenge of validating such a system, using real world data.

FUSEnet is currently a government system hosted by ORNL that stores unclassified, copyright-protected scientific information and provides remote access for approved users to analyse the stored data within a cloud computing environment to satisfy the research objectives of the IARPA FUSE Program. A key tenet within FUSEnet is that data integrity and availability is maintained. An ORNL developed “data diode” embedded within FUSEnet gateways allows access to protected data, but prevents data removal by users. As necessary, a mechanism for approved data export is built into the system architecture. Also by design, the activities and work products of individual user teams are segregated from each other in the cloud computing virtual environment.

FUSEnet Capabilities

The FUSEnet computing environment is based on the Cloud service model. These models are usually described by a three layer classification called SPI for **S**aaS, **P**aaS, and **I**aaS (Tian & Zhao, 2015) and adapted as follows:

- SaaS – Software as a Service: applications that are available on-demand.
- PaaS – Platform as a Service: refers to a computing platform of software components and middleware that are used by end-users to develop and manage their cloud applications. Typically, cloud providers at this layer offer databases, web servers, development environments, and application monitoring tools.
- IaaS – Infrastructure as a Service: physical or virtual machines with access to data storage and other operating system services. The cloud user is typically expected to install and maintain operating-system images.

The unique processing capabilities of FUSEnet are in the SaaS and PaaS levels. The IaaS capabilities were established with off-the-shelf software and hardware solutions as a result of understanding the operational needs of FUSEnet users, big data analytics, and optimizing central processing unit (CPU) and input/output (I/O) performance. One of the major challenges with the computing environment is with moving large volumes of data (terabytes) to and from the disk storage to the CPUs for processing. This challenge is met with ever increasing improvements and replacements for the IaaS without having any operating impact on the SaaS or PaaS layers. FUSEnet demonstrated this with an improvement in the data I/O transfer by replacing the disk storage system over its earlier version. Further, FUSEnet SaaS and PaaS software can be hosted on commercial IaaS platforms that meet the requirements for its intended usage.

A summary of the FUSEnet benefits and capabilities that support repeatability of big data experiments includes:

- An organized repository of 100 million published scientific and patent documents,
- Technical in-house expertise for maintenance of data pertaining to integrity and availability, pedigree, and version control,
- Reliable data sources including data provided by, Thomson Reuters, Lexis-Nexis, Elsevier, Institute of Electrical and Electronics Engineers (IEEE), Nature Publishing Group, PubMed Central, and others,
- Technical expertise with the format and details of the data, and
- Four analytical software applications with evidentiary traceability and indicators for assessing repeatability:
    - Assess and forecast technical research and technology developments,
    - Reverse-search the events contributing to a technology or development,
    - Drill down the evidence supporting the assessment and forecast,
    - Remote end-user workspaces ready-to-run the applications and the analytics platform,
    - Multiple analytics capabilities including Natural Language Processing (NLP), Parts-of-Speech (PoS) detectors, deduplication, belief network modelling, and machine learning,
    - Operation of the system with 24/7 and 99.8% availability within domain-specific expertise with the current ORNL technical staff,
    - Rapid custom development to meet unique end-user analytics requirements, and
    - Immediate data protection for the repository and custom end-user data.

The FUSEnet SaaS Level

At the SaaS level, four unique software applications perform automated technical assessments for supporting the detection and forecasting. Each of these applications process and analyse published scientific and engineering papers that are made available in the FUSEnet data repository. Unlike previous approaches to detecting emergence, which are based on the citation analysis of papers and patents (Bettencourt et al., 2008; Huang et al., 2014), the following application systems extract information from the text of publications and patents, identifying authors, their affiliations, addresses, as well as classifying types of organizations and publications. Although these applications have the same objectives, their analytical techniques are uniquely different and hence provide different insights into the organization and search of the data (Babko-Malaya et al., 2013). These analysis techniques include: feature extraction (Michaelis et al., 2012), time series analysis, sentiment and network analysis (Fürstenau & Rambow, 2012), and emergent detection and prediction (Brock et al., 2012), among others. The four main applications developed within the FUSEnet system are ARBITER from BAE Systems, Copernicus from SRI International, Emerge from BBN, and DETAiLS from Columbia University.

The FUSEnet PaaS Level

The aforementioned SaaS applications use a variety of tools and libraries at the PaaS level. While the SaaS level in FUSEnet is the automated assessment, the FUSEnet PaaS computing platform can best be described as a "Network Analysis" (Otto & Rousseau, 2002) and text analytics platform. Text analysis uses statistical pattern learning to find patterns and trends from text data (in our case, scientific literature and patents). A summary of several key tools that FUSEnet provides are in Table 1. A selection of software libraries for network analysis and text analysis in FUSEnet, available for ensuing that experiments can be repeated, is shown in Table 2.

The FUSE Program licensed and installed a large number of scientific papers and patents from several suppliers in multiple languages including English and Chinese. The data sets include bibliographic citations of journal articles (108+ million), full text journal articles (5+ million), patent backfile records (14+ million at beginning of 2013 for the US and China), and updates to the patent backfile records, (51+ million for the US and China). A backfile is a single file containing the original patent application data plus all updates to the patent (both by the originator and by the patent office) up to the time the backfile was created.

Fig. 1 illustrates the large increase in scientific journal articles and patent applications as included in the FUSE research system during the past two decades. The number of Chinese patent applications is increasing dramatically and has now surpassed the number of US patent applications. Also, the number of Chinese journal articles is increasing at a rate faster than the rest of the world.

Table 1. FUSEnet PaaS support software packages.

	FUSEnet PaaS Analytics Tool	*Technical Usage*	*SaaS application that uses it*
1	MySQL[2]	SQL[3] database typically used to store document, term, and author data.	Emerge, ARBITER
2	MongoDB[4]	Document-oriented, NoSQL database used to store extracted entities and indicator-specific data.	Emerge, Copernicus
3	MALLET	Machine Learning and NLP[5] Toolkit for Java. Provides topic modelling for document clustering.	Emerge
4	Sofia-ml	Fast incremental machine learning algorithm. Provides clustering of documents from topic models generated by MALLET.	Emerge
5	Lucene IR system	Used for its indexing engine.	Emerge
6	Scikit-learn	Machine learning models.	Emerge
7	Tomcat/Solr Web Server	Used for Term indexing.	ARBITER
8	Apache ActiveMQ[6]	Messaging and integration patterns.	ARBITER
9	Cassandra	NoSQL database.	ARBITER
10	Virtuoso	RDF[7] triple storage.	ARBITER
11	OpenRDF/Sesame	RDF processing including parsing, storing, reasoning and querying.	ARBITER
12	Spring Framework	Used for Integration using JMS.	ARBITER
13	Lucene/Solr	Document level information search, retrieval and storage engine.	ARBITER, DETAiLS
14	Open NLP	Machine learning based toolkit for processing natural language text.	ARBITER
15	Netica	Used for working with belief networks and influence diagrams.	ARBITER
16	Elasticsearch	Extension on Lucene that provides search and analytics.	Copernicus
17	Hadoop 2+	Used for extract, transform, and load (ETL) and de-duplication processing.	Copernicus
18	Berkeley Parser	Sorts and assigns words in sentences into subjects, verbs, and objects.	DETAiLS
19	Duke	Deduplication engine written in Java operating with Lucene.	DETAiLS
20	Stanford Chinese Word Segmenter	Split Chinese text into a sequence of words.	DETAiLS
21	Stanford Part-of-Speech (POS) Tagger	Reads text and assigns parts of speech to each word (noun, verb, adjective, etc.).	DETAiLS
22	UIMA	Unstructured Information Management Architecture (UIMA) is a general framework for analysis of unstructured information and its integration with search technologies.	DETAiLS
23	Weka	Machine learning software written in Java for data analysis and predictive modelling.	DETAiLS

[2] MySQL is a well-known relational database manager used in a wide variety of systems, including Twitter, Wikipedia, Facebook, Google, Wordpress, and countless more websites and other applications.
[3] SQL (Structured Query Language) is a special-purpose programming language designed for managing data held in a relational database management system (RDBMS),
[4] MongoDB is a document-oriented, NoSQL database.
[5] NLP is Natural Language Processing where algorithms are used to derive meaning from human language.
[6] Apache ActiveMQ is an open source message broker written in Java together with a full Java Message Service (JMS) client.
[7] RDF is Resource Description Framework and is used to express data in subject-predicate-object expressions.

Table 2. Subset of FUSEnet software libraries for social network and text analysis.

	Library/Package	*Description*	*SaaS application that uses it*
1	Arpack	Linear algebra routines for Java	Emerge
2	JDOM	XML processing library for Java	Emerge
3	Jwnl	Java WordNet library	Emerge, ARBITER
4	Matrix-toolkits-java	Linear algebra data structures for Java	Emerge
5	BLAS	Linear algebra subroutines	Emerge
6	LAPACK	Linear algebra data structures and subroutines	Emerge
7	Libquadmath	High-precision math libraries	Emerge
8	Beanshell	Scripting for Java	Emerge
9	Trove4j	High-performance data structures for Java	Emerge
10	JGrapht	Graphical data structures and algorithms for Java	Emerge
11	JUNG	Java Universal Network/Graph Framework	ARBITER
12	R	Development environment for statistical computing and graphics	ARBITER

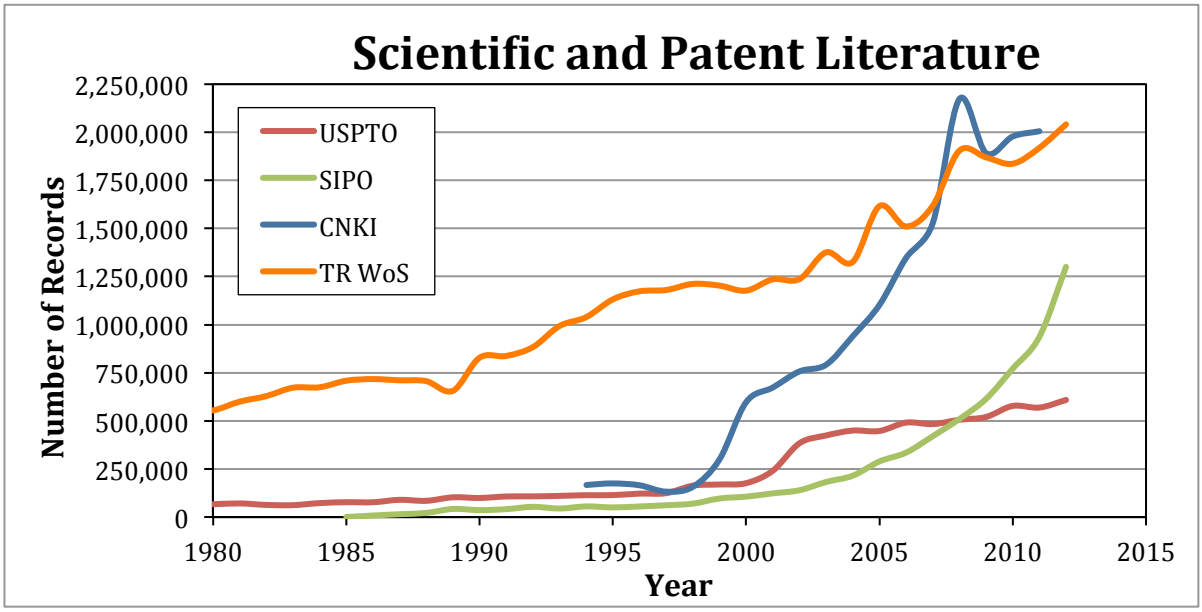


Figure 1. Number of records per year for the four largest datasets in the FUSEnet collection including patent records from the US (USPTO) and Chinese (SIPO) patent offices (i.e. number of backfile records at the beginning of 2013) and journal article citations from China (CNKI) and Thomson Reuters' Web of Science (TR WoS).

The FUSEnet IaaS Level

The deployed second generation FUSEnet at ORNL has the following summary specifications:

- 770 gigaFLOPS[8] of maximum performance,
- 16 blade servers (plus 2 support blades), each with 2 CPUs, each with 6 cores, totalling 192 cores, or processors; additional blade with USB 3.0 for dedicated data transfer/export,
- 3.07 TB of RAM w/ 192 GB per node,
- Disks:
    - EMC Isilon: 340 TB (useable; includes 6.4 TB SSD) running NFS over 10 Gb/s Ethernet,
    - HP LeftHand: 260 TB of effective disk storage; will be reconfigured for backup and
    - Isilon disk I/O up to 1 gigabyte/sec per blade,
- Networking: Flex-10 modules totalling 160 Gbits/sec bandwidth per enclosure x 2 enclosures (theoretical maximum),
- Virtualized computing space through VMware[9],
- Access and control policies enforced by ORNL Computing Data Center, and
- Call Center and metrics for service quality.

Table 3. Characteristics of cloud providers and applicability to FUSEnet requirements.

	Vendors	*Cloud Offering Overview*	*Applicability to FUSEnet*
1	Amazon Web Services	Overall market leader offering virtual servers, MapReduce (Hadoop) for search engine, large data storage, SQL databases, NoSQL databases, mobile integration, business applications including email, payment systems, and workflow.	PaaS (databases), IaaS
2	Google Cloud Platform	App Engine web application platform (PaaS), virtual machines, file storage, SQL databases, NoSQL, big dataset support, mobile integration.	PaaS (databases, web apps), IaaS
3	IBM SmartCloud	SaaS including data warehousing and analytics, business analytics engine, business process management, financial modelling, payment systems, medical analysis, social media analysis, transportation management, medical analytics, SQL databases, NoSQL databases, mobile integration.	SaaS (social media analysis), PaaS (databases, web apps), IaaS
4	Microsoft Azure	Windows or Linux virtual machines, messaging, scheduling, SQL databases, NoSQL databases, mobile integration.	PaaS (databases), IaaS
5	Rackspace Cloud	High bandwidth networking, virtual machines, data storage, process load balancing.	IaaS

Analysis of Technical Requirements and Alternatives versus Commercial Cloud Providers

Representative current cloud solution offerings from commercial vendors include but are not limited to the following: Amazon Web Services (AWS), IBM SmartCloud, Microsoft Azure, Google Cloud Platform, and Rackspace Cloud Servers. Considering the data management, experimentation requirements and the strategic issues, the question arises, "Are the IaaS and

[8] In computing, FLOPS (for FLoating-point Operations per Second) is a measure of computer performance, useful in fields of scientific calculations that make heavy use of floating-point calculations. For such cases, it is a more accurate measure than the generic instructions per second. Computers capable of performing greater than 1 Giga FLOPS are termed as supercomputers.

[9] VMware, Inc. is a software company that provides cloud and virtualization software and service.

PaaS from these selected vendors sufficient for hosting and maintaining the FUSEnet SaaS and PaaS?" A summary of the cloud providers and the offering are described in Table 3.

Analysis of SaaS Technical Alternatives

FUSEnet consists of four unique technical emergence software applications. Current cloud providers are not in the business of providing this niche capability. Cloud providers offer more general SaaS services such as Enterprise Resource Planning (ERP), general accounting, medical, and financial applications for managing business administration operations. If FUSEnet were to be employed on a 3rd party cloud, unique, domain-specific expertise would be required to operate and manage the FUSEnet software applications.

Analysis of PaaS Technical Alternatives

FUSEnet consists of several framework and middleware solutions combined with math-based libraries that are unique to network and text analysis. With the exception of IBM SmartCloud, current cloud providers are not in the business of exclusively providing this niche capability. Cloud providers offer more general PaaS software such as databases, email, and web servers. The features of the network and social analytics tools in SmartCloud should be further evaluated.

Analysis of IaaS Technical Alternatives

FUSEnet is operated in a secured, cloud environment at the Data Computing Center at ORNL. It currently operates on the hardware infrastructure described above. This FUSEnet hardware was performance tested to determine its disk I/O (input/output) throughput under various load conditions. Software programs were used to perform these tests at a low level or 'raw' I/O set of read and write tests and at the application layer with tests that simulated application disk usage. From these initial test results and further repeated testing, the FUSEnet disk I/O was optimized for handling the volume and type of data used in the system. Further tests were performed to compare FUSEnet with another commercial cloud offering, which demonstrated similar or better performance for FUSEnet depending on the operating conditions selected. Currently, the FUSEnet storage system is in its second generation as a result of these performance tests and evaluations. The FUSEnet software and data can be operated on 3rd party (IaaS) environments that can meet the overall system requirements as follows:

- Handle big data that is mixed structured and unstructured and continuously growing.
- Protect selected data and apps (commercial, proprietary) that remain in the cloud.
- Rapidly deploy software solutions to the data.
- Provide virtualization for operating systems including common Linux distributions, Windows and Mac OS.
- Rapidly ingest data into the system.
- Provide the computing performance involving big data analytics software services.
- Provide an easy-to-use big data analytics platform.
- Provide high-performance big data storage and retrieval up to 500 TBs and continue to scale.
- Provide robust, state-of-the-practice cyber security.

In general, commercial firms are advised to consider strategic issues with regards to cloud scope, service levels, and deployment needs. For the FUSEnet environment, Table 4 summarizes these strategic concerns.

The overall need for a secured FUSEnet environment involves the capability to employ software services, such as the analytics described earlier, that uses the data within the FUSEnet cloud, but cannot copy the data out of the cloud. FUSEnet is equipped with custom

middleware software within the PaaS called a Data Diode that monitors activities and prevents the exfiltration of data. Thus, the commercial and proprietary data is protected from being taken outside the FUSEnet enclave (Abercrombie, MacIntyre, & Schlicher, 2011). The Data Diode involves a change to the Linux distro (distribution)[10] so that an IaaS provider must approve the customer to host their own virtualized and configurable operating system (MacIntyre, Paul, & Schlicher, 2011).

Table 4. Strategic issues for the FUSEnet environment.

	Strategic Issue	*Description*	*Assessment for FUSEnet*
1	Cloud Scope – what is the design to meet the need?	Identifies the availability, performance, and security needs; sufficient and planned computing power, storage, and bandwidth.	FUSEnet is monitored daily and reported monthly with the current operational stats: Availability: 99.8%; CPU usage: 12-18%; Memory usage: 56-65%; Storage usage: 69%. FUSEnet is installed with a Data Diode that protects against data exfiltration of its repository. FUSEnet is a virtual environment with separated computing enclaves. Each user or user group within an enclave has the freedom to compose and perform their needed computational research without directly impacting other users.
2	Service Levels	Identifies the expected workload, admin support, service delivery needs, timing and I/O response.	FUSEnet Test and Evaluation (T&E) simulates heavy end-user loading. This is measured to be an increase of 5-10% of the daily load. For its initial usage, FUSEnet could simultaneously host 3-4 heavy end-users loading. The Admin support is at two levels: operating system and the virtual layer through VMware.
3	Deployment Needs	Identifies the integration needs with infrastructure services.	FUSEnet operates on VMware that isolates the PaaS from dependencies on the hardware and the Operating System. The current FUSEnet system, including the number of cores, performance of the cores, memory, and the Isilon storage, is a proven baseline for simultaneously hosting 3-4 heavy end-user loading.

Discussion and Conclusions

This paper addresses science policy with a method and a technique to assess research, increasing its value to the US national scientific community by making available a computing environment to support repeatable scientific big data experimentation of world-wide scientific literature. The computational capability ensures the integrity, availability and confidentiality of new technologies and new technical knowledge. This will position scientific investigators (academic, commercial, and government) with an advantage to address the technical and political challenges all three entities face. FUSEnet offers this unique capability and this paper describes a computing environment necessary to support repeatable experimentation, and recommends a system that is housed at the ORNL Data Center in order to provide value to investigators from a variety of sources while adhering to recently mandated Data Management Planning.

[10] A Linux distribution (often called a distro for short) is an operating system made as a collection of software based around the Linux kernel and often around a package management system

Acknowledgments

We thank colleagues and other reviewers for their assistance and helpful comments. This research was supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Energy (DOE). This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOE, ORNL, or the U.S. Government.

References

2 CFR 215 - Uniform Administration Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (OMB Circular A-110) (2012).

Abercrombie, R. K., MacIntyre, L. P., & Schlicher, B. G. (2011). *Protection of Data in Virtual and Physical Computing Environments (Invention Disclosure Number: 201102659, DOE S-Number: S-124,217)*. Oak Ridge: Oak Ridge National Laboratory.

Azoulay, P. (2012). Research efficiency: Turn the scientific method on ourselves. *Nature, 484*(7392), 31-32.

Babko-Malaya, O., Hunter, D., Amis, G., Meyers, A., Thomas, P., Pustejovsky, J., et al. (2013, May 8-10). *Characterizing Communities of Practice in Emerging Science and Technology Fields.* Paper presented at the 2013 International Conference on Social Intelligence and Technology (SOCIETY).

Bettencourt, L. A., Kaiser, D., Kaur, J., Castillo-Chávez, C., & Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics, 75*(3), 495-518.

Brock, D. C., Babko-Malaya, O., Pustejovsky, J., Thomas, P., Stromsten, S., & Barlos, F. (2012, November 2-4). *Applied Actant-Network Theory: Toward the Automated Detection of Technoscientific Emergence from Full-Text Publications and Patents.* Paper presented at the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposiurm Social Networks and Social Contagion, Arlington, VA.

Collins, D. (2014). *Knowledge Article: Data Management*. Oak Ridge: Oak Ridge National Laboratory.

Data management plan. (2014). Retrieved June 21, 2015 from: http://en.wikipedia.org/wiki/Data_management_plan

Dhar, V. (2013). Data Science and Prediction. *Commununications of the ACM, 56*(12), 64-73.

Dissemination and Sharing of Research Results. (2010). Retrieved June 21, 2015 from:http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Fürstenau, H., & Rambow, O. (2012). *Unsupervised induction of a syntax-semantics lexicon using iterative refinement*. Paper presented at the First Joint Conference on Lexical and Computational Semantics.

Huang, M.-H., Huang, W.-T., Chang, C.-C., Chen, D.-Z., & Lin, C.-P. (2014). The Greater Scattering Phenomenon Beyond Bradford's Law in Patent Citation. *Journal of the Association for Information Science and Technology, 65*(9), 1917-1928.

Intangible property - Code of Federal Regulations 2 CFR 200.315(2014).

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (3rd ed.). Chicago: University of Chicago Press.

MacIntyre, L. P., Paul, N. R., & Schlicher, B. G. (2011). *Data Diode (Copyright Document Number 90000008)*. Oak Ridge: Oak Ridge National Laboratory.

Michaelis, J. R., McGuinness, D. L., Chang, C., Luciano, J. S., & Hendler, J. (2012). *Applying Multidimensional Navigation and Explanation in Semantic Dataset Summarization*. Paper presented at the 11th International SemanticWeb Conference (ISWC 2012).

Murdick, D. A. (2011). Foresight and Understanding from Scientific Exposition (FUSE). Retrieved June 21, 2015 from: http://www.iarpa.gov/index.php/research-programs/fuse

Otto, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*(6), 441-453.

Readron, S. (2014). Text-mining offers clues to success. *Nature, 509*, 410.

Statement on Digital Data Management. (2014). *DOE Office of Science Funding Opportunities* Retrieved June 21, 2015 from: http://science.energy.gov/funding-opportunities/digital-data-management/

Tian, W., & Zhao, Y. (2015). Chapter 1 - An Introduction to Cloud Computing. In W. Tian & Y. Zhao (Eds.), *Optimized Cloud Resource Management and Scheduling* (pp. 1-15). Boston: Morgan Kaufmann.

Is Italy a Highly Efficient Country in Science?

Aparna Basu[1]

[1]*aparnabasu.dr@gmail.com*
Formerly at CSIR-NISTADS CSIR National Institute of Science Technology and Development Studies, New Delhi (India)

Abstract
In an earlier study on measuring national efficiencies in the production of scientific papers and patents of several developed and developing countries (Basu, 2013; 2014a), we found that Italy has the highest efficiency in the production of papers. While this has not gone unnoticed in the literature (Daraio and Moed, 2011) they have taken it as an 'overcompensation effect' and an indication of decline. By examining the work of several authors, we find instances where the information put forward, when taken together, support our findings – that Italy has a high efficiency in scientific publication but only an average efficiency in patenting. We note that Italy's profile along a host of parameters is quite distinct with respect to the OECD average (DeJaeger, 2012). Using a typology of countries based on their publication and patenting efficiencies (Basu, 2014b) we infer that Italy is not one of the countries that have shifted national priorities from publications to patents, like USA, Japan, Germany, or Korea.

Conference Topic
Science policy and research assessment

Introduction
According to Hollanders and Soete, investment on R&D (GERD) is a correlate of development (Hollanders and Soete, 2010). Developed countries have higher GERD shares as compared to GDP shares, the Gross Expenditure on R&D (GERD) being the expenditure on the creation of new knowledge. Countries that have increased R&D expenditures, such that GERD share/GDP share tends to or exceeds unity, are on the path of development. How do increased investments of resources translate into outputs? Do developed countries make more efficient use of their resources? Efficiency of scientific productivity at the national level has been considered earlier by several authors (May, 1997; Rousseau, 1998; King, 2004, Vinkler, 2005, 2008; Shelton, 2008; Leydesdorff & Wagner, 2009; Wendt et al., 2012), who also point out difficulties in making cross-national comparisons. Primarily, they have dealt with publications and citations as compared to research expenditure or GNP and have considered mostly European countries, the US, Japan and China. Rousseau has considered both publications and patents. More recently, Shelton and Leydesdorff have also considered outputs such as patents and number of graduates in addition to papers, using regression models to predict outputs for a given set of inputs (Shelton and Leydesdorff, 2011). Some papers that have used different techniques such as Data Envelopment Analysis (DEA) to study national research productivity and efficiency are Rousseau (1998), Sharma and Thomas (2008) and Lee (2005). According to Hu et al., who used the distance function approach, intellectual property rights protection, technological cooperation among business sectors, knowledge transfer between business sectors and higher education institutions, agglomeration of R&D facilities, and involvement of the government sector in R&D activities significantly improve national R&D efficiency (Hu, et al., 2014)
In our earlier study on the efficiencies of nations in the production of scientific outputs with respect to inputs such as manpower and expenditure in science, we found significant variation in their efficiencies (Basu, 2013, Basu, 2014a). In particular, we noted that the efficiency of production of papers with respect to both expenditure on R&D (GERD) and manpower were the highest for Italy. This fact has not gone unnoticed the literature on Italian science. Daraio

and Moed (2011) did an extensive study on manpower, research expenditure, publications and citations to compare Italy with other productive EU countries. They called Italy "a Cathedral in the desert", but at the same time chose to focus on other factors to argue that Italian science was in decline. Our attempt here is to see if there were other indications in the literature which could have pointed to the fact of Italy's high efficiency, but were missed at the time.

Data and Methodology

Data on scientific papers and patents is taken from the SCI-Expanded and USPTO for the years 2008 and 2007. (The data and analysis are from our earlier papers (Basu, 2013, 2014a) and reproduced here for convenience.) Restricting to the USPTO, the United States Patent Office, gives a bias in favour of the USA termed as the 'home advantage'. Ideally data from some of the other major patent databases such as the European Patent Office EPO should be included in the analysis. However for this preliminary study we have only considered the USPTO.

The Gross Domestic Product GDP and Gross Expenditure on Research and Development GERD for the years 2002 and 2007, are both adjusted to Purchasing Power Parity (PPP) in order to make local investments comparable across countries. Manpower is measured in terms of Full Time Equivalents (FTEs) engaged in R&D. Data is obtained from the UNESCO Science Report 2010 (UNESCO, 2010).

The share of GERD and the share of GDP are shown for a selected set of developing and developed countries Table 1. The GERD/GDP share is an indicator of development (Hollanders & Soete, 2010).

Table 1. GERD and GDP shares of selected countries (2002 and 2007).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Country	GDP share 2002	GDP share 2007	GERD share 2002	GERD share 2007	GERD share/ GDP share 2002	GERD share/ GDP share 2007	GERD share/ GDP share 2007-2002
EU	25.3	22.5	26.1	23.1	1.03	1.03	0.00
USA	22.5	20.7	35.1	32.6	1.56	1.57	0.01
China	7.9	10.7	5	8.9	0.63	0.83	0.20
Japan	7.4	6.5	13.7	12.9	1.85	1.98	0.13
Germany	4.9	4.3	7.2	6.3	1.47	1.47	0.00
India	3.8	4.7	1.6	2.2	0.42	0.47	0.05
France	3.7	3.1	3.9	3.4	1.05	1.10	0.04
UK	3.7	3.2	3.9	3.4	1.05	1.06	0.01
Italy	3.3	2.8	2.2	1.9	0.67	0.68	0.01
Brazil	2.9	2.8	1.6	1.8	0.55	0.64	0.09
Russia	2.8	3.2	2.0	2.0	0.71	0.63	-0.09
Mexico	2.1	2.3	0.5	0.5	0.24	0.22	-0.02
Korea	2.0	1.9	2.8	3.6	1.40	1.89	0.49
Canada	2.0	1.9	2.4	2.1	1.20	1.11	-0.09
Australia	1.3	1.2	1.3	1.4	1.00	1.17	0.17

Table 2 shows the manpower and GERD figures (in FTE's and billion $ PPP) together with the output of papers in the Science Citation Index-Expanded using fractional counts, and patents in the USPTO.

Table 2. Manpower, GERD, Papers and Patents for selected countries.

Country	GERD $bnPPP	Manpower (FTE's)	Papers SCI-E	Patents USPTO
Australia	15.36	87,140	28,313	1,516
Brazil	20.20	133,266	26,482	124
Canada	23.96	139,011	43,539	3,806
China	102.40	1,423,380	104,968	7,362
France	42.89	215,755	57,133	3,631
Germany	72.24	290,853	76,368	9,713
India	24.79	154,827	36,261	741
Italy	22.12	96,303	45,273	1,836
Japan	147.90	709,974	74,618	33,572
Korea	41.30	221,928	32,781	6,424
Mexico	55.90	37,930	8,262	81
Russia	23.40	451,213	27,083	286
Spain	19.34	130,896	35,739	363
UK	41.04	261,406	71,302	4,007
USA	398.00	1,425,550	272,879	81,811

Definitions

To define efficiency we have considered some inputs and outputs in the science system, and their ratio ouput/input. The inputs have been taken as the expenditure and manpower in research. The outputs are scientific patents and papers published by the nations. For two inputs and two outputs there are four possible components of efficiency (Basu, 2013).

The efficiency for paper production for each country has two values *EE(Pap)* and *ME(Pap),* defined for expenditure and manpower as,

$$\textit{Expenditure Efficiency EE(Pap)} = \textit{Papers/GERD} \quad (1)$$

$$\textit{Manpower Efficiency ME(Pap)} = \textit{Papers/Manpower} \quad (2)$$

where GERD is the national expenditure on R&D (in PPP), and the manpower is in terms of full time equivalents in R&D (FTE's).

The efficiency for patent production also has two values *EE(Pat)* and *ME(Pat),*

$$\textit{Patent Expenditure Efficiency EE(Pat)} = \textit{Patents/GERD} \quad (3)$$

$$\textit{Patent Manpower Efficiency ME(Pat)} = \textit{Patents/Manpower} \quad (4)$$

While papers and patents are homogeneous entities, GERD is made up of several components such as HERD, BERD, GOVERD, which are the expenditures on the Higher Education sector, the business sector and the government sector. Each of these components contributes in a different way to output of papers and patents. For example, expenditure in the business sector is expected to give rise to patents rather than papers, Higher education and government expenditures give rise to primarily papers, while defence expenditure, which is part of expenditure in the government sector does not produce many papers or patents. While this indicates that questions of efficiency are more complex than what has been considered here, in the present study we will use GERD as a single homogeneous entity.

Analysis

In Table 1 we see the inputs made by a set of selected countries in the years 2002 and 2007 to R&D (GERD), expressed as a share. A country is taken to be a developed country if its share of GERD is higher than its share of GDP (GERD share/GDP share >1; Hollanders & Soete, 2010). Using this criterion we see from Table 1 that in both 2002 and 2007 the EU as a whole, USA, Japan, Germany, France, UK, Australia and Korea had GERD share/GDP share >1, and would be termed developed countries. We note that Italy is missing from this list, although it is a part of the EU. It is listed along with China, India, Brazil, Mexico, Russia for which GERRD/GDP is less than 1. The data indicates that expenditure on R&D in Italy is lower than would be expected for a developed country.

A plot of Expenditure efficiency and Manpower efficiency in the production of scientific papers shows that Italy has the highest efficiency in both directions (Fig. 1). This implies for the amount of money invested and manpower deployed in the R&D system, Italy has the highest efficiency. This observation makes Italy and its science system an interesting object of study.

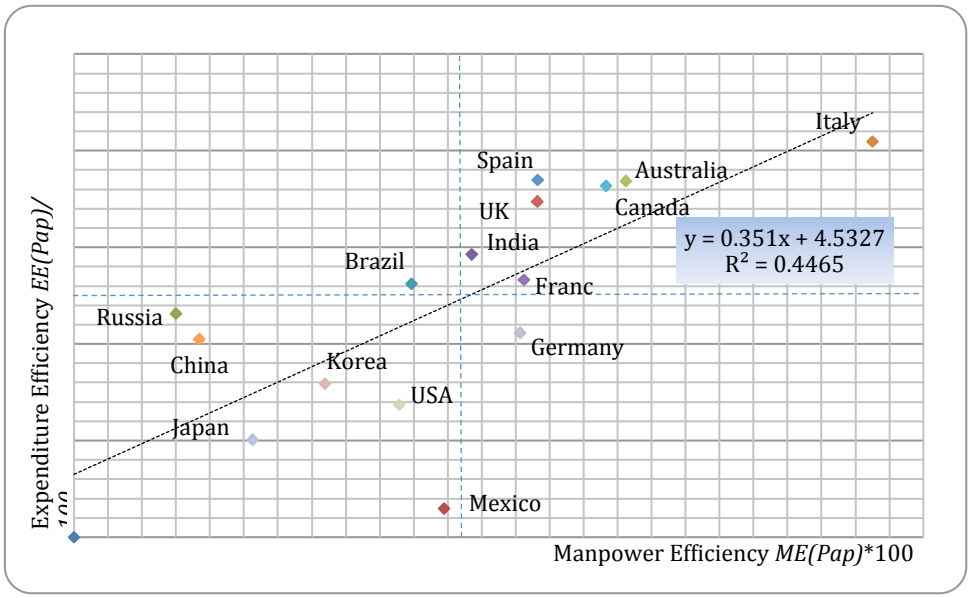


Figure 1. Efficiency of paper production with respect to expenditure EE(Pap) and manpower ME(Pap). Note that Italy scores very high on both dimensions.

For patent production we have the corresponding quantities EE(Pat) and ME(Pat) calculated using Eqns 3 and 4, and plotted in Figure 2. Here we note that USA, Japan are at the highest level in patenting efficiency, while Germany, Korea and Canada are at a medium level. UK and Australia are just above average and Italy and France are somewhat above the average (blue dotted lines) on manpower efficiency ME(Pat) but below average on expenditure efficiency EE(Pat). China, India, Spain, Mexico, Brazil and Russia are below average in patenting efficiency.

The high degree of collinearity (R^2=0.9) in the graph suggests that manpower and expenditure are correlated, which is not surprising since a large fraction of the expenditure usually goes toward salaries. This is also true to some extent of the efficiencies of paper production (Fig. 1).

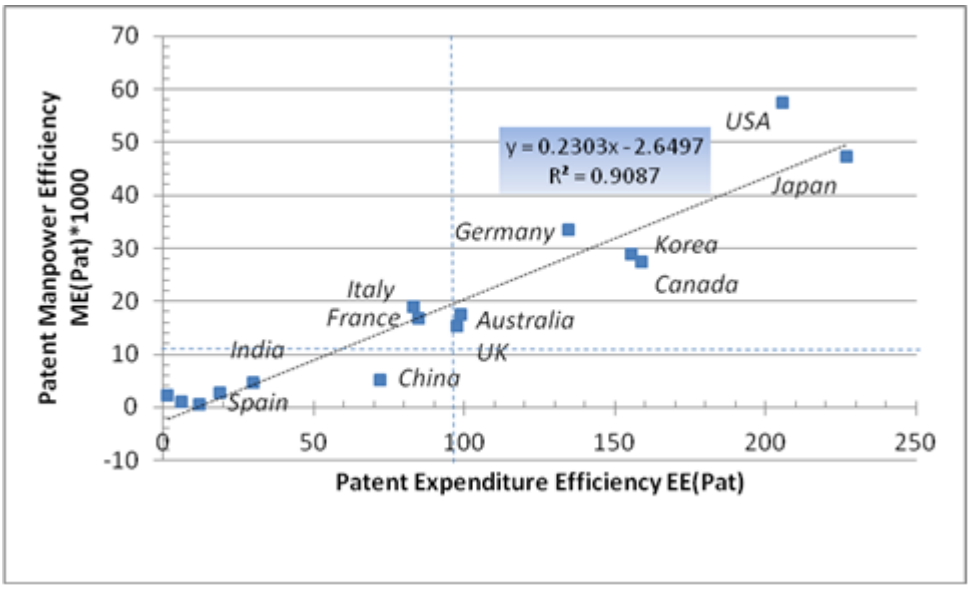


Figure 2. Efficiency of patent production with respect to expenditure EE(Pat) and manpower ME(Pat). Countries in left corner are Russia, Mexico and Brazil.

It should be emphasized that since there are 4 dimensions, two-dimensional graphs give only a partial picture of the similarity of profiles of the different countries.

The case of Italy

The case of Italy is somewhat unusual because of the very high values of efficiency of paper production with respect to both manpower and expenditure (Fig. 1). While this has not been explicitly stated in the literature, it is possible that there were indications of it in the work of others (Daraio & Moed, 2011; Foland & Shelton, 2010). Our attempt will be to trace such instances that support our finding. Firstly, we consider expenditure and recall that Italy had GERD share/GDP share less than unity, which categorises it with developing countries (Table 1). In Figure 3 we look at the GERD values of some countries (OECD data, 2012). Among a set of European countries together with US and Japan, Italy has the lowest value of the input GERD as a percentage of GDP. Since efficiency is the ratio of output to input, a low value of input raises efficiency. Spain also has a low value of expenditure, which makes its publication efficiency with respect to expenditure high. However its publication efficiency with respect to manpower is low (Fig. 1)

In terms of the business component of GERD (BERD) and the Government expenditure (GOVERD) the same trend prevails (Figs. 4 & 5) showing that Italy has almost the lowest values among these countries. This has also been noted in Daraio and Moed (2011).

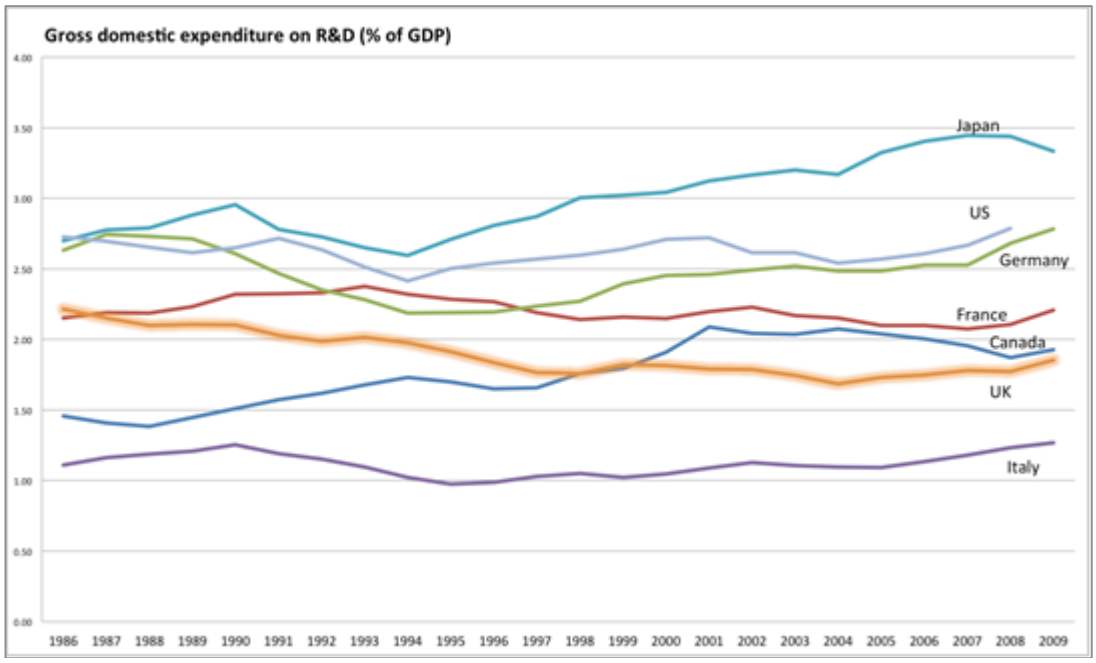


Figure 3. Gross domestic Expenditure on R&D (Source: OECD data, 2012).

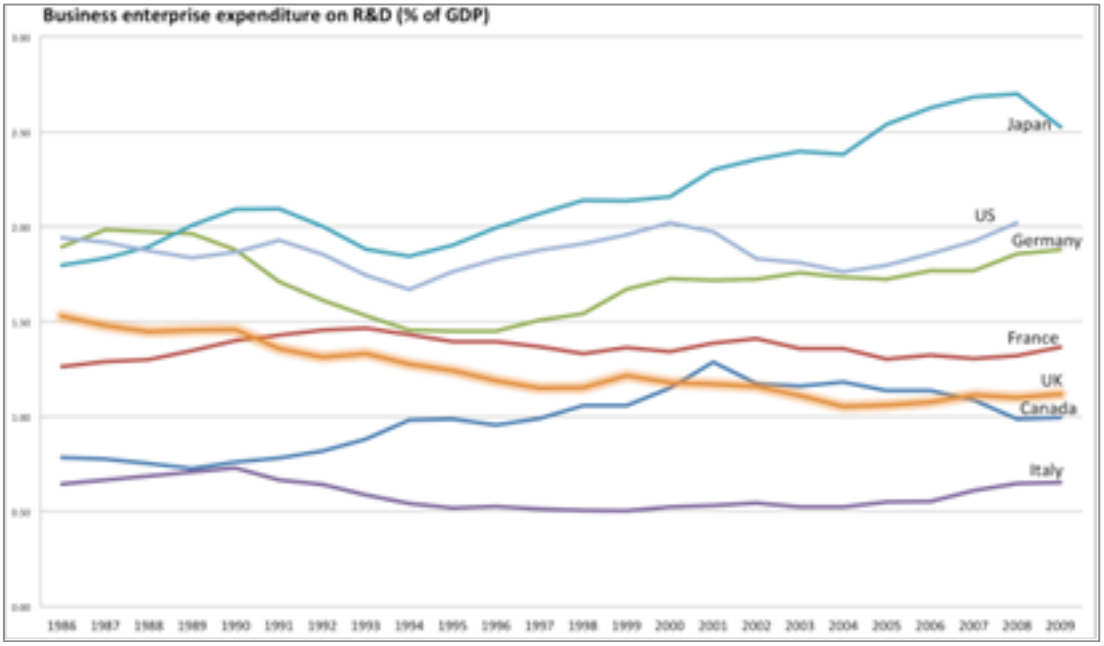


Figure 4. Business Enterprise data BERD (Source: OECD data, 2012).

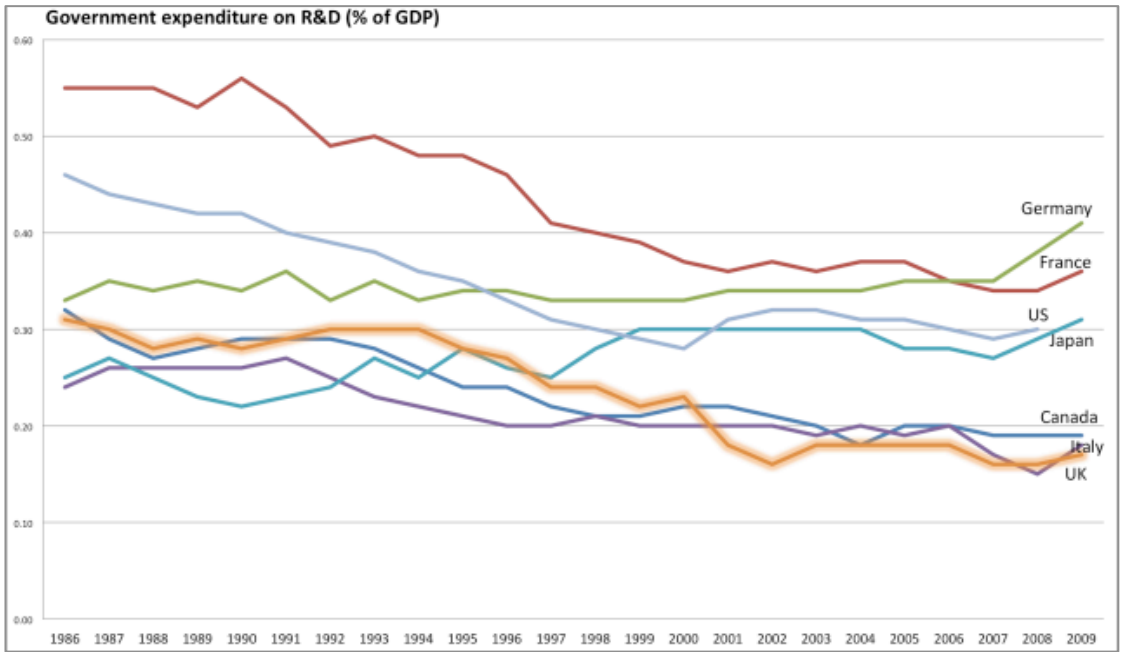


Figure 5. Government Expenditure on R&D, GOVERD (Source: OECD data, 2012).

Figures. 3-5 show that Italy has one of the lowest values of R&D expenditure as a share of GDP among all the countries shown. It also had the lowest expenditure on military R&D spending, a sector not expected to produce many papers or patents, as seen from Figure 6 reproduced from Foland and Shelton (2010).

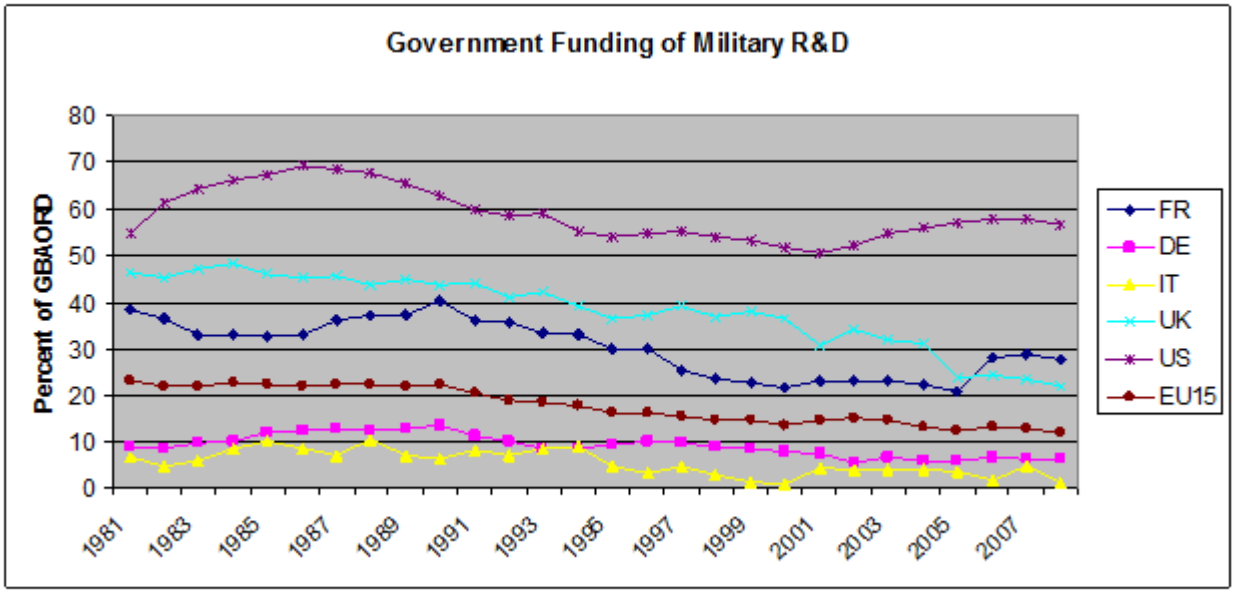


Figure 6. Government Funding of military R&D, showing that Italy has one of the lowest military spending (Source: Foland and Shelton, 2010).

At the same time, in a graph by the same authors showing growth rates of published papers for different countries, it is clearly seen that Italy had the highest growth rate over two successive decades (Fig. 7). Thus it would appear that there has been an efficiency increase with respect to expenditure for Italy, both due to lowered expenditure on R&D as well as increases in publication output.

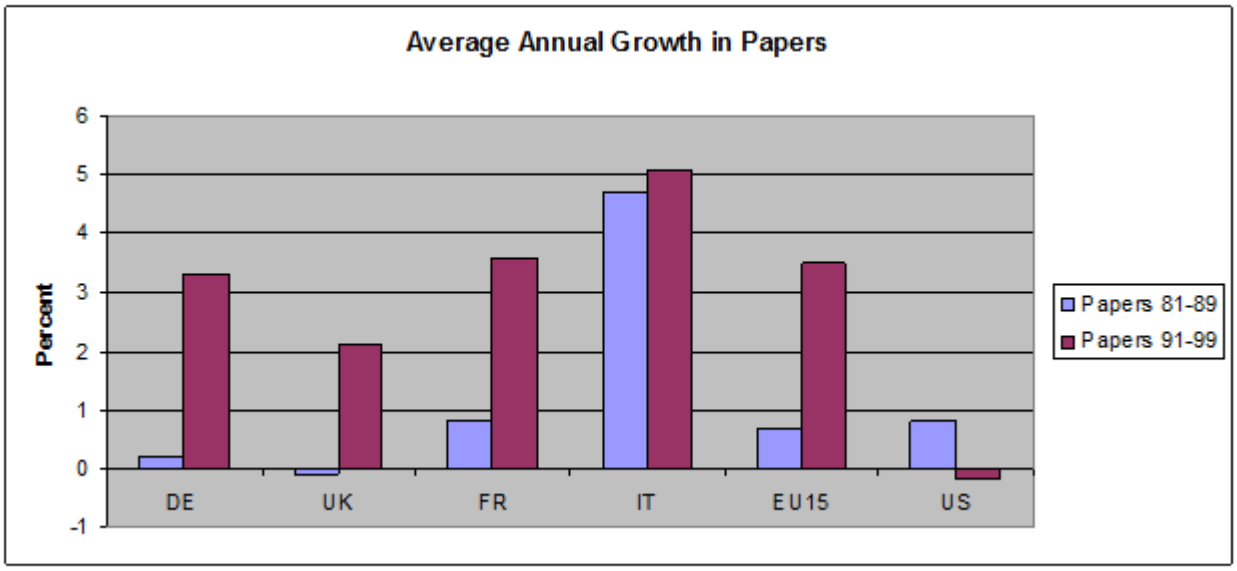


Figure 7. Average annual growth in papers, for a few selected countries (Source: Foland & Shelton, 2010).

Finally, we find more detailed information in a series of country profiles created by DeJaeger (2010) for 39 OECD countries and some developing countries. From Italy's profile a comparison of Italy's outputs with other countries shows some interesting points. DeJaeger's profile of Italy is reproduced below (Fig. 8).

The GERD is low, about 1.26% of GDP, about half the OECD average and more in line with the R&D intensity of emerging economies, as seen earlier.

The manpower values are also lower than OECD average. At the same time the output of papers is on par with the average output of the group of OECD countries. This would give Italy a higher efficiency of publication with respect to manpower as compared to the average. Daraio and Moed (2010) also note in their paper that Italy's publications grew in the period 1980-2009, till it had the highest publication output per researcher amongst other European countries (*see Figure 8 in Daraio and Moed; they however, they prefer to use papers per thousand population as an index instead, and predict a decline for Italy based on a lack of correlation between citation impact and manpower values.*)

In brief, while the number of researchers per thousand total employment is low compared to the average, Italy's output of papers per million population is on par with the average of the other countries, making its efficiency high for publications (Fig. 8). Triadic patents per million population is very low compared to other countries (Fig. 8), which coupled with low values of expenditure and research manpower lead to a medium value for patenting efficiency (Fig. 2).

Another point of interest is the high percentage of foreign funding in GERD as compared to other countries. DeJaeger (2012) notes that internationalization in Italy is high. About 41% of scientific articles and 13% of PCT patents were produced with international collaboration. In 2009, industry funded 44% of GERD, Government funding was 42% and 9% was funded from abroad. Regarding international collaboration Daraio and Moed find that Italy's share of internationally co-authored bilateral papers is lower than other OECD countries and their role (vis a vis first authorship) is like the developing countries (Fig. 4 in Daraio & Moed, 2011). From Figure 8 we also see that Italy has a higher number of foreign co-inventors as compared to other countries. It is possible that foreign funds apply to these sectors.

In summary Italy appears to be a country, which has achieved a high efficiency of publication of papers funded with low funds a substantial part of which is from foreign sources. Its

expenditure in the business sector is also low, but its patenting is close to average again indicating medium efficiency.

One limitation of our study is that citations have not been considered in the definition of efficiency. Even though Italy's citations appear to be favorable in some studies (Aspen Report, 2012, Dario & Moed, 2011), it is possible that considering citations would give a different picture. Other caveats common to most bibliometric studies refer to the use of publications as homogeneous units without reference to disciplinary biases in productivity and efficiency, difficulties in comparing expenditures (should one use Purchasing Power Parity, PPP $?), as well as manpower due to differing conventions in different countries (Wendt, et al, 2011).

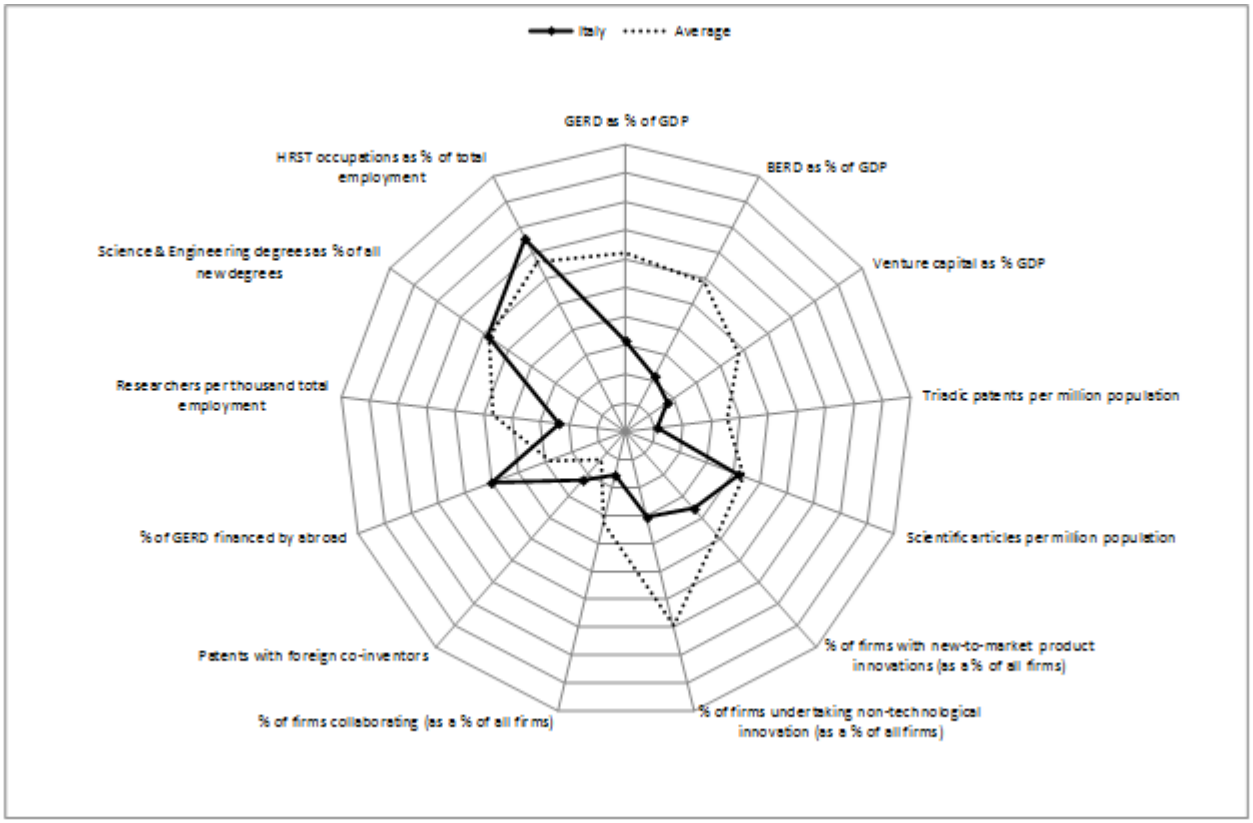


Figure 8. Italy's position vis-à-vis OECD countries on several parameters related to science (Source: De Jaeger, 2012).

Discussion

Efficiency of different countries in the production of papers and patents with respect to manpower and expenditure were calculated by us to obtain a national comparison of R&D efficiency. Unlike many earlier studies on efficiency that included only OECD and other developed countries and Japan and China, we have included several developed and developing countries (see also Basu, 2014, a, b). It was found that Italy had the highest efficiency in the production of papers as compared to the developed and developing countries (Fig. 1). We concluded that Italy has an unusual profile which though noticed in the literature, has not been further investigated (Aspen Report, 2012; Daraio & Moed, 2011). In Italy, research expenditure as a fraction of GDP was found to be low, not only in comparison with other OECD countries, but actually in line with developing countries as noted by us here. At the same time scientific articles per million population are on par with the average OECD value (Fig 8). Italy's expenditure on the military R&D sector is also low (Foland & Shelton; 2010). This may be contrasted with the US where 50% of the government expenditure goes to

the military. Since the defence sector is one that does not produce papers or patents, this gives an advantage to Italy in the computation of efficiency in terms of scientific publication output. In other words, Italy spends much less of its GDP on R&D as compared to other developed countries, at the same time achieving the same rate of publications per million population as other OECD countries (OECD figures; Figure 8). Dario and Moed (2011) refer to this as the 'Çathedral in the desert'.

Research manpower as a proportion of total employees is also much lower than the average OECD value, but science degrees are at the average OECD value (Fig. 8). We note here that the OECD makes a distinction between researchers and human resources in S&T (HRST) where the latter would include technical staff. HRST figures as a proportion of total employees in Italy are much higher than average OECD values (Fig. 8). The possible implication of this is that in Italy, the mix of research staff (academic, research, technical) may be different compared to other countries, with a higher component of technical staff. Since a large part of research expenditure goes towards salaries, and technical staff is likely to be less well paid, this may be a contributing factor toward economy in research expenditure. This conjecture needs to be validated by further research.

All of these features where output is average but inputs are low contribute to high efficiency, which is what we have observed in the case of Italy. In case it should appear that high efficiency in the case of Italy is only because of low inputs, it should be pointed out that growth in the output of papers was the highest for Italy over two successive decades (Foland & Shelton, 2010; Daraio & Moed, 2011). Another possible factor in achieving higher levels of publication than expected from low investments in R&D could be international funding and high collaboration. A substantial part of GERD in Italy comes from foreign sources (Fig. 6).

However, the number of patents are low, not only in the USTPO as seen in our study but also for Triadic patents as seen in the country profiles by De Jaeger. Since the expenditure outlay is also low in the business sector which contributes more to patents (BERD; Figure 4), the efficiency in patenting given by their ratio is close to average (Fig. 2). At the same time the number of foreign co-inventors is high, almost double the OECD value (Fig. 6).

In addition to the observations above regarding possible explanations for the high efficiency in science and relatively lower efficiency in patenting in Italy, we refer to our recent paper on a typology of countries based on research efficiency (Basu, 2014b). According to Basu, as national priorities shift from publications to patents as they appear to have done, fuelled by large increases in the business component of GERD, countries have witnessed a fall in publications (not only through the "displacement effect" due to the rise of China) coupled by a rise in patent efficiency. Countries that have moved in this direction are the USA, Japan, and Germany. Italy apparently has not made this transition, and is characterized by very low levels of investments by the business sector and low efficiency in patenting, but a high efficiency in publication. (Shelton and Leydesdorff have used expenditure in the government and business sectors and shown their relation to different outputs, Shelton & Leydesdorff 2011).

While Shelton and Ali (2011) have noted other countries like Turkey, Greece, Poland and Slovakia as being scientifically efficient, Italy appears to have been missed. Daraio and Moed (2011) in their detailed study 'Is Italian science declining?', observed that Italy had the highest productivity per researcher, and among the lowest levels of R&D expenditure for a selected set of EU countries, (for the period around 2007-2008), but instead of regarding it as efficiency, they argued on the basis of lower levels of foreign collaboration and publication output per 1000 inhabitants and detailed policy analysis that Italy was on the verge of a decline in science. They attributed the performance to an 'overcompensation effect', and state that the "the productivity of the system is often used in the political debate to justify a further cut in spending", underlining their apprehensions.

In summary, it appears that Italy has produced over 3% of the world's papers and shown the highest growth rate in two decades (amongst EU countries) with a modest outlay (in line with less developed countries), both in terms of expenditure and manpower in a demonstration of high efficiency in basic science. Of greater concern is the fact that Italy is only average in patenting efficiency, and falls below OECD averages in BERD, venture capital, technological firms undertaking innovative activities or with technological products to market. On the international front, it has much higher contribution to GERD from foreign funds and has almost twice as many co-inventors as compared to other OECD countries.

Acknowledgements

The author acknowledges a grant under the Emeritus Scientist Scheme of the Council of Scientific and Industrial Research, New Delhi (2010-2014), and thanks anonymous referees for their comments.

References

Albuquerque, E. (2005). Science and Technology systems in less developed countries. In H.Moed, W. Glanzel, U.Schmoch (Eds.) *Handbook of Quantitative Science and Technology Research* (pp. 759-778) Kluwer Academic Publishers.

Aspen Institute Italia. (2012). Research in Italy, Strengths and Weaknesses.

Basu, A. (2013). Efficiencies in national scientific productivity in terms of manpower and funding in science, in *Proceedings of the 14th International Society for Scientometrics and Informetrics (ISSI) Conference*, Vienna, July 15-19, 2013: (pp. 1954-1956) http://www.issi2013.org/Images/ISSI_Proceedings_Volume_II.pdf

Basu A. (2014a). The Albuquerque model and efficiency indicators in national scientific productivity with respect to manpower and funding in science, *Scientometrics*, *100*(2), 531-539.

Basu, A. (2014b). A typology of countries based on efficiency of publication and patenting with respect to manpower and expenditure, in Noyons, E. (Ed.)Context Counts – Pathways to Master Big and Little Data, Proceedings of the Science and Technology Indicators Conference, Leiden

Daraio, C. & Moed, H.F. (2011). Is Italian science declining? *Research Policy*, *40*(10), 1380-1392.

DeJaeger, Nils. (2010). OECD Science, Technology and Industry Outlook 2010. http://www.oecd.org /sti/inno/oecdsciencetechnology andindustryoutlook2010.htm

DeJaeger, Nils. (2012). OECD Science, Technology and Industry Outlook, p. 328 http://www.oecd.org /sti/oecdsciencetechnology andindustryoutlook2010.htm

Eurostat. (2012). Statistics Explained, European Commission. http://epp.eurostat.ec.europa.eu /statistics_explained/index.php/R_%26_D_expenditure

Foland, P. & R.D. Shelton (2010). Why is Europe so efficient at producing scientific papers, and does this explain the European Paradox? *11th International Conference on S&T Indicators, Leiden, Sept. 10*, 2010.

Hollanders, H. & L. Soete (2010). The growing role of knowledge in the global economy. In *UNESCO Science Report 2010,* UNESCO Publishing.

Hu, J.-L., Yang, C.-H., & Chen, C.-P. (2014). R&D Efficiency and the national innovation system: an international comparison using the distance function approach. *Bulletin of Economic Research*, 66, 55-71. doi:10.1111/j.1467-8586.2011.00417.x

King, D.A. (2004). The scientific impact of nations, *Nature*, *430*, 311-316.

Lee, H. Y. & Park, Y. T. (2005). 'An international comparison of R&D efficiency: DEA approach', *Asian Journal of Technology Innovation*, *13*, 207–22.

Leydesdorff, L. & Wagner, C. (2009). Macro-level indicators of the relations between research funding and research output, *Journal of Informetrics*, *3*(4), 353–362.

May, R.M. (1997) The scientific wealth of nations, *Science*, 7 February 1997: 793-796.

OECD (2011). "Business R&D", in OECD Science, Technology and Industry Scoreboard 2011, OECD Publishing. http://dx.doi.org/10.1787/sti_scoreboard-2011-18-en

Rousseau, S. & Rousseau, R. (1998). The scientific wealth of European nations: Taking effectiveness into account, *Scientometrics*, *42*(1), 75-87.

Sharma, S. & Thomas, V.J. (2008). Inter-Country R&D Efficiency Analysis: An Application of Data Envelopment Analysis, *Scientometrics*, *76*, 483-501.

Shelton, R.D. (2008). Relations between national research investment and publication output: Application to an American Paradox. *Scientometrics, 74*(2), 191-205.

Shelton R.D. & Leydesdorff, L. (2011). Publish or patent: Bibliometric evidence for empirical trade- offs in national funding strategies, *Journal of the American Society for Information in Science and Technology*, *63*(3), 498-511.

UNESCO Science Report (2010). UNESCO Publishing.

Vinkler, P. (2005). Science indicators, economic development and the wealth of nations. *Scientometrics, 63*, 417-419.

Vinkler, P. (2008). Correlation between the structure of scientific research, scientometric indicators and GDP in EU and non-EU countries. *Scientometrics, 74*(2), 237-254.

Wendt K., Aksnes, D.W, Sivertsen, G., et al. (2012). Challenges in Cross-National Comparisons of R & D Expenditure and Publication Output, In *Proceedings of 17th International Conference on Science and Technology Indicators*, 2(0167) 826-834.

Performance Assessment of Public-Funded R&D Organizations Working on Similar Research Streams: A Multinational Study

Debnirmalya Gangopadhyay[1], Santanu Roy[2] and Jay Mitra[3]

[1] *debn4u@gmail.com*
National Institute of Science Technology and Development Studies (NISTADS), K.S. Krishnan Marg, New Delhi- 110012 (India)

[2]*sroy@imtdubai.ac.ae*
Institute of Management Technology (IMT), Dubai International Academic City, Dubai (United Arab Emirates)

[3]*jmitra@essex.ac.uk*
Essex Business School, University of Essex Wivenhoe Park, Colchester, Essex CO4 3SQ (United Kingdom)

Abstract

The subject of deriving a measure of efficiency of public-funded organizations (primarily not-for-profit organizations) and of ranking these efficiency measures have been major subjects of debate and discussion. In the present study, the methodology of data envelopment analysis (DEA) has been used to analyze the relative performances of public funded R&D organizations across multiple countries working in similar research streams with multiple measures of inputs and outputs. The keywords highlighting the major research areas in the field of non-metrology conducted by National Physical Laboratory (NPL) in India were utilized to select the global comparators working in similar research streams. These global comparators were three R&D organizations located in the USA and one each located in Germany and Japan. The relative efficiencies of the organizations were assessed with variables such as external cash flow (ECF) earned, technologies transferred, publications and patents as outputs and grants received from the parent body and scientific personnel as inputs. The study indicates suggested measures and a set of targets to achieve the best possible performance for NPL and other R&D organizations.

Conference Topic

Science Policy and Research Assessment

Introduction

Public funded research and development (R&D) organizations utilize public money either through government-supported research programs or other public supported activities. These organizations carry out scientific research, deliver technological services to the society and play a fundamental role in an increasingly knowledge-based society ushering in innovations necessary for the development of a competitive industrial system. Research and innovation have become strategic resources and assets to foster competitive national economies (Coccia, 2005). The ability to attract, develop and retain high quality scientific and technical manpower as well as self-sustenance by means of minimizing its dependence on state funding assume vital importance as it impacts delivery that not only addresses national needs but also ensures traction on a global scale.

Globally, public R&D organizations are currently striving to improve their performance as a result of enhanced competition due to liberalization and globalization, increasing demands on the existing resources and being accountable for optimum allocation of these resources. As the R&D process utilizes scarce resources, it becomes crucial to assess the efficiency of this process (Sharma & Thomas, 2008). In the recent past government efficiency concerns have increased, more so in the light of diminishing funds (Gupta et al., 2000). The emerging demand for evaluating the performance of R&D organizations is the result of relentless growth in global competition (Tassey, 2009). However, the provision of quality information

to decision makers through a performance measurement system assumes criticality in such a scenario (Cook et al., 1995).

One major problem in evaluating the efficiency of public institutions is the lack of a good estimate of the production function. The breakthrough came in the research work undertaken by Charnes, Cooper and Rhodes (1978), the first paper using the technique of data envelopment analysis (DEA), even though they never named it that way. The present study makes an attempt to assess the relative efficiency of the National Physical Laboratory (NPL), a constituent establishment of the Council of Scientific and Industrial Research (CSIR), India, with five selected global comparators working in the same research streams located in three countries - the USA, Japan and Germany. Finally, suggesting measures have been proposed highlighting a set of targets to achieve the best possible performance for those R&D organizations, which are less efficient.

Literature Review

It is difficult to measure the performance of an R&D organization because the nature of these organizations and the functions these organizations perform are complex, risky, and uncertain. As opined by Chiesa and Masella (1996), Bremser and Barsky (2004), Loch and Tapper (2001), Brown and Svenson (1998), and Jain and Triandis (1997), it is difficult to identify, measure and compare the performance of R&D organizations. Further, researchers have found it difficult to identify the various outputs/inputs as multiple parameters are involved in the system. As per the existing literature, there exists only a few studies that have been conducted on performance measurement of R&D organizations (Roy, Mitra & Debnath, 2013; Garg et al*.,* 2005).

R&D Output

Considering individual firms as the sample of their study, Pandit, Wasley and Zach (2011) consider R&D as an input to the innovation process and measures the productivity of a firm's innovative activities in terms of the number and the quality of patents. They argue that both of these variables are measures of innovation output or success, and proxy for the economic value of innovation. Chen, Hu and Yang (2011) suggest a multi-dimensional measurement schema including patents, royalties and licensing fees and journal articles. In their study on R&D and the national innovation system, Hu, Yang and Chen (2014) compare R&D efficiency among 24 nations during 1998-2005. In their multiple input-output framework, the input variables are R&D expenditure stock and R&D manpower and the output variables are patents, scientific journal articles, and royalty and licensing fees. Considering public research institutes, Matsumoto et al. (2010) have carried out case studies on market-impact creation outputs from the National Institute of Advanced Industrial Science and Technology, and have modelled R&D output generating economic impact along four stages – R&D output, technology transfer, commercialization, and market impact. This is in line with Roy et al.'s (2003) earlier study where a model to measure the effectiveness of research units was conceptualized. Likewise, research carried out by Laliene and Sakalas (2014) and Agostino et al. (2012) refer to the development of conceptual frameworks for R&D productivity assessment in public research organizations. Lee et al. (2011) have presented an R&D performance monitoring, evaluation and management system for national R&D to mirror not only short-term but also long-term R&D outcomes.

Methodology

Data envelopment analysis (DEA) as developed by Charnes et al. (1978) and extended by Banker et al. (BCC) (1984) has opened up new possibilities in evaluating the performances of many different kinds of entities (referred to as decision making units, DMU), engaged in

different activities and contexts (Cooper et al., 2004). DEA has been used widely to evaluate the performances of countries and regions (Rousseau and Rousseau, 1997, 1998), banks (Brockett et al., 1997), US air force wings (Charnes et al., 1985a), universities (Reichmann, 2004), Japanese manufacturing firms (Goto & Suzuki, 1989), journals (Lozano & Salmeron, 2005), R&D funding on education (Garg et al., 2005), etc. Publications and patents are used extensively to measure R&D efficiency and innovation (Pavitt, 1985). Evaluation of R&D efficiency could be advantageous to identify the better performers for benchmarking and choose better ways to improve efficiency highlighting areas of weakness (Sharma & Thomas, 2008). Charnes et al. (1985) have characterized a unit as influential if it is frequently used in the calculation of efficiency scores.

Researchers who have adopted the DEA methodology to evaluate performances of public research institutes include Rama Mohan (2005) and Roy, Mitra and Debnath (2013). Kim and Oh (2002) conducted a study on designing an R&D measurement system for Korean researchers. Wang et al. (2005) have developed extensive evaluation criteria for multidisciplinary R&D projects in China for ranking and rewarding. Roy et al. (2007) have earlier carried out a study on CSIR exploring the impact of age, research area, and rank on its scientific productivity, again using DEA as one of the methodologies.

Contextual Background of the Study

National Physical Laboratory (NPL), a premier institute of the Council of Scientific and Industrial Research (CSIR), India, has had a commendable track record of contributions and accomplishments since its inception and its scientists have received recognition for their contributions. Though maintenance and up-gradation of national standards of measurements remains the statutory responsibility of the organization, it is also involved in advanced non-metrology related research activities including engineering and electronic materials, material characterization, radio and atmospheric sciences, superconductivity and cryogenics.

A participatory workshop was conducted to diagnose NPL's R&D operations and to focus on aspects related to R&D performance. A particular research area (non-metrology) was selected for the purpose of the current analysis, and accordingly, the keywords, highlighting the organization's major research areas in this field, were utilized to shortlist global comparators. The keywords were searched in the SCOPUS database for a five-year period and global R&D organizations working on similar research streams were shortlisted. Five public R&D organizations were selected based on higher number of publications. These global comparators were the following:

1) National Institute for Materials Science, Japan (NIMS-JP, DMU-A),
2) National Renewable Energy Laboratory, USA (NREL-US, DMU-B),
3) Fritz Haber Institute of the Max Planck Society, Germany (FHI-DE, DMU-C),
4) National Centre for Atmospheric Research, USA (NCAR-US, DMU-D), and
5) Oak Ridge National Laboratory, USA (ORNL-US, DMU-E).

Data structure

The data regarding the inputs and outputs were collected for each DMU including NPL for a five-year period and are presented in Table 1. To ensure confidentiality, the exact period of the data cannot be revealed. Input variables considered in this study were: (1) grants received from the parent body, and (2) the number of scientific personnel (SP) whereas the output variables were: (1) business generated from the industry *i.e.,* external cash flow (ECF) earned, (2) technologies transferred (TT), (3) publications, and (4) number of patents filed.

The methodology to compare performance of any set of research institutes as suggested by Rama Mohan (2005) has been adopted in the present study. To illustrate the results on

efficiency assessment of public R&D organizations including NPL, one input variable and two output variables were considered at the same time.

Table 1. Input and output of different public R&D organizations (five year data).

Public R&D Organization	*Input*		*Output*			
	Grants (Million USD)	*Scientific Personnel (No.)*	*Technologies Transferred (no.)*	*Publication (No.)*	*Patents (No.)*	*ECF (Million USD)*
NIMS-JP - A	94	675	95	7480	195	20
NREL-US - B	141	307	53	2012	99	15
FHI-DE - C	72	206	1	1225	6	3
NCAR-US - D	185	310	5	2345	14	17
ORNL-US - E	107	1075	83	9144	90	23
NPL, India	47	216	3	1024	13	4

The DEAOS software was used for analysis. It analyzes relative performance of business units performing similar functions with an easy to use interface. It provides numerical and graphical output for easy interpretation and communication of results. Some of the key features of DEAOS are:

- The possibility to deal with 25 to 'unlimited' decision making units.
- Flexible facilities – importing from Excel file and direct entry of the data.
- Provides flexible input data management - possibility of addition and deletion of DMUs as well as rows and columns.
- Model input/output orientation selection.
- Provides a tabular scores report (with a variety of sorting methods) and a graphical summary.

Results

ECF generated and technologies transferred vs. scientific personnel

Ratios were calculated for each organization (Table 2) along two dimensions viz., ECF generated per scientific personnel and technologies transferred per scientific personnel. Figure 1 clearly shows that NREL-US (DMU-B) and NCAR-US (DMU-D) are the best performers exhibiting 100% relative efficiency. The efficient frontier, which envelops NIMS-JP (DMU-A), FHI-DE (DMU-C), ORNL-US (DMU-E) and NPL, represents relative efficiency of those organizations. It is observed that NIMS-JP, FHI-DE, ORNL-US and NPL exhibited relative efficiencies of 82, 28, 45 and 36 % respectively. To enhance efficiency from 36 to 46%, NPL is assumed to increase the input-output ratios from the current level of 0.86 to 1.10 (ECF/scientific personnel) and 0.014 to 0.018 (technologies transferred/scientific personnel). An improvement target of 10 %, keeping input (scientific personnel) constant, can be achieved during the next year, if NPL is in a position to increase its ECF to 1.6 M USD and transfer at least 1 technology (Table 3).

Table 2. External cash flow (ECF) and technologies transferred vs. scientific personnel.

Public R&D Organization	*ECF / Scientific Personnel*	*Technology Transferred / Scientific Personnel*
NIMS-JP - A	1.32	0.14
NREL-US - B	2.13	0.17
FHI-DE - C	0.69	0.00
NCAR-US - D	2.44	0.02
ORNL-US - E	0.94	0.08
NPL, India	0.86	0.01

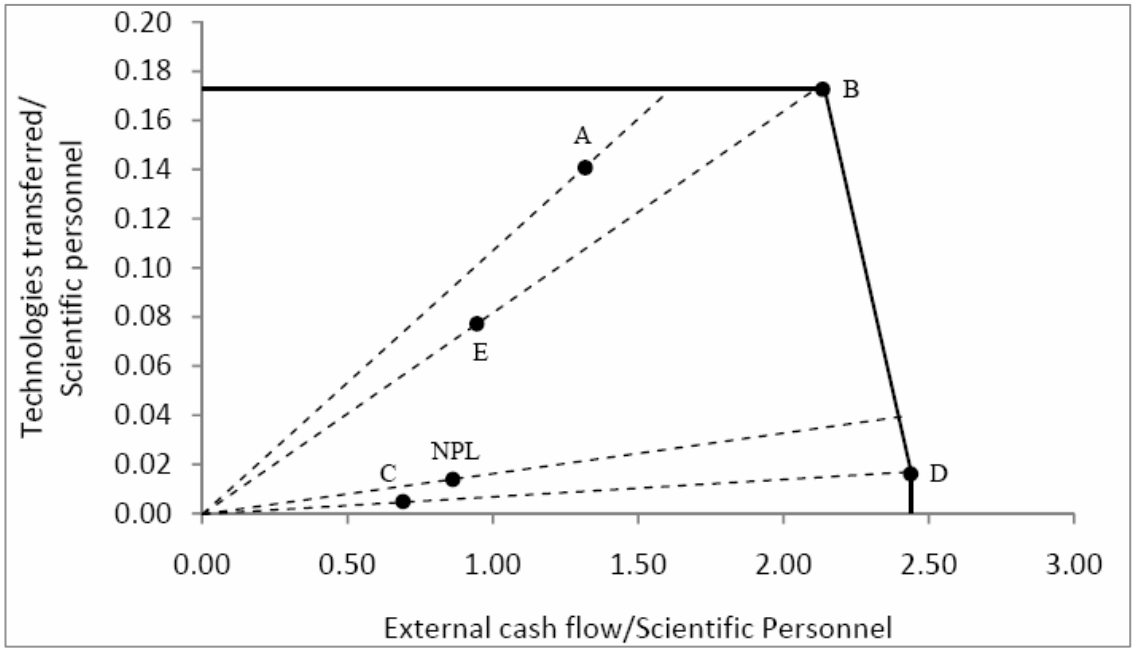


Figure 1. ECF generated and technologies transferred vs. scientific personnel.

Publications and patents vs. scientific personnel

To assess the relative performance of the R&D organizations, publications per scientific personnel and patents per scientific personnel were calculated (Table 4) and graphically represented in Figure 2. NIMS-JP (DMU-A) and NREL-US (DMU-B) show best performance exhibiting 100% efficiency in generating sufficient number of publications and patents per scientific personnel. Performance was found higher in case of ORNL-US (DMU-E) (77%) and NCAR-US (DMU-D) (67%) whereas FHI-DE (DMU-C) (54%) and NPL (43%) perform moderately. However, NIMS-JP is the reference laboratory all the organizations. To achieve improved targets by 10% during the next year, NPL and FHI-DE each would require to publish 240 and 230 papers and 9 and 12 patents respectively (Table 5).

Table 3. Targets for the R&D organizations to improve efficiency by 10%

(Scientific personnel count remaining constant)

Public R&D Organization	*ECF to earn (Million USD)*	*Technology to transfer*
NIMS-JP - A	6.8	12
FHI-DE - C	1.1	0.4
ORNL-US - E	5.1	19
NPL, India	1.6	0.8

Table 4. Pubclications and patents vs. scientific personnel

Public R&D Organization	*Publications / Scientific Personnel*	*Patents / Scientific Personnel*
NIMS-JP - A	11.08	0.29
NREL-US - B	6.55	0.32
FHI-DE - C	5.95	0.03
NCAR-US - D	7.44	0.04
ORNL-US - E	8.51	0.08
NPL, India	4.74	0.06

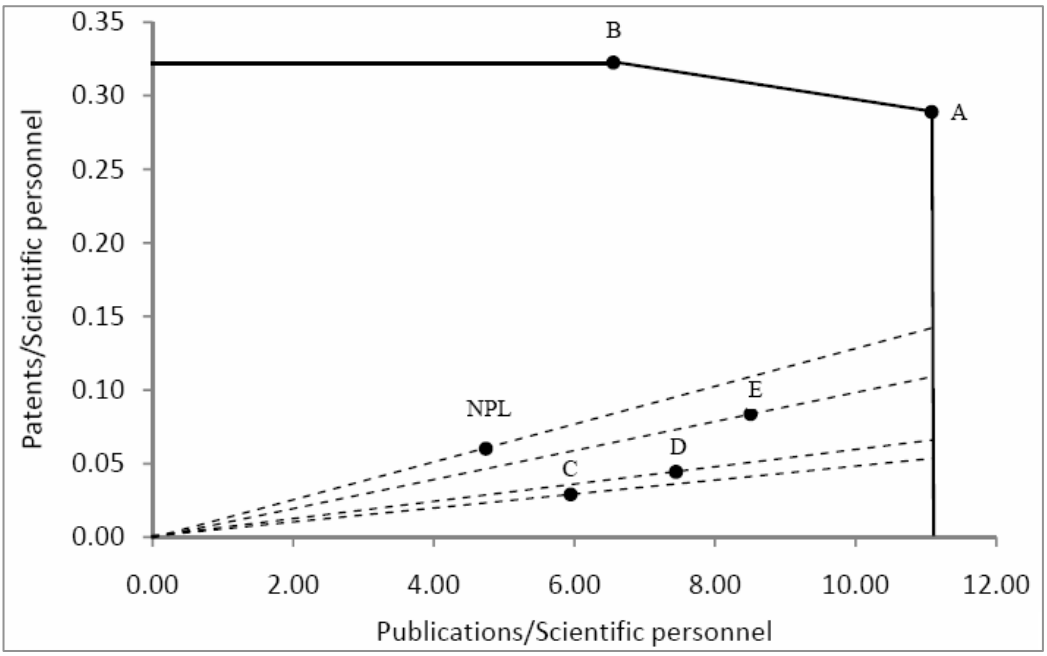


Figure 2. Publications and patents vs. scientific personnel.

Table 5. Targets for the R&D organizations to improve efficiency by 10% (Scientific personnel count remaining constant).

Public R&D Organization	*Publications*	*Patents*
FHI-DE - C	230	12
NCAR-US - D	347	23
ORNL-US - E	1204	96
NPL, India	240	9

ECF generated and technology transferred vs. grants

Next, relative efficiencies of the R&D organizations have been calculated along two outputs (ECF generated and technologies transferred) and one input (grants received from the parent body), (Table 6) and plotted in Figure 3. NIMS-JP (DMU-A) and ORNL-US (DMU-E) show best performance exhibiting 100% efficiency in generating sufficient amounts of ECF and number of technologies transferred per grants received. All the other organizations have ORNL-US in their reference set. To achieve efficiency by 10% during the next year, FHI-DE has to earn 1.5 M USD ECF and to transfer 7 technologies (Table 7).

Table 6. ECF earned and technologies transferred vs. grants received from parent body.

Public R&D Organization	*ECF / Grants*	*Technologies Transferred / Grants*
NIMS-JP - A	0.21	0.02
NREL-US - B	0.10	0.01
FHI-DE - C	0.04	0.00
NCAR-US - D	0.09	0.00
ORNL-US - E	0.21	0.02
NPL, India	0.09	0.00

Table 7. Targets for the R&D organization to improve efficiency by 10% (Grants received from the parent body remaining constant).

Public R&D Organization	*ECF to earn (Million USD)*	*Technology to transfer*
NREL-US - B	3	11
FHI-DE - C	1.5	7
NCAR-US - D	3.9	24
NPL, India	0.8	5

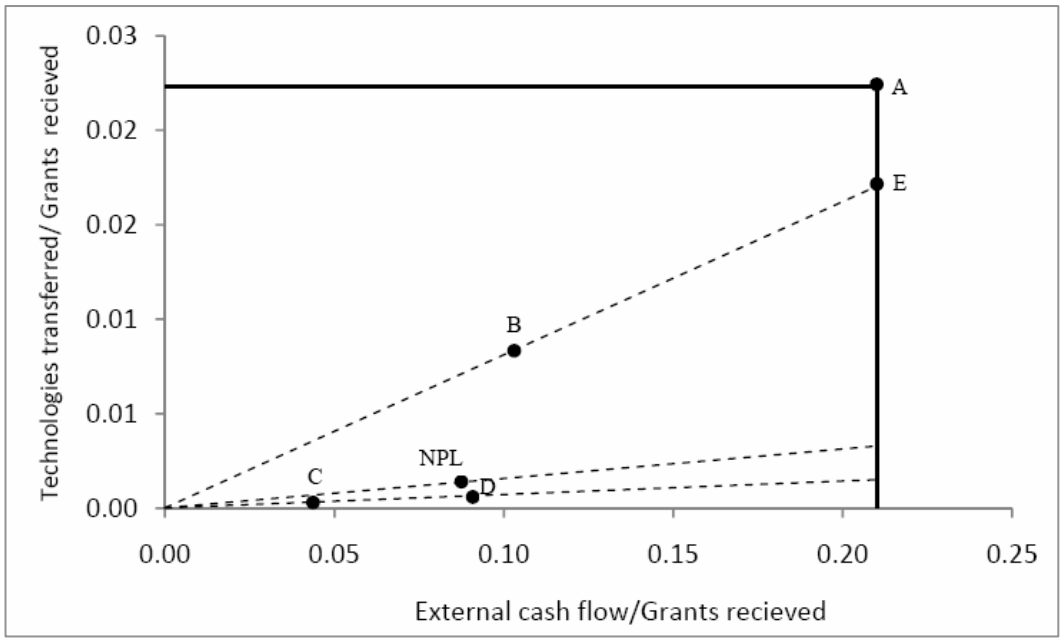


Figure 3. ECF generated and technologies transferred vs. grants received.

Publications and patents vs. grants

To assess the relative performance of the R&D organizations, ratios were calculated for publications per grants received and patents per grants received (Table 8) and graphically represented in Figure 4. NIMS-JP (DMU-A) and ORNL-US (DMU-E) show the best performance exhibiting 100% efficiency. NPL has both NIMS-JP and ORNL-US in its reference set whereas FHI-DE (DMU-C) and NCAR-US (DMU-D) relate only to ORNL-US whereas NREL-US (DMU-B) has only NIMS-JP in its reference set. To achieve efficiency by 10% during the next year, FHI-DE, NCAR-US and NPL have to increase their number of patents by a count of 7, 17 and 5 respectively from the current level (Table 9).

Table 8. Publications and patents vs. grants received from parent body.

Public R&D Organization	*Publication / Grants*	*Patent / Grants*
NIMS-JP - A	1.77	0.05
NREL-US - B	0.32	0.02
FHI-DE - C	0.38	0.00
NCAR-US - D	0.28	0.00
ORNL-US - E	1.89	0.02
NPL, India	0.48	0.01

Publication, patents, ECF generated and technology transferred vs. scientific personnel & grants

The relative efficiencies of R&D organizations on multi-input-multi-output six dimensional model keeping two inputs (*viz.,* scientific personnel & grants received) and four outputs (*viz.,* publication, patents, ECF generated and technology transferred) data have been calculated and the performance of each R&D organization under study is compared with that of every other one following the output oriented measure of efficiency at constant return to scale (CRS), variable return to scale (VRS) along with scale efficiencies (SE). The empirical analysis has been given in Table 10.

Table 9. Targets for the R&D organization to improve efficiency by 10 % (Grants received from the parent body remaining constant).

Public R&D Organization	*Publications*	*Patents*
NREL-US - B	1397	29
FHI-DE - C	617	7
NCAR-US - D	1575	17
NPL, India	399	5

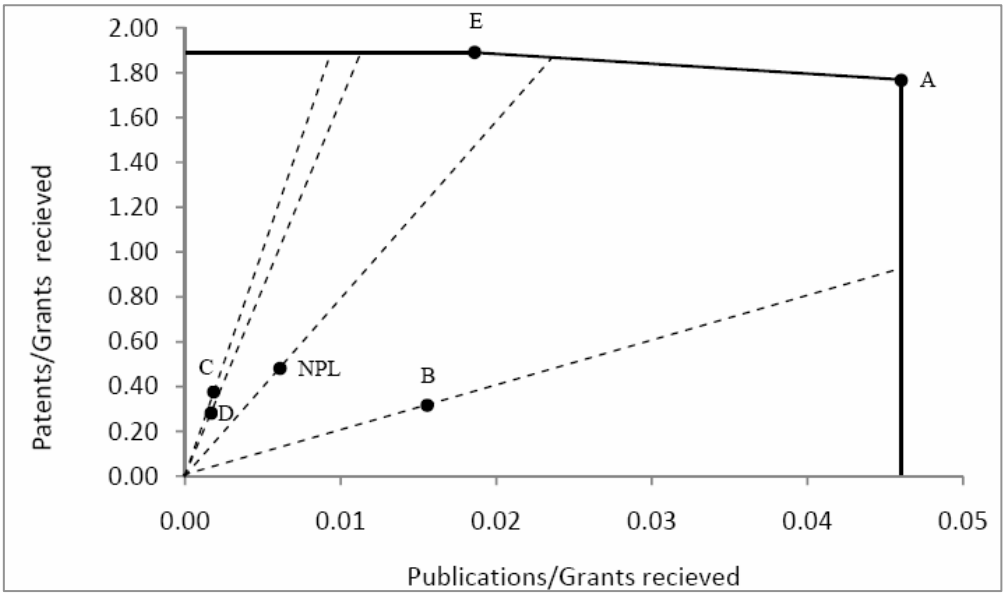


Figure 4. Publications and patents vs. grants received.

Table 10. Relative efficiency percentage of different public R&D organizations.

Public R&D Organization	*ECF & TT/SP*			*Pub. & Pat./SP*			*ECF & TT/Grants*			*Pub. & Pat./Grants*			*Pub., Pat., ECF & TT/SP & Grants*		
	CRS	*VRS*	*SE*	*CRS*	*VRS*	*SE*	*CRS*	*VRS*	*SE*	*CRS*	*VRS*	*SE*	*CRS*	*VRS*	*SE*
NIMS-JP - A	82	100	0.82	100	100	1.00	100	100	1.00	100	100	1.00	100	100	1.00
NREL-US - B	100	100	1.00	100	100	1.00	49	65	0.75	34	51	0.67	100	100	1.00
FHI-DE - C	28	100	0.28	54	100	0.54	21	25	0.84	20	27	0.74	54	100	0.54
NCAR-US- D	100	100	1.00	67	90	0.74	43	74	0.58	15	26	0.58	100	100	1.00
ORNL-US - E	45	100	0.45	77	100	0.77	100	100	1.00	100	100	1.00	100	100	1.00
NPL, India	36	94	0.38	43	85	0.51	42	100	0.42	26	100	0.26	57	100	0.57

Note: CRS: constant return to scale, VRS: variable return to scale SE: scale efficiency; (SE=CRS/VRS)

Technical efficiencies estimated under the CRS model are found to be less than the technical efficiencies coming from the more flexible VRS model. Under the CRS assumption, less average efficiency is found in case of FHI-DE (DMU-C) (54%) followed by NPL (57%) while under VRS, it was found that average technical efficiency score for all the DMUs is 100%, which implies that on an average DMUs could have used resources judicially to produce the same amount of output. However, under the scale efficiency (SE), the average score is found to be 0.54 in case of FHI-DE and 0.57 in case of NPL, which indicate that on an average the actual scale of production has diverged from the most productive scale size. In SE, the score 1 indicates that the DMU is operating at the most efficient scale or optimal size whereas SE less than 1 would be due to decreasing returns to scale (over production) or increasing returns to scale (under production).

Discussion and Conclusions

Over the past three decades, a variety of approaches, parametric and non-parametric, have been developed to investigate the failure of producers to achieve the same level of efficiency (Kalirajan and Shand, 1999). DEA which offers a non-parametric alternative to parametric frontier production function analysis has two advantages over the econometric one in measuring productivity change (Grosskopf, 1986). First, it compares the states to the 'best' practice technology rather than 'average' practice technology as is done by econometric studies. Second, it does not require the specification of an ad hoc functional form or error structure. In DEA, the less-performing units need more inputs to produce the same amount of output (Andersen & Petersen, 1993). DEA produces a piecewise empirical extreme production surface which in economic terms represents the revealed best-practice production frontier (Charnes et al., 1994).

In this study, the performance of each R&D organization (here the DMU) under study is compared with that of every other one following the output oriented measure of efficiency at constant return to scale (CRS), variable return to scale (VRS) along with scale efficiencies (SE). DEA has been used to analyze the relative efficiencies of the public funded R&D organizations keeping one input and two outputs at a time and results have been demonstrated in four possible dimensions. Secondly, the relative efficiencies of R&D organizations on multi-input-multi-output six dimensional model keeping two inputs and four outputs data have also been calculated. Comparatively less efficiency of NPL (0.57) that is a cause for concern might be due to its lower efficiency in generating sufficient amounts of external cash flow, number of technologies assumed to be transferred to the industry per scientific personnel as well as number of papers published and patents filed per grants received from the parent body.

The significance of the work presented in the paper stems from the fact that this is perhaps the first multinational study of relative performance assessment of R&D organizations, all of whom work on similar research themes. Relative performance assessment of different R&D organizations have been ascertained in the past (Roy, Mitra & Debnath, 2013) but the R&D organizations in question were working on diverse research streams. The focus of the current study, therefore, seems much more relevant as absolute comparators were first identified and thereafter assessed in terms of their performance characteristics. The present work has opened up new avenues for further research in this area.

References

Agostino, D., Arena, M., Azone, G., Molin, M.D., & Masella, C. (2012). Developing a performance measurement system for public research centres. *International Journal of Business Science and Applied Management*, *7*(1), 43-60.

Andersen, P. & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science*, *39*, 1261-1264.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data development analysis. *Management Science*, *30*, 1078-1092.

Bremser, W.G. & Barsky, N. P. (2004). Utilizing the balanced scorecard for R&D performance measurement. *R&D Management, 34*(3), 229-237.

Brockett, P. L, Charnes, A., Cooper, W. W., Huang, Z. M., & Sun, D. B. (1997). Data transformations in DEA cone ratio envelopment approaches for monitoring bank performances. *European Journal of Operational Research*, *98*, 250-268.

Brown, M.G. & Svenson, R.A. (1998). Measuring R&D productivity. *Research and Technology Management, 41*(6), 30-35.

Charnes, A., Clark, T., Cooper, W. W., & Golany, B. (1985). A developmental study of data envelopment analysis in measuring the efficiency of maintenance units in the U.S air forces. *Annals of Operations Research*, *2*, 95-112.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, *2*, 429-444.

Charnes, A., Cooper, W. W., Lewin, A. Y., & Seiford, L. M. (1994). *Data Envelopment Analysis: Theory, Methodology and Applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Chen, C., Hu, J., & Yang, C. (2011). An international comparison of R&D efficiency of multiple innovative outputs: the role of the national innovation system. *Innovation Management, Policy and Practice*, 13(3), 341-360.

Chiesa, V. & Masella, C. (1996). Searching for an effective measure of R&D performance. *Management Decision, 34*(7), 49-57.

Coccia, M. (2005). A scientometric model for the assessment of scientific research performance within public institutes. *Scientometrics, 65*, 307-321.

Cohen, W. M. & Levinthal, D. A. (1989). Innovation and learning: the two faces of R&D. *The Economic Journal*, *99*, 569-596.

Comanor, W. S. & Scherer, F. M. (1969). Patent statistics as a measure of technical change. *The Journal of Political Economy*, *77*, 392-398.

Cooper, W. W., Seiford, L. M., & Zhu, J. (2004). Data envelopment analysis: history, models and interpretations. In W. W. Cooper, L. M. Seiford & J. Zhu (Eds.) *Handbook on Data Envelopment Analysis,* (pp. 1-39). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Garg, K. C, Gupta, B. M., Jamal, T., Roy, S., & Kumar, S. (2005). Assessment of impact of AICTE funding on R&D and educational development. *Scientometrics*, *65*, 151-160.

Goto, A. & Suzuki, K. (1989). R&D capital, rate of return on R&D investment and spillover of R&D in Japanese manufacturing industries. *The Review of Economics and Statistics*, *71*, 555-564.

Grosskopf, S. (1986). The role of the reference technology in measuring productive efficiency. *The Economic Journal*, *96*, 499-513.

Gupta, A. K., Bhojwani, H. R., Kaushal, R., & Kaushal, H. (2000). Managing the process of market orientation by publicly funded labs: Case of CSIR. *R&D Management, 30***,** 289-296.

Hall, B. H., Griliches, Z., & Hausman, J. A. (1986). Patents and R&D: Is there a lag? *International Economic Review*, *27*, 265-283.

Hu. J. L., Yang, C.H., & Chen, C.P. (2014). R&D efficiency and the national innovation system: an international comparison using the distance function approach. *Bulletin of Economic Research*, *66*(1), 55-71.

Jain, R.K. & Triandis, H.C. (1997). *Management of Research and Development Organizations: Managing the Unmanageable.* 2nd Ed., New York: John Wiley & Sons.

Kalirajan, K. P. & Shand, R. T. (1999). Frontier production functions and technical efficiency measures. *Journal of Economic Surveys*, *13*, 149-172.

Laliene, R. & Sakalas, A. (2014). Conceptual structure of R&D productivity assessment in public research organizations. Economics & Management, *19*(1), 25-35.

Lee, H., Kim, M.S., Yee, S.R., & Choe, K. (2011). R&D performance monitoring, evaluation and management system: a model and methods. *International Journal of Innovation and Technology Management*, *8*(2), 295-313.

Loch, C.H. & Tapper, U.A.S. (2001). Implementing a strategy-driven performance measurement system for an applied research group. *The Journal of Product Innovation Management, 19*, 185-98.

Lozano, S. & Salmeron, J. L. (2005). Data envelopment analysis of OR/MS journals. *Scientometrics*, *64*, 133-150.

Matsumoto, M., Yokota, S., Naito, K. & Itoh, J. (2010). Development of a model to measure the economic impacts of R&D outputs of public research institutes. *R&D Management*, *40*(1), 91-100.

Pandit, S., Wasley, C.E., & Zach, T. (2011). The effect of research and development input and output on the relation between the uncertainty of future R&D performance and R&D expenditures. *Journal of Accounting, Auditing and Finance*, *26*(1), 121-144.

Pavitt, K. (1985). Patent statistics as indicators of innovative activities: possibilities and problems. *Scientometrics*, *7*, 77-99.

Rama Mohan, S. (2005). Benchmarking evaluation of performance of public research institutes using data envelopment analysis. *Journal of Scientific and Industrial Research*, *64*, 403-410.

Reichmann, G. (2004). Measuring university library efficiency using data envelopment analysis. *Libri*, 54, 136-146.

Rousseau, S. & Rousseau, R. (1997). Data envelopment analysis as a tool for constructing scientometric indicators. *Scientometrics, 40*, 45-56.

Rousseau, S. & Rousseau, R. (1998). The scientific wealth of European nations: taking effectiveness into account. *Scientometrics, 42*, 75-87.

Roy, S., Mitra, J., & Debnath, R.M. (2013). Ranking R&D institutions of India: an application of DEA. *International Journal of Business Development and Research*, *1*(2), 49-66.

Roy, S., Nagpaul, P.S., & Mohapatra, P.K.J. (2003). Developing a model to measure the effectiveness of research units. *International Journal of Operations & Production Management*, *23*(12), 1514-29.

Sharma, S. & Thomas, V. J. (2008). Inter-country R&D efficiency analysis: an application of data envelopment analysis. *Scientometrics, 76*, 483-501.

Tassey, G. (2009). Methods for assessing the economic impacts of government R&D. http://www.nist.gov/director/prog-ofc/report03-1.pdf.

Outlining the Scientific Activity Profile of Researchers in the Social Sciences and Humanities in Spain: The Case of CSIC

Adrián A. Díaz-Faes[1], María Bordons[1] Thed van Leeuwen[2] and Mª Purificación Galindo[3]

[1] *adrian.arias@cchs.csic.es, maria.bordons@cchs.csic.es*
Quantitative Analysis in Science & Technology Group (ACUTE), IFS, Spanish National Research Council (CSIC), Albasanz 26-28, Madrid 28037 (Spain)

[2] *leeuwen@cwts.leidenuniv.nl*
CWTS-Centre for Science and Technology Studies, Leiden University, PO Box 905 2300 AX Leiden (the Netherlands)

[3] *pgalindo@usal.es*
Statistics Department, Salamanca University, Alfonso X El Sabio s/n, Salamanca 37007 (Spain)

Abstract
Scientific activity of Social Sciences and Humanities researcher's comprises an assorted set of publication channels such as books, book chapters and national and international journal articles. Since knowledge dissemination in the field is characterised by a greater use of national journals and local languages, international bibliographic databases do not offer a suitable coverage. This work pursues to draw a comprehensive picture of the publication behaviour of CSIC researchers in the Social Sciences and Humanities from a micro-level perspective. For this purpose, Web of Science and an internal CSIC database called 'ConCiencia' were used along with a set of indicators describing the activity profile of researchers as well as the prestige of publication channels. Differences in the publication pattern of researchers in SSH were explored, and the relationship between their research performance and personal features such as age, gender and professional rank were analysed. In the Humanities, researchers with higher academic rank and age showed greater activity in books and non-WoS articles, whereas in the Social Sciences, higher rank was related to internationally-oriented scientific publications and a more collaborative activity. Considering only WoS articles would shrink meaningfully the visibility of CSIC researchers.

Conference Topic
Science policy and research assessment

Introduction
Outlining the scholarly work of researchers in the Social Sciences and Humanities (SSH) is often regarded as a challenge in bibliometrics, since the predominant publication types in these fields are not well covered by large bibliographic databases such as Web of Science or Scopus (Hicks, 2004). At this point, it is quite clear that dealing with journal publications, it is not enough for the SSH (Archambault et al., 2006; Sivertsen & Larsen, 2012) remaining books and books chapters as a major communication channel, chiefly in the Humanities. Moreover, due to the more local orientation of research in the SSH, knowledge dissemination in the field is characterized by a greater use of national journals and local languages (van Leeuwen, 2013). On the other hand, even though there has been a certain trend to consider SSH as a whole, different behavior between both communities can be expected (Mañana-Rodríguez & Giménez-Toledo, 2013).

The aforesaid factors hinder the potential capacity of the traditional bibliometric analyses to provide a reliable picture of the scientific activity of the SSH researchers and the development of national or regional databases to obtain full coverage of publications in the SSH has been suggested (Martin et al. 2010). This type of database has been developed in some countries such as Norway, Denmark, Finland and Belgium (Flanders), motivated by the need to monitor the performance of university scholars and in line with the development of performance-based

funding of university research (Sivertsen, 2010). Studying the activity of SSH researchers in Spain is difficult, because there is not such a full coverage national bibliographic database, but it can be addressed at the institutional level because many institutions collect the scientific output of their researchers, mainly with evaluative purposes.

This study focuses on the scientific activity of SSH researchers at the Spanish National Research Council (CSIC), the largest public institution dedicated to research in Spain which makes up more than 4,000 researchers and 125 institutes spread all over the country. This work pursues to draw a comprehensive picture of the publication behaviour of CSIC researchers in SSH from a micro-level perspective. An assorted set of publication channels such as books, books chapters, international and national journal articles are considered and specific indicators to assess the prestige of the different publication channels are introduced. Differences in the publication pattern of researchers in SSH are explored, and the relationship between their research performance and personal features such as age, gender and professional rank are analyzed.

Methodology

This study analyses the scientific output of 268 active researchers in 2007 in the SSH area affiliated to the Spanish National Research Council (CSIC) and comprises both permanent researchers and postdoctoral research fellows. The time span under analysis is 2007-2011. Publications were collected from two different sources: Web of Science (WoS) (SSCI+AHCI+SCIE), which was used to download the more international articles; and an internal CSIC database called 'ConCiencia', to obtain other publication types not covered by WoS (books, books chapters and non-WoS journal articles). To cope with names inconsistencies and achieve a proper allocation of the publications to the researchers, different algorithms were used. A manual revision of the output collected, especially for the 'ConCiencia' database, was done. Based on the information retrieved, the following indicators were computed:

a) Activity profile of researchers

- % Books: proportion of books published by a researcher with regard to its total number of publications. In the same way, the next three indicators were calculated.
- % Book chapters.
- % WoS articles.
- % Non-WoS articles.
- Sum of publications: the total number of publications published by each researcher, including books, chapters in books and journal articles.
- Average number of authors/paper: this indicator measures the average number of authors per publication for the total output of a given researcher (WTI2, 2014).
- % International collaboration: share of the total output of each researcher co-authored with researchers affiliated with one or more foreign institutions.
- % English: proportion of a researcher´s output published in English.

b) Prestige of publication channels

- Top books and chapters (pptop10% Books & Chapters): proportion of books and chapters of a given researcher published by the top 10% publishers according to the Scholarly Publisher Indicators Project (SPI) (Giménez-Toledo, Tejada-Artigas & Mañana-Rodriguez, 2013). This project describes the Indicator of Quality of Publishers according to Experts (ICEE), which is based on a quality assessment of publishers rated by Spanish researchers in a national survey.

- Proportion of papers in first quartile journals (Q1): share of papers published in the top 25% journals of the impact factor journal ranking by subject category (source: Journal Citation Reports).
- Proportion of papers in top non-WoS journals (pptop10% non-WoS articles): % of non-WoS papers published in top journals according to the Integrated Scientific Journal Classification (CIRC) (Torres-Salinas et al. 2010). CIRC is a proposal for a categorization of journals in SSH developed by a group of experts in bibliometrics in Spain. It distinguishes four categories of journals (A, B, C and D) according to their visibility measured integrating the results of different journal classifications and assessments tools. For the purposes of this paper, "top journals" are those included in the categories "A" and "B".

Table 1. Impact indicators for the different types of publication channels.

Type of publication channel	*Indicators of impact/prestige*
WoS articles	Impact factor (25% top journals by impact factor)
Non-WoS articles	CIRC (categories A and B)
Books/Book chapters	SPI (10% top publishers by expert opinion)

c) Personal data: age, professional rank (P=postdoctoral research fellow, TS=tenured scientist, RS=research scientist and RP=research professor) and gender of researchers were provided by CSIC.

A preliminary inspection of the similarity between variables was explored by means of Multidimensional Scaling (MDS). Non-linear Principal Component Analysis (NLPCA) was used to explore the relationship between personal features of researchers and their performance. Statistical analyses were performed with SPSS (v.20).

Findings

A total of 268 researchers had at least one publication in the period 2007-2011. In the whole SSH area, men represented 59% of all researchers, average age of researchers was 50 years old, and half of the researchers were in the lowest scientific category (tenured scientist). Postdoctoral research fellows accounted for only 7% of researchers in the area. Small differences between the Humanities and Social Sciences can be observed in Table 2.

Table 2. Personal features and scientific rank of researchers in SSH.

		Humanities (N=192)		*Social Sciences (N=76)*		*Total (N=268)*	
Gender	Men	115	60%	42	55%	157	59%
	Women	77	40%	34	45%	111	41%
Rank 2007	Post-doc	12	6%	6	8%	18	7%
	Tenured scientists	98	51%	42	55%	140	52%
	Research scientists	46	24%	13	17%	59	22%
	Research professors	36	19%	15	20%	51	19%
Age		50 ± 9 (28-70)		49 ± 10 (32-70)		50 ± 9 (28-70)	

Note: age expressed as average ± standard deviation (min-max).

A total of 3,004 documents were published by CSIC researchers in SSH during 2007-2011. Differences between Humanities and Social Sciences in the main publication types used are observed: WoS articles predominate in the Social Sciences while book chapters are the most frequent publication channel in the Humanities (Table 3).

Table 3. Share of publication channels by area.

	Books	*Chapters*	*Non-WoS Articles*	*WoS Articles*	*Total*
Humanities	14% (397)	47 % (1,313)	26% (717)	13% (352)	2,779
Social Sciences	8% (65)	27% (214)	29% (227)	36% (289)	795
Total	13% (462)	43% (1,527)	26% (944)	18% (641)	3,574

Note: the total is higher than 3,004, because the publication count is made at the individual level.

Publication profile of researchers

A MDS was applied to the set of variables which make up the activity profile of researchers to reveal their underlying structure. In terms of similarity, the plot gives away greater levels of international collaboration and English-written publications for WoS articles. The patterns for the remaining publications types (books, chapters and non- WoS articles) seems to be mainly related to higher levels of productivity and being written in national languages (Figure 1).

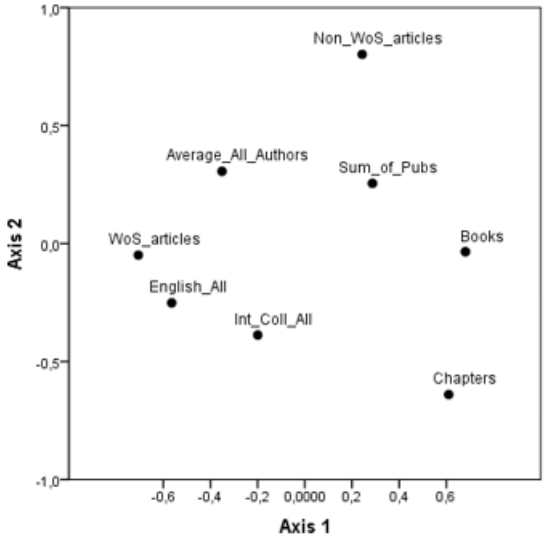


Figure 1. MDS for the scientific activity profile.

The diversity of publication channels in the output of researchers is the norm in SSH. Around 1/3 of the researchers presented output of the four different types considered: articles covered by WoS, non-WoS articles, books and book chapters. Three and two types of publication channels were observed in 40% and 17% of the researchers respectively, while only 12% of researchers had results of a single type. Several differences between Social Sciences and Humanities can be put forward: researchers who disseminate research among the four different types of publication channels considered are more frequent in Humanities (36% vs 24%), while using only WoS-covered journals is more common among Social Sciences researchers (16% vs 4). Finally, it is interesting to remark that around 22% of Social Sciences researchers and 41% of those in the Humanities may remain invisible in Web of Science-based studies since they do not show any publication covered by this database.

Research performance of scientists

Main statistics concerning research performance of scientists in SSH are shown in Table 4. A higher number of total publications is observed for researchers in the Humanities (15.1 vs 10.8), especially due to their high number of book chapters. Researchers in the Humanities exhibit a higher use of top publishers for books and chapters, while Social Sciences researchers present a greater share of articles in high impact factor journals.

Table 4. Description of the research performance of researchers in SSH.

	Humanities		*Social Sciences*	
	Mean	*SD*	*Mean*	*SD*
No. Books	2.1	2.5	0.9	1.0
No. Chapters	7.1	5.7	2.9	3.3
No. WoS Articles	1.9	4.3	3.9	4.1
No Non-WoS Articles	3.9	4.7	3.1	3.7
Sum of Publications	15.1	12.2	10.8	7.5
pptop10%_Books & Chapters	35.9	26.5	23.7	28.9
pptop10%_Non_WoS_Articles	32.8	35.3	37.3	37.9
% Q1 WoS Articles	12.9	29.7	33.4	36.5
Average number authors/publication	1.7	1.4	2.6	1.1
% International. collaboration	16.9	23.0	24.1	29.8
% English	14.0	19.0	38.6	32.7

To explore the possible relations between personal features of researchers and their performance NLPCA was used, which allows reducing a large number of variables to a smaller number of uncorrelated non-linear combinations of these variables with miminum loss of information (principal components). Two different studies are conducted, since researchers in Social Sciences and Humanities are analysed separately. Preliminary results concerning the plots of component loadings (two-dimensional solution) are shown in Figure 2.

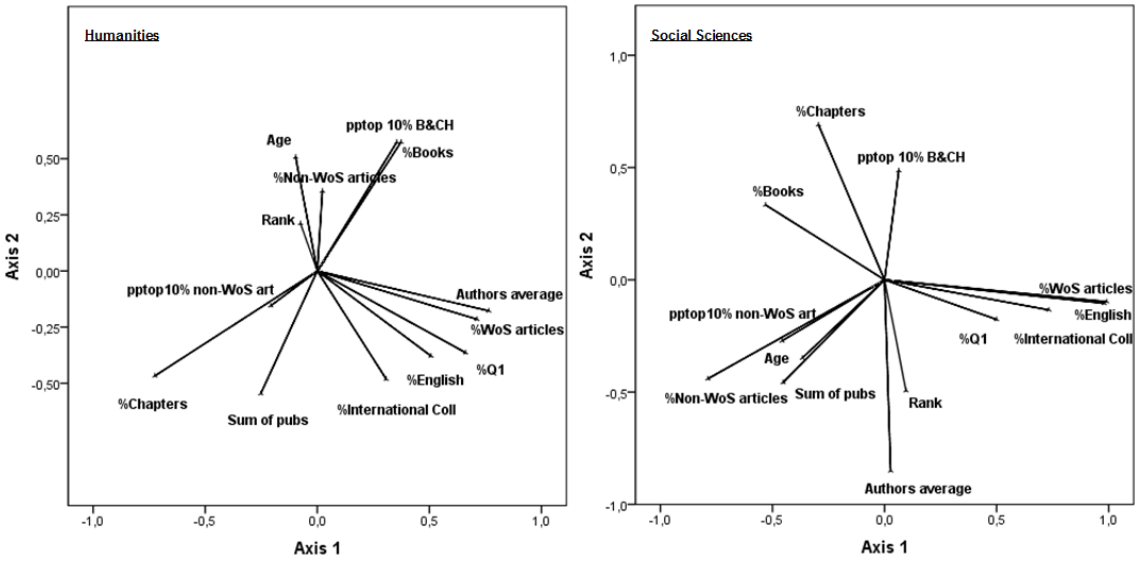


Figure 2. Component loadings in: a) Humanities; b) Social Sciences.

Note: only researchers with 2 or more publications considered

Discussion and conclusions

At this point, some preliminary results can be pointed out in an attempt to provide a comprehensive picture of the activity of CSIC researchers in SSH from a micro-level perspective:

- Taking into account only WoS articles would shrink meaningfully the visibility of CSIC researchers in SSH, in particular in the Humanities.
- Different constraints of the 'ConCiencia' system are identified. More rigour in the input of data (carried out by researches themselves) as well as in the cleaning and validation processes (by the institution) would be advisable.

- In the Humanities, researchers who hold a higher rank and age present greater activity in books and non-WoS articles. However, a high number of total publications is apparently not associated to a higher rank.
- In the Social Sciences, a higher academic rank is associated to internationally-oriented scientific publications (high share of WoS articles) as well as a high productivity (high number of publications) and collaborative activity (high number of co-authors).
- Differences between the Social Sciences and Humanities are observed, but even within each of these fields different typologies of researchers according to their publication pattern, collaboration practices and international/national orientation may exist. These factors are being explored at present.
- Although our study focuses on four different types of academic output, it is still not comprehensive, since it does not consider the non-scholarly literature, which may have an important societal impact.

Acknowledgments

Adrián A. Díaz-Faes is granted with a JAE predoctoral fellowship by the Spanish National Research Council (CSIC).

References

Archambault, E., Vignola-Gagne, E., Cote, G., Lariviere, V. & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. Scientometrics, 68(3), 329–342. doi:10.1007/s11192-006-0115-z.

Giménez-Toledo, E., Tejada-Artigas, C. & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: results of a survey. *Research Evaluation*, 22(1), 64-77.doi:10.1093/reseval/rvs036

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 473–496). Dordrecht, The Nederlands: Kluwer Academic. doi:10.1007/1-4020-2755-9_22.

van Leeuwen, T.N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship? *Bibliometrie - Praxis und Forschung*, 2013, 1-18. Retrieved June 15, 2015 from: http://www.bibliometrie-pf.de/article/viewFile/173/215

Mañana-Rodriguez, J. & Giménez-Toledo, E. (2013). Scholarly publishing in social sciences and humanities, associated probabilities of belonging and its spectrum: a quantitative approach for the Spanish case. *Scientometrics*, 94, 893-910. doi:10.1007/s11192-012-0838-y

Martin, B., Tang, P., Morgan, M., Glanzel, W., Hornbostel, S., Lauer, G., et al. (2010). Towards a bibliometric database for the social sciences and humanities—A European scoping project. Research report produced for DFG, ESRC, AHRC, NWO, ANR and ESF.

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions. *ISSI Newsletter*, 6(1), 22–28.

Sivertsen, G. & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics*, 91(2), 567–575. doi:10.1007/s11192-011-0615-3.

Torres-Salinas, D, Bordons, M., Giménez, E., Delgado, E., Jiménez, E. & Sanz, E. (2010). Clasificación integrada de revistas científicas (CIRC): propuesta de categorización de las revistas en ciencias sociales y humanas. *El Profesional de la* Información, 19(6), 675-683.

WTI2 (2014). Scholarly publication patterns in the social sciences and humanities and their relationship with research assessment. *Science, Technology & Innovation Indicators 2014. Thematic paper 2*, 1-26 Retrieved June 15, 2015 from: http://dialogic.nl/documents/other/sti2_themepaper2.pdf

A Bibliometric Assessment of ASEAN's Output, Influence and Collaboration in Plant Biotechnology

Jane G. Payumo[1] and Taurean C. Sutton[2]

[1] jane.payumo@kaust.edu.sa
4700 King Abdullah University of Science and Technology, Thuwal, 23955-6900 (Saudi Arabia)

[2] taurean.sutton@wsu.edu
1610 NE Eastgate Blvd., Suite 650, Washington State University, Pullman, WA, (United States)

Abstract

This research uses 10-year (2004-2013) publication and citation data related to plant biotechnology to assess the research performance, impact, and collaboration of member states of the ASEAN in plant biotechnology. Findings indicate increased scientific output of ASEAN countries in plant biotechnology as well as increased research collaborations by individual member states and with international partners throughout the 10-year period. The nature of collaboration by ASEAN is linked with the status of economic development of each country. Domestic and international collaborations are strong and are increasing through the years, regional collaboration on the other hand is found to be limited. This limited regional partnership can be a concern for the region's goal of economic integration. Further studies using bibliometric data analysis is suggested for policy diagnosis in plant biotechnology cooperation, knowledge flows, and effect of plant biotechnology research in economic development between ASEAN countries.

Conference Topic

Bibliometrics and research evaluation

Introduction

The Association of Southeast Asean Nations (ASEAN) has declared biotechnology as the main area of cooperation in science and technology. ASEAN, a regional association composed of 10 countries namely: Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar (Burma), Philippines, Singapore, Thailand, and Vietnam, considers plant biotechnology as the next pillar of regional economic growth (Hautea & Escaler, 2004; Erbisch & Maredia, 1998) and the answer to their food security needs. If ASEAN will continue to invest in plant biotechnology in the next years, it will be beneficial to have information on the current state of research and collaboration for strategic direction setting. This research drawing on bibliometric data, hence, will add to understanding the level and nature of collaboration, including research performance of ASEAN countries in plant biotechnology. This is relevant for ASEAN policy makers in charge of setting direction and designing strategies for research cooperation, and planning research investments, especially on biotechnology, at the country and regional levels.

Methodology

This research is based on 2004-2013 publications in plant biotechnology authored and co-authored by 10 member states of ASEAN. The data were extracted from Elsevier's Scopus database, the world's largest abstract and citation database of peer-reviewed literature (Elsevier B.V., 2014). Different keyword combinations were used to locate plant biotechnology-related publications guided by the glossary of biotech terms by the U.S. National Institute for Food and Agriculture (NIFA, 2014) and the National Agricultural Library Agricultural Thesaurus (National Agricultural Library, 2014). Additional filter was then set according to affiliation country to include only the publications published by the 10

ASEAN countries. No filter was set for the type of publication; all document type, namely: article, review, conference paper, short survey, note, editorial, letter, book chapter, book, and article in press were included. This research also highlights the use of a home-grown open-source 'publication parser' tool (Sutton, 2013); this tool was useful in parsing extracted files from Scopus for analysis of various indicators of interest at the country, institutional, and individual levels. The methodology, including interpretation of the different indicators, builds on best practices on indicators research that have been developed throughout the years (Moed, Glänzel, & Schmoch, 2004).

Results and Discussion

Publication output and citation impact

During the 10-year period (2004-2013), ASEAN researchers produced an overall total of 7,907 papers related to plant biotechnology; this output has increased 15% per year. These publications were written by more than 13,000 unique authors. The number of researchers producing knowledge for the region has increased steadily throughout the years with numbers reaching close to 8,000 authors in 2013 compared to less than 2,000 authors in 2004. Interestingly, ASEAN's plant biotechnology publications have mostly been published in open source journals such as Plos One. ASEAN's plant biotechnology publications have been cited more than 117,000 times with the highest citation count observed in 2007. The average citation per publication for plant biotechnology publications of ASEAN (19.81) is more than twice higher than the average CPP of all ASEAN publications (8.4) indicating higher influence of plant biotechnology publications than publications in other research areas.

Country output and ASEAN research investments

We then classified the 10 ASEAN countries into three groups based on expenditures on research and development (R&D) (UNESCO Institute for Statistics, 2015): (1) high income countries (HIC) with R&D spending more than 1% of gross domestic product (GDP); (2) middle income countries (MIC) with R&D spending of 0.1 to 0.9% of GDP; and (2) lower middle-income countries (LMIC) with R&D spending of 0.0 to 0.09% of GDP. A significant difference on the publication output in plant biotechnology of HICs with larger R&D investments was noted compared with that of LMICs with less research investments (Table 1). Thailand produced the most number of publications (n = 2489). Malaysia and Singapore are the other top three ASEAN producers with more than 150 PPY and CAGR of 29% and 9%, respectively. Philippines with a CAGR of 8% and Vietnam with a CAGR of 19% produced an average of 75 and 41 PPY, respectively. LMICs, namely Brunei Darussalam, Cambodia, Laos, and Myanmar experienced no growth during the ten-year period and have only produced an average of 1-2 papers per year. Interestingly, Indonesia despite its low R&D investments, hence, classified as a LMIC here, was able to produce 61 PPY and is growing at 12% CAGR. The number of authors contributing to ASEAN publications except the LMICs namely: Brunei Darussalam, Cambodia, and Laos, is growing. An increase in the number of contributing authors was especially noted for Malaysia; the country's number of authors from 2004 to 2013 has increased almost 15 fold.

HICs with higher number of publications received more total citations than lower income countries. Singapore is the most highly cited in plant biotechnology followed by Thailand, Malaysia, and Philippines. With the exception of Indonesia, other LMICs received the least amount of citations for their plant biotechnology publications during the last two decades.

Table 1. Comparison of 2004 and 2013 article output, CAGR, and citation count for ASEAN.

Country	Country classification	Publication output	2004	2013	CAGR	No. of authors	Citation count
Malaysia	MIC	2,199	39	510	29%	10,511	14,584
Vietnam	MIC	418	14	83	19%	2,474	3,957
Thailand	MIC	2,489	108	377	13%	12,688	27,863
Indonesia	LMIC	611	33	104	12%	3,421	7,208
Myanmar	LMIC	23	1	3	12%	100	180
Singapore	HIC	1,594	101	234	9%	10,953	49,094
Philippines	MIC	757	46	104	8%	4,444	14,492
Cambodia	LMIC	6	1	0	-100%	64	135
Brunei	LMIC	35	0	0		30	157
Laos	LMIC	10	0	3		136	186
Total		7,907					117,856

Note: CAGR of Cambodia and Brunei resulted in undefined values and left blank in this table. Source: Scopus

The topmost institution publishing plant biotechnology-related articles in the region are mostly local public research universities (e.g. University Brunei (Brunei), Bogor Agricultural University (Indonesia), National University of Laos (Laos), University of Malaya (Malaysia), Yezin Agricultural University (Myanmar), National University of Singapore (Singapore), and Mahidol University (Thailand). For Cambodia, Vietnam and Philippines, the top producers of publications on plant biotechnology were research institutions and include Cambodian Agricultural Research and Development Institute, Institute of Biotechnology, and International Rice Research Institute (IRRI). The two former institutions are national leading research institutions in bioscience and plant biotechnology while IRRI is an international research organization.

Collaboration

Guided by a decision tree adapted from Lan (2014), we distinguished four types of research collaboration: (1) domestic - in which all authors are in the same country; (2) regional – in which one ASEAN author co-authored with another ASEAN country; and (3) international – in which authors in the ASEAN countries published together with at least one author from another country besides the ASEAN countries. Single authorship and publications that involved intra-institutional co-authorship are not classified as collaboration in this research.

Single author publications and publications that involved intra-institutional co-authorship for ASEAN is very limited; they only constitute 15% of ASEAN's total publications in plant biotechnology. Eighty five percent of ASEAN's total publications in plant biotechnology, on the other hand, involved research collaboration, growing at a CAGR of 15%. Interestingly, the most active institutions that engaged in collaborations in ASEAN are the public universities and institutions of higher education; these institutions have also been noted earlier to be publishing most and the active generators of knowledge for ASEAN. These results confirm observation that plant biotechnology research in ASEAN countries is increasingly conducted now by a group of collaborating researchers rather than by a single researcher (Katz & Martin, 1997; Glänzel, 2001).

The region's co-authored publications that involved domestic partnership are growing at a CAGR of 15%. Six ASEAN members were engaged in domestic collaborations with

Malaysia, Thailand, and Singapore having the highest % shares of domestic collaborations at 42%, 37%, and 20%, respectively. Brunei Darussalam, Cambodia, Laos, and Myanmar have no record of domestic collaborations.

ASEAN publications that involved regional collaboration are very limited with less than 1% of the total collaborations of ASEAN. The highest number of publications that involved regional collaborations was recorded in 2013 (n = 21); there was no regional collaboration noted for 2007 and 2008. Ironically, 2007-2008 were the early years of the adoption of ASEAN's Economic Blueprint, which serve as the guide for the establishment of the ASEAN Economic Community. All the higher income countries have co-authored with another ASEAN country although numbers are quite limited (Figure 1). Philippines and Thailand have collaborated mostly with all of the ASEAN countries except Brunei Darussalam. Laos and Myanmar are two of the most active in regional collaborations despite their late membership to the regional association. Both countries have strong regional collaborations with Thailand, their closest ASEAN neighbor; Laos and Thailand used to belong to one country (Siam) and have basically the same language. Brunei has no record of collaborations with any of the ASEAN members.

The region has a very high rate of international collaboration in plant biotechnology research during 2004-2013 at 65% and the rate of collaboration is growing at a CAGR of 11 %. Similar with domestic and regional collaborations, the highest number of publications that involved international collaborations was recorded in 2013 (n = 227) while the least was recorded in 2004 (n = 717). ASEAN has partnered with 115 countries that are in varying stages of economic development. U.S. remains to be the main international research partner of choice among ASEAN countries. ASEAN is also tapping into the research expertise and resources of other Asian nations like Japan, China, South Korea, and India and advanced countries like United Kingdom, France, Germany, Canada, and The Netherlands. Arunachalam and Doss (2000) had the same observation and stated that Asian countries are fast increasing their share of worldwide international collaboration in science and expanding its collaboration beyond the traditional collaboration with advanced nations such as the United States.

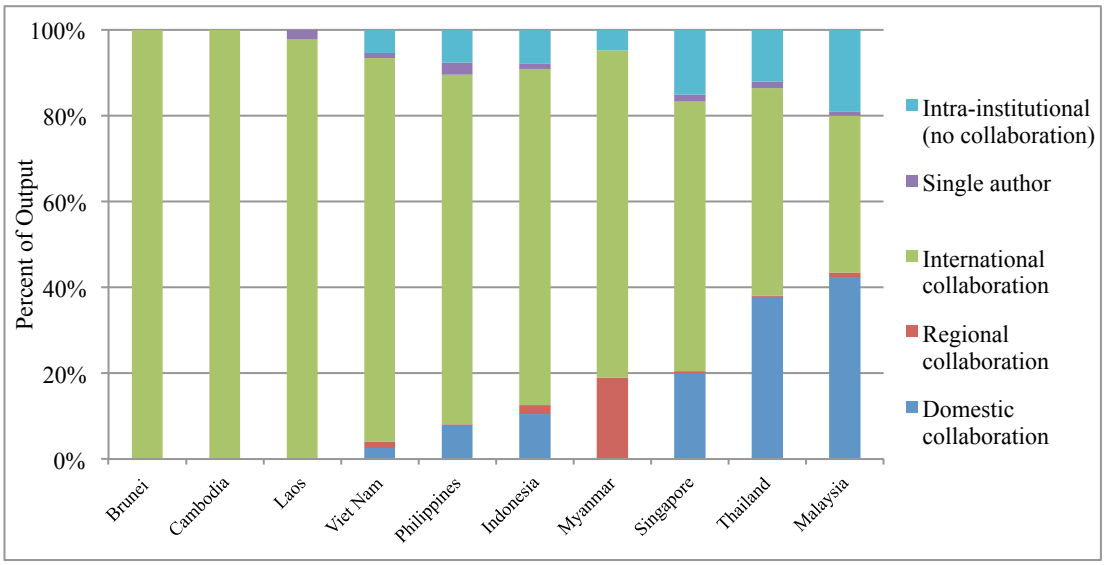


Figure 1. Percentage of different types of collaboration for individual ASEAN countries in plant biotechnology, 2004-2013. Source: Scopus

Brunei Darussalam, Cambodia and Laos are particularly noted for very high international collaboration. There are many justifications for this high collaboration rate and may include

the need for complementary and synergistic research expertise, greater visibility in the international plant biotechnology arena, and greater research output despite limited research investments. Interestingly, the higher income countries and the top ASEAN producers, namely Malaysia, Thailand, and Singapore have lower scientific output with the international community compared with other ASEAN countries, which validates observation that these countries have now higher domestic research capability, hence, would not need as much international collaboration as lower income countries. As expected, ASEAN publications that involved international partnerships received the highest citation count (n = 86,423) supporting earlier research while publications that involved regional collaborations received the least citation count (n = 547). It is interesting to note that despite the regional collaborations involving more authors and one or more ASEAN countries, the citation count was lower compared to single authored publications. This can indicate the less quality and influence of publications resulting from regional partnerships.

Conclusion and Recommendations

Using bibliometric data for the period 2004-2013 sourced from the research abstract database, Scopus, and deconstructed through a non-commercial home-grown publication parser tool, this paper investigates ASEAN's research output, influence and research collaboration in the area of plant biotechnology. Analysis of the 10-year period indicated an increase in ASEAN plant biotechnology-related scientific output. The publication activity obviously varies from country to country but evident that it is linked with R&D investments: higher income countries such as Singapore produced more publication than lower middle-income countries such as Brunei Darussalam. Most of the knowledge producers of ASEAN were from local research institutions, which are a good indication of improvements in domestic research capability and increase knowledge generation activity among this group. The relatively stable trend of publication generation and increasing R&D investments in countries such as Singapore, Thailand and Malaysia, likewise, provides a good indication that more research output can be expected from these countries. The growth of the publication records especially of Indonesia and Vietnam supports the increasing commitment of these countries and their researchers to contribute in advancing the plant biotechnology field. Philippines need to push and incentivize its local research and academic institutions to produce more and increase their scientific output and not rely on international institution to boost the country's scientific productivity. Brunei Darussalam, Cambodia, Laos, and Myanmar need to improve their research infrastructure and level up their research investments to catch up with other ASEAN countries.

The increasing number of collaborative research teams and number of contributing authors based on co-authorship data in ASEAN publications over the course of the 10-year period, however, is an encouraging result. It represents an increase in the pool of researchers and a change in the balance of research focused more on collaborative research teams among ASEAN researchers and their partners and not on lone scientist.

All the 10 ASEAN countries are actively engaged in research collaboration in plant biotechnology although in varying degrees. The publication output by countries in terms of the collaboration types: domestic, regional and international, differ and is also noted to be linked with status of economic development. Domestic collaborations are very strong for higher income countries with higher R&D investments while lower income countries with lower research investments tend to publish more with their international counterparts. There is more preference for collaboration with more advanced nations but at least the region has expanded its collaboration beyond the United States.

Regional partnerships are, however, very limited, and can be a concern for ASEAN's goal of integration. ASEAN regional collaboration still lag behind in terms of productivity and

quality research in plant biotechnology, which is very evident from the region's low research output and citation count for publications co-authored among ASEAN researchers. Higher regional collaboration rate is only observed to countries that are in close proximity to each other, with common language, and with historical links. Kumar, Rohani, & Ratnavelu (2014) found the same scenario after doing bibliometric work in the field of economics. The low regional collaboration was also mentioned in one of the latest reports by the Asian Development Bank, Regional Cooperation and Cross-Border Collaboration in Higher Education in Asia: Ensuring that Everyone Wins (Asian Development Bank, 2012). Hence, it remains to be seen whether regional collaboration will serve as an important platform for continuing to modernize plant science in ASEAN and sharing knowledge in plant biotechnology. More investments in research cooperation, funding mechanisms for regional plant biotechnology research, and other regional incentives need to be setup so ASEAN can realize the goal of its regionalization agenda. Regular quantitative monitoring of inputs and outcomes of research in ASEAN is likewise encouraged to monitor research performance and help in developing research management and science policies, particularly in economic development. Additional research focused on mapping of research collaboration network among ASEAN researchers and their global partners, and a brain circulation study can be done to understand the mobility of ASEAN researchers and whether such movement helps in increasing regional productivity and collaborations and whether such benefits flow back to ASEAN. Furthermore, a qualitative study that would determine other factors that influence an ASEAN researcher to collaborate with another ASEAN researcher or a global partner is suggested.

References

Arunachalam, S. & Doss, M.J. (2000). Mapping international collaboration in science through coauthorship analysis. *Current Science*, 79(5), 621-628.

Asian Development Bank. (2012). *Regional Cooperation and Cross-Border Collaboration in Higher Education in Asia: Ensuring that Everyone Wins*. Retrieved June 15, 2015 from: http://www.adb.org/sites/default/files/publication/29931/regional-cooperation-higher-education-asia.pdf

Elsevier. (2014, December 26). *Scopus*. Retrieved June 15, 2015 from: http://www.scopus.com/

Erbisch, F. & Maredia, K. (1998). *Intellectual Property Rights in Agricultural Biotechnology.* Wallingford: CAB International.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, 51(1), 69-115.

Hautea, R. & Escaler, M. (2004). Plant biotechnology in Asia. *AgbioForum*, 7(1 and 2), 2-8.

Katz, J. & Martin, B. (1997). What is research collaboration? *Research Policy*, 26(1), 1-18.

Kumar, S., Rohani, V.A., & Ratnavelu, K. (2014). International research collaborations of ASEAN Nations in economics, 1979–2010. *Scientometrics*, 101(1), 847-867.

Moed, H., Glänzel, W., & Schmoch, U. (2004). *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems.* New York, Boston; London, Moscow: Kluwer Academic Publishers.

National Agricultural Library. (2014, December 22). *National Agricultural Library Agricultural Thesaurus Library*. Retrieved June 15, 2015 from: http://agclass.nal.usda.gov/agt.html

NIFA. (2014, December 6). *Glossary of Biotechnology Terms: NIFA*. Retrieved June 15, 2015 from: http://www.csrees.usda.gov/nea/biotech/res/biotechnology_res_glossary.html

Sutton, T.C. (2013). Publication Parser Tool for Scopus. Version 2.0, Pullman, WA.

Science and Technology Indicators In & For the Peripheries. A Research Agenda

Ismael Rafols[1,2], Jordi Molas-Gallart[1] and Richard Woolley[1]

i.rafols@ingenio.upv.es
[1]Ingenio (CSIC-UPV), Universitat Politècnica de València, València (Spain)
[2]SPRU (Science and Technology Policy Research), University of Sussex, Brighton (UK)

jormoga@ingenio.upv.es, ricwoo@ingenio.upv.es
[1]Ingenio (CSIC-UPV), Universitat Politècnica de València, València (Spain)

Abstract
This paper aims to propose a research agenda that explores the problems that emerge when S&T indicators are used in peripheral contexts, that is, in geographical or social spaces that are somehow marginal to the centres of scientific activity. In these situations evaluators and decision-makers are likely to use indicators that were designed to reflect variables relevant in the dominant social and geographical contexts --i.e. in the leading countries, languages, disciplines, etc.--, but that are usually not adequate in peripheral contexts. We propose to examine various dimensions of periphery. First, the cognitive dimension: areas of research, such as the humanities that capture less attention (and resources) than the more prestigious disciplines, such as molecular biology. Second, the geographical dimension: e.g. global south vs. global north, regions vs. metropolises. Third, the social group dimension: women, the poor, or perhaps the elderly have social needs that are different from those of richer or more powerful groups --and the problems affecting the former tend be less researched than those of the later. The research agenda proposed would investigate the mechanisms by which performance indicators tend to be biased against the peripheries (e.g. bias in language, journal or topic coverage in conventional databases). We suggest how these biases may suppress scientific diversity and shift research towards a higher degree of homogeneity.

Conference Topic
Science policy and research assessment

Introduction
Science and technology indicators are becoming increasingly used over a wide variety of contexts as research activities become prominent in a larger range of countries, a broader set of organisations, and over a wider range of disciplines or topics (Sa, Kretz et al., 2013). Given that the indicators used in new contexts are often the same, or close adaptations of the indicators used in the traditional disciplines, elite universities and dominant scientific countries, one may wonder about their validity (i.e. adequacy of the indicator to the concept/object is supposed to measure) and their robustness (or sensitivity to contingency in the measuring conditions) (Gingras, 2014).

In this work-in-progress contribution, we propose that many of the new contexts where indicators are used constitute what we call the peripheries or the margin of the research system: spaces that have less visibility, less prestige and/or less resources. As peripheries, these spaces have not had the capacity or influence to develop home-grown indicators suited for their activities -- and are instead relying on indicators borrowed from the central or dominant disciplines and/or countries. For example, it is a recurrent debate in policy to which extent scientometric indicators can be used in the social sciences and humanities (Martin, Tang and Morgan, 2010). Another recurrent example is the case of peripheral countries such as Brazil, where studies have showed that publication practices and citations differed significantly from those in the leading scientific nations, given that they "are significantly influenced by factors "external" to the scientific realm and, thus, reflect neither simply the

quality, influence nor even the impact of the research work referred to." (Velho, 1986, p. 71; see also Velho & Krigge, 1984).

In this contribution we explore dimensions in which the use of indicators in peripheral contexts may be problematic, providing misleading information for research assessment or strategy development. In these contexts, we propose that alternative methods should be explored and potentially developed to create new indicators that are fit for purpose.

This exploration will be developed into the central research agenda for a joint conference of the networks RICYT (the Ibero-American network of Science and Technology Indicators, http://www.ricyt.org) and ENID (the European Network of Indicators Designers, http://enid-europe.org) to be celebrated in Valencia between 14 to 16 December 2014.

A relational and multidimensional conceptualisation of periphery

The Oxford English Dictionary defines "periphery" as

"The region, space, or area surrounding something; a fringe, margin. Now chiefly: the outlying areas of a region, most distant from or least influenced by some political, cultural, or economic centre."

Its cousin, the Oxford Dictionary of English provides a slightly different definition:

"A marginal or secondary position in, or aspect of, a group, subject, or sphere of activity."

There is already a long history of grappling with the question of peripheries in relation to global social and economic change and development (Prebisch, 1949). Science studies in Latin America have long discussed their peripheral situation and how it meant that their scientific knowledge was dependent, "transplanted" and thus often not properly adapted to their domestic needs -- rather the needs of the Northern countries exploiting their economic resources. For example Vessuri (2004, p. 174) explains that:

> *"Irrespectively of their capabilities, these scientific thinkers were "peripheral" in three senses: in their marginal position in the outer ridges of European culture; in their partial commitment to the scientific endeavour (forced by the immediate pressures for survival in the middle of often unstable contexts, and the economic and political urgencies of new nations); and in their role as agents for the exploitation of natural resources of economic interest for the European centres of power, who gave them legitimacy and support." (Our translation from Spanish)*

A noticeable characteristic of this description is the multidimensional nature of the "sense" or spaces of the peripheries of Latin American scientists: culturally (or cognitively), institutionally (partial commitment), in economic terms (unstable resources and dependent on European funding) and in the topics addressed (those of interest to the centres of power).

These definitions suggest two important traits of the notion of periphery, as illustrated by Vessuri's quote above. First, it refers to a situation that is somehow marginal, far from the centre, and where, consequently, less attention is paid. The periphery is therefore *always defined in relation to a centre* where the main locus of the relevant activity resides.

Second, the concept can *relate to many different dimensions* (political, cultural, economic, different "spheres of activity"). In turn these dimensions may or may not be linked with a geographic location; for instance a centre of economic activity will be a specific geographic location. Geographic locations tend to be centre (or periphery) for a variety of dimensions: it is common for political, economic and cultural activities to cluster around geographical centres of power and influence. Similarly, peripheral regions will be peripheral along several dimensions and so the application of the term peripheral to a region has come to indicate a situation of structural disadvantage with broad economic, political and social implications. Developing countries were long ago described as "the" periphery, but within every geographical region we can also encounter peripheral zones (Southern European and Eastern European countries as peripheral to the European Union, or relatively poor regions as

peripheral within their country). Yet, not all dimensions will be correlated for a specific locality. Cambridge is a geographic centre of learning and research (a centre in a cognitive dimension) but, as a city, it is not a centre of political power, although the social group of Cambridge alumni, lecturers and researchers are part of both a political and a cognitive centre.
Also, not all relevant dimensions need to have a geographical expression. One can think for instance of social dimensions like gender or class that can be interpreted under the lenses of centre and periphery but are not associated with specific geographic localities. We can therefore refer to peripheral social groups (the disenfranchised, the poor…) whose economic and social needs will be different from those of richer or more powerful communities, even when part of this groups may be located in centres of political power (e.g. the poor neighbourhoods in Washington DC).
Similarly, cognitive dimensions are not necessarily associated with geographic locations; for instance, cognitive peripheries would include areas of research that do not capture the attention of mainstream politicians and receive more limited resources. From this perspective, many fields in the humanities could be considered a peripheral field of knowledge when compared to mainstream natural or engineering sciences.

How conventional indicators are problematic in the peripheries

As we have seen, the notion of a periphery is thus fundamentally a relational one. A periphery is always constituted in relation to a centre, or core. From an indicator perspective, the same entity may thus be peripheral or central depending on the frame of analysis. A particular region may be the centre of nanomaterials research in a particular country, but peripheral in relation to global nanomaterials research, for example. Whether the region is depicted as periphery or centre depends on the frame of comparison. A problem with the *use of indicators* is thus the risk of inappropriate comparisons that can render important activities as relatively trivial.
A second problem relates to whether what is being measured about a particular entity is relevant knowledge in terms of the needs, objectives or valued activities of that entity. The application of an indicator constructed to reflect the needs, objectives or valued activities of another entity may not produce useful information – only a mismatched comparison. A problem with the *content of indicators* is thus the risk of inappropriate comparisons that can
render important activities as relatively invisible or lacking in impact. The use of indicators can thus play a role in *constituting peripheries*.
Our goal in this section is to analyse how indicators developed to assess policies and activities related to Science and Technology address peripheral spaces and whether they have constitutive (intended or unintended) effects on these peripheries. We therefore need to identify the dimensions that are relevant to the conduct of S&T.
Each periphery faces its own knowledge generation and application context and may be better analysed using specific, tailored indicators. Yet, by and large they need to rely on indicators, and analytical models developed for the studies of "centre" spaces. Evaluators and decision-makers are likely to use indicators that were designed to reflect variables relevant in the dominant social and geographical contexts --i.e. in core regions, languages, disciplines, etc.--, but that are usually not adequate in peripheral spaces.
Let us see some examples of dimensions where use of indicators in the periphery is problematic.

Language

Language has long been known to be a major problem for performance measures, given that non-English articles tend to be much less cited. Van Leeuwen et al. (2001) showed that the inclusion or not of non-English publications in the analysis of citation impact has a major influence in the outcomes of indicators. Van Raan et al. (2011) showed that this also had

major effects in university rankings. Vasconcelos et al. (2008) showed that language proficiency is highly correlated with citation impact and h-index of researchers. This means that for the purposes of comparison, non-English publication should be excluded in most analysis.

Gender

In many fields of science, women tend to publish less and accrue less citations than men. However, various studies have consistently found that women tend to do more interdisciplinary research (e.g. Leahey, 2007; Van Rijnsoever and Hessels, 2011). Hence, the effect of gender on performance depends on the indicators choice: if publications and citations are taken as a measure of the value of a contribution, the indicators will tend to disadvantage female researchers.

Basic vs. applied vs. research

Applied studies tend to cite fundamental studies more than the reverse. As a result, fundamental research tends to appear as more central in global science maps (Rafols, Porter and Leydesdoff, 2010). This is possibly a perception bias without serious repercussions. The serious problem is that even within a given scientific field as defined by conventional classifications such as Web of Science Categories, applied research tends to be significantly less cited than fundamental research (van Eck et al., 2013).

Interdisciplinary research

Interdisciplinary research can be thought of as peripheral to the extent that it is published in areas outside the disciplinary cores. It turns out that interdisciplinary research tends to be published in journals with lower rating in journal rankings and, within a field, with journals with a lower Journal Impact Factor (Rafols et al., 2012). As a result interdisciplinary research tends to be in a disadvantage when using this type of journal-based indicators (with citation indicators, the effect may vary as it depends on relative citation rates between fields that are being cross-fertilised).

Conclusions

S&T indicators tend to be biased against organisations, countries or disciplines in the periphery. This is possibly due to the fact that indicators were not initially designed for the peripheries. At the same time, the use of these indicators in assessments linked to the distribution of resources can have constitutive effects, reinforcing for instance the peripheral character of a region or discipline. These remain unresolved problems for S&T indicators and their use in evaluation. In this contribution we shed light on this bias in multiple dimensions, in order to foster critical awareness of the problems caused by biases as well as the development of context sensitive indicators (Lepori & Reale, 2012).

Acknowledgments

Ismael Rafols thanks Diego Chavarro (who is doing a thesis on the Journal Indexing Systems in Latin America) for many illuminating discussions on indicators in developing countries.

References

Gingras, Y. (2014). Criteria for evaluating indicators. In B. Cronin & C. Sugimoto (Eds.), *Beyond Bibliometrics:Harnessing Multidimensional Indicators of Scholarly Impact* (pp. 109–126). Cambridge, MA and London, UK: MIT Press.

Martin, B. R., Tang, P., Morgan, M. et al. (2010). *Towards a Bibliometric Database for the Social Sciences and Humanities – A European Scoping Project* (A report for DFG, ESRC, AHRC, NWO, ANR and ESF). Brighton, UK: SPRU.

Leahey, E. (2007). Not by Productivity Alone: How Visibility and Specialization Contribute to Academic Earnings. *American Sociological Review*, *72*(533-561).

Lepori, B. & E. Reale (2012). S&T indicators as a tool for formative evaluation of research programs. *Evaluation 18*(4): 451-465.

Prebisch, R. (1949). Interpretación del proceso de desarrollo latinoamericanoen. *CEPAL*, Santiago de Chile, Chile.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinarity. The case of innovation studies and business and management. *Research Policy*, *41*(7), 1262–1282.

Sá, C. M., A. Kretz, et al. (2013). "Accountability, performance assessment, and evaluation: Policy pressures and responses from research councils." *Research Evaluation 22*(2): 105-117.

Van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLoS ONE*, *8*(4), e62395. doi:10.1371/journal.pone.0062395

Van Leeuwen, T. N., Moed, H. F., Tijssen, R. J., Visser, M. S., & Van Raan, A. F. (2001). Language biases in the coverage of the Science Citation Index and its consequencesfor international comparisons of national research performance. *Scientometrics*, *51*(1), 335-346.

Van Raan, A. F., Van Leeuwen, T. N., & Visser, M. S. (2011). Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. *Scientometrics*, *88*(2), 495-498.

Van Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. Research Policy, *40*(3), 463–472.

Vasconcelos, S. M., Sorenson, M. M., Leta, J., Sant'Ana, M. C., & Batista, P. D. (2008). Researchers' writing competence: a bottleneck in the publication of Latin-American science?. *EMBO reports*, *9*(8), 700-702.

Velho, L., & Krige, J. (1984). Publication and citation practices of Brazilian agricultural scientists. *Social Studies of Science*, *14*(1), 45-62.

Velho, L. (1986). The "meaning" of citation in the context of a scientifically peripheral country. *Scientometrics*, *9*(1), 71-89.

Vessuri, H. (2004). La hibridización del conocimiento. La tecnociencia y los conocimientos locales a la búsqueda del desarrollo sustentable. *Convergencia*, *35*(May-August), 171–191.

Patterns of Internationalization and Criteria for Research Assessment in the Social Sciences and Humanities

Gunnar Sivertsen

gunnar.sivertsen@nifu.no
Nordic Institute for Studies in Innovation, Research and Education
(NIFU) P.O. Box 5183 Majorstuen, N-0302 Oslo (Norway)

Abstract

This paper investigates the developments during the last decades in the use of languages, publication types, and publication channels, in the social sciences and humanities (SSH). The purpose of the study is to develop an understanding of the processes of internationalization and to apply this understanding in a critical examination of an often used criterion in research evaluations in the SSH: Coverage in Scopus or Web of Science is seen in itself as an expression of research quality and of internationalization. This extrinsic 'coverage criterion' is beyond the control of academia and without support in analysis of how research quality and relevance is achieved through scholarly publishing in the SSH. It needs to be replaced by intrinsic criteria based on the SSH's own concepts of field-specific research excellence and societal relevance. The study will demonstrate this by using data from scholarly publishing in the SSH that go beyond the coverage in the commercial data sources by giving a more comprehensive representation of the SSH.

Conference Topic

Science policy and research assessment

Introduction

The presence of publications in Scopus or Web of Science (WoS) has increasingly become a criterion in evaluations of research in the social sciences and humanities (SSH). Some countries have even installed protocols for research evaluation or performance-based funding models where publications that are indexed by the commercial databases are treated separately in indicators of "internationalization" and "research quality". In other countries, there is a general belief that research quality can be promoted in the SSH by expecting more publications in the limited number of international journals that have been selected for indexing. Consequently, for several years already, Elsevier and Thomson Reuters have experienced a pressure from researchers in the SSH to have more journals indexed. Both providers have responded by increasing the coverage of journals and book series, and, recently, even of books in the SSH. However, the coverage of the scholarly publication output in the SSH is still limited (Sivertsen, 2014). The shortage is mainly due to the more heterogeneous scholarly publication patterns in the SSH where publishing in international journals is supplemented by book publishing and the use of journals in the native languages (Hicks, 2004; Archambault et al, 2006; Engels, Ossenblok & Spruyt, 2012; Sivertsen & Larsen, 2012; Sivertsen, 2014).

Just as with the abuse of Journal Impact Factors in research assessment in science, technology and medicine (STM), the 'coverage criterion' in the SSH represents an artefact which is external to and beyond the control of the scholarly norms and standards that it is sought to represent. It creates unnecessary tensions between fields in the SSH with different degrees of coverage in the databases. It also creates debates about what will happen to the use of books and native languages in the SSH. In these debates, the general development towards publishing in journals covered by Scopus or Web of Science is often perceived as "inevitable" and driven by new evaluation regimes, not by internal scholarly standards. In this study, I will develop an understanding of the processes of internationalization in the SSH which is

independent of the 'coverage criterion' and instead related to concepts of field-specific research excellence and societal relevance in the SSH.

Methods

For the purpose of this study, data are needed that give a complete representation of scholarly publishing it the SSH, also of publications in books, series and journals not covered by Scopus or Web of Science. In 2005, Norway was the first country to establish a national information system with complete quality-assured bibliographic data covering all peer-reviewed scholarly publishing in the total higher education sector (Schneider, 2009; Sivertsen, 2010). This national system, which is now called CRISTIN (Current Research Information System in Norway) and has been expanded beyond the higher education sector, provides the main source of data for this study.

The methodology of the bibliographic data collection in the Norwegian CRISTIN database (www.cristin.no) has been published earlier (Sivertsen, 2010; Sivertsen & Larsen, 2012; Sivertsen, 2014). Scientific and scholarly publications of all fields are covered completely according to an agreed definition. Among other criteria, the definition demands originality and scholarly format in the publication and peer-review in its publication channels. All publication channels (journals, series, book publishers) and publication types (see below) are standardized in the database.

Humanities is defined in our study as the disciplines included in this major area in the OECD Field Classification.[1] The *Social Sciences* are defined in the same way with the exception of Psychology, which we have not included in this study. Note that Law and Educational Research are classified as social sciences by OECD.

Two supplementing data sets (A, B) will be used, each of them for a more specific purpose:

A. For the analysis of publication patterns in the SSH down to the level of individual researchers, we use data from the above-mentioned CRISTIN system which cover the four years 2010-2013. The unit of analysis is publications per researcher within a variable of three publication types (articles in journals or series with ISSN; articles in books; books) and a dichotomous variable of languages (Norwegian (the native language); International languages). The data include 1,895 unique researchers in the humanities with 7,145 unique publications, and 3,229 unique researchers in the social sciences with 11,817 unique publications.
B. For the analysis of the development of publication patterns in the SSH over time, we use data that are defined and collected in the same way as in data set A, but aggregated at the level of disciplines. The data cover the years 2005-2011. The unit of analysis is publication per discipline (and major area) with the same variables of publication types and languages as in data set A. Data set B includes 14,558 unique publications in the humanities and 19,450 unique publications in the social sciences.

Results, Part I: Characteristics of the Publication Patterns in the SSH

As seen in *Table 1*, publications in journals and series represent a little more than half of the publications in the humanities and two thirds of the publications in the social sciences, indicating that book publishing is important as well, especially in the form of articles in books (edited volumes). There are, however, just as wide differences *within* each of the two major areas: Only 45 per cent of the publications in History are in journals, compared to 61 per cent in Linguistics. In Sociology, only 46 per cent of the publications are in journals, compared to 75 per cent in Economics.

[1] OECD: REVISED FIELD OF SCIENCE AND TECHNOLOGY (FOS) CLASSIFICATION IN THE FRASCATI MANUAL, version 26-Feb-2007, DSTI/EAS/STP/NESTI (2006)19/FINAL.

Table 1. Number and percentage publications per publication type. Based on data set A.

	Humanities N	*Humanities %*	*Soc Sci N*	*Soc Sci %*
Books	328	4.6 %	273	2.3 %
Articles in books	2,861	40.0 %	3,640	30.8 %
Articles in journals or series	3,956	55.4 %	7,904	66.9 %
Total	7,145	100.0 %	11,817	100.0 %

The scholarly publication types in the SSH are often discussed as if they represent alternatives to each other: Is the use of one of the publication types increasing at the cost of the others? Are monographs becoming obsolete in the SSH? Before we study the trends, we shall observe an indication that the publication types are supplementing each other rather than competing with each other. As seen in Table 2, the numbers and percentages of *the researchers* that actually use a certain publication type are significantly higher than in Table 1, indicating that more than one publication type is often present in the publishing profile of an individual researcher. As an example, although less than a third of the publications in the social sciences are articles in books, more than half of the researchers are using this publication type.

Table 2. Number and percentage of the researchers using a publication type within four years. Based on data set A.

	Humanities N	*Humanities %*	*Soc Sci N*	*Soc Sci %*
Books	297	15.7 %	273	8.5 %
Articles in books	1,187	62.6 %	1,676	51.9 %
Articles in journals or series	1,537	81.1 %	2,775	85.9 %
Total (unique researchers)	1,895		3,229	

Table 3 demonstrates to what degree the publishing profiles of individual researchers include more than one publication type. Even in the social sciences, where journal articles represent two thirds of the output, almost half of the researchers who publish these articles also use other publication types.

Table 3. Number and percentage of the researchers using a publication type that also uses another publication type within four years. The percentages are related to the numbers (N) in Table 2. Based on data set A.

	Humanities N	*Humanities %*	*Soc Sci N*	*Soc Sci %*
Books	265	89.2 %	250	91.6 %
Articles in books	891	75.1 %	1,275	76.1 %
Articles in journals or series	930	60.5 %	1,291	46.5 %

So far, we can conclude that book publishing and journal publishing seem to supplement each other rather than represent alternatives in the SSH. We will return to a possible explanation for this in the discussion at the end.

We now turn to another dimension in the publication patterns of the SSH – the language dimension. In non-English speaking countries, the use of the native language in scholarly

publications is an indication that the publication is mainly oriented at a national or regional audience of readers in which not only peers, but also students, policy makers, professionals, media and a wider public may be reached as well. Since scholarly publications in the native languages are relatively frequent in the SSH, publishing in an international language is, on the other hand, not the normal situation, as in the sciences, but a clear expression of an ambition to reach an international audience of experts in the field.

We proceed as with the publication types and start with an overview of the use of language in publications in Table 4. In both the humanities and the social sciences, the majority of scholarly publications are in the international languages. However, publications in the native language are much more frequent than in the sciences, indicating that such publications have a specific role in the SSH.

Table 4. Number and percentage publications per language type. Based on data set A.

	Humanities N	*Humanities %*	*Soc Sci N*	*Soc Sci %*
International language	4,368	61.1 %	8,666	71.7 %
Norwegian language	2,777	38.9 %	3,418	28.3 %
Total	7,145	100.0 %	11,817	100.0 %

Again, the question may be raised: Are the native and international languages supplementing each other, or are they competing as alternatives? By going down to the level of individual researchers, we can observe in Table 5 that high proportions of the researchers combine both types of languages in their publication practice. While a majority of researchers publish in the international languages, there is *no minority of researchers* publishing in the native language only. Researchers in the SSH are *normally bilingual* in their publication practice (if their native language is not English).

Table 5. Number and percentage of the researchers using international and native languages in their scholarly publications within four years. Based on data set A.

	Humanities N	*Humanities %*	*Soc Sci N*	*Soc Sci %*
International language	1,482	78.2 %	2,687	83.2 %
Norwegian language	1,228	64.8 %	1,725	53.4 %
Total (unique researchers)	1,895		3,229	

A more general conclusion from the results so far, is that although the *majority of publications* in the SSH are published in journals and in international languages, *the majority of researchers* are publishing in books and in the native language as well. Is this picture changing?

Results, Part II: Developments in the Publication Patterns in the SSH

To study the developments, we use data set B, by which it is possible to cover a longer period of time. The general picture is that the publication patterns in the SSH are quite stable, both with regard to publication types (Figure 1) and the use of international versus native languages (Figure 2). In relative shares, the uses of international languages and of journals are increasing, but not by a high rate. In absolute numbers, there is no in reduction book publishing or the use of the native language, since in data set B, which we are using here, there was an increase in the total number of publications by more than 50 per cent between 2005 and 2011.

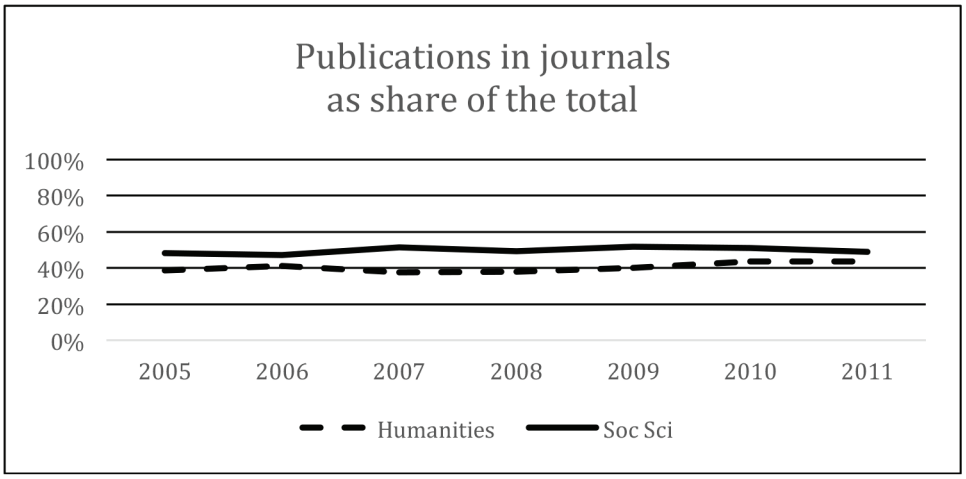


Figure 1. Scholarly publications in journals as a percentage of the total, which also includes articles in books and books. Based on data set B.

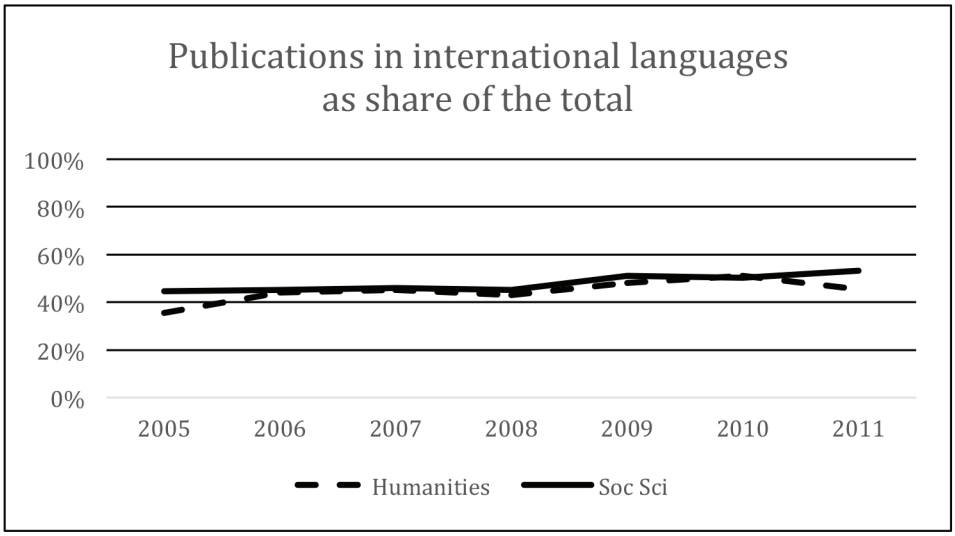


Figure 2. Scholarly publications in international languages as a percentage of the total, which also includes publications in the native language. Based on data set B.

Discussion and Conclusions

The normal publication practice in the SSH, in which both types of languages, and books as well as journals, are used for scholarly publishing by the majority of researchers, seems to prevail during a period of internationalization. The stability of the publication patterns, as well as their differences *within* the SSH (Sivertsen & Larsen, 2012; Ossenblok, Engels & Sivertsen, 2012), indicate that the choice of language and publication type is not just a question of new trends versus old traditions. Publication patterns are more deeply rooted in scholarly norms, methods and practices. The monograph, the edited book and the journal article represent different methodologies that may all need to be used at different times. The choice of language depends on the international scholarly relevance of the research versus the societal relevance for the culture and society being studied. One and the same research project may well contribute with different parts to both dimensions. The SSH would lose their *raison d'être* and societal impact by disconnecting from the surrounding culture and society and mainly communicating in international journals that are only read by peers abroad. At the same time, publishing in those specialized journals on the international level is necessary in

order to be confronted with and inspired by the scholarly standards, critical discussions and new developments among other experts in the field.

In the context of criteria for research evaluation in the SSH, there is a need to accept that none of the alternatives in the two dimensions of the scholarly publication patterns that have been described here – language and publication type - can be regarded as more valuable alternatives. All of them contribute – with different roles and connected to different methodologies, audiences and feedbacks – to research excellence and societal relevance of the SSH. The coverage in Scopus or the Web of Science of the scholarly publishing pattern in the SSH is far from complete (Sivertsen, 2014). Hence, *coverage in a commercial indexing service* should not be used as a criterion for research quality or an indicator of internationalization in the SSH.

References

Archambault, E., Vignola-Gagne, E., Cote, G., Lariviere, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics*, 68(3), 329-342.

Engels, T.C.E., Ossenblok, T.L.B., & Spruyt, E.H.J. (2012). Changing publication patterns in the social sciences and humanities 2000-2009. *Scientometrics*, 93(2), 373-390.

Hicks, D. (2004). The four literatures of social science. In Moed, H., Glänzel, W., & Schmoch, U. (Eds.) *Handbook of Quantitative Science and Technology Research*. Kluwer Academic Publishers.

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration, *Science, Technology & Human Values*, (17), 101-126.

Ossenblok, T.L., Engels, T.C., & Sivertsen, G. (2012). The representation of the social sciences and humanities in the Web of Science – a comparison of publication patterns and incentive structures in Flanders and Norway (2005–9). *Research Evaluation*, 21(4), 280-290.

Schneider, J.W. (2009). An Outline of the Bibliometric Indicator used for Performance-based Funding of Research Institutions in Norway, *European Political Science*, 8, 364–78.

Sivertsen, G. (2010). A performance indicator based on complete data for the scientific publication output at research institutions, *ISSI Newsletter*, 6, 22–8.

Sivertsen, G. & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, 91(2), 567-575.

Sivertsen, G. (2014). Scholarly publication patterns in the social sciences and humanities and their coverage in Scopus and Web of Science. In *Proceedings of the Science and Technology Indicators Conference 2014 Leiden*, ed. Ed Noyons, 598-604. Leiden: Centre for Science and Technology Studies.

Looking for a Better Shape: Societal Demand and Scientific Research Supply on Obesity

Lorenzo Cassi[1], Ismael Rafols[2], Pierre Sautier[3] and Elisabeth de Turckheim[4]

[1] *lorenzo.cassi@uni-paris1.fr*
Observatoire des Sciences et Techniques (HCERES-OST) and CES University of Paris 1 Pantheon-Sorbonne, Paris (France)

[2] *i.rafols@ingenio.upv.es*
Ingenio (CSIC-UPV), Universitat Politècnica de València, València (Spain),
SPRU (Science and Technology Policy Research), University of Sussex, Brighton (UK),
and Observatoire des Sciences et Techniques (HCERES-OST), Paris (France)

[3] *pierre.sautier@obs-ost.fr*
Observatoire des Sciences et Techniques (HCERES-OST), Paris (France), and Ingenio (CSIC-UPV), Universitat Politècnica de València, València (Spain)

[4] *elisabeth.deturckheim@obs-ost.fr*
Observatoire des Sciences et Techniques (HCERES-OST), Paris (France), and INRA, Délégation à l'évaluation, Paris (France)

Abstract
As science policy shifts towards an increasing emphasis in societal problems or grand challenges, new scientometric tools are required to inform decision-makers. However, while traditional bibliometrics could focus on the knowledge production side (the science supply), grand challenges also demand to investigate the articulation of societal needs. In this paper, we present an exploratory investigation of the grand challenge of obesity -an emerging health problem with enormous social costs. We illustrate a potential approach, showing: (a) how scientific publication can be used to describe existing science supply by using topic modelling based on publication abstracts; (b) how question records in the French parliaments can be used as an instance of social demand; and (c) how the comparison between the two may show (mis)alignments between societal concerns and scientific outputs.

Conference Topic
Science policy and research assessment

Introduction
Tackling complex global problems or grand challenges – such as climate change, food security, poverty reduction, risk of global pandemics – requires not only to increase R&D expenditure, but also the exploration and eventually the coordination of a variety of stakeholders with different areas of expertise and pursuing diverse research avenues. Typically these challenges benefit from the understanding of the physical and biological phenomena underlying a challenge (e.g. the virus and its genes), but also demand an understanding of the environmental and social contexts in which they occur, and the policy networks and instruments available in those contexts (Ely, Van Zwanenberg & Stirling, 2014).
Science policy funding schemes for societal problems or grand challenges seek to align science supply with social problems or needs. Although science is conducted in conditions of incomplete knowledge, it is well documented that certain particular research options are much

better aligned to certain outcomes (Sarewitz, 1996, pp. 31–49). It is, for example, very unlikely that astrophysics be useful for improving health care in malaria. Historically, several lines of inquiry in science policy have explored the alignment between research options and social outcomes, namely related to priority-setting and evaluation of research, but also to broader considerations related to the "supply" and "demand" of policy-relevant science. For this reason, a suitable interpretation of the alignment issue should be based on our understanding of the current state of the science (the *supply*) and what is required to achieve social goals (the *demand*) (Sarewitz & Pielke, 2007). The "demand" side must therefore consider not only the plurality of outcomes, but also various ways of articulating specific science or technology -driven pathways for achieving them. This in turn can refer to a process of public deliberation whereby different outcome preferences or divergent underlying values are made explicit by stakeholders. Similarly, the "supply" side is not just about how much "high-risk, high-return" research should be undertaken, but also about what type of outcomes are more or less *likely* to result from a given line of research. In this article, one the one hand, we apply the concept of *research landscape* (Wallace & Rafols, 2014) in order to map the scientific research on obesity.

On the other hand, we symmetrically map one of the interpretations (representations) of social needs (demand) on obesity. The supply-demand schema can be represented as in Figure 1. Here societal demand and scientific supply are not related directly in one single way. Instead they can relate via a variety of interpretation/representations of the "obesity" social needs. These representations shape science policy and affect actions that may reconcile supply and demand.

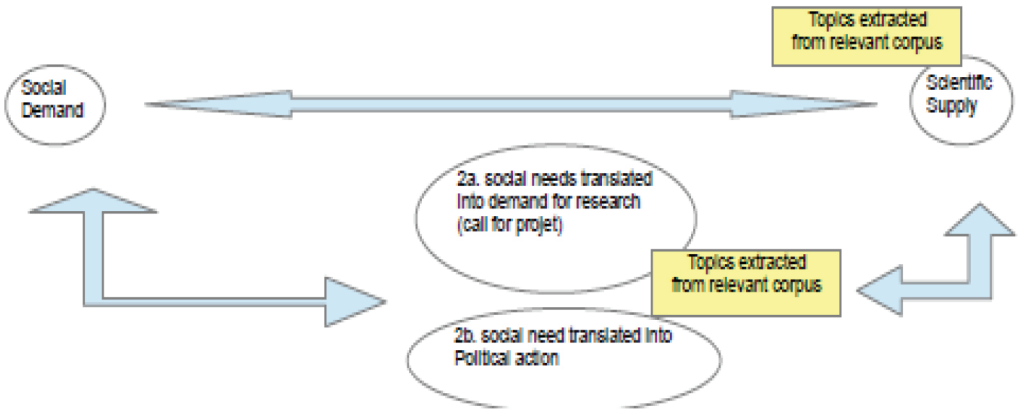


Figure 1. Social demand – scientific supply and political discourse as an example of intermediary representation.

In this paper we investigate the alignment (or lack thereof) between science supply and social demand by mapping, first, the scientific supply via the research landscape of obesity as defined by topic modelling of publication abstracts, and, second, social demand according to political discourse in the French parliaments. These maps of both supply and demand are specific and partial representations used in this preliminary and exploratory study -- other, complementary representations would be possible. For example, supply could be represented by a topic modelling of grants abstracts (as Talley et al., 2011 did for the US National Institutes of Health). And demand could be mapped using newspaper articles, among many other sources.

Data and Methods

Data

In order to define the relevant corpus for obesity, we follow a two-step method. First, we retrieve all publications with indexed MeSH term matching the search *obes** in MEDLINE/PubMed during the 2002-2013 period. This search was performed on October 16, 2014 and it returned 87,315 records.

Then, we launched *medlineR*, a routine based on the R language that allows the user to match data from Medline/PubMed with records indexed in the ISI Web of Science (WoS) database (Rotolo & Leydesdorff, 2015). The routine identified 71,055 WoS records (WoS core collections), with 'article' or 'review' as document types.

Second, we used Leiden's classification system to identify clusters of publications related to obesity. The classification system is constructed at the level of individual publications and clustering is based on direct citations (Waltman & van Eck, 2012) for the period 2000-2013. Obesity publications appear in 4,718 micro-clusters (in which at least one publication is tagged obesity), out of 32,466 micro-clusters for the whole WoS corpus. All the publications from clusters with at least 25% of publications tagged as 'obesity' were considered to be relevant for the study. This threshold of 25% is arbitrary and exploratory. Further explorations will use a lower threshold to test the robustness of this choice. The obesity corpus thus obtained contains 54,424 publications.

Topic modelling

Topic modelling provides a suite of algorithms to discover hidden thematic structure in large collections of texts. A topic model takes a collection of texts as input and it discovers a set of topics (recurring themes that are discussed in the collection) and the degree to which each document exhibits those topics.

Latent Dirichlet Allocation (LDA) is the simplest topic model. The intuition behind LDA is that documents exhibit multiple topics. LDA is a statistical model of document collections that tries to capture this intuition. It is most easily described by its generative process, the assumed random process. A topic is defined as a distribution over a pre-defined vocabulary. Moreover, it is assumed that the topics are specified before data have been generated (technically, the model assumes that the topics are generated first, before the documents). Now for each document in the collection, we generate the words in a two-stage process:

1. Randomly choose a distribution over topics.
2. For each word in the document
   (a) Randomly choose a topic from the distribution over topics in step #1.
   (b) Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the idea that each document contains multiple topics. Each document exhibits the topics with different proportion (step #1); each word in each document is drawn from one of the topics (step#2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).

The goal of topic modelling is to automatically identify the topics from a collection of documents. The documents themselves are observed, while the topic structure (the topics, per-document topic distributions and the per-document per-word topic assignments) is a hidden structure.

Results on Science Supply

For this study, we fitted a 100-topic model to the 54,424 publications of the obesity corpus. We perform LDA with the R package “topicmodels” and visualize the output using LDAvis.

Figure 2 shows a map of these 100 topics. Topics are located close to one another if they are similar in terms of distributions of the words belonging to the selected dictionary. The measure of topic similarity is the matrix of Jensen-Shannon divergences between topics, considered as distributions over words, into two-dimensional coordinates and is represented in a 2d space through multi-dimensional scaling (i.e., principal coordinates analysis).
In addition, a clustering technique is used to cluster topics into research areas. We applied k-means clustering to the topics as a function of their two-dimensional locations in the global topic view with k=10. Labels are assigned to clusters. These labels are obtained by extracting the most relevant terms for each cluster of topics, where the term distribution of a cluster of topics is defined as the weighted average of the term distributions of the individual topics in the cluster.

Figure 2. Map of topics of publications on obesity (2003-2013).

Results on the Societal Demand

The same approach has been used to map the social demand. In order to define one possible interpretation, we refer to the questions that the members of the French Parliament (i.e. Assemblée Nationale or Senate) can ask to the government. Deputies and senators publicly question the members of the Government in different ways. The question can be asked during a Parliament session to the government or be written and a Parliament session is not necessary and addressed to one of the ministers. We retrieved this information and built up two datasets.

First, we selected all the questions asked by the Senators where the word *obes** was reported in the public database - with records from 1985 on - which is now available. We got 242 questions from 1992 - year of the first occurrence of 'obesity' in these questions - to 2014. Second, we collected oral and written questions asked by members of the Assemblée Nationale in the last three legislatures, getting: 422 questions (2002-2007), 870 (2007-2012) and 380 (2012 – 2014). The output of the 10-topic model is shown below for the Senate questions.

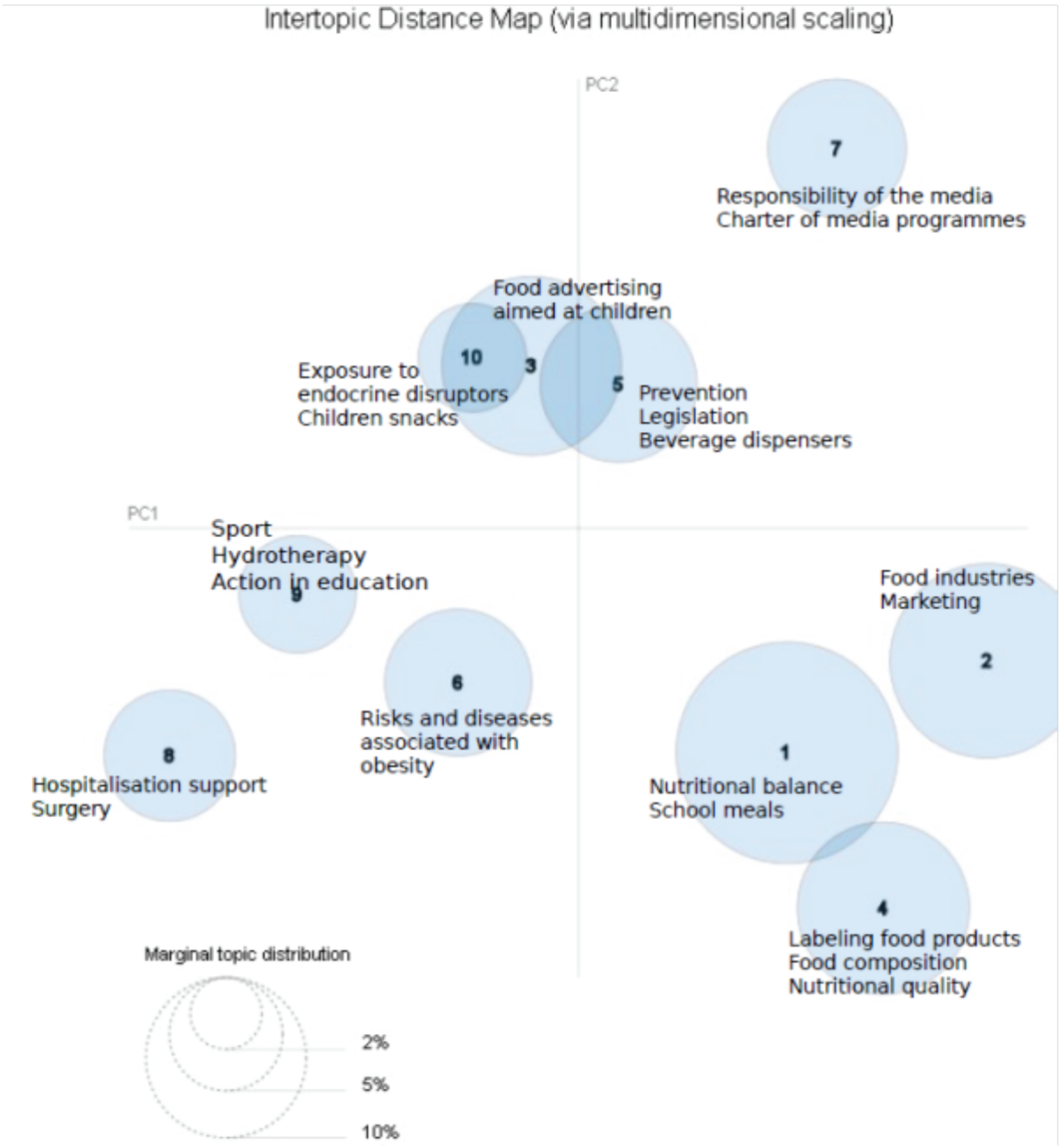


Figure 3. Map of topics for questions in the French Senate (1992-2014).

Discussion

In the centre of the Figure 2, we have a cluster of topics concerning *Physical activity, psychology and quality of life*, then turning around clockwise we find *Food consumption and diet* and then two clusters concerning mainly topics linked to biology research and further four clusters related to medical and surgery issues. The clusters of topics identified in the research landscape are mainly concerning medical and biological issues and only two clusters seem to deal with social and behavioural determinants of obesity, respectively *Obesity & ethnicity* and *Food consumption and diet*. The political discourse (Figure 3) seems to be organised around topics different from the research landscape. Among *the ten topics defined*

three main groups are reasonably identified. The first one, on the top part of the graph (i.e., topics number 3, 5, 7 and 10), is concerned mainly with children nutrition and the role of media as in advertising. A second group of topics, on the bottom right of the graph (i.e., topics number 1, 2 and 4), deals with food industry, marketing, and labelling issues. Finally, a third group, at the bottom left (i.e., topics number 6, 8 and 9) is concerned by medical and surgery issues. Only three out of ten topics of political discourse seem to find a counterpart in the research landscape. A preliminary analysis therefore suggests that, while research is concerned about the biophysical mechanisms that lead to obesity, many of the political questions are about the social mechanisms that favour obesity, such as advertisement, beverages, marketing, etc. This may suggest insufficient research regarding the social origin of obesity.

Acknowledgments

We would like to thank Tommaso Ciarli for suggesting to us the use of Parliament database as one of the possible representations of social needs. We thank Ludo Waltman for sharing the article level classification system.

References

Ely, A., Van Zwanenberg, P., & Stirling, A. (2014). Broadening out and opening up technology assessment: Approaches to enhance international development, co-ordination and democratisation. *Research Policy*, *43*(3), 505–518. doi:10.1016/j.respol.2013.09.004

Rotolo, D., & Leydesdorff, L. (2014). Matching MEDLINE/PubMed data with Web of Science (WoS): A routine in R language. *Journal of the Association for Information Science and Technology* (Forthcoming).

Sarewitz, D. (1996). *Frontiers of Illusion: Science, Technology and the Politics of Progress*. Philadelphia: Temple University Press.

Sarewitz, D., & Pielke, R. A. (2007). The neglected heart of science policy: reconciling supply of and demand for science. *Environmental Science & Policy*, *10*(1), 5–16.

Talley, E. M., Newman, D., Mimno, D., Herr, B. W., Wallach, H. M., Burns, G. A. P. C., & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods*, *8*(6), 443–444. doi:10.1038/nmeth.1619

Wallace, M. L., & Rafols, I. (2014). Research portfolios in science policy: moving from financial returns to societal benefits. *http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2500396*.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. doi:10.1002/asi.22748

Does Quantity Make a Difference?

Peter van den Besselaar [1] & Ulf Sandström [2]

[1] *p.a.a.vanden.besselaar@vu.nl*
VU University Amsterdam, Department of Organization Sciences & Network Institute

[2] *ulf.sandstrom@oru.se*
Royal Institute of Technology, INDEK & Orebro University, Business School, Orebro

Abstract
Do highly productive researchers have significantly higher probability to produce top cited papers? Or does the increased productivity in science only result in a sea of irrelevant papers as a perverse effect of competition and the increased use of indicators for research evaluation and accountability focus? We use a Swedish author disambiguated dataset consisting of 48,000 researchers and their WoS-listed publications during the period of 2008-2011 with citations until 2014 to investigate the relation between productivity and production of highly cited papers. As the analysis shows, quantity does make a difference.

Conference Topic
Indicators; Science policy; Research assessment

Introduction
One astonishing feature of the scientific enterprise is the role of a few extremely prolific researchers (Price, 1963). Thomson Reuters call them *Highly Cited Researchers* and they are listed and recognized per area. Based on another dataset, Scopus publications, Klavans & Boyack (2015) call them "superstars" and use them for large-scale studies of publication behaviour, thereby showing that superstars publishes less in isolated areas (retrieved using a clustering procedure), in dying areas, or in areas without an inherent dynamics. Highly productive and cited researchers tend to look for the new opportunities. Obviously, the highly productive researchers have to be taken into consideration for many reasons, both for science policy and for scholarly understanding of how the science system works.
Within bibliometrics there is a discussion on how to measure and to identify the superstars. Many current papers discuss the correlation between the various indicators developed for performance measurement. One of the stable outcomes is that there is a high correlation between the numbers of papers a researcher has published and the number of citations received (Bosquet & Combes, 2013). From that perspective, both indicators tend to measure the same attribute of researchers, as is actually materialized in the introduction of the H-index (Hirsch, 2005). Parallel, the discussion about impact has shifted from counting (field normalized) numbers of citations to more qualified types of citations and publications. As the progress of science rests on the huge amount of effort and publications, the number of real discoveries and path breaking new ideas is rather small. This has led to a different focus. Instead of counting publications and citations, the decisive difference is whether a researcher contributes to the small set of very highly cited papers. Different thresholds are deployed, from the top 1% or 10% of the highly cited papers or with the CCS method proposed by Schubert & Glänzel (1988). Only when reaching into these select set of papers that qualifies for citations above the x% level one can be considered as really having distinctive result that contributes to scientific progress. Increasingly, performance measures take this selectivity into account, and when calculating overall productivity and impact figures for researchers, papers (productivity) and citations (impact) are weighted differently depending on the impact percentile the paper belongs to (Sandström & Wold, 2015).
Of course, the question now comes up what a good publication strategy is – given this way of performance evaluation. Is publishing a lot the best way – or does that generally lead to normal

science, with low impact papers? The total number of citations received may still be large, but no top papers may have been produced. This is also the underlying idea of emerging movements in favour of 'slow science' like e.g., in the Netherlands; there the 'science in transition' movement (Dijstelbloem et al., 2014) was able to convince the minister of science and the big academic institutions to remove productivity as a criterion from the guidelines for the national research assessment (SEP). The underlying idea is that quality and not quantity should dominate – and that with all the emphasis on publications this has become corrupted.

However, others seem to see this differently. In his important work on scientific creativity, Simonton (2004) has extensively argued that (i) having a breakthrough idea is a low probability event that happens by chance, and therefore that (ii) the more often one tries, the higher the probability to have a 'hit' so now and then. There are also other contextual factors that may improve the chance for important results, but overall, the number of tries (publications) is the decisive variable. This also explains why Nobel laureates have so many more publications than normal researchers (Zuckerman, 1967; Sandström & Van den Besselaar, forthcoming). The more often you try (publish), the higher the probability that there is something very new and relevant, and atypical for the scientific community (Uzzi et al., 2013).

This brings us to the question whether there is a strong positive, or a negative relation between overall output (number of publications) and high impact papers. The answer of this question may inform our understanding of knowledge production and scientific creativity, but is also practically relevant for selection processes, and as explained above for research evaluation procedures: is high productivity a good thing, or a perverse effect and detrimental to the progress of science?

Methods and Data

In order to investigate this, we use the 74,000 WoS-publications 2008-2011 (with citations until 2014) of all researchers with a Swedish address using the following document types in databases SCI-E, SSCI and A&HCI: articles, letters, proceeding papers and reviews.

For identifying authors and keeping them separate we use a combination of automatic and manual *disambiguation* methods. An algorithm for disambiguating unique individuals was developed by Sandström & Sandström (2009), based on Soler (2007) and Gurney, Holdings & van den Besselaar (2012) and was found to proceed fast, although with minor manual cleaning methods. The deployed method takes into account surnames and first-name initials, the words that occur in article headings, and the journals, addresses, references and journal categories used by each researcher. There is also weighting for the normal publication frequency of the various fields.

As indicated, the data covers 74,000 articles and 195,000 author shares that have been judged to belong to Swedish universities or other Swedish organisations. In a few cases, articles from people who have worked both in Sweden and in one or more Nordic countries have been kept together, and articles have thus been included even if they came into being outside Sweden (the process of distinguishing names is thus carried out at Nordic level).

All articles by each researcher are ranked, based on received citations and according to the about 260 subject categories as specified in the Web of Science, and the articles are divided into CSS (Characteristic Scores and Scales) classes (0, 1, 2, 3). While measures based on percentile groups (e.g. top1% etc.) are arbitrarily constructed, CSS have some advantages concerning the identification of outstanding citation rates (Glänzel & Schubert, 1988). The CSS method is a procedure for truncating a sample (e.g., a subfield citation distribution) according to mean values from the low-end up to the high-end. Every group created using this procedure helps to identify papers that fulfil the requirements for being cited above the respective thresholds. In this paper we will use two levels, level CSS1 and CSS3, which in the

former case cover the 20%-25% most cited papers, and in the latter case the about 2%-3% of most cited papers: the "outstandingly cited papers" (Glänzel, 2011).
In this paper we will investigate the relation between quality and quantity in several different ways. We proceed in this way, as from a methodological perspective different options are open, without a convincing argument which one would be the better. By using a variety of methods, we avoid to produce results as artefacts of the method deployed.
(i) Firstly, we calculate the probability to have one, two or three and more top cited papers, given the productivity level. We calculate this for the health, i.e. medical sciences (about 15,000 researchers), where we classify these authors in several productivity classes. Class 1 has one publication in the four years period under study, class 2 has two, class 3 has three to four, class 4 has five to eight, class 5 has nine to sixteen, class 6 has seventeen to 31 publications, and finally class 7 covers researchers with 32 or more publications. Publications are integer counted, but citations are field normalized.
(ii) Secondly, we do a simple regression with the total number of (integer counted, IC) publications as the independent variable, and the (also integer counted) number of top cited publications in terms of one of the definitions as discussed above. Also, here citations are field normalized. We have here all researchers, without normalizing for field based productivity figures. As the total set of researchers is dominated by life and medical sciences and by natural sciences, and as these groups have comparable average publications and citations, we assume that this does not really influence the results. Under point four below, we introduce a way of taking field differences in productivity into account.
(iii) Thirdly, we do the same analysis as described above, but use fractional instead of full counting. This helps to investigate the effect of different ways of counting on the relations under study.
(iv) Fourthly, we move to the field-normalized (fractional counted) productivity, and calculate the relation between in this way defined productivity and having at least one publication in CSS1 respectively in CSS3. In the last analysis, we can provide an integrated analysis of all researchers across all fields, as we produced field normalized output counts. This is done with a method – Field Adjusted Production (FAP) based on Waring estimations – as initially developed by Glänzel and his colleagues (Braun, Glänzel & Schubert, 1990; Koski, Sandström & Sandström, 2011) during the 1980s. FAP is further explained and tested in Sandström & Sandström (2009). Basically, the method is used in order to compensate for differences between research areas concerning the normal rate of scholarly production. For this all journals in the Web of Science have been classified according to five categories (applied sciences, natural sciences, health sciences, economic & social sciences, and arts & humanities). Categorisation of journals into macro fields is based on Science Metrix classification of research into five major domains. Note that in some of the following analysis we will refrain from applying the Waring method, consequently, instead the analysis will be performed per scientific macro fields (for further information, see < http://science-metrix.com/en/classification>).

Results

(i) Does the probability of highly cited papers increase with productivity?

We calculated the number of top cited papers (CSS3) for each of the seven productivity classes. From this, Figure 1 was created. Clearly, the probability increases with productivity, and this is the case for 1, 2 and 3 or more papers in the CSS3 class. In fact, the relation is slightly different for the three criteria. The higher the criterion, the larger the effect is at the high end of the productivity distribution.

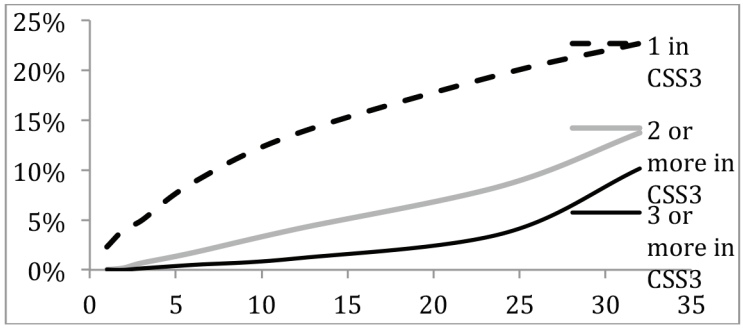


Figure 1. Share of papers in the CSS3 top cited class by productivity class.

(ii) What is the effect of productivity on the number of highly cited papers?

We have done a regression analysis with highly cited papers as dependent variable, and productivity as independent variable. We did the analysis for the various top cited classes. In the three figures below, we show the regression results. For papers in the top 1% of the cited papers (Figure 1) the correlation is about 0.5. For the CSS3, the top 10% of the cited papers, and the CSS1 classes, the correlations are 0.58, 0.78 and 0.88. The correlations are fairly high.

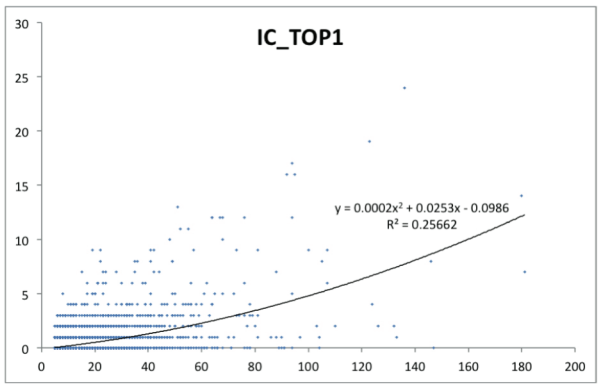


Figure 2. Top 1% of cited papers by total number of papers.

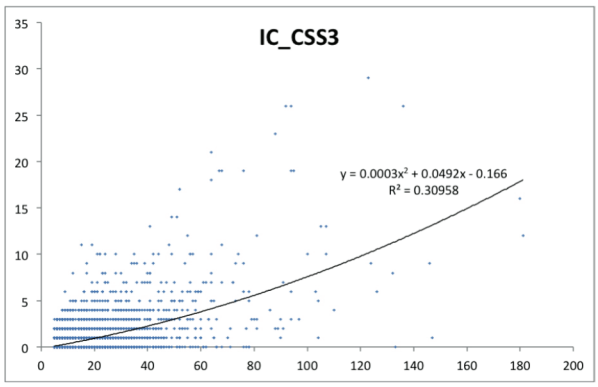


Figure 3. CSS3 cited papers by total number of papers.

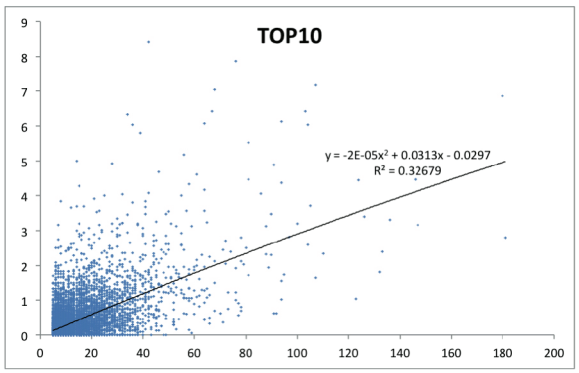


Figure 4. Top 10% of cited papers by total number of papers.

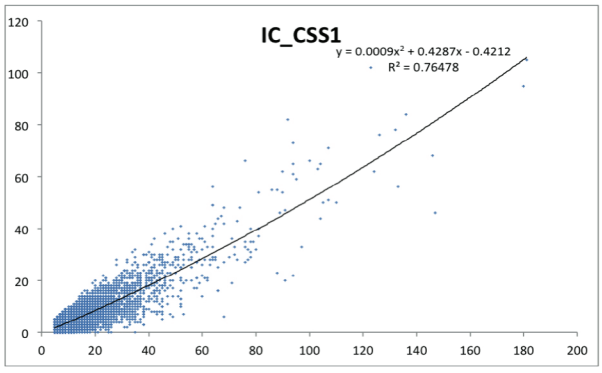


Figure 5. CSS1 cited papers by total number of papers.

Interestingly, the correlation becomes higher the lower the citation threshold. Why this is the case is not yet investigated. A possibility is that high productive researchers with top papers always have co-authors of these high cited papers who themselves are not highly productive. In that sense one also expects top cited authors in the lower productivity segments, reducing the explained variance. So probably, one should only include PIs in the analysis to avoid this effect. This could be the topic for a subsequent study.

One should realize that a small share of all authors produces most of the papers and of the highly cited papers. The 6.3% of most productive researchers (everybody above eleven publications in four years) are responsible for 37% of all papers and for 53% of the top 1% of the cited papers. Also this supports the idea that quantity makes a difference.

(iii) And the effect of fractional counted productivity on the number of highly cited papers?

We did the above analysis also using fractional counting of productivity. The patterns are the same, but the correlations are about .15 to .20 lower than in the full counted model. How this can be explained will be addressed in a coming paper. But also here, the 6.3% of the most productive authors are decisive: they have 46.8% of the fractional counted top 1% of the cited papers.

(iv) What is the effect of field adjusted production counting?

The relation between having at least one paper in CSS1 and total field normalized output is plotted in Figure 6, and as becomes obvious, the correlation is fairly high (r = 0.79), and not much smaller than in the above four where we did not use the field adjusted production (0.90, see Figure 5). The results here suggest that indeed the more papers someone publishes, the higher the probability of having a paper in the group of fairly good papers cited above the threshold of CSS1.

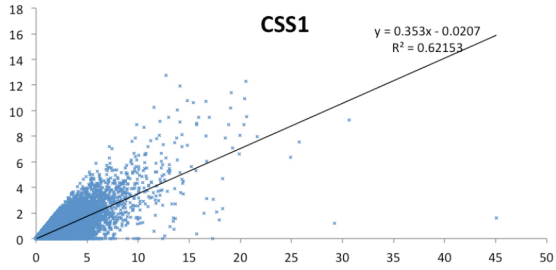


Figure 6. Fractionalized CSS1 by field adjusted production (all areas of science).

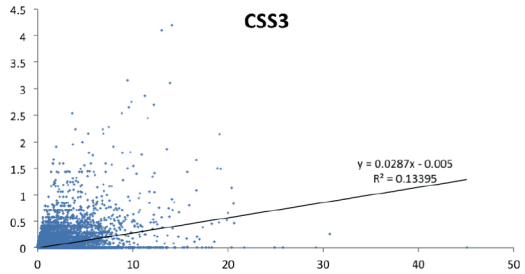


Figure 7. Fractionalized CSS3 by field adjusted production (all areas of science).

We also plot the relation between having at least one paper in the CCS3 (Figure 7), so in a much more narrow defined top, and field-normalized productivity, and although correlation is lower here, it is still considerable (r = 0.37). However, in the CSS3 case, the correlation when applying FAP is lower than the correlation without applying FAP (Figure 3), namely is 0.58. These differences need some further exploration.

The underlying distribution for the fields of Natural sciences and Medical and Life sciences are given in Table 1, which shows for seven distinct productivity categories the percentage of Swedish researchers in that category, the average number of papers published in a four-year period, the average fraction of paper production, and of course the percentage of researchers with at least one paper in CCS3.

As 'field adjusted' production (FAP) might be a rather abstract concept, we have translated it below for the various disciplines into 'normal papers'. So, what is the relation between the number of papers produced (in a period of four years) and the probability of having a 'top cited paper' (in the top 2%-3% cited papers CSS3 class) during the period 2008-2014? This is a more sophisticated version of the analysis presented in section (i) above. As we clearly see in Table 2, the higher the number of papers, the more likely that one has a paper that ends up to be an outstandingly cited paper. Actually, the increase is rather steep and one may say that in most disciplines only with some ten papers in the period under consideration, there is a good chance of having a top paper. The humanities have a different pattern, as with a production of five papers one has the highest chance of reaching the top.

Table 1. CSS3 papers by production levels, Health sciences and Natural sciences

	Medical and life sciences				Natural sciences			
Category	researchers	Mean P	Frac P	CSS3	researchers	Mean P	Frac P	CSS3
1 (1 paper)	40.8%	1	0.2	0.03	9,0%	1	0.2	0.02
2 (2 papers)	16.92%	2	0.4	0.06	16,3%	2	0.5	0.05
3 (>2-4)	17.08%	3.4	0.7	0.10	17,4%	3.4	0.9	0.10
4 (>4-8)	13.36%	6.1	1.3	0.21	13,7%	6.1	1.6	0.21
5 (>8-16)	7.23%	11.6	2.4	0.44	8,3%	11.5	2.8	0.40
6 (>16-32)	3.36%	22.3	4.4	1.05	4,1%	22.0	4.7	0.87
7 (>32)	1.18%	50.5	8.8	3.45	1,2%	47.6	9.8	2.68
Average		4.3	0.9	0.17		4.6	1.1	0.17

Data for this table is built on publications from 37,114 researchers.

Table 2: Probability of one outstanding paper (CSS3) at different levels of production.

Average # of		Discipline				
publications	Class	Natural	Health	Applied	Ec &Soc	Hum
1	1	5%	7%	7%	6%	9%
2	2	11%	13%	13%	13%	8%
3	3	20%	21%	21%	24%	25%
6	4	31%	34%	33%	34%	33%
11	5	49%	54%	53%	55%	33%
20	6 / 7	61%	80%	66%	83%	
38	7			88%		
46	7	83%				
49	7		93%			

Note: Data for this table consist of ≈190,000 article shares with <40 authors per paper. The numbers of publication are the field-specific averages per productivity class (for more information, see Table 1).

Conclusions

As the above results show, there is not only a strong correlation between productivity (number of papers) and impact (number of citations), that also holds for the production of high impact papers: the more papers, the more high impact papers. In that sense, increased productivity of the research system is not a perverse effect of output oriented evaluation systems, but a positive development, as it strongly increases the occurrence of breakthroughs and important inventions (c.f. Uzzi et al., 2013). The currently upcoming discussion that we are confusing quality with quantity therefore lacks empirical support. As we deployed a series of methods, with results all pointing in the same direction, the findings are not an artefact of the selected method.

The analysis also gives an indication of the output levels that one may strive at when selecting researchers for grants or jobs.

We also plan some future work: Firstly, we plan to extend the analysis to some other countries, which of course requires large-scale disambiguation of author names. Secondly, we will in a next version control for number of co-authors, and for gender. The former relates to the discussion about team size and excellence, the latter to the ongoing debate on gender bias and gendered differences in productivity. Thirdly, the aim is to concentrate on principle investigators, and remove the incidental co-authors with low numbers of publications, as they may seem to be high impact authors at the lower side of the performance distribution. This all should lead to a better insight in the relation between productivity and impact in the science system.

References

Bosquet, C. & Combes, P-P. (2013). Are academics who publish more also more cited? Individual determinants of publication and citation records. *Scientometrics*, 97: 831-857.

Braun, T., Glänzel, W. & Schubert, A. (1990). Publication productivity: from frequency distributions to scientometric indicators. *Journal of Information Science,* 16: 37-44.

Dijstelbloem, H., Huisman, F., Miedema, F. & Mijnhardt, W. (2014). Science in Transition Status Report: Debate, Progress and Recommendations. http://www.scienceintransition.nl/wp-content/uploads/2014/07/Science-in-Transition-Status-Report-June-2014.pdf.

Glänzel, W. (2011). The application of characteristic scores and scales to the evaluation and ranking of scientific journal. *Journal of Information Science,* 37(1): 40-48.

Glänzel, W. & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. *Journal of Information Science,* 14: 123-127.

Gurney, T., Horlings, E. & van den Besselaar, P. Author disambiguation using multi-aspect similarity indicators. *Scientometrics,* 91: 435-449.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. PNAS, 102(46): 16569-16572.

Klavans, R. & Boyack, R.W. (2015). Scientific superstars and their effect on the evolution of science. http://www.enid-europe.org/conference/abstract%20pdf/Klavans_Boyack_superstars.pdf.

Koski, T., Sandström, E. & Sandström, U. (2011). Estimating research productivity from a zero-truncated distribution. Paper to the 2011 ISSI Conference in Durban.

Price, D.J.S. (1963). *Little Science, Big Science*. New York: Columbia University Press.

Sandström, U. & Sandström, E. (2009). The field factor: towards a metric for academic institutions. *Research Evaluation*, 18(3): 243–250

Sandström, U. & van den Besselaar, P. (2015). *Before the prize: Nobel Prize laureates recognition by their scientific community*. Manuscript in preparation.

Sandström, U. & Wold, A. (2015). Centres of Excellence: reward for gender or top-level research? In B. Bjorkman & B. Fjaestad (Eds.). *Thinking Ahead: Research, Funding and the Future* (pp. 69-91). Stockholm, Makadam Publ.

Simonton, D.K. (2004). *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. New York: Cambridge Univ Press. [Reprinted 2008].

Soler, J-M. (2007). Separating the articles of authors with the same name. *Scientometrics* 72 (2): 281–290. DOI: 10.1007/s11192-007-1730-z.

Uzzi, B., Mukherjee, S., Stringer, M. & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342, 468-472.

Zuckerman, H. (1967). Nobel laureates in science: Patterns of productivity, collaboration, and authorship. *American Sociological Review*, 32 (3): 391-403.

On Decreasing Returns to Scale in Research Funding

Philippe Mongeon[1], Christine Brodeur[1], Catherine Beaudry[2] and Vincent Larivière[3]

[1] *philippe.mongeon@umontreal.ca;christine.brodeur@umontreal.ca;*
Université de Montréal, École de bibliothéconomie et des sciences de l'information,
C.P. 6128, Succ. Centre-Ville, Montréal, QC. H3C 3J7 Canada

[2]*catherine.beaudry@polymtl.ca*
École Polytechnique de Montréal, Département de mathématiques et de génie industriel,
C.P. 6079, Succ. Centre-Ville, Montréal, QC. H3C 3A7 Canada
Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), CP 8888, Succ. Centre-Ville, H3C 3P8 Montreal, Qc. (Canada)

[3] *vincent.lariviere@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, H3C 3J7 Montreal, Qc. (Canada) and
Université du Québec à Montréal, Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Observatoire des Sciences et des Technologies (OST), CP 8888, Succ. Centre-Ville, H3C 3P8 Montreal, Qc. (Canada)

Abstract
In most countries, basic research is supported through governmental research councils that select, after peer review, the individuals or teams what will receive funding. Unfortunately, the number of grants these research councils can allocate is not infinite, and many researchers (45% in Quebec) are not able to obtain any funding. A small minority of those who do get funded account for the majority of the available funds. However, it is unknown whether or not this is an optimal way of distributing available funds. The purpose of this study is to measure the relation between the amount of funds given to 14,103 individual Quebec's researchers over a fifteen year period (1998-2012) and the total outcome of their research in terms of output and impact from 2000 to 2012. Our results show that both in terms of the quantity of papers produced and of their scientific impact, the concentration of research funding in the hands of a so-called 'elite' of researchers generally produces diminishing returns.

Conference Topic
Science policy and research assessment

Introduction
In most countries, basic research is supported through governmental research councils that select, after peer review, the individuals or teams that will receive funding. Unfortunately, the number of grants these research councils can allocate is not infinite. For example 20% to 45% of Quebec's researchers, depending on the discipline, had no external funding between 1999 and 2006 (Larivière et al., 2010). National scientific agencies, including the National Science Foundation (NSF – United States) and Natural Science and Engineering Research Council (NSERC – Canada), also tend to give fewer grants of a higher value, which leads to high rejection rates (Joós, 2012; NSERC, 2012; NSF, 2013). In Canada, 10% of the researchers funded by the Social Sciences and Humanities Research Council (SSHRC) accumulate 80% of available funds, 10% of those funded by the Canadian Institutes of Health Research (CIHR) obtain 50% of the funds, and 10% of those funded by the NSERC accumulate 57% of the funds.[1] The situation is similar in Quebec where we combine funding from the national

[1] Data compiled by the Observatoire des Sciences et Technologies (OST) using results of competition for each of the councils, and the *Almanac of Post-Secondary Education in Canada, of the Canadian Association of University Teachers*.

and provincial agencies: 20% of the researchers getting 80% of the funds in social sciences and humanities (SSH), 50% of the funds in health, and 57% of the funds in natural sciences and engineering (NSE) (Larivière et al., 2010). With a few researchers receiving most of the funds available and many not receiving any, it seems legitimate to ask whether this concentration of funds leads to better collective gains than funding policies that promote a more even distribution of funding. The aim of this study is to provide a partial answer to this question, by linking the amount of funding obtained by Quebec's scientists with their research productivity and impact.

Even though the funding of science theoretically plays a substantial role in scientific discoveries, its relation to outcomes has not been extensively researched. McAllister and Wagner (1981) observed a linear relationship between funding and output at the institution level. A few years later, Moed et al. (1998) found that departments of Flemish universities with the most funding actually had a decrease in publications. Other studies (e.g., Heale et al., 2004 and Nag et al., 2013) investigated the relation between the amount of funding and the research output of individual researchers. They reported that one of the strongest determinants of the number of publications was the amount of funding, although an increase in funds did not yield a proportional increase in the number of articles. Thus, there are decreasing returns to scale. Others have found that productivity is only weakly related to funding (Fortin & Currie, 2013), and that publications do not increase linearly with the amount of funding but rather appears to reach a plateau (Berg 2010). On the whole, while most studies—unsurprisingly—found a positive relationship between inputs and outputs, very few have looked at decreasing returns to scale associated with the concentration of research funding. Nicholson and Ioannidis (2012) found that only a minority (about 40%) of all researchers eligible to NIH funding who published highly cited articles (1000 citations or more) actually received such funding. Previous studies found that funded researchers publish more (Gulbrandsen & Smeby, 2005) and are more cited (Zhao, 2010; Jowkar, 2011; Campbell et al., 2010; van Leeuwen et al., 2012) than those who do not receive any funding.

This study aims to contribute to this debate, by analyzing the research output and impact of all of Quebec's researchers from all disciplines over a period of 15 year. More specifically, it aims at answering two questions: 1) how does the research productivity and scientific impact of individual researchers vary with the amount of funding they receive? 2) Is this variation similar in the three general fields of science that are health, natural sciences and engineering, and social science and humanities?

Methods

Data on funding for all Quebec's academic researchers from 1998 to 2012 were obtained from the Information System on University Research, an administrative database from the Quebec provincial government that covers all funded research in Quebec's universities. Researchers were divided in three broad research disciplines: Social Sciences and Humanities (SSH), Natural Sciences and Engineering (NSE) and Health according to the discipline of their university department. Some were put in two different disciplines (N=169), and those for whom the discipline was not known and not found were excluded (N=263). The number of researchers in each field is shown in table 1. For each researcher, we calculated the total amount of funding received from the three main funding agencies in Quebec (FRQSC [SSH], FRQNT [NSE] and FRQS [health]) and Canada (SSHRC [SSH], NSERC [NSE] and CIHR [Health]). The total funds attributed for each projects were divided equally by the number of researchers on the application, each of them receiving an equal share. Other sources of funding were not taken into account. Publication data for each researcher from 2000 to 2012 were obtained from Thomson Reuters' Web of Science. Since citations take time to accumulate, they were counted up to the end of 2013.

Table 1. Number of Quebec's researchers by field

Field	*Number of researchers*	*Funded*		*Not funded*	
		N	*%*	*N*	*%*
SSH	6,229	3,869	62.1%	2,360	37.9%
NSE	3,244	2,647	81.6%	597	18.4%
Health	4,630	2,666	57.6%	1,964	42.4%
Total	14,103	9,182	65.1%	4,921	34.9%

Similarly to Berg (2010), we divided researchers in bins of equal size (50 researchers per bin), except for the bin regrouping researchers who did not receive any funding (see table 1 for the number of researcher in each field who did not receive funding). For each bin, we calculated the average and median amount of funding received. Then we calculated the average and median of four indicators used to measure the research outcome: the total count of articles, the fractional number of articles, the total number of citations and the average relative citations (ARC).

Results

Figure 1 and Figure 2 provide the mean and median number of papers of researchers, using both full (Figure 1) and fractional counting (Figure 2), as a function of total funding received. For each bin for each discipline and each indicator, the average is higher than the median, implying a skewed distribution of the data. The high values of R^2 in both figures indicate that the number of publications is strongly linked to the amount of funding received by researchers. The best fit line for each domain is a quadratic equation which suggests diminishing returns. For example, the median number of publications of researchers in NSE who received about \$5 million is about 72 (and 19 for fractional count), while those who receive \$2.5 million published a median number of 47 articles (13 for fractional count). Thus, doubling the funding does not seem to double the output. In Health, the most funded bin received almost three times more funding than the second most funded one, but published only two times more articles. Furthermore, in health, the apex is reached within the data range, which shows that a decline in production could be associated with higher levels of funding. On the whole, the correlation between funding and publications appears to be strong in all fields with values of R^2 higher than 0.91, but for each domain and calculation method, a rapid growth in the number of publications is observed for smaller amounts received and is followed by a slower growth as funding increases. However, this effect is less apparent for the total number of publications in SSH.

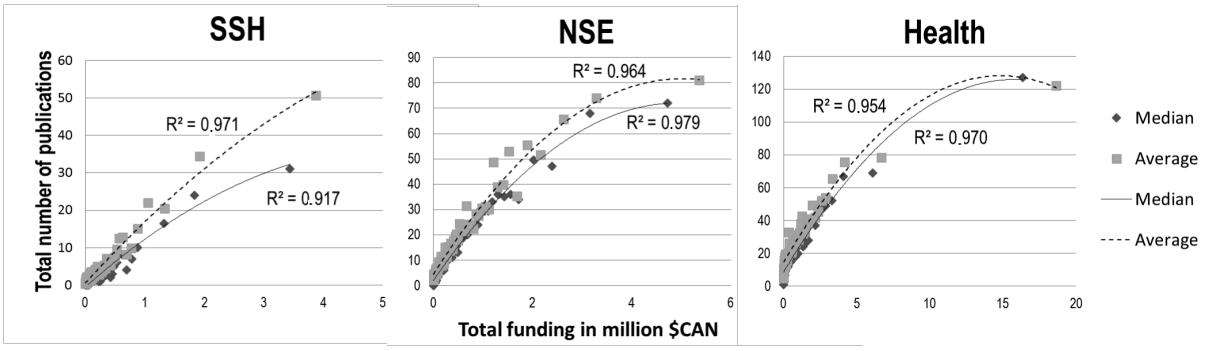


Figure 1. Full number of publications as a function of the amount of funding received.

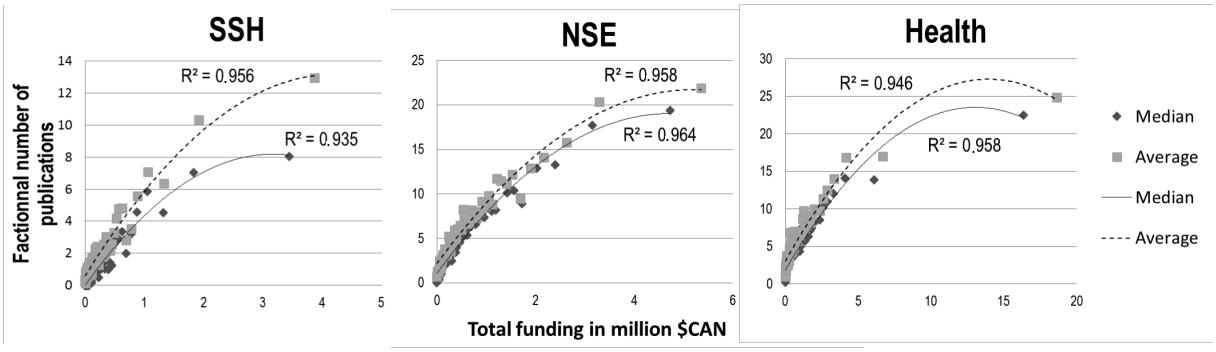


Figure 2. Fractional number of publications as a function of the amount of funding received.

Figure 3 shows the relationship between raw citations and funding received; the best-fit line is also a quadratic equation suggesting decreasing returns to scale in scientific impact. Similar to publications, the relation between the average of relative citations and the amount of funding (Figure 4) is weaker than for the previous indicators, with R^2 between 0.4 and 0.9. The nature of the relation is also different, the best-fit line being a power function, except for the median in SSH and the average in NSE, which are quadratic function. The power function indicates decreasing returns: the average relative of citations keeps increasing when increasing the total of funding, but not proportionally. For both impact indicators, we observe a trend similar to that observed for the number of publications. While the impact of papers published increase rapidly for funding of less than approximately $2 million in NSE and $5 million in health, the total number citations increase at a much slower pace once this threshold is met. Here, SSH are the exception, with the total number of citation seemingly increasing more rapidly for highly funded researchers. For field-normalized citations, the impact remains almost the same for all fields after a threshold of approximately $1 million is met.

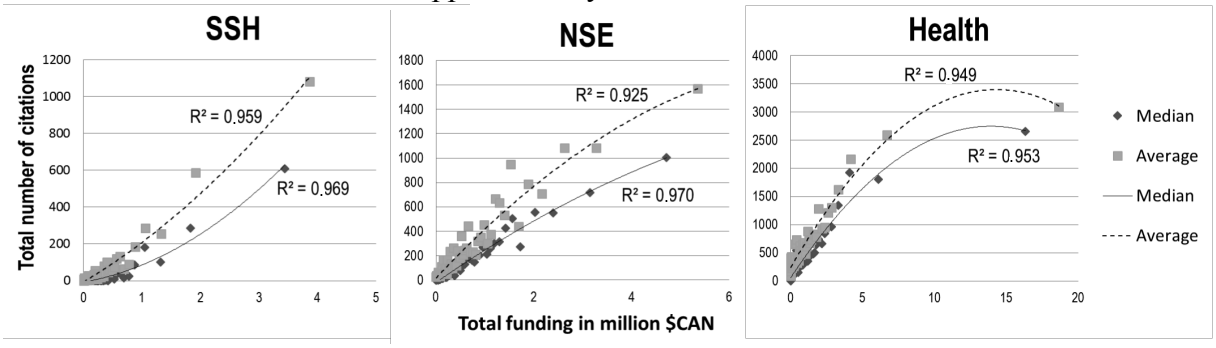


Figure 3. Total number of citations as a function of the amount of funding received.

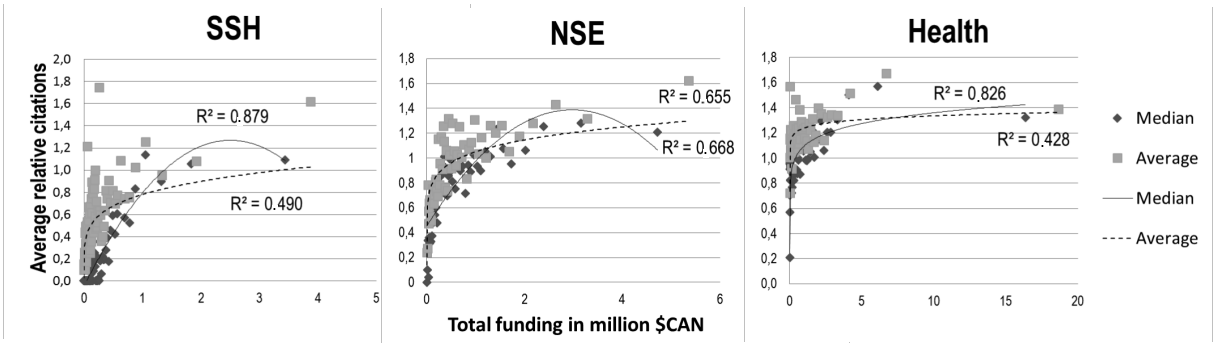


Figure 4. Average relative citations as a function of the amount of funding received

Discussion and conclusion

Based on our observations, funding is strongly linked to productivity and impact of individual researchers, but there are decreasing returns to scale for all of the indicators measured, except for the total citation count in SSH. This suggests that, even though more funding does in general lead to a higher number of publications, giving bigger grants to fewer individuals may not be optimal. If maximum output is the objective, then giving smaller grants to more researchers seems to be a better policy. In terms of scientific impact, the quickly reached plateau indicates that increasing funding has a very small impact on relative citations. Again, if the goal of research funding is to generate research that has a greater impact, giving grants to more researchers seems to be a better decision.

According to our results, SSH seem to be an exception, showing very little decreasing returns to scale. However, this could be explained by the fact that some research specialties in SSH (e.g., psychology and geography) have publication practices that are similar to those in NSE or Health. A closer look at the data shows that some researchers in psychology and geography tend to be both more funded – since they are often funded by the health and natural sciences funding agencies respectively – and more prolific than those in other field. Twenty-three (23) of the 50 most funded researchers and 33 of the 50 most prolific researchers are in those two fields, while they were 10 out of 50 in a randomly selected bin of researchers with less funding. Thus, the lower decrease in return of research funding in SSH could potentially be explained by an overrepresentation psychology and geography researchers in the highly funded bins, and their underrepresentation in less funded ones.

One of the many potential explanations for these decreasing returns is the high cost of equipment and infrastructures. Some research projects may simply not be possible without these initial investments, which do not necessarily lead to more output. Furthermore, while receiving funding does provide researchers with the means to carry on their research projects, it does not guarantee that they will succeed at achieving publishable results. Research grants are sometimes used as a performance indicator, which encourages researchers to apply for more grants (Hornbostel, 2001) that they might not necessarily need. This could lead to an inefficient use of the funds received (Sousa, 2008). Another explanation could be that researchers receiving larger grants may not participate directly on all the work funded with those grants (Boyack & Jordan 2011)

Some limitations of this study should be acknowledged. We did not control for other factors that can have an impact on a researcher's productivity (e.g., team size, academic age or gender), so further research may want to take into account such factors, as well as sources of funding other than government grants. Also, some of the potential outcome of funding and research cannot be measured with bibliometric indicators (e.g., the number of students trained and social outcomes). The funding received is sometimes linked to a particular project, and further research could aim at comparing outcomes of funded projects specifically. Another limit might be the lower coverage of SSH publications in the Web of Science, since researchers in SSH tend to publish in local journals or to publish books. Finally, as discussed above, the division of researchers in three broad disciplines might be problematic, especially for SSH. A more precise clustering of researchers based on research topic could provide better results and a clearer understanding of the phenomenon of decreasing returns of research funding.

In sum, both in terms of the quantity of papers produced and of their scientific impact, the concentration of research funding in the hands of a so-called 'elite' of researchers generally produces diminishing returns. In a context where financial resources devoted to research are declining in constant dollars, it is important to ask whether the way funding is allocated is optimal. Our numbers show that it is not the case: a more egalitarian distribution of funds would yield greater collective gains. It should be understood that the main determinant of

scientific production is not so much the money invested, but, rather the number of researchers' at work and, by funding a greater number of researchers, we increase the overall research productivity. Research policies that concentrate financial resources also seem to forget that there is a certain degree of serendipity associated with scientific discoveries, and by funding the work of many researchers as possible, we increase the likelihood that some of them make major discoveries.

Acknowledgments

This work was funded by the Canada Research Chairs program, and by the Social Sciences and Humanities Research Council of Canada.

References

Berg, J. (2010). Measuring the scientific output and impact of NIGMS grants. *NIGMS Feedback Loop Blog*. Retrieved June 15, 2015 from:http://loop.nigms.nih.gov/2010/09/measuring-the-scientific-output-and-impact-of-nigms-grants/

Boyack, K. W, & Jordan, P. (2011). Metrics Associated with NIH Funding: A High-Level View. *Journal of the American Medical Informatics Association*, 18(4), 423–431.

Campbell, D., Picard-Aitken, M., Côté, G., Caruso, J., Valentim, R., Edmonds, S., & Archambault, E. (2010). Bibliometrics as a performance measurement tool for research evaluation: the case of research funded by the National Cancer Institute of Canada. *American Journal of Evaluation*, 31(1), 66-83.

Fortin, J. M. & Currie, D. J. (2013). Big science vs. little science: how scientific impact scales with funding. *PLoS ONE*, 8(6), e65263.

Gulbrandsen, M. & Smeby, J.C. (2005).Industry funding and university professors' research performance. *Research Policy*, 34(6), 932-950.

Heale, J.P., Shapiro, D. & Egri, C.P. (2004). The determinants of research output in academic biomedical laboratories. *International Journal of Biotechnology*, 6(2-3), 134-154.

Joós, B. (2012). NSERC's discovery grant program: disquieting changes & why they matter to Canadian science. *CAUT Bulletin*, 59(1).

Jowkar, A., Didegah, F. & Gazni, A. (2011). The effect of funding on academic research impact: a case study of Iranian publications. *Aslib Proceedings*, 63(6), 593-602.

Larivière, V., Macaluso, B., Archambault, É. & Gingras, Y. (2010). Which scientific elites? On the concentration of research funds, publications and citations. *Research Evaluation*, 19(1), 45-53.

McAllister, P. R. & Wagner, D. A. (1981). Relationship between r-and-d expenditures and publication output for United-States colleges and universities. *Research in Higher Education*, 15(1), 3-30.

Moed, H. F., Luwel, M., Houben, J. A., Spruyt, E. & Van Den Berghe, H. (1998). The effects of changes in the funding structure of the Flemish universities on their research capacity, productivity and impact during the 1980's and early 1990's. *Scientometrics*, 43(2), 231-255.

Nag, S., Yang, H., Buccola, S. & Ervin, D. (2013). Productivity and financial support in academic bioscience. *Applied Economics*, 45(19), 2817-2826.

Nicholson, J. M. & Ioannidis, J. P. A. (2012). Research grants: Conform and be funded. *Nature*, 492(7427), 34-36.

NSERC. (2012). *2012 Competition Statistics Discovery Grants Program*. Retrieved June 15, 2015 from: http://www.nserc-crsng.gc.ca/_doc/Funding-Financement/DGStat2012-SDStat2012_eng.pdf.

NSF. (2013). *Summary Proposal and Award Information (Funding Rate) by State and Organization*. Retrieved June 15, 2015 from: http://dellweb.bfa.nsf.gov/awdfr3/default.asp.

Sousa, R. (2008). Research funding: less should be more. *Science*, 322(5906), 1324-1325.

van Leeuwen, T.N. & Moed, H.F. (2012). Funding decisions, peer review, and scientific excellence in physical sciences, chemistry, and geosciences. *Research Evaluation*, 21(3), 189-198.

Zhao, D. Z. (2010). Characteristics and impact of grant-funded research: a case study of the library and information science field. *Scientometrics*, 84(2), 293-306.

How Many is too Many?
On the Relationship between Output and Impact in Research

Vincent Larivière[1] and Rodrigo Costas[2]

[1] *vincent.lariviere@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, H3C 3J7 Montréal, Qc. (Canada) and Université du Québec à Montréal, Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Observatoire des Sciences et des Technologies (OST), C.P. 8888, Succ. Centre-Ville, H3C 3P8 Montreal, Qc. (Canada)

[2] *rcostas@cwts.leidenuniv.nl*
Leiden University, Center for Science and Technology Studies, Wassenaarseweg 62A, 2333 AL Leiden (The Netherlands)

Abstract
Over the last few decades, the massification of quantitative evaluations of science and their institutionalisation in several countries has led many researchers to aim at publishing as much as possible. This paper assesses the potential adverse effects of this behaviour by analysing the relationship between individual researchers' productivity and their proportion of highly cited papers. In other words, does the share of an author's top 1% most cited papers increase, decrease or remain stable, as her number of total papers increase? Using a large dataset of disambiguated researchers (N= 25,994,021) over the 1980-2012 period, this paper shows that the higher the number of papers a researcher publishes, the more likely they are amongst the most cited in their domain. This relationship was stronger for older cohorts of researchers, while decreasing returns to scale were observed in some domains for more recent cohorts. On the whole, these results suggest that at the macro-level, the culture of publishing as many papers as possible did not yield to adverse effects in terms of impact, especially for older researchers. For such researchers, who have had a long period of time to accumulate scientific capital, there can never be too many papers.

Conference Topic
Science Policy and Research Assessment

Introduction
In the second half of the 20th Century, but even more so over the last few decades, evaluations have become widespread in various spheres of society (Dalher-Larsen, 2011). Although scientific research has long been exempt from external evaluations thanks to Vannevar Bush and post WWII non-interventionist science policy, it has always been assessed internally through peer review. These means of evaluating research and researchers have, however, slowly changed since the 1980s, when researchers and administrators became aware of the roles that bibliometric analyses could play in such evaluations. Quantitative publication and citation analyses gained even more importance in the 2000s (Cameron, 2005), when tools for assessing individual researchers' output and impact became widespread. While in some cases, these methods have been developed to complement peer review in the allocation of research funding—such as the BOF-key in Flanders (Belgium) (Debackere & Glänzel, 2004), the Research Assessment Exercise/Framework in the UK—in other settings, these quantitative evaluations of research have become the main mean through which research is assessed and funded (Sörlin, 2007). Various publication-based and citation-based funding models can be found in Australia, Norway, Denmark, Sweden and Finland—and translates as the currency through which academic exchanges of tenure, promotion and salary raises are made (e.g. Fuyono & Cyranoski, 2006).
While there has always been subliminal bibliometrics performed through peer evaluation—as reviewers were skimming through reviewees' CVs through the process—the massification of

evaluations and their institutionalisation led many researchers and institutions to put large emphasis on the number of papers they published. This has led to adverse effects (Binswanger, 2015; Frey & Osterloh, 2006; Haustein and Larivière, 2014; Weingart, 2005). Indeed, like any social group, researchers are prone to change their behaviour once the rules of the games become explicit or what is expected from them; phenomenon that could be referred to as the Hawthorne effect (Gillespie, 1993), or to Goodhart (1975) or Campbell's laws (1979). As most evaluations and rankings are first based on numbers of published papers, this has created incentives for researchers to *author as many papers as possible.* In Australia (Butler, 2004), where publications counts were used without differentiating between publication venue or citations received, researchers have been found to increase their numbers of publications in journals with high acceptance rates and lower impact. Along these lines, the h-index, which together with the Impact Factor, is likely the most popular bibliometric indicator in the scientific community, is largely determined by numbers of papers published than on citations (Waltman & van Eck, 2012).

Within this context, researchers have adopted many publication strategies. While some researchers focus on publishing few, high-quality papers—e.g. 'selective' (Costas & Bordons, 2007) or 'perfectionists' (Cole & Cole, 1973)—others publish as many papers as possible, without not all of them necessarily being of high quality—e.g. 'prolific scientists' (Cole & Cole, 1973) or 'big producers' (Costas & Bordons, 2008)). However, little is known on the publication strategy that yields the highest results in terms of impact. In order to better understand the relationship between productivity and impact, this paper compares, for a large dataset of disambiguated researchers (N= 25,994,021), their total number of papers with the proportion of these papers that made it to the top 1% most cited of their field. Thus, this paper aims at answering the following key question: Does an authors' share of top papers start to decrease with a certain number of papers published? Or is it stable, as production and impact are two distinct dimensions of scientific activity. In other words, *how many is too many?* What is the probability for an author to publish top cited papers relate to the number of papers published? A good analogy for this is archery: if an archer throws one arrow, what is the probability that it hits the center of the target? Does an increase in the number of arrows thrown leads to an increase in the proportion of arrows hitting the center of the target?

Two opposite hypotheses could be made. The first one would be that authors with just 'average' production—rather than low or high production— are the ones more likely to publish top cited papers, as these authors, perhaps, focus more on the 'quality' of their output than just on quantity (i.e. selective scholars). The second hypothesis would be that, it is the authors with very high number of papers who, on average, publish the highest proportion of top cited papers. This hypothesis would be on agreement with the theory of Merton's cumulative advantages (Merton, 1968), and supported by empirical work in the sociology of science (Cole & Cole, 1973). Similarly, in a Bourdieusian framework, the main goal of a researcher is to increase its rank in the scientific hierarchy and gain more scientific capital (Bourdieu, 2004). If publishing a high number of scientific papers and being abundantly cited are the ways through which researchers can reach this goal, then they will adapt their behaviour to reach these evaluation criteria.

This focus on publishing as many papers as possible—often referred to as 'salami slicing'—has been long discussed (e.g. Abraham, 2000; Jefferson, 1998). However, only a few authors have analysed the effect of 'salami slicing' on papers' citations. For instance, Bornmann and Daniel (2007) have shown, for a small sample of PhD research projects in biomedicine (N=96), that an increase in the number of papers associated with a project lead to an increase in the total citation counts of papers associated with the projects. However, they do not show whether the impact of each paper taken individually increases with the number of papers published. Similar to this study, Hanssen and Jørgensen (2015) analysed the effect of

'experience' on papers' citations; experience being defined as the author's previous number of publications. Drawing a sample of papers in transportation research (N=779) they show that experience is a statistically significant determinant of individual papers' citations, although this increase becomes marginal once a certain threshold is met in terms of previous papers published.

Methods

This paper uses Thomson Reuters' Web of Science (WoS) for the period 1980-2012. Only journal articles are included. Given that the units analysed in this paper are individual researchers, we used the disambiguation algorithm developed by Caron & van Eck (2014) to identify the papers of individual researchers. On the whole, the algorithm managed to attribute papers to 25,994,021 individuals, which were divided into seven cohorts based on the year of their first publication (Table 1).

Table 1. Number of disambiguated researchers per cohort

Year of first publication	Number of researchers
<=1985	3,574,667
1986-1990	2,733,002
1991-1995	3,282,421
1996-2000	3,810,652
2001-2005	4,310,886
2006-2011	6,930,289
>=2012	1,352,104

As we want to assess researchers' contribution to research that has the highest impact, we isolated for each discipline the top 1% most cited papers published each year (normalized by WoS subject categories). Citations are counted until the end of 2013, and exclude self-citations. The broad disciplines used are those of the 2013 Leiden ranking which are based on the assignment of WoS Subject Categories to five main domains (CWTS, 2013). Figures in the paper presents classes of numbers of papers in which there are at least 100 researchers.

Results

Figure 1 presents, for the oldest cohort studied—researchers who have published their first paper before 1985—the relationship between the number of papers throughout their career and the proportion of those papers that made it to the top 1% most cited. For any specific number of papers, the expected value of top 1% papers is, as one might expect, 1%. Researchers for all five domains have one thing in common: authors with very few papers are, on average, much less likely to publish high shares of top 1% most cited papers. For *Biomedical and health sciences* and for *Social sciences and humanities* we observe a continuous increase in authors' proportion of top papers as their overall number of papers increases. For *Life and earth sciences* the share of papers does increase with the number of papers, until about 10 papers where they starts to oscillate, although in general an increasing pattern is still observed, especially after 40 papers. Perhaps the most deviant pattern is found in *Mathematics and computer science* where for just for the very low levels of production there is an increase in the share of highly cited publications, but this share decreases between

4 and 20 papers. It then starts to increase again for higher numbers of papers, despite important fluctuations. Natural sciences and engineering follow a similar pattern, with a decrease in the share of top papers between 6 and 30 papers, followed, in this case, by a clear increase until very high levels of productivity.

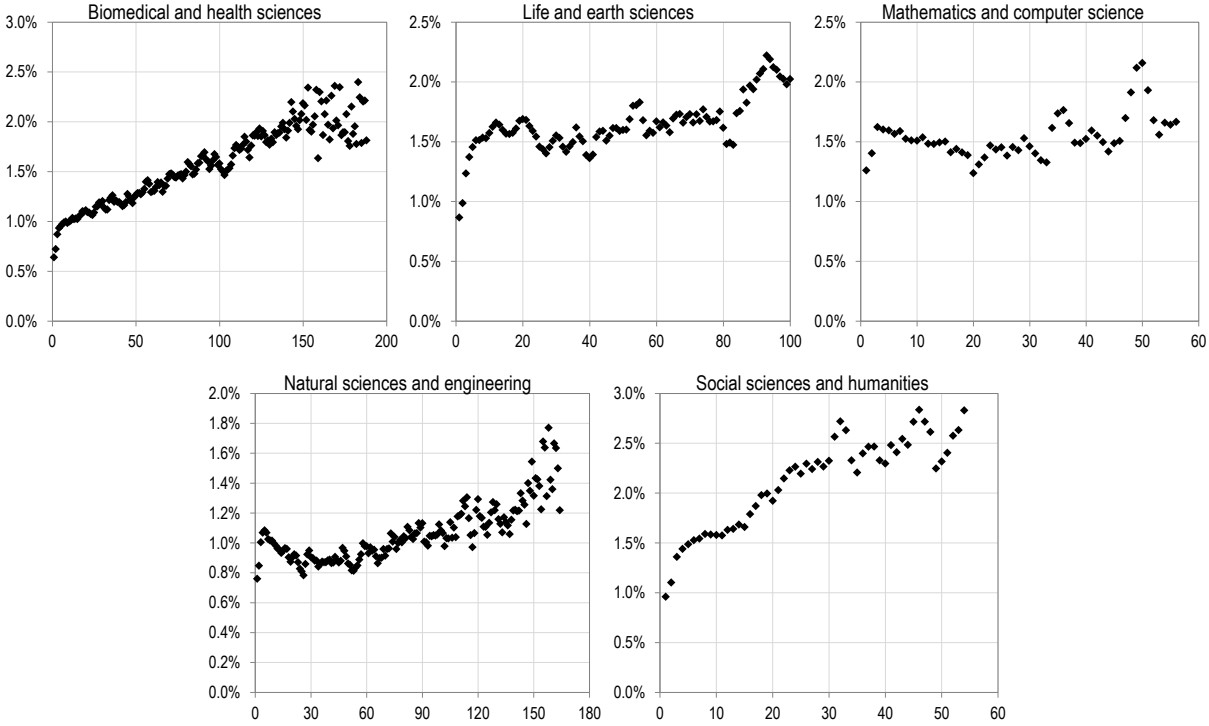


Figure 1. Proportion of top 1% most cited papers (*y* axis), as a function of the number of papers published (*x* axis), for the cohort of researchers who have published their first paper before 1985, by domain. Only classes of numbers of papers with 100 researchers or more are shown.

When researchers who have published their first paper between 2006 and 2011 are considered, different pattern are observed (Figure 2). For *Biomedical and health sciences* there is an increase in the share of highly cited publications up to around 15 publications, when some important fluctuations—or certain decreasing returns to scale—start to appear. A similar pattern is observed for the *Life and earth sciences* with the variability starting from levels of production of around 10 publications although, in this case, a decrease is clearly observed. For the other domains the pattern tends to be clearly increasing, although oscillations are also observed for the higher levels of production, which could also be seen as decreasing returns to scale. For the other three domains, there is clearly an increase in the share of top papers as the number of papers increases. However, we also observe for these three fields a decrease at very high levels of productivity.

An important characteristic of this cohort is that it got socialized to research recently—when the evaluation culture was more present—which might explain why they might be more prone to try to publish as much as possible. However, the drop in the share of top papers observed in each domain—although at different levels of productivity—suggests that these academically-younger scholars struggle to keep impact high once a certain threshold is met. This might be due to the fact that these scholars have not yet secured permanent or tenure positions and, thus, might feel that they cannot be as selective as older scholars who might choose their collaborators more easily.

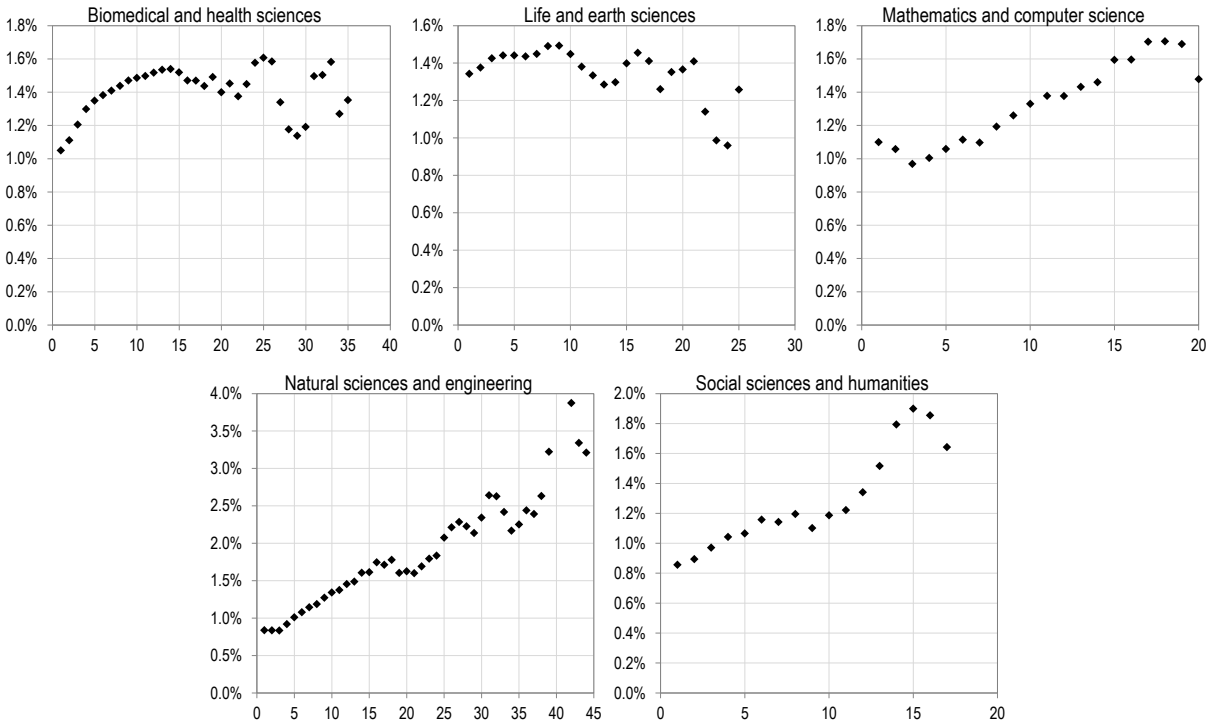


Figure 2. Proportion of top 1% most cited papers, as a function of the number of papers published, for the cohort of researchers who have published their first paper between 2006 and 2011, by domain. Only classes of numbers of papers with 100 researchers or more are shown.

Discussion and Conclusion

Previous research has shown that, in many contexts, the focus on indicators in research evaluation has had adverse effects, especially in terms of papers published (e.g. Binswanger, 2015). This paper aimed to provide an original analysis of one of these adverse effects, which is to aim to *publish as much as possible*. Our results have shown that, especially for older researchers, the higher the number of papers published throughout their careers, the higher the share of these papers ends up being amongst the top cited papers of their fields. This effect was higher for *Biomedical and health sciences* and for *Social sciences and humanities*, but in all fields the most active group of researcher was also having a higher share of top cited papers. A general exception to this trend was found in academically-younger researchers working in the field of *Life and earth sciences*, where higher scientific output was associated with lower impact than low-to-mid scientific output. Decreasing returns to scale were also more common for more junior researchers than senior ones.

These results conform to the Mertonian theory of cumulative advantages (Merton, 1968): the higher the number of papers an author contributes to, the more he or she gets known and, hence, is likely to attract citations. In Bourdieusian terms, the more an author publishes and accumulates citations in a domain, the more this capital will yield additional papers and citations. The relationship could also be in the other direction, as highly cited authors might have more opportunities to contribute to papers, given the scientific capital they have accumulated. Still, the results show that top cited authors do not only contribute on average to more papers, but also to more *highly cited* papers. On the whole, these results suggest that, at the macro-level, the culture of publishing *as many papers as possible* did not yield to adverse effects in terms of impact, especially for senior researchers. For such researchers, who have had a long period of time to accumulate scientific capital, there can never be *too many papers*.

References

Abraham, P. (2000). Duplicate and salami publications. *Journal of Postgraduate Medicine, 46*(2), 67.

Binswanger, M. (2015). How nonsense became excellence: Forcing professors to publish. In Welpe, I.M. et al. (eds). *Incentives and Performance* (pp. 19-32). Switzerland: Springer International Publishing.

Bornmann, L., & Daniel, H. D. (2007). Multiple publication on a single research study: does it pay? The influence of number of research articles on total citation counts in biomedicine. *Journal of the American Society for Information Science and Technology*, *58*(8), 1100-1107.

Bourdieu, P. (2004). *Science of Science and Reflexivity*. Cambridge, UK: Polity Press.

Butler, L. (2003). Modifying publication practices in response to funding formulas. *Research Evaluation*, 12(1), 39-46.

Cameron, B. D. (2005). Trends in the usage of ISI bibliometric data: Uses, abuses, and implications. *portal: Libraries and the Academy*, *5*(1), 105-125.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*(1), 67-90.

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In *19th International Conference on Science and Technology Indicators. Context Counts: Pathways to Master Big Data and Little Data* (pp. 79-86). CWTS-Leiden University Leiden.

Cole, J. R., & Cole, S. (1973). *Social Stratification in Science*. Chicago: University of Chicago Press.

Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, *1*(3), 193–203.

Costas, R., & Bordons, M. (2008). Is g-index better than h-index? An exploratory study at the individual level. *Scientometrics*, *77*(2), 267–288.

CWTS. (2013). *Leiden Ranking 2013 Methodology*. http://www.leidenranking.com/Content/CWTS%20Leiden%20Ranking%202013.pdf

Dahler-Larsen, P. (2011). *The Evaluation Society*. Palo Alto, CA: Stanford University Press.

Debackere, K., & Glänzel, W. (2004). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics, 59*(2), 253-276.

Frey, B. S., & Osterloh, M. (2006). *Evaluations: Hidden Costs, Questionable Benefits, and Superior Alternatives*. Institute for Empirical Research in Economics, University of Zurich.

Fuyuno, I., & Cyranoski, D. (2006). Cash for papers: putting a premium on publication. *Nature*, *441*(7095), 792-792.

Gillespie, R. (1993). *Manufacturing Knowledge: A History of the Hawthorne Experiments*. Cambridge, New York: Cambridge University Press.

Goodhart, C.A.E. (1975). *Problems of Monetary Management: The U.K. Experience*. Papers in Monetary Economics (Reserve Bank of Australia).

Hanssen, T. E. S., & Jørgensen, F. (2015). The value of experience in research. *Journal of Informetrics, 9*(1), 16-24.

Haustein, S., & Larivière, V. (2015). The use of bibliometrics for assessing research: possibilities, limitations and adverse effects. In Lempe, I.M. et al. (eds.) *Incentives and Performance* (pp. 121-139). Switzerland: Springer International Publishing.

Jefferson, T. (1998). Redundant publication in biomedical sciences: Scientific misconduct or necessity? *Science and Engineering Ethics*, *4*(2), 135-140.

Merton, R. K. (1968). The Matthew effect in science. *Science, 159*(3810), 56-63.

Sörlin, S. (2007). Funding diversity: performance-based funding regimes as drivers of differentiation in higher education systems. *Higher Education Policy*, *20*(4), 413-440.

Waltman, L., & Van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, *63*(2), 406-415.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, *62*(1), 117-131.

Research Assessment and Bibliometrics: Bringing Quality Back in

Michael Ochsner[1] and Sven E. Hug[2]

[1] *ochsner@gess.ethz.ch*
ETH Zurich, D-GESS, Mühlegasse 21, 8001 Zürich (Switzerland) and FORS, c/o University of Lausanne, Géopolis, 1015 Lausanne (Switzerland)

[2] *sven.hug@gess.ethz.ch*
ETH Zurich, D-GESS, Mühlegasse 21, 8001 Zürich (Switzerland) and University of Zurich, Evaluation Office, Mühlegasse 21, 8001 Zürich (Switzerland)

Introduction

Bibliometric indicators are used to compare research performances and also to assess and evaluate research performance (see, e.g. Gimenez-Toledo et al., 2007; Lane, 2010). However, recently scholars voice protest against bibliometric assessments (see, e.g., Lawrence, 2002; Molinie & Bodenhausen, 2010; Drubin, 2014). The arguments put forward are manifold. For example, the application of the impact factor, which is often used, but not meant, to evaluate individual researchers, is criticized (DORA, 2013). Then, there are myriads of perverse or unintended effects, like focus on high impact journals and mainstream topics, focus on review articles and short communications, strategic behavior, or lack of replication because of the low reputation of replication studies (e.g., Butler, 2007; Lawrence, 2003; Mooneshinghe et al., 2007). Furthermore, scholars from the social sciences and humanities (SSH) criticize that that bibliometric indicators cannot capture quality (e.g., Plumpe, 2009).

The authors of this paper were involved in a project to develop quality criteria and indicators for humanities research (see http://www.psh.ethz.ch/crus). Here, we argue that while bibliometric indicators and methods are powerful tools to describe research practices and, to some extent, scientific impact, there are some problems when they are readily used as quality indicators in research assessments. We feel that also other disciplines can learn from the critique of humanities scholars on simplistic quantitative assessments and from the findings of the research on quality in the humanities.

Notions of quality

The aim of the project "Developing and Testing Research Quality Criteria in the Humanities" was to find quality criteria and indicators that were at the same time accepted by the humanities scholars and implementable in different linguistic, cultural, and disciplinary settings. Analyzing the humanities scholars' critique, we found that the development of criteria must take into account the disciplinary research practices, that the measurement must be transparent and consensual, and that the notions of quality must be made explicit (Hug et al., 2014). We used the Repertory Grid technique to make the notions of quality explicit and base the development of quality criteria on the actual research practices. We found that there are two different conceptions of quality, a more traditional one, which can be described with individual, ground-breaking research that opens up new paradigms, and a more modern conception that can be described as interdisciplinary, project-focused, and public-oriented. Both kind of research can be good as well as bad (Ochsner et al., 2013). Hence, interdisciplinarity, for example, differentiates between two different ways of doing research but is not an indicator of quality (interdisciplinarity can point to good research, when it merges different theories and methods, but it can equally point to bad research that uses interdisciplinarity only for getting funding or for the career). Therefore, notions of quality should be taken into account in research evaluations. They might shed light on gaming strategies as well as on problems with indicators that are not linked to research practices or research quality.

Catalogue of quality criteria

Using the notions of quality, we developed a catalogue of quality criteria that are linked to the research practices in the humanities. Humanities scholars then rated these criteria as well as indicators measuring those criteria. We found that a broad range of quality criteria and aspects must be taken into account to adequately assess research quality (Hug et al., 2013) and that only about 3% to 32% of the scholars' notions of quality can be quantified adequately, depending on the discipline. Furthermore, we found that there is a mismatch between the quality criteria put forward by the scholars and the quality criteria used in evaluation procedures (Ochsner et al., 2012). Hence, current evaluation procedures do not measure research quality in the humanities adequately. This does not mean that the existing evaluation procedures and criteria are useless (e.g., societal impact is not necessarily linked to research quality but is a legitimate criterion in evaluations), but it shows that

a very important dimension of research assessment is not reflected adequately: quality of research.

The humanities, so what?!

Our research bases on the humanities. What is the relevance of this research to the rest of academia? First, we argue that humanities scholars, while not specialised in quantification, are experts in critical thinking. Hence, their critique of evaluation procedures often points to the consequences of the instruments on research practices. This is what increasingly also happens in the natural sciences (e.g., DORA, 2013; Drubin, 2014) because some perverse effects start to become apparent. Hence, a focus on research practices in assessments could help minimise negative impact of indicators. Second, when we presented the criteria at conferences and workshops, also natural scientists were present. They surprisingly often said that the criteria we presented made also sense to them with a few exceptions. Hence, what could be learned from the case of the humanities would be the following: base evaluation procedures on research practices; be aware that the indicators used will affect the research practices; formulate quality criteria in a way that makes sense to the scholars; involve as many stakeholders as possible in the definition of quality criteria.

Bringing quality back in

While the bibliometric community is well aware of the possible drawbacks of bibliometric indicators, the most common reaction by the research evaluation community is to look for other sources of the same kind of indicators and altmetrics. We think that the problem is not a technical one but a conceptual. At the beginning of any research evaluation and science policy should be a reflection on the goals. Do we want scholars to use most of their time to feed Twitter, comment on Research Gate, or 'pimp' their statistics in Google Scholar?
We think that research evaluation should bring quality back in. Evaluation and assessments should not solely *judge* the merits of scholars but help them to *enhance* their impact by fostering research quality. Hence, bibliometrics and altmetrics are powerful instruments to describe certain impacts, visibility, networks etc. But research assessments should also make clear statements about other aspects of research quality. Therefore, the disciplinary community should have a say in what criteria are applied in their assessments. New ideas of research evaluation based on research practices should lead scientific discussion much more than technical issues vaguely related to research quality.

Acknowledgments

This paper is based on work that was supported by the Rectors' Conference of the Swiss Universities (CRUS) within the framework of the SUK B-05 Innovation and Cooperation Project "Mesurer les performances de la recherche". Matching funds were provided by the University of Zurich.

References

Butler, L. (2007). Assessing university research: a plea for a balanced approach. *Science and Public Policy, 34*(8), 565-574.

DORA. (2013). San Francisco Declaration on Research Assessment. http://am.ascb.org/dora/

Drubin, D. (2014). Time to discard the metric that decides how science is rated. *The Conversation*. http://theconversation.com/time-to-discard-the-metric-that-decides-how-science-is-rated-27733

Gimenez-Toledo, E., Roman-Roman, A. and Alcain-Partearroyo, M. D. (2007). From Experimentation to Coordination in the Evaluation of Spanish Scientific Journals in the Humanities and Social Sciences. *Research Evaluation, 16*(2), 137–48.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2013). Criteria for assessing research quality in the humanities: a Delphi study among scholars of English literature, German literature and art history. *Research Evaluation, 22*(5), 369–383.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2014). A framework to explore and develop criteria for assessing research quality in the humanities. *International Journal for Education Law and Policy , 10*(1), 55–64.

Lane, J. (2010). Let's Make Science Metrics More Scientific. *Nature, 464*(25), 488–489.

Lawrence, P.A. (2002). Rank injustice. The misallocation of credit is endemic in science. *Nature, 415*, 835-836.

Lawrence, P.A. (2003). The politics of publication. Authors, reviewers and editors must act to protect the quality of research. *Nature, 422*, 259-261.

Molinie, A. & Bodenhausen, G. (2010). Bibliometrics as Weapons of Mass Citation. *Chimia **64***(1-2), 78-89.

Mooneshinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most published research findings are false - but a little replication goes a long way. *PLOS Medicine, 4*(2), e28.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2012). Indicators for research quality in the humanities: opportunities and limitations. *Bibliometrie - Praxis und Forschung, 1*(4).

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2013). Four types of research in the humanities: setting the stage for research quality criteria in the humanities. *Research Evaluation, 22* , 79–92.

Plumpe, W. (2009). Stellungnahme zum Rating des Wissenschaftsrates aus Sicht des Historikerverbandes. In C. Prinz & R. Hohls (Eds.), Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaft? (pp. 121–126). Historisches Forum. Berlin: Clio-online.

Under-reporting research relevant to local needs in the global south. Database biases in the representation of knowledge on rice

Ismael Rafols,[1,2] Tommaso Ciarli[2] and Diego Chavarro[2]

[1] *i.rafols@ingenio.upv.es*
Ingenio (CSIC-UPV), Universitat Politècnica de València, València, Spain,
SPRU (Science and Technology Policy Research), University of Sussex, Brighton, UK,
and Observatoire des Sciences et Techniques (HCERES-OST), Paris, France

[2] *t.ciarli @sussex.ac.uk, diego.chavarro@sussex.ac.uk*
SPRU, Science Policy Research Unit, University of Sussex, Brighton, UK

Introduction

There is an increasing demand for science to help in addressing grand challenges or societal problems, such as tackling obesity, climate change or pandemics. In this context, it becomes important to understand what different sciences can offer to tackle these problems, and towards which directions scientific research should be developed. A useful starting point is to investigate what is the existing science supply, and which research options are better aligned to address grand challenges and societal demands (Sarewitz & Pielke, 2007). In order to map the science supply, we need a representation of the knowledge on research topics relevant for a problem.

Bibliometrics can provide very helpful tools for developing knowledge representations. However, these representations are highly dependent on the data and methods used. As a result, bibliometric tools or indicators often reproduce the biases in the data collection and treatment. For example, it has been shown that conventional bibliometric analyses are biased against non-English languages (Van Leeuwen et al., 2001), developing countries (Velho & Krige, 1986), applied science (Van Eck et al., 2013), the social sciences and humanities (Martin et al., 2010) and interdisciplinary research (Rafols et al., 2012). The aim of this paper is to investigate the biases introduced by available databases in the representation of research topics.

In a previous study on rice research, we showed that the bibliographic database CAB Abstracts (CABI) – which is focussed on agriculture and global health – has a larger coverage of rice research for most low income countries than Web of Science (WoS) or Scopus (Ciarli, Rafols & Llopis, 2014). For example, India has twice the number of publications in CABI on rice compared to Scopus and about 4 times those in WoS. In this study, we present evidence that shows that this unequal coverage distorts significantly the knowledge representation of rice research, globally and for different countries. Such bias may have policy effects, in particular for a societal issue such as rice production.

As shown in Figure 1, we find that the journal coverage of the bibliometric databases WoS and Scopus under-represent some of the more application oriented topics (namely: i) production, productivity and plant nutrition (top left); ii) plant characteristics (top center); and iii) diseases, pests and plant protection (center).

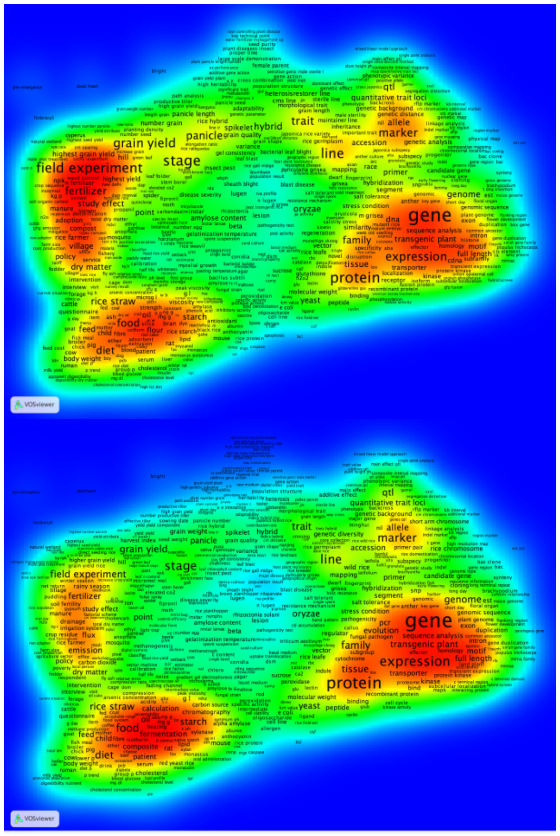


Figure 1. Publication density for rice research in CABI (top) and in WoS (bottom). The top left and top right areas under-report in WoS are related to production and seed characteristics.

Given that these are issues relevant to small farmers, producing for the local market, and with no access to the seeds developed with molecular biology techniques (GM – bottom left), we pose the

question whether the inadvertent effect of the biases in the dominant database is to under-represent, the type of research that has most chances of being relevant for improving their wellbeing, without introducing the use of the highly contested GM seeds.

Figure 2 illustrates that under-representation of research on production, pest and seed characteristics is particularly acute in some countries with molecular biology research (related to GM), but with a focus on research to address food security and local farming needs (in this case Iran). Rice research in these countries tends to be more focused on increasing crop yield, precisely the topic under-represented in WoS and Scopus.

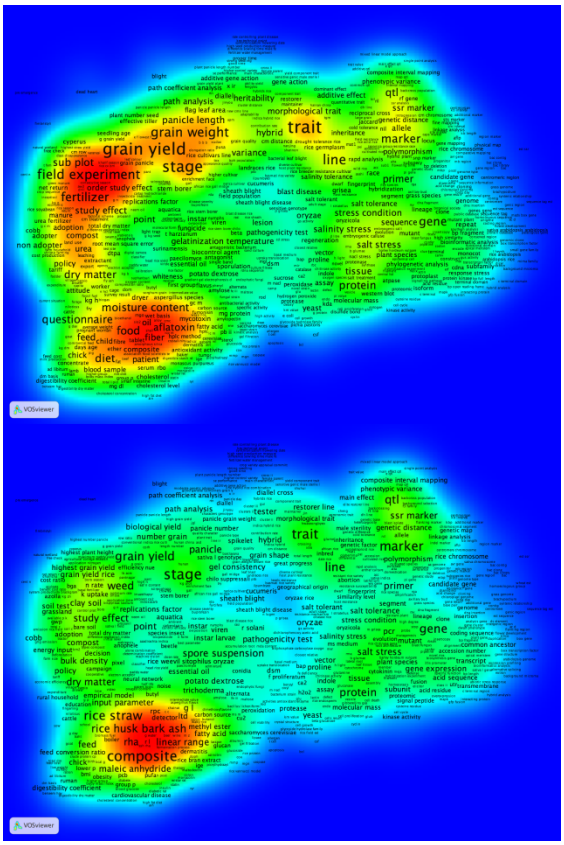


Figure 2. Publication density for rice research in Iran for CABI (top) and WoS (bottom).

Conclusions

Since knowledge representation can play a significant role in framing research strategies, policy and technological development, in this ignite talk we want to draw attention to the topic bias in the dominant bibliometric databases. From a technical point of view, few bibliometric and science policy experts will be surprised to hear that WoS and Scopus, are under-representing low income countries and more applied research. Given these results, we pose the question whether such conceptual biases may result in strategies that do not take into account knowledge and techniques which may be developed in closer connection to farmers and consumers local needs. This study does not answer this question, but it shows that it is a meaningful and important issue for bibliometrics to address: bibliometric exercise that use dominant databases may have a negative effect on policies relevant to important social issues, particularly in developing countries.

Information on methods and data

Publications on rice for the period 2003-2012 were downloaded from the WoS (including SCI-Expanded, SSCI, A&HCI, CPCI-S i CPCI-SSH) searching “rice” or “oryza” in the field “topic”. Scopus records were downloaded searching in title, abstract or keywords, i.e. TIT-ABS-KEY ("rice" OR "oryza"). Similarly, documents with “rice” or “oryza” were searched in title and abstract of the database CAB Abstracts. The records of the different databases were matched with multiple matching algorithms. The analysis was carried out using Vantage Point, the statistical package R and the visualisation programme VOSviewer.

Acknowledgments

We acknowledge support from the EU (Marie Curie Integration fellowship to IR), the UK ESRC (RES-360-25-0076) and the US NSF (Award #1064146). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the funders.

References

Ciarli, T., et al. (2014). The under-representation of developing countries in the main bibliometric databases. *Proceedings of the S&T Indicators Conference* (97–105). Leiden.

Martin, B. R., Tang, P., Morgan, M., & al. (2010). *Towards a Bibliometric Database for the Social Sciences and Humanities – A European Scoping Project* (A report for DFG, ESRC, AHRC, NWO, ANR and ESF). Brighton, UK: SPRU.

Rafols, I., et al. (2012). How journal rankings can suppress interdisciplinarity. *Research Policy*, *41*(7), 1262–1282.

Sarewitz, D., & Pielke, R. A. (2007). The neglected heart of science policy: reconciling supply of and demand for science. *Environmental Science & Policy*, *10*(1), 5–16.

Van Eck, N. J. et al. (2013). Citation Analysis May Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLoS ONE*, *8*(4), e62395. doi:10.1371/journal.pone.0062395

Van Leeuwen, T. N., et al. (2001). Language biases in the coverage of the Science Citation Index. *Scientometrics*, *51*(1), 335-346.

Velho, L., & Krige, J. (1984). Publication and citation practices of Brazilian agricultural scientists. *Social Studies of Science*, *14*(1), 45-62.

Network DEA approach for measuring the efficiency of University-Industry Collaboration Innovation: Evidence from China

Yu Yu[1] , Qinfen Shi[2] and Jie Wu [3]

[1] *yuyu0801@139.com*
Hohai University, Business School, No.8 Focheng Road West, 211100 Nanjing (China)

[2] *shiqf@njupt.edu.cn*
Nanjing University of Posts and Telecommunications, School of Management, No.9 Wenyuan Road, 210023 Nanjing (China)

[3] *0511wujie@163.com*
Jiangsu University of Science and Technology, School of Economics and Management, No.2 Mengxi Road, 212003 Zhenjiang (China)

Introduction

Collaborative innovation is a trans-disciplinary approach for developing the wholeness synergy to improve the competitiveness of an organization through holistic, competitive and complementary interactions between and among innovation participants in a specific environment (Bommert, 2010; Swink, 2006). The collaborative innovation system essentially consists of three sectors: industry, universities, and the government, with each sector interacting with the others, while at the same time playing its own role. Collaborative innovation system is a complex conglomerate of interacting independent parties. The network of institutional relations among universities, industries, and governmental agencies has been considered as a Triple Helix (TH). Collaborative innovation system (CIS) is based on a multi-input, multi-output transformation relation. It is an important issue to investigate the performance related to the transformation process of limited innovation resources for improving collaborative innovative outputs. Previous studies have been done to evaluate the performance of collaborative innovation. However, those studies failed to consider the complexity of the collaborative innovation system. Data envelopment analysis (DEA) is a method for measuring the efficiency of peer decision making units (DMUs). Recently network DEA models been developed to examine the efficiency of DMUs with internal structures. The internal network structures range from a simple two-stage process to a complex system where multiple divisions are linked together with intermediate measures. In this study, we propose a network DEA with parallel production systems to measure the efficiency of University-Industry Collaborative Innovation. The purpose of the present study is to construct a complete measurement framework characterizing the CIS' production framework from original S&T investment to final outputs, and measure the CIS' process-oriented technical efficiency, which is implemented in China's context. It is hoped that this study will benefit China's collaborative innovation policy-making.

Network DEA model

We propose a network DEA with parallel production systems in this section. Assume that there are n DMUs, and each DMU has two *sub-DMUs*. Figure 1 depicts the visual structure of the DEA model.

The part of inputs is consumed by SDMU1 and SDMU2 together, and part of DMU output is co-produced by SDMU1 and SDMU2. Besides, some inputs and outputs are consumed or produced by SDMU1 or SDMU2 alone. Variables are defined as follows: $X_1 = (x_{1j}^1, \mathrm{K}\ , x_{mj}^1)$ represent m separate inputs which are consumed by SDMU1; $X_2 = (x_{1j}^2, \mathrm{K}\ , x_{hj}^2)$ represent h separate inputs which are consumed by SDMU2; $X_s = (x_{1j}^s, \mathrm{K}\ , x_{lj}^s)$ represent l inputs consumed by SDMU1 and SDMU2 together. The vector of $Y_1 = (y_{1j}^1, \mathrm{K}\ , y_{sj}^1)$ are s outputs produced by SDMU1; the vector of $Y_2 = (y_{1j}^2, \mathrm{K}\ , y_{tj}^2)$ are t outputs produced by SDMU2; the vector of $Y_s = (y_{1j}^s, \mathrm{K}\ , y_{uj}^s)$ are u outputs produced by SDMU1 and SDMU2 together.

For analytical tractability, we use $X_{s1} = (x_{1j}^{s1}, \mathrm{K}\ , x_{lj}^{s1})$, $X_{s2} = (x_{1j}^{s2}, \mathrm{K}\ , x_{lj}^{s2})$, $Y_{s1} = (y_{1j}^{s1}, \mathrm{K}\ , y_{uj}^{s1})$ and $Y_{s2} = (y_{1j}^{s2}, \mathrm{K}\ , y_{uj}^{s2})$ to represent the shared inputs and outputs of SDMU1 and SDMU2 in each subsystem, and $X_s = X_j^{s1} + X_j^{s2}, Y_s = Y_j^{s1} + Y_j^{s2}$.

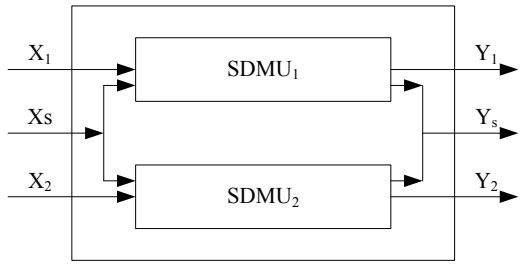


Figure 1. Parallel system structure.

In this study, we choose new product sales as independent output in Industry sub-system, the number of universities' published papers as independent output in universities sub-system. Patent applications in IU collaboration innovation system mainly come from both industry and universities subsystems; therefore the number of patent applications is seen as a shared output in the system.

According to DEA parallel production system efficiency evaluation model proposed by Kao (2009), parallel production system efficiency of the DMU under constant returns to scale (CRS) can be represented as follows:

$$\overline{\theta}^*_{CRS} = \min \quad \theta$$

$$s.t.$$

$$\sum_{k=1}^{2}\sum_{j=1}^{n} \lambda_j^k y_{rj}^{sk} \geq y_{ro}^s \quad r = 1,\mathrm{K}\ ,u$$

$$\sum_{j=1}^{n} \lambda_j^1 y_{rj}^1 \geq y_{ro}^1 \quad r = 1,\mathrm{K}\ ,s$$

$$\sum_{j=1}^{n} \lambda_j^2 y_{rj}^2 \geq y_{ro}^2 \quad r = 1,\mathrm{K}\ ,t$$

$$\sum_{k=1}^{2}\sum_{j=1}^{n} \lambda_j^k x_{ij}^{sk} \leq \theta x_{io}^s \quad i = 1,\mathrm{K}\ ,l$$

$$\sum_{j=1}^{n} \lambda_j^1 x_{ij}^1 \leq \theta x_{io}^1 \quad i = 1,\mathrm{K}\ ,m$$

$$\sum_{j=1}^{n} \lambda_j^2 x_{ij}^2 \leq \theta x_{io}^2 \quad i = 1,\mathrm{K}\ ,h$$

$$\sum_{j=1}^{n} \lambda_j^1 = \sum_{j=1}^{n} \lambda_j^2$$

$$\lambda_j^k \geq 0 \quad k = 1,2;\ j = 1,\mathrm{K}\ ,n$$

The main data in this paper are all selected in the "*China Statistical Yearbook of Science and Technology*". Considering the time lag in innovation activities, we select the data in 2009 as input data and the data in 2010 as output data in this paper. This study excludes all provinces that have missing data. Finally, this study evaluates 30 observations of Chinese provinces.

Table 1 summarizes three efficiency scores under constant returns to scale (CRS), variable returns to scale (VRS) and non-increasing returns to scale (NIRS).

Table 1. Three Efficiencies of Chinese provinces.

Province	$\overline{\theta}^*_{CRS}$	$\overline{\theta}^*_{NIRS}$	$\overline{\theta}^*_{VRS}$
Beijing	0.5903	1.0000	1.0000
Tianjin	0.9412	1.0000	1.0000
Hebei	0.6656	0.6656	0.6692
Shanxi	0.3089	0.3089	0.3189
Inner Mongolia	0.4715	0.4715	0.4974
Liaoning	0.4605	0.4605	0.4636
Jilin	1.0000	1.0000	1.0000
Heilongjiang	0.3869	0.3869	0.3882
Shanghai	0.8232	1.0000	1.0000
Jiangsu	0.8229	1.0000	1.0000
Zhejiang	0.8769	0.8791	0.8791
Anhui	0.6534	0.6546	0.6546
Fujian	0.5968	0.5968	0.6002
Jiangxi	0.5474	0.5474	0.5491
Shandong	0.6453	1.0000	1.0000
Henan	1.0000	1.0000	1.0000
Hubei	0.6291	0.8497	0.8497
Hunan	0.6651	0.6667	0.6667
Guangdong	0.8773	1.0000	1.0000
Guangxi	0.7016	0.7016	0.7095
Hainan	0.9648	0.9648	1.0000
Chongqing	0.9698	0.9903	0.9903
Sichuan	0.4845	0.5530	0.5530
Guizhou	0.6488	0.6488	0.6661
Yunnan	0.5810	0.5810	0.6081
Shaanxi	0.6860	0.6860	0.6861
Gansu	0.8782	0.8782	0.8828
Qinghai	0.3233	0.3233	0.8972
Ningxia	0.5769	0.5769	0.6545
Xinjiang	0.7036	0.7036	0.7416

Results

The average efficiency under constant returns to scale of University- Industry collaborative innovation in China is 0.7642. However, the efficiencies of some provinces are less than the average efficiency. By the view of economic region, the efficiencies of UI collaborative innovation in eastern, northern and southern coastal China are higher than other areas in China.

Acknowledgments

The work is supported by National Natural Science Foundation of China (No. 71471091, 71271119).

References

Bommert, B. (2010). Collaborative innovation in the public sector. *International Public Management Review, 11*(1), 15-33

Kao, C. (2009). Efficiency measurement for parallel production systems. *European Journal of Operational Research*, *196*(3), 1107-1112.

Swink, M. (2006). Building collaborative innovation capability. *Research-technology Management, 49*(2), 37-47.

Promotions, Tenures, and Publication Behaviours: Serbian Example

Dejan Pajić and Tanja Jevremov

dpajic@ff.uns.ac.rs; tanja.jevremov@uns.ac.rs
University of Novi Sad, Faculty of Philosophy, Department of Psychology,
Dr Zorana Đinđića 2, 21000 Novi Sad (Serbia)

Introduction

Bibliometric indicators became a common tool for evaluating universities (Geuna & Martin, 2003). Furthermore, individual academics and researchers are also evaluated, promoted, and tenured based on their productivity, particularly the one visible in international databases such as the Web of Science (WoS). This methodology is widely accepted even in non-English speaking countries (Pajić, 2014).

Growing emphasis on bibliometric indicators is followed by a continuing debate on their suitability for the evaluation in social sciences and humanities (SS&H) (Nederhof, 2006). Secondary importance of journals and the prevalence of monographs are usually identified as the key features of "publication behaviour" in SS&H (Hicks, 2012). Economics and psychology are often considered to be more similar to sciences (Engels, Ossenblok, & Spruyt, 2012).

This paper presents initial results on the scientific productivity of professors promoted and tenured at the University of Novi Sad (UNS). The main goal was to analyse publication patterns in SS&H and their implications for the evaluation of individuals.

Data and method

UNS is the second largest state university in Serbia. It consists of 14 faculties and 2 research institutes. Presented analysis was focused on the production of professors promoted or tenured in 2009-2013 at 6 UNS faculties in SS&H. Data were taken from the reports publicly available on the UNS website[1]. Each report contained bibliography provided by the candidate and was verified by the corresponding committee of at least three members.

The sample included 297 professors in language and literature (99), education (62), economics (32), psychology (27), law (26), history (19), sociology (12), philosophy (10), and science (e.g. professors of chemistry at teachers colleges) (10). The total of 9007 publications were extracted and categorized according to the origin (national, international), and type (books, journal articles, proceedings, other). In order to balance the differences in the publication counts among the researchers of different academic rank, only publications from the last promotion period of 5 years were taken into account.

[1] http://www.uns.ac.rs/sr/izborZvanje/bilteni.html Reports were removed during the preparation of this paper and are no longer available online, but are available from the authors.

Since this is a preliminary analysis, it was mostly based on descriptive statistics. Because of skewed distributions, non-parametric tests were used to test the basic differences among disciplines.

Results and discussion

Kruskal-Wallis test indicates significant differences in scientific productivity among researchers from different fields: $H(8, 297) = 22.99$, $p < .01$ (Figure 1). It is difficult to draw a solid conclusion, mainly because of highly skewed distributions and large individual differences, but clearly psychology and sciences have the highest median values, while the lowest scientific activity is that of the researchers in the field of law. The most pronounced individual differences were observed in the fields of language and literature, and educational sciences.

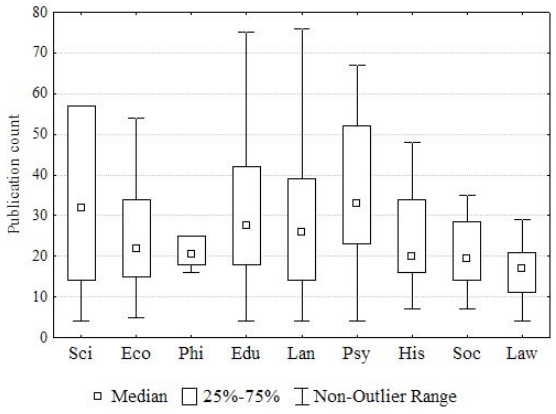


Figure 1. Differences in scientific productivity among researchers in nine scientific fields.

Distributions of the major types of publications among scientific fields differ significantly: χ^2 (16, 8492), $p < .01$ (Figure 2). The share of articles is somewhat unusually high in humanities, and ranges around 40% in all fields. Contrary to usual beliefs, psychology and sciences have the lowest proportion of journal articles within the total number of publications. On the other hand, the highest proportion was detected in the field of law where journal articles account for almost 2/3 of all publications. However, the list of the most frequent journal titles revealed that more than half of the articles were from a journal published by the same faculty where the candidates were promoted or tenured.

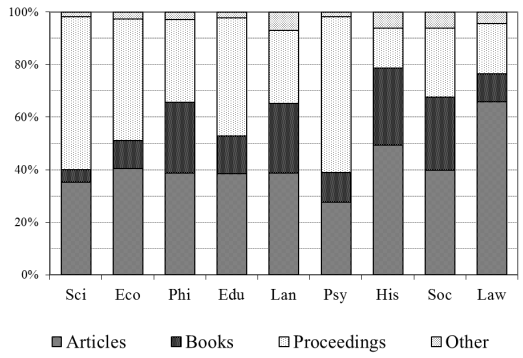


Figure 2. Proportions of different types of publications in nine scientific fields.

Our results have confirmed the importance of book chapters and monographs in humanities, although this type of publication is not predominant in any of the fields. Conference abstracts and proceedings are the most frequent type of publication in four out of nine analyzed fields.

Figure 3 shows the proportions of (inter)national publications across scientific fields. The strongest focus on international sources is noticeable in the sciences, and the lowest in history, sociology, and law. The results that are not in line with the usual beliefs are rather nationally oriented publication behavior of Serbian psychologists, and a relatively high ratio of international sources in philosophy.

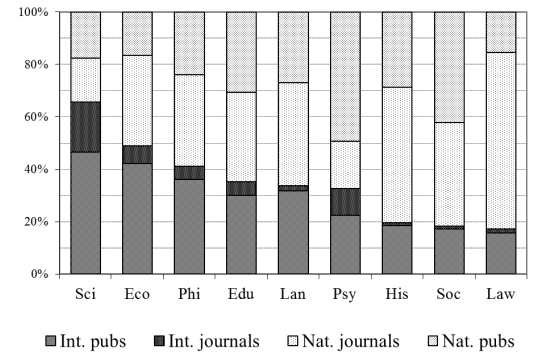


Figure 3. Proportions of national and international publications in nine scientific fields.

Professors at the faculties in Serbia are required to have one to three papers published in WoS journals prior to promotion or tenure. Table 1 shows the list of the 15 most common (allegedly) WoS journals reported in 297 reports. The majority of journals are actually national or regional WoS journals with the rather low impact factor values (IF). The disturbing fact is that several professors were promoted based on their articles published in journals of dubious quality, those that were dropped from WoS because of academic malpractice (e.g. *HealthMED*, *TTEM*, *Metalurgia Int*) or were never indexed by WoS nor any major international bibliographic database (e.g. *Brit Amer Stud*). In addition, 12 other journals were falsely reported as top ranked WoS titles.

Table 1. Most common (allegedly) WoS journals listed in 297 promotion and tenure reports.

Journal title	%	Country	IF
Psihologija	17.50	SRB	0.188
TTEM	5.83	B&H	drop.
HeathMED	5.13	B&H	drop.
Croat J Educ	3.03	CRO	0.034
Roman J Eng Stud	2.30	ROM	-
Med Sport	2.30	ITA	0.125
Vojnosan pregl	2.10	SRB	0.269
New Edu Rev	1.63	POL	drop.
Filoz istraživanja	1.63	CRO	AHCI
Brit Amer Stud	1.40	ROM	-
Panoeconomicus	1.16	SRB	0.778
Riječ	1.16	CRO	-
Didactica Slov	0.93	SLO	drop.
ICCCC	0.93	ROM	0.694
Metalurgia Int	0.93	ROM	drop.

drop. - dropped from WoS

Conclusion

Our results have shown that SS&H are clearly more nationally oriented compared to sciences. However, journals as knowledge dissemination channels seem to be equally important across all fields. Apart from the conference proceedings, journal articles are the most common type of publications. It's obvious that the current promotion and tenure rules affect the professors' publication behaviour. Such patterns are not determined simply by the characteristics of a discipline, but in some cases by the ease of access to particular sources, e.g. journals having a rather lenient editorial policy.

Science policy institutions should be aware that the evaluation is a dynamic process that must combine both the rules and the means to assess the effects of those rules and to monitor their implementation.

References

Engels, T.C.E., Ossenblok, T.L.B., & Spruyt, E.H.J. (2012). Changing publication patterns in the Social Sciences and Humanities, 2000-2009. *Scientometrics*, *93*(2), 373-390.

Geuna, A., & Martin, B R. (2003). University Research Evaluation and Funding: An International Comparison. *Minerva*, *41*(4), 277-304.

Hicks, D. (2012). One size doesn't fit all: on the co-evolution of national evaluation systems and social science publishing. *Confero*, *1*(1), 67-99.

Nederhof, A.J. (2006). Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review. *Scientometrics*, *66*(1), 81-100.

Pajić, D. (2015). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics, 120*(3), 2131-2150.

The Serbian Citation Index: Contest and Collapse

Dejan Pajić

dpajic@ff.uns.ac.rs
University of Novi Sad, Faculty of Philosophy, Department of Psychology,
Dr Zorana Đinđića 2, 21000 Novi Sad (Serbia)

The Past

Ten years ago, a poster titled *The Serbian Citation Index: context and content* was presented at the ISSI conference held in Stockholm (Šipka, 2005). *Serbian Citation Index (SCIndeks)* was at the time a pioneering effort to build a comprehensive, open access citation index of Serbian scientific journals with three missions: local *dissemination* of research findings in the open access mode, global *promotion* of the Serbian science, and objective *evaluation* of national journals, institutions, and researchers.

Started as an ambitious project of the group of enthusiasts and volunteers in 1990s, SCIndeks has become truly embraced nationally during the 2000s. In the period when Serbia was represented in the Web of Science (WoS) with only three journals, SCIndeks was recognized as a tool to enhance the public accountability, visibility, and quality of local journals. *Centre for Evaluation in Education and Science (CEES),* SCIndeks developer and publisher, started receiving full financial support from the Serbian Ministry of Science (SMS), both for the maintenance of SCIndeks and for publishing the *Journal Bibliometric Report (JBR)*. The report is published annually and contains the national impact factor and almost 20 other bibliometric indicators for over 300 journals covered by SCIndeks. JBR is used for journal rankings and, indirectly, as a data source for the evaluation of individual researchers, their promotions, and tenures.

The Contest

The role and importance of a national citation index cannot be evaluated outside the global scientific information market. The first test for SCIndeks was the recognition and perception of Serbian journals by the major international database providers. After Elsevier's Scopus and Google's Scholar appeared in 2004, Thomson Reuters' indexing policy has also changed radically. The question was whether the CEES efforts to improve the visibility and quality of local journals would result in increased number of titles accepted for indexing in WoS and Scopus. Figure 1 shows the number of journals published in Serbia and three neighbouring countries indexed in WoS and Scopus. All countries have managed to improve their visibility in international databases, but the Serbian progress is only slightly ahead of Bulgarian and far behind Romanian and Croatian. Neither Bulgaria nor Romania has national citation index or a repository of national journals. On the other hand, Croatian journals are presented in the *Portal of Scientific Journals of Croatia* and the *Croatian Scientific Bibliography*, both funded by the government, but having limited functionality compared to SCIndeks, especially regarding the support for journal editors, evaluators, and science policy institutions. It seems that the mission to promote journals through SCIndeks has failed or at least has not succeeded in lowering a potential bias in inclusion policies of the major database providers.

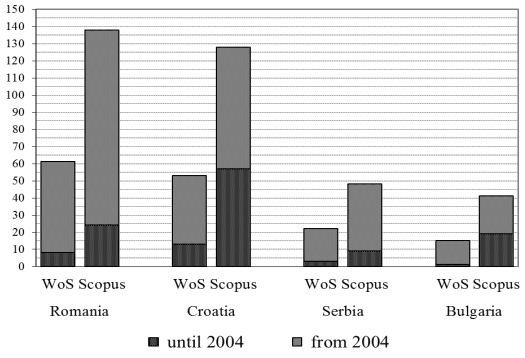


Figure 1. Growth in the number of WoS and Scopus journals published in Serbia and three neighbouring countries.

Another, and perhaps the more important contest, was carried out at the local (political) level. Every assessment brings the risk of conflict of interest. If such an assessment influences the allocation of funds and promotion and tenure decisions, the risk is even higher. Although the government supported CEES financially, it did not fully uphold the practical implementation of CEES reports on the quality of national journals (Šipka, 2014). Journal rankings based on impact measures and SCIndeks data were often altered by the ministerial committees in order to favour the very journals whose editors were members of those committees. In some cases, worst ranked national journals were given the status of international ones. At the level of individuals, it would mean that a candidate for promotion would earn points sufficient for a position of assistant professor by publishing two articles in a bottom-ranked local journal or a journal that was not even accepted for indexing in the national citation index.

The Collapse

In 2014, SMS has ceased to finance both the JBR and SCIndeks. In 2015, the effects of that decision have become visible in the form of significantly reduced SCIndeks coverage. A large amount of data were taken offline and became inaccessible to the users of SCIndeks and other web services, such as Google Scholar. Table 1 shows the amount of this "information market disturbance".

Table 1. SCIndeks data available online before and after the cut of funding.

No. of	*Apr. 2008*	*Apr. 2014*	*Apr. 2015*
journals	357	411	56
abstracts	82.876	151.027	19.900
full texts	23.421	58.068	12.172
references	917.567	2.078.642	335.344

As a response to the CEES' "strategic move", SMS has decided to continue using SCIndeks data for evaluation purposes and to finance JBR after all. However, all journals are now required to pay the indexing fees, including some additional costs for options like the full-text availability, cited reference search and cross-linking within SCIndeks. In short, a communication failure between CEES and SMS anticipates the start of a "natural selection" process for the majority of Serbian academic journals and the collapse of the open science idea in Serbia.

One aspect of this collapse is the fact that tens of thousands of papers written by the authors from Serbia are no longer available online and that additional costs are required for them to reappear. Another equally relevant issue is the profile of journals currently accessible through (what was) the national citation index. All of those journals are willing (or able) to pay the indexing fees, but just a few of them were previously classified as leading national journals. An example of this obvious compromise is the fact that although the diversity of affiliations within journal issues was strongly encouraged by both the national regulations and earlier SCIndeks inclusion guidelines, CEES indexes several journals with the majority of papers written by the authors affiliated with the journal's publishing institution.

The Future

Under the current circumstances, SCIndeks can no longer be considered to be the national citation index. The question is who should be concerned with the fact that it has become a mere commercial product with the special status at SMS. The state is surely a loser in this scenario being unable to claim and protect at least the metadata whose production it financed for several years. As for the Serbian scientific community, its future reactions are maybe not that hard to predict. A certain segment of this community has already expressed their opinion on this matter through the acts of various interest groups opposing the implementation of evaluation methodology based on SCIndeks data. On the other hand, an increasing number of researchers from Serbia are shifting the focus towards international journals, both when publishing and citing journal articles (Pajić & Jevremov, 2014). The evaluation of national science is hence being either spurned or entrusted to the international publishers and their reviewers. In this context, national citation index is becoming a costly repository whose functionalities will not be missed much by researchers or journal editors. More than 300 Serbian journals are now available online and none of them relies solely on SCIndeks when it comes to the visibility. Although some editors are satisfied with the combination of journal's personal website and free Google Scholar services, the growing number of Serbian journals are also being available through other databases and repositories, such as the *Directory of Open Access Journals*, *ERIH PLUS* or *EBSCO* databases. What was conceived as a joint effort to truly promote Serbian science has turned into an "every man for himself" strategy ten years after.

Conclusion

The basic idea of a national citation index was fully justified in the period of domination of Thomson Reuters' citation indices. But this domination is not nearly as strong as it was before, mainly due to the emergence of Scopus and Scholar. We can consider *SciELO* (now hosted by WoS) as an example of a successfully realized "peripheral" citation index. If this was achieved by covering some 1,200 journals from 12 different countries, then SCIndeks and its 400+ journals tell us how justified is the idea of a national citation index and how ambitious it should be. SCIndeks and its fate is the fate of any self-sufficient and rigid science policy institution, but also the fate of any scientific community that is simply too confined and too small. Too small to neglect the inevitable globalization of science, too small to rely on the integrity of its own members to ensure the quality control, and finally too small to satisfy its own ambitions.

References

Pajić, D., & Jevremov, T. (2014). Globally national - locally international: Bibliometric analysis of a SEE psychology journal. *Psihologija*, *47*(2), 263–277, doi:10.2298/PSI1402263P

Šipka, P. (2005). The Serbian citation index: context and content. In *Proc. of ISSI 2005, Stockholm, Sweden, July 24-28, 2005* (pp. 710–711).

Šipka, P. (2014). Methods of evaluation of scientific journals: use and abuse [in Serbian]. In Lj. Vučković-Dekić & N. Arsenijević (Eds.), *Vrednovanje nauke i naučnika* (pp. 9–30). Kragujevac: Fakultet medicinskih nauka & Beograd: Akademija medicinskih nauka.

Selecting Researchers with a Not Very Long Career - The Role of Bibliometrics

Elizabeth S. Vieira[1] and José A. N.F. Gomes[2]

[1] *elizabeth.vieira@fc.up.pt,* [2] *jfgomes@fc.up.pt*
REQUIMTE/Departament of Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal

Introduction

The scientific community has developed many institutionalized forms of evaluation where peer review has an important role, but recently, bibliometric methods have been gaining some acceptability to assess the scientific performance.

The two techniques have been related to one another in different ways: 1) bibliometric methods have been used to analyze the peer review processes (Moed, 2005, chapters 19 and 20); 2) the peer review process uses bibliometric parameters as an auxiliary instrument (Moed, 2005 chapter 18, p. 233-234); and 3) peer reviewers are called in to validate and correct the results of some bibliometric process (e.g. Norris & Oppenheim, 2003; Rinia, van Leeuwen, van Vuren, & van Raan, 1998). There are some national scientific systems that use bibliometric techniques or a mix of bibliometric techniques and peer review to decide the allocation of funding (e.g. Excellence in Research for Australia (ERA); *Valutazione della Qualità della Ricerca* (VQR)). Taking into account the advantages and limitations of bibliometric techniques and the intensive use, recently, there is a growing interest in its potential in helping peers to prepare the final decisions and therefore several studies have been made on the subject (e.g. Vieira, Cabral, & Gomes, 2014a, 2014b, Bornmann & Leydesdorff, 2013). In this study, we exploit the usability of bibliometrics as support tool this time in selecting candidates that had been awarded their PhD's more than 6 and less than 12 years ago and had worked as independent researchers for less than 6 years. We deem this study important as: (1) there is a growing use of bibliometric indicators and it is important to know their caveats and strong points at the different levels; and (2) the use of bibliometric indicators is more controversial when applied to individual researchers, especially at initial steps of their careers.

Methodology

This study considers the applicants to the development grants of the opening *Investigador FCT* carried out in Portugal since 2012. The publications indexed in the Web of Science Core Collection of the 120 applicants from the Engineering and Technology (28), Natural Sciences (23), Exact Sciences (48) and Medical and Health Sciences (21) were used to calculate a set of bibliometric indicators that are intended to describe the scientific performance. Bibliometric techniques are not used in a formal way in the opening. However, we are looking for indicators that may be implicit in peer judgments. A set of 17 indicators was determined: *TD* (number of documents); *TDC* (number of cited documents); *NDF* (number of documents after fractionation by the total number of authors); *PA* (% of articles); *PP* (% of proceedings papers); *PR* (% of reviews); *PAP* (% of documents as articles and proceedings papers simultaneously); *PDAC* (% of documents as corresponding author); *h* index, h_{nf} index (Vieira & Gomes, 2011); $SNIP_m$ (median of all the *SNIP*s of the journals where the applicant has published, Moed, 2010); SJR_m (median value as in the $SNIP_m$, Gonzalez-Pereira, Guerrero-Bote, & Moya-Anegon, 2010); *PTDIF* (% of documents published in journals with Impact Factor- *IF*); *PQ1* (% of documents published in journals in the first quartile in its scientific domain, according to the *IF*); *HCD* (% of documents highly cited in the top 10%); *NI* (average number of citations per document after normalization); *DIC* (% of documents with international collaboration). There is a huge number of bibliometric indicators and we tried to select those that describe the several dimensions of the scientific production. Nevertheless other indicators could be used.

Using as dependent variable the decision of the peers panel (selected-1; not selected–0) and the bibliometric indicators as independent variables we applied binary logistic regression aimed at determining those indicators that can be used to predict the final decisions made by the peers.

Results

The model

The application of the binary logistic regression lead to the following model:

$$P_i = \frac{e^{-1.88+1.116SJR_m+0.064HCD}}{1+e^{-1.88+1.116SJR_m+0.064HCD}}$$

where P_i is the probability of the applicant i to be selected by the peers for funding. The SJR_m and the *HCD* are the indicators that were found to be able to represent the decisions made by the peers panel.
The sensitivity determined for this model was 73.2%, the percentage of false positives obtained was 35% and 70% of the cases are predicted correctly by the model. The probability of the forecasted probability by the model for a selected applicant to be higher than that of a non-selected one is 75.3% (ROC curve).

Forecasts

The predictions given by the model are useful in preparing the decisions to be taken by the peers, but the use can be increased if complemented with some type of uncertainty measure. Here, this is shown using the margins concept. Margins are being used in bibliometrics at the individual level for the first time as far as we know.
In Figure 1 is shown the probability of a given applicant to be selected for funding as we increase the value of the *HCD* and SJR_m, respectively, and maintaining the average value of the other variable. For each predicted value is also shown the confidence interval at 95%, working as the uncertainty measure. All this information can be used by the peers to improve the decision making process.

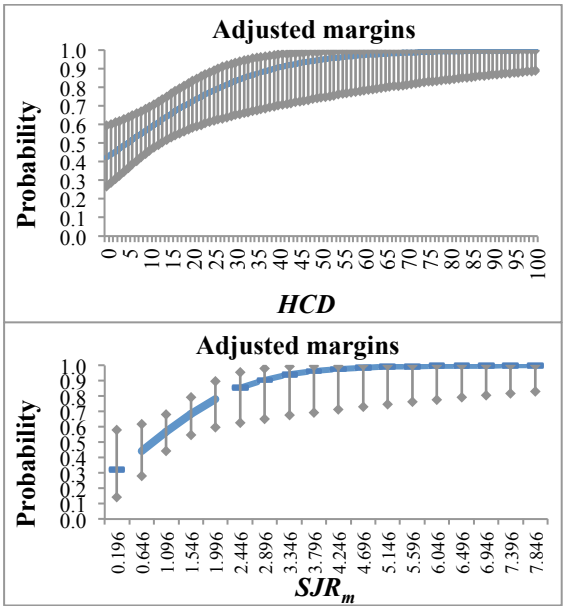


Figure 1. Predicted probabilities complemented with confidence intervals (95%). The dashed zone represents values with a few observations.

Conclusions

From this study some findings can be drawn:

- ✓ The bibliometric indicators are useful in describing the performance of applicants with PhD's earned 6 to 12 years ago.
- ✓ A composite indicator (*HCD* and *SJRm*) when used by the peers will have a positive impact on the final decision.
- ✓ Bibliometric indicators can be used, for example, as input tool helping peers panel in their decision making process as the indicators can give consistent and objective information.
- ✓ The *HCD* is a serious candidate as tool in support decisions of peer evaluations as it was also found to be useful in describing the final decisions in other types of openings (Vieira et al., 2014a, 2014b).

Acknowledgements

Elizabeth Vieira wishes to acknowledge the financial support from FCT (Foundation of Science and Technology), Portugal, through a grant SFRH/BPD/99246/2013. Data on applicants and results were kindly made available by *FCT*.

References

Bornmann, L., & Leydesdorff, L. (2013). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from incites and f1000. *Journal of Informetrics, 7*(2), 286-291.

Gonzalez-Pereira, B., Guerrero-Bote, V.P., & Moya-Anegon, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics, 4*(3), 379-391.

Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. The Netherlands: Springer

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics, 4*(3), 265-277.

Norris, M., & Oppenheim, C. (2003). Citation counts and the research assessment exercise V - archaeology and the 2001 RAE. *Journal of Documentation, 59*(6), 709-730.

Rinia, E. J., van Leeuwen, T. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria - evaluation of condensed matter physics in the Netherlands. *Research Policy, 27*(1), 95-107.

Vieira, E. S., Cabral, J. A. S., & Gomes, J.A.N.F. (2014a). Definition of a model based on bibliometric indicators for assessing applicants to academic positions. *Journal of the Association for Information Science and Technology, 65*(3), 560-577.

Vieira, E. S., Cabral, J. A. S., & Gomes, J.A.N.F. (2014b). How good is a model based on bibliometric indicators in predicting the final decisions made by peers? *Journal of Informetrics, 8*(2), 390-405.

Vieira, E. S., & Gomes, J.A.N.F. (2011). An impact indicator for researchers. *Scientometrics, 89*(2), 607-629.

Differences by Gender and Role in PhD Theses on Sociology in Spain

Lourdes Castelló Cogollos[1]; Rafael Aleixandre Benavent[2], Rafael Castelló Cogollos[3]

[1] *lourdes.castello@uv.es*
UISYS-Universitat de València. Plaça Cisneros, 4. 46003-València (Spain)

[2] *rafael.aleixandre@uv.es*
INGENIO (CSIC-Universitat Politècnica de València). UISYS-Universitat de València. Plaça Cisneros, 4. 46003-València (Spain)

[3] *rafael.castello@uv.es*
Departement de Sociologia i Antropologia Social. Facultat de Ciències Socials. Universitat de València Av. Tarongers, 4b. 46021 València (Spain)

Introduction

In recent years, there has been a growth in the number of papers that synthesize empirical research studies on gender and sex inequalities in academic statements. Furthermore, these studies can comply with European requirements of equalities since the Treaty of Amsterdam of 1999 enacted that equality between men and women should be included in all policies (Fernández Álvarez, 2014).

Theses are the research papers by excellence and a good indicator to elucidate the lines and research trends in a field of science, since this work must be original and specialized and are subject to a rigorous academic assessment (Delgado López Cózar et al., 2006).

Our objective is to analyse the differences in gender representation in the Spanish sociological theses focusing on three actors involved in the process: PhD students, supervisors and academic assessment boards.

Method

Records were obtained from TESEO, the governmental database of the Spanish Ministry of Education, Culture and Sport, which includes the Spanish theses defended and approved after evaluation. The search was limited to theses indexed by UNESCO codes related to Sociology (code 63) and to theses from the departments of Sociology of Spanish universities. A relational database was created to analyse and compare results.

Results

The total number of theses defended was 3,413. In the role of the PhD student, men presented 253 more theses than women did, while in the role of supervisor and academic assessment board, the differences were much greater: 1,004 and 1,159, respectively (Table 1).

Table 1. Number of PhD theses by gender and role.

Role	Male	Female	Total
PhD student	1,833	1,580	3,413
Supervisor	1,593	589	2,182
Assessment board	1,824	665	2,489

The percentage difference between males and females for PhD students is of 7 points, while for supervisors is of 47 points in favour of males, and for academic assessment boards this difference is of 47 points (Figure 1). The highest percentage of difference occurs in the role of academic assessment board, where 73.3% of board members were of males (Figure 1).

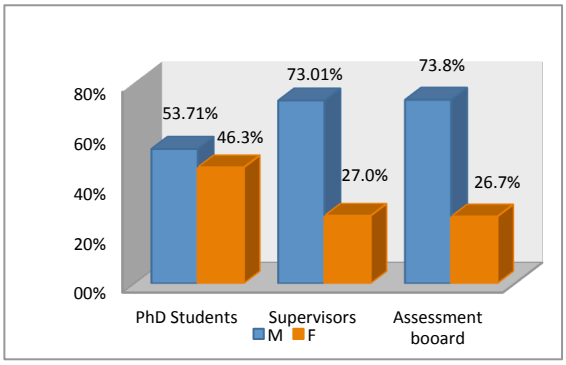


Figure 1. Percentage differences in PhD theses by gender and role

In the annual evolution of the percentages in the roles of supervisor and academic assessment board, men remain between 70% and 80% and women between 20% and 30%. On the contrary, from 2006-2010 period, women-PhD students reach parity (50%) and even surpass men in conducting thesis, ranking 57.8% in the last five-year period analysed (2011-2013) (Figure 2).

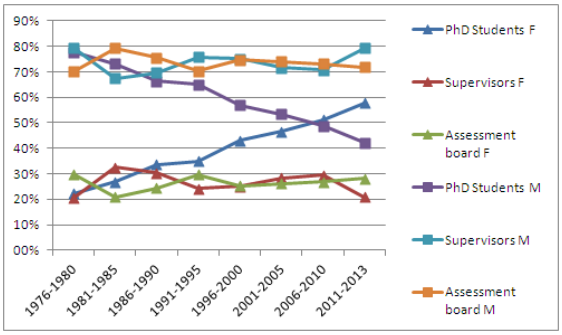


Figure 2. Five-year evolution of PhD theses by gender and role (1976-2013).

Discussion and Conclusions

Although a century has elapsed since the first woman enrolled in a Spanish university and its presence in several strata of the university has greatly improved, the percentage of women compared to men remain far from achieving parity in some roles.

The participation of women at the Spanish universities has increased steadily and its consolidation as PhD students today is a reality (Bermudez et al., 2011). However, from this stage, the academic careers of women slow down and the number of women who leave after doctorate is large (Bordons et al., 2003; Villarroya et al., 2008). Consequently, the percentage of female lecturers in Spain is between 30% and 35%, and the female professors between 14% and 20%. Therefore, it is noteworthy the existing great inequality in the Spanish universities as a professional field and that even though women are more numerous and better prepared than men at all levels of education, this is not reflected in prestigious academic positions (González Alcaide et al., 2009).

In conclusion, the promotion of women to positions of great academic responsibility is slow and is not in line with the number of women who obtained his doctorate in Sociology in Spain. Future research could explore other variables and behaviours, for example, if students of one gender tend to have supervisors from other different gender, as well as these trends in other fields and countries.

Acknowledgments

This work has benefited from assistance by the National R+D+I of the Ministry of Economy and Competitiveness of the Spanish Government (CSO2012-39632-C02-01) and Prometeo Program for excellent research groups of Generalitat Valenciana (GVPROMETEO2013-041).

References

Bordons, M., Morillo, F., Fernández, M.T., & Gómez, I. (2003). One step further in the production of bibliometric indicators at the micro level: Differences by gender and professional category of scientists. *Scientometrics*, *57*(2), 159-173.

Bermúdez, M.P., Guillén Riquelme, A., Gómez García, A., Quevedo Blasco, R., Sierra, J., & Buela Casal, G. (2011). Análisis del rendimiento del doctorado en función del sexo. *Educación XXI, 14*(1), 17-33.

Delgado López Cózar, E., Torres Salinas, D., Jiménez Contreras, E., & Ruiz Pérez, R. (2006). Análisis bibliométrico y de redes sociales aplicado a las tesis bibliométricas defendidas en España (1976-2002): Temas, escuelas científicas y redes académicas. *Revista Española de Documentación Científica, 29*(4), 493-524.

Fernández Álvarez, O. (2014). The gender perspective in managing knowledge through cross-curricular studies in higher education. *Procedia - Social and Behavioral Sciences*, *161*(19), 269-274.

González Alcaide, G., Agulló Calatayud, V., Valderrama Zurián, J.C., & Aleixandre Benavent, R. (2009). Participación de la mujer y redes de coautoría en las revistas españolas de Sociología. *Revista Española de Investigaciones Sociologicas, 126*, 153-166.

Villarroya, A, Barrios, M., Borrego, A., & Frías, A. (2008). PhD theses in Spain: A gender study covering the years 1990–2004. *Scientometrics*, *77*(3), 469–483.

The Trends to Multi-Authorship and International Collaborative in Ecology Papers

João Carlos Nabout[1], Marcos Aurélio de Amorim Gomes[2], Karine Borges Machado[3], Barbbara da Silva Rocha[4], Meirielle Euripa Pádua de Moura[5], Raquel Menestrino Ribeiro[6], Lorraine dos Santos Rocha[7], José Alexandre Felizola Diniz-Filho[8] and Ramiro Logares[9]

[1] *joao.nabout@ueg.br,* [7] *lo.rrane@hotmail.com*
State University of Goiás, Br 153, 3105, Fazenda Barreiro do Meio, CP 459, CEP 75132-903, Anápolis, GO (Brazil)

[2] *marcos.bioamorim@gmail.com,* [5] *meirielle-euripa@hotmail.com,* [6] *raquel.menestrino@gmail.com*
State University of Goiás, PPG Recursos Naturais do Cerrado, Br 153, 3105, Fazenda Barreiro do Meio, CP 459, CEP 75132-903, Anápolis, GO (Brazil)

[3]*karineanjos06@hotmail.com,* [4] *barbbararocha@hotmail.com*
Federal University of Goiás, PPG Ecologia e Evolução, Campus Samambaia, Goiânia, GO (Brazil)

[8] *diniz@ufg.br*
Federal University of Goiás, Campus Samambaia, Goiânia, GO (Brazil)

[9] *ramiro.logares@gmail.com*
Institute of Marine Sciences, CSIC, Barcelona ES-08003 (Spain)

Introduction

The global number of papers published in different areas has increased over the years (King, 2004). Moreover, the science has experimented changes in academic production scenarios, such as decreased number of solo and increased team authors over the years (Nabout et al., 2015). For many a researcher the number of authors is one measure of collaborations (Price, 1958).

In fact the collaboration has promoted strong changes in science, and there are different reasons for collaboration: increased publication quality (Padial et al., 2010), and sharing costs and ideas (Vermeulen, Parker & Penders, 2013). For Ecology, complex questions such as global climate change, conservation plans of biodiversity among others, have promoted collaboration between scientists (Nabout et al., 2015). Moreover, there are different possible levels of colaboration and an important paper of Katz & Martin (1997) addresses this issue. For these authors, collaboration is: "Thus, a 'research collaboration' could be defined as the working together of researchers to achieve the common goal of producing new scientific knowledge." (Katz & Martin, 1997)

In general, the collaboration can be inter- or intra- at different spatial scales (e.g. national or international; intra or interinstitutional). This variation indicates levels of collaboration. Therefore, collaborations can occur between researchers from the same institution, between institutions of the same country and between different countries (Katz & Martin, 1997). Several methods have been proposed to measure the collaboration and using different units (researchers, institutes).

The aim of this study is to investigate the temporal trends of number of authors in Ecology journals between 1945 until 2014. Moreover, we will investigate the influence of level of collaboration (intra-institution - II; between-institutions - BI and between-countries - BC) in scientific quality (i.e. number of citation of paper). Our hypothesis is that collaborative papers (BC) generate more citations.

Data

To assess the number of authors and level of collaboration in Ecology papers, we selected all journals listed in category "Ecology" in Web of Science (www.isiknowledge.com, searched in February of 2015). We selected for this study only original articles (type of document), excluding notes, reviews, errata and others. We adopted this strategy to control the influence of type of document in the number of authors (Padial et al., 2010). The selection of papers considered all periods available in the Web of Science database (1945-2014). For collaboration analysis we consider only recent papers (2012-2014). For each paper, the following data were obtained: i) number of authors, ii) number of citations, iii) year of publication, and iv) the level of collaboration. For this last variable, papers were categorized according to the number of institutions of the authors and co-authors and their location. Therefore, authors affiliated with the same

institution were classified as intra-institutional collaboration (II); between-institutional in same country (BI) or between institution in different countries (BC).

Temporal Trends of Number of Authors

We found a total of 333,214 articles published in journals in the Ecology of Thomson-ISI between the years 1945 and 2014. The investigation of the number of authors per paper demonstrated a strong decay in the numbers of single-authored papers. In the early years, about 80% of papers in Ecology were single-authored. In 2014 this value is 4.8%. Statistical models suggest that in 2030 only 0.01% of papers will be single-authored (see Nabout et al., 2015). In addition, the number of papers with two authors have also declined from the beginning of the '90s. Therefore, recently there has been observed the increment in the number of papers with four and five authors, which enhances the tendency of multi-authored papers in Ecology. This trend has been observed in many other areas of science (Abt, 2007).

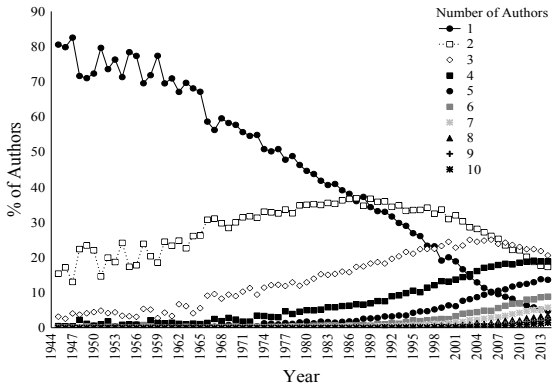


Figure 1. Temporal trends of the proportion of number of authors in Ecology Papers.

Levels of Collaboration

The papers of the years 2011, 2012 and 2013 exclusively of Ecology, totaling 10,457, were classified according to the level of collaboration (II, BI or BC). The Kruskal-Wallis (H) one-way analysis of variance by ranks was performed to assess if the number of citations is affected by the level of collaboration. We found a strong statistically significant difference ($P<0.01$), suggesting that collaborative papers written by authors from different countries received more citations Figure 2). This result reinforces the importance (and a recent trend) of international collaboration.

Using the same analysis we observed that the number of authors differs significantly between the levels of collaboration. In other words, BC papers have higher number of authors than those of SI and BI papers ($H = 1868$, $P < 0.001$). Therefore, the number of authors can also be an indication of the level of collaboration.

Finally, our work shows an increase in the number of multi-authored papers in Ecology. This is probably due to the complexity of questions in ecology which promotes collaboration between researchers. In addition, international collaborations have promoted papers with more citations (see Glänzel, 2001). Thus, the reduction of travel costs and the internet has allowed greater exchange between countries. In addition, governmental strategies can help in the exchange of researchers, such as the Program Science Without Border in Brazil. Thus, we encourage collaboration between researchers seeking to improve the ecological research of countries.

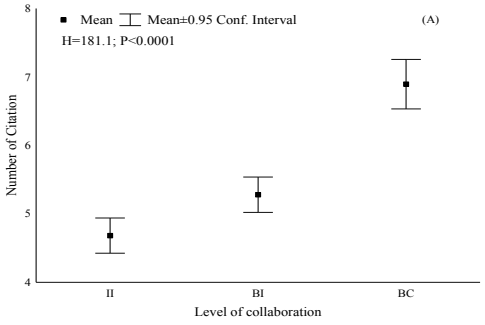


Figure 2. Number of citations for each one of level of collaboration.

Acknowledgments

Our work on Scientometrics and Ecology has been continuously supported by different grants FAPEG, CNPp and CAPES.

References

Abt, H.A. (2007). The future of single–authored papers. *Scientometrics*, *73*(3), 353–358.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics, 51*(1), 69-115.

Katz, S., & Martin, B.R. (1997). What is research collaboration? *Research Policy*, *26*(1): 1–18.

King, D.A. (2004). The scientific impact of nations. *Nature*, *430*, 311-316.

Nabout, J.C., Parreira, M.R., Teresa, F.B., Carneiro, F.M., Cunha, H.F., Ondei, L.S., Caramori, S.S. & Soares, T.N. (2015). Publish (in a group) or perish (alone): the trend from single- to multi-authorship in biological papers. *Scientometrics*, *102*, 357-364.

Padial A.A., Nabout, J.C., Siqueira T., Bini, L.M., Diniz-Filho, J.A.F. (2010). Weak evidence for determinants of citation frequency in ecological articles. *Scientometrics*, *85*,1-12.

Price, D.J.S. (1963). *Little Science, Big Science*. New York: Columbia University Press.

Vermeulen, N., Parker, J. N., & Penders, B. (2013). Understanding life together: A brief history of collaboration in biology. *Endeavour*, *37*(3), 162–171.

A Bootstrapping Method to Assess Software Impact in Full-text Papers

Erjia Yan[1] and Xuelian Pan[2]

[1] *ey86@drexel.edu*
Drexel University, College of Computing and Informatics, 3141 Chestnut Street, Philadelphia, PA 19104 (U.S.A.)

[2] *panxuelianmail@gmail.com*
Nanjing University, School of Information Management, Nanjing, 210093 (P.R. China)

Introduction and Motivation

There is a concerted effort to study science of science in multiple spheres. However, a clear gap exists in how to incorporate digital outputs, such as software, as an integral component in scholarly communication. This tension has become aggravated in recent years because software can be the end products in many scientific inquiries. Therefore, there is the need to build a framework to assess the impact of software in science. One cornerstone in the framework is the design of text-based methods to identify software entities in full-text corpora because these entities are largely mentioned in the text rather than formally cited in the way as their publications counterpart. This research-in-progress paper will serve this purpose by the development and evaluation of a bootstrapping method to automatically extract software entities from a full-text data set.

Despite the effort of indexing digital outputs such as Thomson Reuters' Data Citation Index or SageCite by University of Bath, U.K., the use of full-text data is necessary to identify patterns of software references because these digital outputs are referenced in unsystematical ways in scientific literature. They can be embedded in documents by digital object identifiers (DOIs), hyperlinks, and featured on dedicated websites or simply be mentioned in paragraphs, footnotes, endnotes, acknowledgements, or supplementary materials. A 2014 citation study on three oceanographic data sets showed that these digital outputs are more likely to be mentioned in the text than formally cited (Belter, 2014). Intuitively, one would think of curating a list of software names; however, it will not be feasible due to the velocity, variety, and volume of software that has been developed and applied constantly. Thus, merely using metadata or static listings is incapable of capturing the full extent of the impact of software. Instead, full-text publication data provide the crucial context for this purpose.

This study will use a bootstrapping method to identify software uses in a full-text data set. It will allow us to expand the impact and attribution mechanism by assessing the impact of software.

Methods

The bootstrapping method is used to extract software entities from full-text papers. It is a self-sustaining technique used to iteratively improve a classifier's performance through seed terms (Riloff & Jones, 1999; Riloff, Wiebe, & Wilson, 2003). The bootstrapping process contains the following steps: (1) Label seed terms or learned entities in the text. Seed terms are used in the first iteration, and learned entities are used in other iterations. (2) Generate contextual patterns of seed terms in the first iteration, and create contextual patterns of learned entities in other iterations. (3) Score these contextual patterns and select top ranked N patterns as candidate patterns. (4) Score entities extracted by candidate patterns and select top ranked M entities as learned entities. (5) Go back to the first step until the system cannot learn any new positive entities.

The calculation of pattern scores and entity scores determine the effectiveness of the bootstrapping method. If a pattern gets a higher score, then it is selected into the candidate pattern pool. Entities extracted by these candidate patterns are considered as candidate entities. To boost the performance, we incorporated three heuristic rules to the calculation of pattern scores. The first feature is an unlabeled entity containing at least one uppercase letter. An entity with this feature gets a score of 1 if it contains one or more uppercase alphabetic letters; otherwise, it gets a score less than 1. The second feature focuses on version numbers. An entity with this feature gets a score of 1 if a version number is collocated. The third and fourth features deal with the presence of trigger words: a score of 1 if the left context (third feature) or right context (fourth feature) of an entity contains trigger words.

Preliminary Results

To construct a corpus that has a good balance between sentences having software entity that mentions and does not mention, we selected 427 sentences that a particular software entity is mentioned from papers published between January 6 and December 29, 2013 in the data set. 573 sentences that do not contain software entities were also included in the corpus. We use this data collection method to attain a balanced experiment set to evaluate several entity extraction methods.

Experiments that use randomly sampled sentences will be pursued as future work. We used nine frequently occurring seed terms in the proposed bootstrapping method, including SAS, SPSS, MotIV, PAML, rGADEM, Limma, PICS, PHYLIP, and Minitab. To prepare the gold standard, we manually labeled software entities in the experiment data set and in total annotated 292 unique entities. The annotations are considered as the gold standard.

Table 1 displays the experimental results of the RlogF metric entity extraction system (Thelen & Riloff, 2002), Stanford Pattern-based Information Extraction and Diagnostics (SPIED), and our software extraction system. All methods in Table 1 used the same sets of seed terms, stop word list, and common word list.

Table 1. Experimental results of software extraction.

System	*Prec*	*Recall*	*F*
RlogF	91%	7%	0.12
SPIED	40%	28%	0.33
OurSystem	80%	62%	0.70

Table 1 shows that our system performed better than RlogF and SPIED based on the F score. Although RlogF has the highest precision, it missed a great number of software entities and resulted in the lowest recall. By comparing the software entities extracted by our system and the gold standard, we found seven of the one-time occurring entities were not identified by our system thus reducing the recall. We speculate that the recall may be improved when more sentences that contain low frequently occurring software entities are added to the data set such that the bootstrapping method will be able to learn their contexts.

Table 2. Popular software use in science.

Freq	*Software entities*
2	Prism, PASW, Vienna RNAfold, survival, Stata, SeqMan, rtracklayer, R2WinBUGS, Quantity One, PyPop, Origin, Microsoft Office Excel, JMP, GeneSpring GX, genefilter, FlowJo, Effective T3, Cytoscape, COMSTAT, CellquestPro, APE, ADE4, MetaMorph Imaging System
3	SigmaPlot, WinBUGS, T3SEpre, Statistica, MetaMorph, TiMAT2, stats, Statistical Package for the Social Sciences, STADEN, limma Bioconductor
4	HyPhy, IRanges, ImageJ, Affy, Vienna RNA
5	SigmaStat, MEGA, Vegan, Geneious
≥6	R, SAS, SPSS, MotIV, Bioconductor, Weka, PAML, rGADEM, Limma, PICS, PHYLIP, Minitab, Cellquest, RNAfold, Image J, GraphPad Prism

Table 2 shows 59 popular software entities in science which occurred more than once in the test corpus based on our extraction method. Statistical software packages are well presented in Table 2; however, we also see some domain-specific open access software tools—future impact assessment may primarily focus on these.

Conclusion and Future Work

The contemporary research landscape is changing: software has increasingly been developed and applied in many data-driven projects. Therefore, there is the need to assess its impact on science and to incorporate software in scientific evaluations. This paper is part of a larger effort to build a scientific assessment framework for digital outputs that include software and data. It has proposed a bootstrapping method to extract software entities in a full-text corpus. Results show that it has successfully extracted software entities with the F score at the 0.7 level which is an improvement over the baseline methods RlogF and SPIED. Future work will involve using the whole *PLOS ONE* full-text set and introducing more advanced features to further enhance the performance of the method. Research will also benefit from integrating the number of full-text software entity mentions with citation- and usage-based metrics to complement the impact assessment of software.

Acknowledgments

Erjia Yan is supported by the National Consortium for Data Science (NCDS) Data Fellows program for the project "*Assessing the Impact of Data and Software on Science Using Hybrid Metrics*". Xuelian Pan was a visiting PhD candidate at Drexel University, supported by China Scholarship Council, when this work was performed.

References

Belter, C. W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLOS ONE, 9*(3), e92590.

Riloff, E., & Jones, R. (1999). Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *Proceedings of* AAAI-99. Menlo Park, CA: The AAAI Press.

Riloff, E., Wiebe, J., & Wilson, T. (2003). Learning subjective nouns using extraction pattern bootstrapping. *HLT-NAACL Association for Computational Linguistics,* (pp. 25-32).

Thelen, M., & Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. *Proceedings of the ACL-02 Association for Computational Linguistics,* (pp. 214-221).

Article and Journal-Level Metrics in Massive Research Evaluation Exercises: The Italian Case

Marco Malgarini[1], Carmela Anna Nappi[1] and Roberto Torrini[1]

{marco.malgarini, carmelaanna.nappi, roberto.torrini}@anvur.it
ANVUR, Via Ippolito Nievo 35, 00153, Rome. (Italy)

Introduction

Article level metrics are usually the preferred choice for research evaluation. However, for recent articles they may be integrated or substituted considering some measure of journal impact (Abramo et al., 2012). The use of journal level metrics is also often considered as particularly appealing for administrative purposes, because of their readily availability, easiness to use and comprehensibility (Bordons et al., 2002). On the other hand, the IF is often criticized on the grounds of its possible biases and lack of methodological consistency (Vanclay, 2012). The aim of our paper is to provide evidence about the effects of the use of journal level metrics on the results of a massive research evaluation exercise like the one that has been performed in Italy with reference to the period 2004-2010 (VQR 2004-2010, see Ancaiani et al., 2015). More specifically, in the following we evaluate the effects of the use of the impact factor (IF) on the ranking of Italian Universities at the aggregate level, at the area level and for individual researchers.

Effect of the use of the Impact Factor at the University level

In order to assess the impact of the use of IF, we calculate two different indicators of research quality, denoted as R_VQR and R_IF. The former is based on the rules used for the VQR, and the latter uses only the Impact Factor in order to evaluate the articles; the analysis is limited to the research products evaluated only with bibliometrics. We then rank the 93 Italian Universities on the basis of those indicators, finding that the Spearman correlation index among the two rankings is equal to 0.92; moreover, the R^2 of a regression of R_VQR over R_IF and a constant is equal to 0.85. Hence, the analysis at the aggregate level shows that the final ranking of Italian Universities based on journal metric alone is very close to that obtained with the VQR algorithm (see also Figure 1).

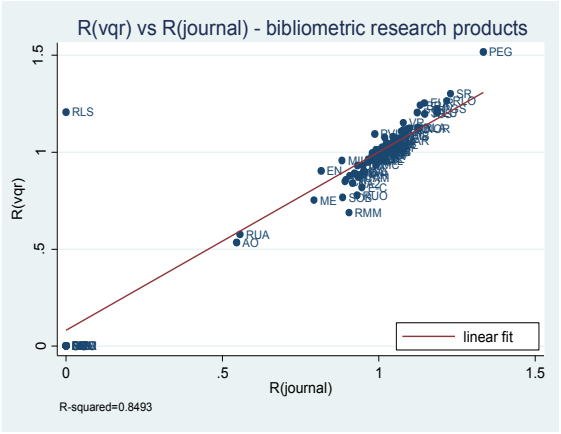


Figure 1 – The relationship among University evaluation performed with different metric.

Effect of the use of the Impact Factor at the Area level

However, it is well possible that the relationship is weaker when we are interested in ranking Universities in each scientific area. In order to shed light on this issue, we repeat the analysis for the 14 areas considered in the VQR (Table 1). Correlation between the two rankings is still above 0.8 in all the Research Areas except for Chemistry. The Spearman correlations among rankings are significant at 5% level in all the research areas.
Table 2 reports the coefficients of the regressions of R_VQR on R_IF (beta) and a constant (alpha); the table also reports the R^2 of the regression (column 3) and the standard deviation (column 4) normalized with respect to the average value of R_{iVQR} in each Area. Standard deviation is pretty low if compared to the average value of R (around 7%) in the Areas of Mathematics, Physics and Industrial Engineering, while in Earth Science, Medicine and Biology the normalized standard deviations grow to 17% of the average level of R in those areas. Similarly, the areas with a low normalized standard deviation are also whose with a higher R^2 and vice-versa. Hence, results confirm that the two evaluation methods bring very similar results also at the area level.

Table 1. Spearman Correlation between Rankings obtained with VQR bibliometric rules and Journal metric (* indicates statistical significance at 5%).

Research Area	Spearman	# Univ.
Mathematics	0.926*	64
Physics	0.825*	65
Chemistry	0.654*	60
Earth Science	0.724*	46
Biology	0.861*	66
Medicine	0.701*	58
Veterinary Sciences	0.876*	50
Construction engineering	0.720*	54
Industrial engineering	0.769*	67
Psychology	0.764*	61

Table 2. Sensitivity of research evaluation to the use of the Journal Impact Factor at the area level.

	(1)	(2)	(3)	(4)
Research Area	α	β	R^2	St. dv.
Mathematics	-0.055	1.039***	0.921	0.058
Physics	-0.13**	1.124***	0.847	0.060
Chemistry	-0.029	0.998***	0.706	0.100
Earth Science	0.180	0.815***	0.478	0.170
Biology	-0.142	1.132***	0.720	0.168
Medicine	0.083	0.894***	0.340	0.167
Veterinary Sciences	-0.004	1.016***	0.787	0.125
Construction engineering	0.186*	0.813***	0.532	0.100
Industrial engineering	-0.014	1.004***	0.675	0.070
Psychology	0.0778	0.916***	0.744	0.155

Effect of the use of the Impact Factor at the individual level

We finally look at how the use of the IF influences evaluation results for each h individual researcher. In this case, we regress individual scores obtained using either citations or the Impact Factor. Results of the estimation are reported in Table 3.

The relationship among the results obtained with the two different metrics is now rather weak: the R^2 of the regression is equal to 0.18 for the whole sample, varying between 0.20 and 1.156 in each year. The constant of the regression is rather high, while the beta coefficient associated with the IF is much lower than in previous estimates. Hence, at the individual level using alternatively only the citations or only the impact factor would imply a rather different outcome.

Table 3. Citations vs Journal Metric scores at individual level.

	Coefficient							
	Whole sample	*2004*	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*
IF	0.488***	0.525***	0.531***	0.521***	0.507***	0.487***	0.507***	0.383***
Constant	0.280***	0.247***	0.232***	0.254***	0.282***	0.301***	0.233***	0.374***
#obs.	76,15	9,23	9,77	10,24	10,88	11,56	12,15	12,31
R^2	0.184	0.201	0.197	0.202	0.197	0.194	0.186	0.156

Conclusions

Overall, results may be considered as supportive of the idea of using two different bibliometric indicators for assessing research quality: on one hand, the use of the IF is not found to bias in a significant way University rankings, both at the aggregate and at the Area level; on the other hand, at the individual level, citations and IF evaluation are found to be rather different, pointing to the need of integrating the two different information in order to obtain a more robust measure of research quality for each individual researcher.

References

Abramo G., D'Angelo C.A., & Costa F. (2012). Citations versus journal impact factor as proxy of quality: could the latter ever be preferable? *Scientometrics*, *84*(3), 821-833.

Ancaiani A. et al. (2015). Evaluating scientific research in Italy: the 2004-2010 Research evaluation exercise. Forthcoming in *Research Evaluation*.

Bordons, M., Fernández, M. T., & Gomez, I. (2002). Advantages and limitations in the use of impact factor measures for the assessment of research performance. *Scientometrics*, *53*(2), 195-206.

Vanclay J.K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification? *Scientometrics*, *92*(2), 211-238.

Accounting for Compositional Effects in Measuring Inter-Country Research Productivity Differences: The Case of Economics Departments in Four European Countries

Giannis Karagiannis[1] and Stelios Katranidis[2]

[1] *karagian@uom.edu.gr*, [2] *katranid@uom.edu.gr*
University of Macedonia, Department of Economics, Egnatia str. 156, 546 36Thessaloniki (Greece)

Introduction

Most of cross country studies on research productivity differences do not take into account compositional differences in academic staff force, such as sex, years of experience, origin of PhD studies, even though there are well documented evidence that (a) males tend to publish more than females (Gupta et al., 1999); (b) junior academic staff tend to publish more and in better outlets than senior stuff (Ben-David, 2010); and (c) academic staff with PhD studies in North America tend to be more productive (Katranidis et al., 2014). These aspects of observed faculty heterogeneity affect research productivity and are expected to have an impact on country average performance (Combes et al., 2003)[1].

Methodology and Data

In this paper we use the pure output or the single constant input DEA model, which is also known in the literature as the Benefit-of-the-doubt (BoD) model, to construct in the first stage a composite indicator of research productivity based on publication and citation counts at the faculty staff level. In particular, the BoD model in its multiplier form is given as (Cherchye et al., 2007):

$$\begin{aligned} I^k = \max_{s_i^k} & \sum_{i=1}^{N} s_i^k I_i^k \\ \text{st} & \sum_{i=1}^{N} s_i^k I_i^j \le 1^j \quad \forall j = 1, \dots, K \\ & s_i^k \ge 0 \qquad \forall i = 1, \dots, N \end{aligned} \tag{1}$$

where I_i^k is the i^{th} sub-indicator of the k^{th} unit, s_i^k are the weights to be estimated, j is used to index units and i to index sub-indicators which in our case correspond to different research outcomes (i.e., publication and citation counts). The BoD model is equivalent to the multiplier form of the input-oriented, constant returns to scale (CRS) DEA model when there is a single constant input that takes the value of one for all evaluated units. Based on this, the dual formulation of the BoD model is given as:

$$\begin{aligned} I^k = \min_{\lambda_j^k} & \sum_{j=1}^{K} \lambda_j^k 1^j \\ \text{st} & \sum_{j=1}^{K} \lambda_j^k I_i^j \ge I_i^k \quad \forall i = 1, \dots, N \\ & \lambda_j^k \ge 0 \qquad \forall j = 1, \dots, K \end{aligned} \tag{2}$$

where λ refers to intensity variables. Then the results at the country level are obtained by using the aggregation rule suggested by Karagiannis (2013), namely:

$$I = \frac{1}{K} \sum_{k=1}^{K} I^k \tag{3}$$

Thus, the aggregate composite performance indicator equals the simple (un-weighted) arithmetic average of the estimated individual composite indicators.

At the second stage we use Ray (1991) regression model to account for several contextual variables such as country dummies, a sex dummy, years of experience, and origin of PhD studies (i.e., overseas, Europe, home country and inbreeding), i.e.:

$$I^k = h\left(z_r^k\right) + e^k, \tag{4}$$

where r is used to index contextual variables and is $e^k < 0$ represents managerial inefficiency pure of (favorable and unfavorable) contextual variables. After taking into account the impact of contextual variables through (4) we re-calculate faculty level research performance scores and country averages. Our interest is to examine if and by how much these country averages differ from the unadjusted ones obtained via (1) or (2), and which countries are affected the most by the contextual variables.

We apply the above methodology to European faculty members in selected departments of Economics. In particular our sample consists of four countries, i.e., Belgium, Denmark, Greece and Portugal and a total of 383 faculty members and 15 departments. The analysis covers the period 1996-

[1] This research is implemented through the Operational Program "Education and Lifelong Learning" and is co-financed by the EU (European Social Fund) and Greek national funds.

2012 and the publication and citation count data come from Scopus database.

Empirical Results

Our main empirical results are summarized in the following tables:

Table 1. Unadjusted Composite indicator vs. efficient and unproductive faculty members.

	Unadjusted Composite indicator	Number of efficient faculty members	Number of unproductive faculty members
Belgium	0.144	1	6
Denmark	0.105	0	10
Greece	0.084	0	9
Portugal	0.062	1	18

Table 2. Number of unproductive faculty members vs. Adjusted Composite Indicator.

	Number of unproductive faculty members	Max value	Standard deviation	Adjusted Composite Indicator
Belgium	6	1	0.18	0.120
Denmark	10	0.588	0.11	0.100
Greece	9	0.667	0.10	0.086
Portugal	18	1	0.13	0.062

According to the unadjusted composite indicator, Belgian faculty members are found to be the more efficient and Portuguese the less efficient. In addition, in these two countries we can find the two fully efficient faculty members we have identified. At the same time these two countries are the ones with the relatively higher heterogeneity in terms of research productivity as indicated by the standard deviation of the unadjusted composite indicator.

When the composite indicator scores are adjusted for the potential impact of the aforementioned contextual variables by means of (4), the resulting efficiency scores change but not as much. They tend to improve a little bit for Belgium, Denmark and Portugal because these countries have a relatively higher percentage of inbred faculty members who in turn perform better compared to other faculty members. On the other hand, Portugal performance is adversely affected by the relatively larger percentage of females (31%) who though publish less than males and this counteract with the positive effect of inbred faculty, resulting in an unchanged national average.

Concluding Remarks

The empirical results indicate that the overall effect of the contextual variables considered is positive for the two northern European countries, i.e. Belgium and Denmark, and negligible for the two southern European countries, i.e., Greece and Portugal. Nevertheless, the two northern European countries perform better than the two southern European countries, regardless of environmental differences.

References

Ben-David, D. (2010). Ranking Israel's economists, *Scientometrics, 82,* 351-364.

Cherchye, L., Moesen, W., Rogge, N. & T. van Puyenbroeck. (2007). An Introduction to "Benefit of the Doubt" Composite Indicators, *Social Indicators Research*, *82*, 111-45.

Combes, P.P. & Linnemer, L. (2003). Where are the economists who publish? Publication concentration and rankings in Europe based on cumulative publications. *Journal of the European Economic Association, 1,* 1250-1308.

Gupta, B.M., Kumar, S., & Aggarwal,B.S. (1999). A comparison of productivity of male and female scientists of CSIR, *Scientometrics*, *45*, 269-289.

Katranidis, S., Panagiotidis, T., & Zontanos, C. (2014). An evaluation of the Greek universities' economics departments. *Bulletin of Economic Research, 66*(2), 173-182.

Karagiannis, G. (2013). On Aggregate Composite Indicators, unpublished manuscript.

Metrics 2.0 for Science 2.0

Isidro F. Aguillo

isidro.aguillo@csic.es
Cybermetrics Lab. IPP-CSIC. Albasanz, 26-28. Despacho 3E14. Madrid 28037 (Spain)

Science 2.0

The concept Science 2.0 is a recent development designed to take advantage of the new sharing technologies and social networks of the Web 2.0 and that it is now strongly linked to the current and future research policies of the European Commission.

According to ideas developed by Ben Shneiderman this Science in Transition can be described according to two groups of actions,

Integrating the whole research cycle and its stakeholders, including all and both activities and people involved in them, far beyond that focusing only on the authors of papers, and

Opening the whole set of data; tools, results and metrics derived from the cited research (and communication) cycle from the very first moment the information is generated.

The urgent need to adapt the current set of quantitative indicators to this new concept is the reason for this poster. We intend to provide a critical analysis of the current status of the bibliometrics y related quantitative techniques for science evaluation and to introduce a new umbrella term, Metrics 2.0, for describing future scenarios for the discipline.

Current Metrics situation

A SWOT analysis is introduced for describing major issues related to bibliometrics and the attitude of bibliometricians and the rest of scientists' attitude regarding the discipline.

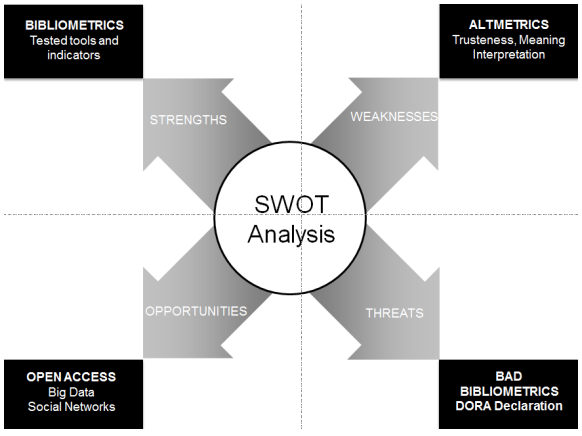


Figure 1. SWOT analysis of bibliometrics.

In recent years the term Informetrics become popular for describing an extended set of disciplines that are closely related to bibliometrics, including patentometrics or webometrics. However the fast development of the Internet, especially regarding the social networks, make this term become obsolete for describing a increasingly complex situation.

Specifically, there are two current developments that are having an impact on the discipline:

Altmetrics 'revolution'. The Web 2.0 tools have used as sources for extracting quantitative data when they are proxies for scholarly communication. Thousands of papers are exploring the capabilities of the different social networks using citation analysis for comparative purposes with mentions, readings or visits to bibliographic units.

Moving beyond Journal-level Metrics. After decades of criticism, and with the recent publication of the Declaration of San Francisco (DORA) the level of analysis is moving from Journal-level to Article-level metrics.

Proposals for Metrics 2.0

Regarding bibliometrics

The most serious problem is related to the way the contribution of each author (and the institution/s to whom is affiliated) is measured in a co-authored document. Traditionally two options were used: Full count (100% of merit for each author) and fractional counting (dividing full merit by the number of authors equally). As the number of authors per paper is growing exponentially, the last option is being discarded in most of the cases. Other alternatives, like identifying in the signature the relative contribution of each author, are still not a feasible option.

Traditionally full count is supported as it favours collaboration, especially international one. But this option is masking relevant phenomena for policy decisions. For example asymmetric collaboration with developing countries provides to their scientists and institutions with output/impact values that are not correlated with their low R&D investment prompting funders to not increase their budgets. Even with symmetric collaboration the full count based results are not able to discover the impact of the current economic crises that reduced

considerably the money invested in scientific research.
Taking into account that is a temporary proposal that intends not to reduce the level of scientific collaboration we suggest using a variant of the full count giving 50% of the merit instead of 100% to each author in papers with two or more authors.
In the case of organizations (and countries) where it is possible to identify the leading institutional author this should be granted the 100% authorship.
Although not a perfect or definitive solution this proposal should be especially useful for solving the problem of 'bad bibliometrics' that spoiled the major university rankings.

Regarding altmetrics

Apart of an ugly name, altmetrics is a confusing tangled set of mixed value tools. A first proposal could be to segregate the field in different subfields according to the tool that is involved. So, twittermetrics is different in both methodology and results interpretation to wikimetrics, for example. But there are two actions that are perhaps far more justified. It is highly recommended to set up a new discipline called Usagemetrics for the analysis of visits, visitors and their behaviour to academic and scientific websites. This is a very rich environment with dozens of candidate variables to build indicators independent from the standard citation motivations. The second moving is related to the tools where mention motivations are close to the citation ones, the most obvious one is Mendeley. In similar cases the proposal is to transfer these tools from altmetrics to the traditional bibliometrics arena.

Regarding Open Data y Big Data

The scientific community is strongly pushing for making openly available the data obtained from the experiments that is used later for preparing papers. Beyond the usefulness of this Open Data for replicating the results or for comparative purposes, the success of the initiative can make available huge amounts of information that could be considered, regarding the size-related challenges they pose, at the same level of the Big Data produced by the so-called Big Science. This is call for the scientific authorities for considering offering Big Data facilities and services for an extended group of scientists.

$$\text{Big Data} = \sum \text{Open Data}$$

Regarding Author Profiles

Until very recently the author-level metrics were technically a complex work when huge numbers of researchers were involved. Now the profiling services offered by several services (ResearcherID, Google Scholar Citations) or the major interests by the own research organizations (CRIS) and supported by disambiguation identifiers (ORCID) are changing completely the situation. In this new scenario, inspired by the results of the EU Project ACUMEN, we propose to set up author profiles with the following characteristics:
Bibliometric indicators from several sources, Non-bibliometric indicators, like those from altmetrics sources; context information like academic age, academic status, gender, levels of funding, networks membership and role, geographical or discipline biases, among others.
Rankings are a valuable tool if context is appropriately included in their elaboration. Relative indicators (percentages, quartiles) are being shown as far more trusted for this kind of classifications. However the use of composite indicators is still an open unresolved question that is still strongly criticised by the experts.

Conclusion

Metrics 2.0 should open and transparent, with data and indicators provided in a rich metadata environment.
Multiple sources and indicators are required, reflecting the diversity of the research activities, counting correctly and exhaustively the results and evaluating the different levels and magnitudes of the visibility and impacts of these results for all the communities, academic or not.
Presentation of the indicators, including friendly visualization of data is also relevant, but it is probably secondary to offer to end-users unrestricted customisation (including exporting in several formats) capabilities.
Summarising, bibliometricians can no longer been accountants able to extract, standardize, group and visualize the records from the Web of Science, but experts in several fields, with strong knowledge of different information sources and professionals capable of understanding specific needs and contexts ready to customise procedures according to the specific situation. Data, methodology, results and reports should be open to third parties in a mandatory way.

Evolution of Research Assessment in Lithuania 2005–2015

Saulius Maskeliūnas[1], Ulf Sandström[2] and Eleonora Dagienė[3]

saulius.maskeliunas@mii.vu.lt, ulf.sandstrom@indek.kth.se, eleonora.dagiene@vgtu.lt
[1]Vilnius University Institute of Mathematics and Informatics, Akademijos g. 4, LT-08663 Vilnius (Lithuania)
[2]KTH, Indek – Department of Industrial Economics and Management,
Lindstedtsvägen 30, 10044 Stockholm (Sweden)
[3]Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius (Lithuania)

Introduction

Traditionally, governmental funding of scientific research has been based on input factors (e.g. student numbers), however since the end of the 1980s most developed countries have introduced assessment systems based on scientific output. Numerous examples of research quality assessment can be named as products of innovation and incremental change (Barker, 2007; Hicks, 2012; RDI Council, 2013). An overview of assessment methods applied in Eastern European countries in the field of Social Sciences and Humanities has recently been presented (Pajic, 2015), but information about Lithuanian assessment of research is lacking. Here, we analyse seven sequential Lithuanian methods of research assessment in the period 2005–2015, their influence and consequences.

Evolution of Lithuanian research assessment methodologies

The methodologies of research assessment in Lithuania have changed very often over the period 2005–2015. There is quite a great difference between assessment of papers in Social Sciences & Humanities (SSH) and papers in Science & Technology (S&T). While SSH researchers should have publications in any peer-reviewed journals (Table 1), S&T papers have higher requirements: to gain scores, they have to be published in journals included in *Thomson Reuters Web of Science Core Collection* (WoS) (Table 2).

The value of each research article published in a journal indexed by WoS in SSH was calculated by the following formula in 2006 only:

$$AIV = PVV\frac{N_{IA}}{N_A}\left(1+\frac{IF_j}{IF_{AIF}}\right) \qquad (1)$$

here: AIV – contribution of institution authors; PVV – [primary] value of unit in points; N_{IA} – number of authors from the institution; N_A – total number of authors, IF_j – journal Impact Factor (Thomson Reuters Journal Citation Reports), IF_{AIF} – Aggregated Impact Factor of the subject category in which this journal is listed or average of Aggregated Impact Factors of all subject categories in case the journal is listed in more than one category in Thomson Reuters Journal Citation Reports.

The value of each research article published in a journal indexed by WoS in S&T (2003–2015) and SSH (2008 and 2015) is calculated by the similar formula:

$$AIV = PVV\frac{N_{IA}\sqrt{N_{IP}}}{N_A}\left(1+k\frac{IF_J}{IF_{AIF}}\right) \qquad (2)$$

here: N_{IP} – number of different foreign affiliations (but, if $N_{IP} > N_A$, then there is considered that $N_{IP} = N_A$); $k = 1$ for evaluation until 2007, and $k = 2$ for evaluation of 2008 and later years;

Significant and frequent changes in the evaluation criteria were caused by the search for most fair distribution of governmental funding for Lithuanian research by the Ministry of Science and Education, in order to encourage the highest-level academic research.

All systems of research assessment since 2006 have encouraged S&T researchers to publish their papers in high impact journals and have urged Lithuanian journals to improve their quality as well as actively seek to be indexed in international databases and especially in Thomson Reuters Web of Science. When Thomson Reuters started the expansion of the Web of Science in 2007–2009, many Lithuanian (LT) journals were included into its databases. But, the methodologies used in 2010 and 2011 were disadvantageous to most LT journals as they didn't fulfil the requirements asking *only* for papers in journals which had more than 20 % of citations from journals (citing side) with an impact factor (IF) higher than the aggregate impact factor (AIF) of the respective subject field. This requirement was probably not field neutral but, instead it seemed to be disadvantageous to some fields of science and created funding for other fields. Consequently, some subject fields were downgraded by this requirement and received no score or low scores. However, this citation requirement was not used for evaluation starting from 2012 and will formally withdrawn in 2015.

Since 2009 for SSH and from 2010 for S&T, expert evaluations (by national experts) of papers and monographs presented by institutions is used in addition to previous bibliometric evaluation. Since 2010 the number of 1st level papers and monographs presented by academic and research institutions for expert evaluation is proportional to number of full time equivalent of PhD researchers in both S&T and SSH (i.e., it could be presented not more than one 1st level publication per 5 full time researchers in a research area, and if the unit has doctoral studies in a research area – it can present 1st level publication not depending on number of researchers).

From 2011 the assessment system is carried out every third year (not annually as before). That helps aca-

demic and research institutions to minimize the drawbacks of productivity fluctuations. The last assessment period was 2009–2011. In 2015, there will be an evaluation of 2012–2014; which will determine the allocation of budgets for 2016–2018 for all universities and governmental research institutions. However it is rather complicated to evaluate the dynamics because of rather frequent changes in evaluation criteria. The benchmarking of Lithuanian research 2009–2013 was run on April 2014 – April 2015 by the Research and Higher Education Monitoring and Analysis Centre (MOSTA), following the methodology prepared by Technopolis Group and involving only international European experts. Here the experts have noticed the need for greater internationalization of Lithuanian Social Science research.

Conclusions

The shift in methodologies for formal assessment of scientific publications produced by Lithuanian higher education and research institutions has urged researchers to communicate their results in international scientific journals, and for the Lithuanian scientific journals to seek inclusion in international databases (especially Thomson Reuters Web of Science, Journal Citation Reports) and to improve their quality. The effect of changes in journals' indicators up until 2012 is the focus of a parallel poster presentation (Dagiene & Sandström, 2015). Whether the introduction of national expert evaluation will change this overall pattern or not is yet to be investigated.

References

Barker, K. (2007). The UK Research Assessment Exercise: the evolution of a national research evaluation system. *Research Evaluation*, *16*(1), 3–12.

Dagiene, E. & Sandström, U. (2015). Dynamics between National Assessment Policy and Domestic Academic Journals. Poster presentation submitted to *ISSI 2015*.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261.

RDI Council (2013). Methodology of Evaluation of Research Organizations and Evaluation of Finished Programmes (valid for years 2013–2015). Retrieved on March 20, 2015 from http://www.vyzkum.cz/FrontClanek.aspx?idsekce=695512

Pajic, D. (2015). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics*, *102*(3), 2131–2150.

Table 1. Shift in criteria used for Lithuanian research papers assessment in Social Sciences and Humanities.

2005		2006		2008		Assessment categories	2009		2010; 2011		2015	
Requirements	Value, points	Requirements	Value, points	Requirements	Value, points		Requirements	Value	Requirements	Value	Requirements	Value
Papers in internationally recognised journals	20 (24#)	*Papers in publications indexed by Thomson Web of Science*	30 (S)* 20 (H)*	*Thomson* [Reuters] *Journal Citation Reports (JCR)* IF ≥ 0	25**	1st level	**National expert evaluation** of papers presented by institutions as highest level	1–10 score	**National expert evaluation** of papers presented by institutions (*proportional to researchers' number*)	1–5 score	**National expert evaluation** of papers presented by institutions (*proportional to researchers' number*)	1–10 score
		Papers in internationally recognised journals	10	Papers in internationally recognised journals	15	2nd level	Papers in peer-reviewed journals	15 points	Papers in peer-reviewed journals & book chapters	3 points	*Thomson* Reuters *JCR* IF ≥ 0	3** points
		Papers in other peer-reviewed journals	5	Papers in other peer-reviewed journals	5						Papers in peer-reviewed journals & book chapters	2 points
Papers in other peer-reviewed journals	10 (12#)	Other papers, etc.	2–4	Other papers, etc.	2							
Other papers	4 (5#)						Other papers, etc.	5 points	Other papers, etc.	1–2 points	Other papers, etc.	1 point

– in research on Lithuanistics; * calculation by formula (1) ** calculation by formula (2)

Table 2. Shift in criteria used for Lithuanian research assessment of research papers in Physical, Biomedical and Technological Sciences (according to Lithuanian science classification).

Assessment categories	2005		2006 and 2008		2009		2010; 2011		2015	
	Req.. for a journal	Value, points	Requirements for a journal	Value, points	Requirements for a journal	Value, points	Requirements for a journal	Value	Requirements for a journal	Value
A-category papers 1st level	*Thomson ISI Master Journal List*	10	*Thomson* [Reuters] *Journal Citation Reports (JCR)* IF ≥ 0	30**	*Thomson Reuters JCR* with IF > 20% AIF	15**	**National expert evaluation** of papers presented by institutions *(proportional to researchers' number)*	1–5 score	**National expert evaluation** of papers presented by institutions *(proportional to researchers' number)*	1–5 score
					Thomson Reuters Web of Science (IF ≤ 20% AIF)	15**	*Thomson Reuters JCR with*: (1) IF > 20% AIF; (2) 20% citations from journals with IF > AIF	3** points	*Thomson Reuters JCR* with IF > 20% AIF	3** points
	90 citations from *Web of Science**	5#	*Thomson* [Reuters] *ISI Proceedings*	6	*Thomson Reuters ISI Proceedings*	15				
	Peer-reviewed journal	1	List of databases by the Research Council of Lithuania	6	Peer-reviewed journal	5				
			Peer-reviewed journal	5						
B-category papers (% of A-cat.) 2nd level	–		Physical sciences: B ≤ 0.1 A Biomedicine: B ≤ 0.2 A Technologies: B ≤ 0.3 A		Physical sciences: B ≤ 0.2 A Biomedicine: B ≤ 0.2 A Technologies: B ≤ 0.3 A					

paper published in any publication cited at least 90 times by journals listed in ISI the Master Journal List. Those citations are calculated since 1990 only. ** Calculation by formula (2).

Research-driven Classification and Ranking in Higher Education: An Empirical Appraisal of a Romanian Policy Experience

Gabriel-Alexandru Vîiu[1], Mihai Păunescu[2], and Adrian Miroiu[3]

[1] *gabriel.alexandru.viiu@snspa.ro,* [2]*paunescu.mihai@gmail.com,* [3]*admiroiu@snspa.ro*
National School of Political and Administrative Studies, Povernei Street 6, 010643 Bucharest (Romania)

Abstract
In this paper we investigate the problem of university classification and its relation to ranking practices in the policy context of an official evaluation of Romanian higher education institutions and their study programs. We first discuss the importance of research in the government-endorsed assessment process and analyze the evaluation methodology and the results it produced. Based on official documents and data we show that the Romanian classification of universities was implicitly hierarchical in its conception and therefore also produced hierarchical results due to its close association with the ranking of study programs and its heavy reliance on research outputs. Then, using a distinct data set on research performance we further explore the differences between university categories. We find that our alternative assessment of research productivity – measured with the aid of Egghe's *g*-index – only provides empirical support for a dichotomous classification of institutions.

Conference Topic
University Policy and Institutional Rankings

Introduction
Since the beginning of the 1980s nationally relevant university research coupled with the pressure for accountability have increasingly shaped the policies and priorities of individual universities (Geuna, 2001). Since then, the growing importance of research has been continually underscored by transnational policy documents such as the EU 2020 Strategy, by implementation of performance-based research funding mechanisms which create new competitive pressures within national university systems (Hicks, 2012) and, perhaps most visibly and controversially, by national and international university rankings which fuel debates surrounding 'world-class universities' (Sadlak & Liu, 2007; Salmi, 2009; Shin & Kehm, 2013). It is now well established that "international rankings of universities have become both popular with the public and increasingly important for academic institutions" (Buela-Casal et al., 2007, p. 351). At the same time rankings have also become "successful as an agenda-setting device for both politicians and for the higher education sector" (Stensaker & Gornitzka, 2009, p. 132).

In this paper we present an empirical exploration of the research-driven ranking and classification processes directed toward the Romanian higher education institutions (henceforth "HEIs") in the policy context of a new Law on National Education. In accordance with the new law a comprehensive process of evaluation was conducted in Romania in 2011 with the dual aim of (1) classifying HEIs (at the global, institutional level) and (2) ranking their constituent study programs. The ranking and classification were conducted using a common methodology that heavily emphasized the research productivity of university staff. Our primary objective is to contribute to a better understanding of the relation between the classification and ranking processes by discussing the methodological outline of the official evaluation and by analyzing its results. To achieve this goal we rely on official documents and on data collected with regard to the actual results of the classification and ranking processes. A secondary objective of our paper is to investigate the consistency of the institutional classification categories used in the official evaluation. To do this we employ an alternative data set on research performance, measured using the *g*-index which – for the set of papers of an individual researcher – represents "the largest rank (where papers are arranged in

decreasing order of the number of citations they received) such that the first *g* papers have (together) at least g^2 citations" (Egghe, 2006, p. 144). Our goal is to investigate whether an alternative assessment of research based on this index confirms the official classification of institutions, which was largely determined by research performance.

Background

Theoretical considerations

Higher education in recent years has witnessed the emergence of numerous university rankings, which have been the focus of comprehensive studies that aimed to investigate their methodological underpinnings, theoretical outlook and practical consequences (e.g., Dill & Soo, 2005; Salmi & Saroyan, 2007; Usher & Medow, 2009; Rauhvargers, 2011). In a more recent study Hazelkorn (2013) noted no less than 10 global rankings and at least 60 countries that have introduced national rankings. All these studies highlight (among other aspects) the fundamental importance that ranking systems generally attach to research performance, the deleterious consequences that rankings may have for institutional diversity and quality and, perhaps most importantly, the methodological caution which should be exercised when undertaking and interpreting rankings.

As more and more rankings have been developed over the years and as concerns have mounted regarding their implications and methodological problems (e.g.: van Raan, 2005; Billaut, Bouyssou & Vincke, 2010; Longden, 2011), the adjacent subject of university classification has also received increased attention (see for example Shin, 2009). This has been the case especially at the broader European level where the international ranking impetus has been critically received by scholars and policymakers and carried forward in a new direction with the introduction of the U-Map and U-Multirank initiatives, which, unlike pre-existing commercial rankings, focus on a user-driven approach and emphasize multidimensionality in evaluation.

Classification of universities has tended to be a much less debated subject than rankings, but these two distinct processes are nonetheless naturally interwoven with each other. On the one hand, due to strictures of comparability "classification is a prerequisite for sensible rankings" (van der Wende, 2008, p. 49). On the other hand, classifications are often interpreted as rankings even though this is clearly against the intentions of the classifying agency. Shulman (2005) and McCormick (2008) provide several examples of how the Carnegie Classification of US HEIs is actually understood as a form of ranking by several types of stakeholders.

A useful analytical distinction made between classifications and rankings involves conceptualizing them in the context of the broader notion of institutional diversity which itself may be divided into vertical diversity and horizontal diversity. According to van Vught (2009), the former refers to differences between higher education institutions owing to prestige and reputation while the latter stems from differences in institutional missions and profiles. In light of this distinction, classifications are "eminently suited to address horizontal diversity" (van Vught & Ziegele, 2011, p. 25) while rankings "are instruments to display vertical diversity in terms of performance by using quantitative indicators" (Kaiser, Faber & Jongbloed, 2012, p. 888).

The Romanian policy of classification and ranking

In 2011, following the provisions of the new law on national education a comprehensive national evaluation was conducted for the first time by the Romanian Ministry of Education with the aim of classifying all accredited HEIs and, additionally, of ranking all accredited study programs offered by the universities. This process was by far the most elaborate evaluation of the Romanian system of higher education and the first one to explicitly

undertake an official classification of HEIs and an official ranking of their study programs on the basis of quantitative indicators.

With regard to the classification process the law stipulated that all universities must be classified as belonging to one of the following three classes: *A – universities focused on education*; *B – universities focused on education and research*; and *C – universities focused on advanced research and education*. This would point toward a functional differentiation with regard to research capacity but the law also stipulated that the allocation of public funding was to be a function of the results of the classification process: universities from class A could only receive public funding for study programs at the bachelor level, those from class B could receive funding for programs at both bachelor and master level, while those from class C were the only ones to receive public funding for all types of programs (including PhD). With regard to the ranking of study programs, the law on education did not contain any detailed provisions. However, a subsequent government decision (789/03.08.2011) established five distinct hierarchical classes A (high quality), B, C, D and E (poor quality). These program ranking classes should not be confused with the university classes.

A detailed methodology for the classification and ranking processes was made public through Ministry of Education Order 5212/ 26.08.2012. This methodology outlined a complex system of criteria, performance indicators, variables and weights. Table 1 provides a simplified account of the evaluation methodology for the particular case of social sciences. At the most general level, four common criteria were used for both classification and ranking purposes: (1) research; (2) teaching; (3) relation to the external environment; and (4) institutional capacity. The most important aspect in the evaluation process was the research performance of the staff working in the universities and/or the study programs under assessment. This is especially significant for our later use of the *g*-index.

Table 1. Criteria, indicators and weights used in the evaluation process for university classification and study program ranking (social sciences).

Criteria and global weights	*Performance indicators and weights within criterion*	*Variables within indicator*
I. Research (weight: 0.50)	Results of scientific research - 0.75	11
	Research funding - 0.10	5
	International recognition - 0.02	2
	PhD programs - 0.13	2
II. Teaching (weight: 0.25)	/	6
III. Relation to external environment (weight: 0.20)	Relation to economic environment - 0.20	2
	Relation to social environment - 0.05	3
	Community development - 0.45	3
	Internationalization - 0.30	9
IV. Institutional capacity (weight: 0.05)	Indicator 1 - 0.34	3
	Indicator 2 - 0.11	3
	Indicator 3 - 0.11	4
	Indicator 4 - 0.11	4
	Indicator 5 - 0.11	4
	Indicator 6 - 0.11	1
	Indicator 7 - 0.11	5

Source: Ministry of Education Order 5212/ 26.08.2012

Within the research criterion four distinct performance indicators were defined but the most important of these four was an indicator dealing with the research output of the staff members employed by the universities. This indicator had a weight of 0.75 while the other three indicators (research funding, international recognition, and PhD programs) had much lower weights (0.10, 0.02, and 0.13). This indicator of research output was itself further broken down into 11 different variables such as the relative influence score of articles, the number of publications in journals indexed in the ISI Web of Knowledge, books, book chapters, etc.

For the ranking of study programs each university reported specific data for all of the distinct programs it operated; then, global indicators were calculated at the level of the study program for the first three criteria listed in Table 1. A separate global indicator was calculated at the university level for the institutional capacity criterion. A further step then entailed the calculation of an overall *aggregated index of ranking* (AIR) based on the four global performance indicators and their attached weights. As a final step in the ranking of a study program, its AIR was compared to the highest one obtained among all the similar study programs and, based on certain predefined intervals, it was finally designated as belonging to one of the five ranking classes.

For purposes of classification a separate *aggregated index of classification* (AIC) was calculated at the global level of each university. The AICs were calculated following a formula which incorporated three factors: (1) the absolute value of the research score obtained at the global level of the HEI; (2) a more complex factor calculated as a sum of the global indicators obtained by each of the study programs organized by the HEIs; and (3) an indicator based on the confidence level given to the HEIs by the Romanian Agency of Quality Assurance in Higher Education (ARACIS) following its periodic evaluations.

Upon calculation of the AICs of all universities the class of a particular HEI could finally be determined. Similar to the process used to establish the ranking classes of study programs, in order to determine a university's class its AIC was compared to the highest one obtained within its category (comprehensive universities were compared to other comprehensive institutions, specialized HEIs were only compared to their counterparts). First, universities were sorted in descending order of their AIC scores. Then, again following predefined intervals, universities were classified in one of the three categories A, B or C.

Without going into further details, it is apparent from even a brief analysis of the methodological outline that the evaluation conducted for purposes of classification actually had the general underpinning of a ranking. This is primarily a consequence of the fact that the classification was based on the composite scores of university performance (the AICs), which were sorted in descending order and clustered in accordance with predefined thresholds. Moreover, the classification relied on the research scores obtained by the constituent study programs of the universities and, therefore, on the partial results of the ranking process of these programs. In effect, research was the object of double counting, once at the individual level of the study programs and once more at the aggregated level of the HEIs. Based only on the analysis of the methodology used in 2011, we may argue that the entire classification process was actually hierarchical in nature and that vertical, not horizontal differentiation was a foreseeable consequence not only at the level of study programs (where ranking was explicit) but also with regard to the more general level of universities (where ranking was disavowed in favour of the more neutral label of 'classification'). However, no empirical analysis has so far been undertaken with regard to the relation between the actual results of the classification and the results of the program rankings. In addition, no independent empirical test of the three classification categories has been conducted, either relying on the performance indicators initially used by the Ministry, or on alternative measures of research performance. In the following paragraphs we will address both issues in an attempt to answer several questions related to the classification and ranking processes.

Research questions

Given the unique nature of the classification and ranking processes undertaken by the Romanian Ministry of Education several important aspects invite questioning and empirical study. We will confine our analyses to the following:

1. Did the overlap in methodology with the program rankings have empirically discernible consequences for the more general process of classification? Is there a significant degree of association between particular classes of universities and particular classes of study programs? If so, which types of programs are more common in which types of university?
2. Since the classification process relied heavily on research outputs, can an alternative assessment of the research productivity of universities confirm the threefold classification? Are there significant differences with regard to the research productivity of faculty members *between* the three university classes? Furthermore, are there significant differences with regard to the research productivity of faculty members *within* the three university classes?

The first set of questions addresses the official classification and ranking processes in tandem and implies an investigation of data on the official results. The second set of questions only addresses the classification process and will be explored using a distinct approach, which will be described in the subsequent section.

Methodology

In order to investigate our first set of research questions we created a comprehensive data set of the results of the ranking process for all the study programs evaluated in 2011. We then added the results of the classification of universities in order to obtain a final data set comprising all the study programs, the ranking class in which they were placed following the evaluation process and the class in which the university managing them was placed following the separate evaluation for classification. This primary data set contains 1056 observations of distinct study programs. To test for the level of association between ranking and classification results we created contingency tables for the occurrence of particular types of study programs (i.e. ranked in class A, B, C, D, E) in the three classes of universities (i.e. class A, class B and class C). Additionally, a chi-squared test was also used to investigate the association between the classification and ranking categories.

To explore the second set of research questions we used a distinct data set composed of information on 1,z385 Romanian faculty members active in the fields of political science, sociology and marketing. Specifically, we used their *g*-index to conduct an alternative assessment of university research output. These staff members represent the full populations of staff employed in political science, marketing and sociology study programs and they are spread out across 64 departments (study programs) and 34 distinct universities. Information on the identity of the staff members was obtained from ARACIS and, for each of the staff members in this second data set, the *g*-index was extracted using Anne Harzing's Publish or Perish software (Harzing, 2007) using a procedure previously employed in Vîiu et al. (2012) in an examination of political science departments. With regard to this secondary data set, the results of the official classification of Romanian HEIs would imply that there are significant differences between the staff employed in the three university classes with respect to their research output. To test this we employ analysis of variance and subsequent Tukey HSD tests to reveal the instances where differences between *g*-indices are significant. We first compare the university classes globally, and then refine our analysis to take into account more granular differences between staff types. We thus compare the four staff types – assistants, lecturers, associate professors and full professors – across the three university classes in order to determine whether or not there is a structural difference between these classes.

Results and Discussion

Relation between official ranking and official classification results

With regard to our first set of research questions a review of Table 2 and Figure 1 indicates that universities classified as being focused on education have a limited number of top-performing study programs (90 ranked in A and B, i.e. 17% of all study programs in this class of universities) but cluster the most programs with middle and low performance (those ranked in classes C, D and E add up to 83% of programs managed within the universities focused on education). On the other hand, universities focused on advanced education and research hold a total of 185 study programs and 121 of these (over 65%) are ranked in class A. Another 39 are ranked in class B (thus, over 86% of the programs in this class of universities are ranked in classes A and B) and only less than 5% belong to the lower performing classes D and E. Universities classified as being focused on both education and research have mixed results: out of a total of 344 study programs managed by these universities 189 (55%) are ranked in classes A and B, 28% are in class C and the remaining 17% are ranked in C, D and E.

Table 2. Contingency table of ranking classes of study programs and university classes.

University class	*A - Education*	*B - Education and research*	*C - Advanced research*	*Row total*
Class of study program in official ranking				
A	22 **4.17%**	60 **17.44%**	121 **65.41%**	203
B	68 **12.90%**	129 **37.5%**	39 **21.08%**	236
C	147 **27.90%**	97 **28.20%**	17 **9.19%**	261
D	112 **21.25%**	16 **4.65%**	3 **1.62%**	131
E	178 **33.78%**	42 **12.21%**	5 **2.70%**	225
Column Total	527 **100%**	344 **100%**	185 **100%**	1056

Chi-Square test of ranking classes of study programs and university classes

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	495.433	8	.000
N of Valid Cases	1056		

A more detailed study of the relationship between observed and expected count values of the different classes of study programs within each of the three university classes is also instructive. This study indicates a negative association between programs ranked in classes A and B and universities from class A. A further negative association can also be observed with regard to programs ranked in classes A, D, and E and universities from class B. Finally, universities from class C are negatively associated with study programs ranked in classes B, C, D, and E. On the other hand, a positive association exists between universities from class A and study programs ranked in classes C, D and E. A further positive association exists between universities from class B and programs ranked in classes B and C. Universities from class C are positively associated only with programs ranked in class A.

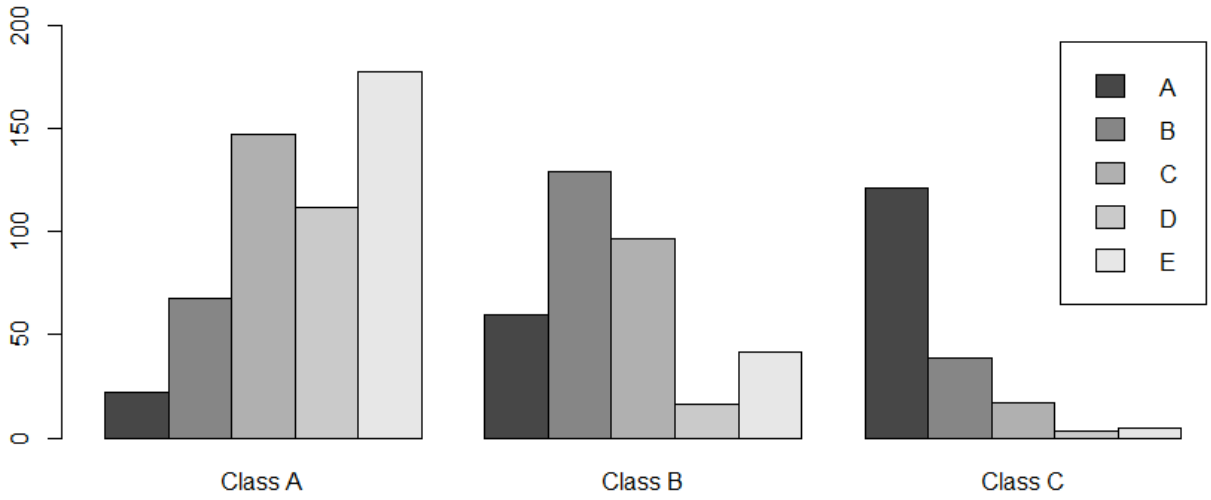


Figure 1. Distribution of study program types across the three university classes.

The results of this analysis paint a rather clear and polarized picture in which universities focused on education generally cluster study programs with poor performance while universities focused on advanced research cluster the programs with high performance. In addition, universities focused on advanced research are fewer and more selective (accounting for a total of only 185 study programs) as compared to universities focused on education (which manage a total of 527 programs). A certain hierarchy is implicit: universities focused on advanced research seem to be 'better' than those focused on both education and research which, in turn, are 'better' than those focused solely on education. However, as we mentioned earlier, these results were to be expected since both the classification and the ranking evaluation relied on a common methodology, which was mostly concerned with research performance. This leads us to our second set of research questions.

Differences in research productivity across and within university classes

We now move to explore whether our secondary data set enables us to distinguish between three university classes. In particular, what we want to see is whether the average *g*-index of all academic staff in class A universities is significantly lower than the average *g*-index of staff in class C universities and also in class B universities. The ANOVA procedure yields the results presented in Table 3. The subsequent Tukey HSD test indicates significant differences between all three means (although the confidence level for the class A – class B distinction is lower, but still above 95%) and therefore seems to provide empirical ground for the threefold classification, which was legally mandated in 2011.

Table 3. ANOVA of *g*-index with regard to university class (N=1,385).

Model summary for ANOVA of g-index with regard to university class

	Sum of Squares	df	Mean Square	F
Between Groups	953	2	476.3	81.62
Within Groups	8065	1382	5.8	Sig.
Total	9018	1384		0.000

Tukey HSD values for ANOVA of g-index with regard to university class

Comparison	Difference	Lower bound	Upper bound	*p*-value
Class A – Class C	-2.119	-2.513	-1.725	0.000
Class B – Class C	-1.714	-2.136	-1.293	0.000
Class B – Class A	0.405	0.054	0.756	0.019

However, the results presented in Table 3 only provide information on the global differences between university classes with regard to the *g*-indices of their entire staff, without further consideration of academic titles. Therefore, in order to test the consistency of the threefold model of classification imposed by the 2011 law, we must explore in greater depth the differences between universities, taking into account not only their classes, but also more granular differences between their academic staff. We thus set out to test not only the global aggregate differences, but also the *structural* patterns of the three classes of universities, taking into account the academic titles of the teaching staff.

In other words, bearing in mind the results of the official evaluation from 2011, we wish to know whether, for example, associate professors from class A universities are significantly different from associate professors in class B universities and from those belonging to class C and, still further, if the associate professors from class B institutions are different from those from class C. Similarly, we also wish to know whether assistants, lecturers and full professors from one class of universities are different from those belonging to the other two classes of universities. Based on such analyses we may draw more general conclusions regarding the degree of structural differentiation that exists between the three classes of universities.

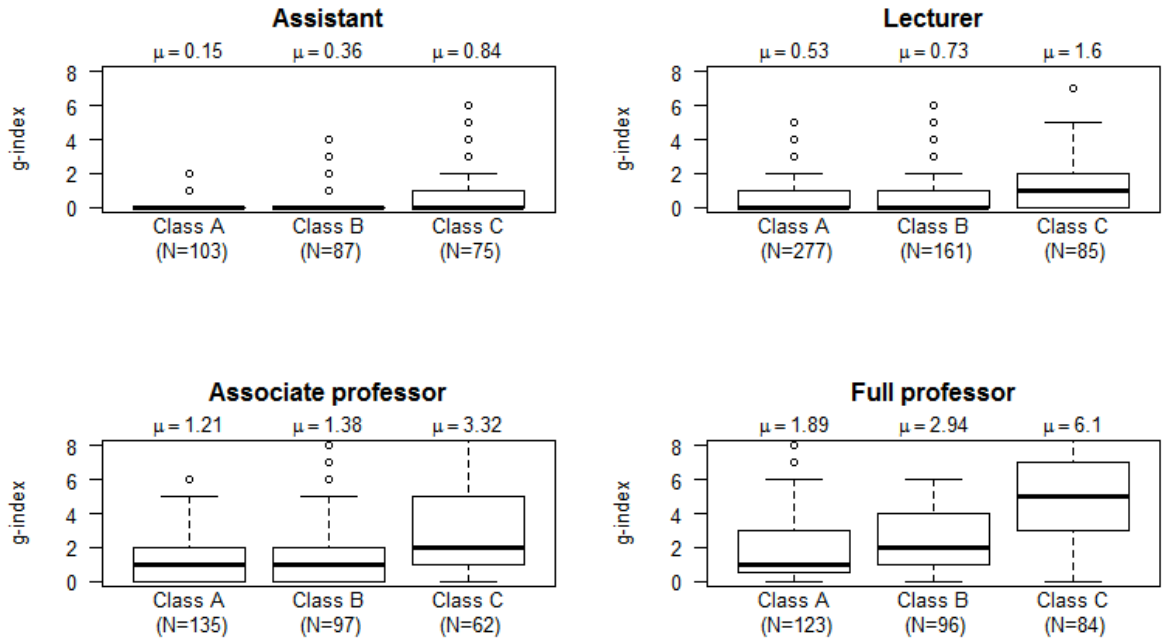


Figure 2. Distribution of g indices by academic title and university class.

Figure 2 illustrates the distribution of academic staff in our secondary dataset with respect to academic titles and also with regard to the class of university they belong to. Mean values are presented in the upper sections as μ. An initial visual inspection of the data would seem to indicate that in the case of assistants, lecturers and even associate professors there are no substantial differences between class A universities and those from class B. On the other hand, all three staff types working in class C universities seem to have substantially different *g*-indices compared to the ones from both class A and class B universities. A somewhat more nuanced picture emerges when looking at full professors. In this case the *g*-indices are more easily distinguishable between university classes and there indeed seem to be differences not only between class C and the other two university classes, but also between these two.

Based on the information contained in Figure 2 and on the ANOVA procedures presented in Appendix 1 we may now answer our secondary research questions. In the case of all staff members (be they assistants, lecturers, associate or even full professors) the parametric

statistical procedures show that universities classified within the official evaluation of 2011 as focused on advanced research (class C) are indeed significantly different from the other two types. In other words, assistants, lecturers, associate and full professors working in these universities focused on advanced research have significantly higher *g*-indices than their counterparts from education-centred universities, as well as from those in universities focused on both research and education. Beyond the clear distinction of staff members working in class C universities, statistical procedures also confirm something that Figure 2 reveals in a more intuitive manner: virtually no statistically significant distinction can be made between class A universities and the universities classified in 2011 as belonging to class B: assistant staff from class A universities are in no way significantly different form assistant staff working in class B universities, lecturers from one are in no way different from lecturers in the other and neither are associate professors. Even the apparent differences described by Figure 2 between full professors from class A universities and those from class B universities do not seem to be statistically meaningful either, as can be observed in Appendix 1. This suggests that a dichotomous classification would fit the data better than the threefold model imposed by the 2011 law.

So far we have argued that the data we have available clearly indicate significant *inter*-university differences (at least insofar as class C universities are made up of staff with higher indices than both class A and B universities). We now turn to *intra*-university differences. We have a reasonable expectation that within research universities there is a greater gap between the four staff types with regard to their scientific productivity. In other words, within class C universities we expect that the *g*-indices of assistants, lecturers, associate and full professors show greater dispersion than the corresponding indices of the equivalent staff that are employed in class A and class B universities. If we review the mean *g*-index values in Figure 2 we can observe that they appear to confirm our expectation. Whereas in the case of class A universities the gap between an average assistant and an average full professor is 1.74 and in the case of class B universities this gap is 2.58, in class C universities the difference is no less than 5.26. This indicates that full professors in research-centred universities have a substantially larger scientific contribution in their fields of study, not only when compared to staff employed in class A and class B universities, but also in comparison to their colleagues from the same university class. This suggests more competitive selection mechanisms of highly qualified academic staff in the research-centred universities compared to the other two university classes. These more competitive selection mechanisms may actually explain the institutional differences.

Concluding Remarks

The boundaries between classification and ranking of higher education institutions are often hard to establish and it is even harder to properly communicate the differences to intended stakeholders. When classification and ranking processes are carried out simultaneously and using common criteria the task of disambiguation becomes virtually impossible and the risk that a classification is perceived as a ranking increases exponentially. In the case of the evaluation conducted in Romania in 2011 the boundaries between classification and ranking were weak from the very inception of these evaluation processes in the law on education. The official methodology for classification and ranking further obscured the differences between the two due to its reliance on common criteria and indicators, most notably the research performance of academic staff employed by the HEIs.

By analysing the official methodology we have shown that the classification of Romanian HEIs carried out in 2011 had the underpinning of a ranking. By further analysing the results of both the classification and ranking processes we have shown that there is a clear association between the outcomes of the global process of classification and those of the more

specific process of program ranking: a polarized landscape thus emerges in which HEIs classified as focused on education cluster the overwhelming part of poor performing programs, while universities classified as focused on advanced research cluster the better part of the top performing programs.

The intermediate class of universities focused on both education and research presents mixed results. However, by conducting an alternative assessment of the research performance of the individual staff employed by Romanian universities in three fields of study we have shown that the threefold classification may not have a sufficiently robust empirical grounding, at least insofar as social sciences are concerned. By using the *g*-index as a concise measure of research performance we have illustrated the fact that the intermediate universities focused on both education and research may not be sufficiently distinct from the universities focused on education and therefore this intermediate class might have a certain degree of redundancy. When looking in our data set of 1385 staff members only at the aggregate results across university classes we do find some empirical grounding for the three classes defined in 2011. However, when analysing in greater detail the structure based on the academic titles and positions, we find less empirical grounds for the threefold classification as most of the staff employed in class A and class B universities are virtually indistinguishable from one another (i.e. assistants, lecturers and associate professors). It is only full professors that seem to make a more substantial difference between class A and class B universities, thus narrowly substantiating a threefold classification, which might otherwise well be a simpler dichotomous one. Previous extensive studies on the quality of Romanian higher education (Păunescu et al., 2012; Vlăsceanu et al., 2011; Miroiu & Andreescu, 2010) revealed the structural isomorphism of the Romanian higher education organizations. The undifferentiated set of standards that all institutions must comply with for purposes of accreditation and public funding led the institutions to adopt similar strategies for achieving these objectives. This is reflected in the poor differentiation and homogeneity of HEIs as shown by their similar scores in the external evaluation of the accreditation agency, similar missions, similar achievements on various performance indicators, etc. While there is empirical support for the vertical differentiation between advanced research universities (usually traditional, older universities) and the rest (more recent, including all private initiatives), the actual structures of the bulk of HEIs, including class A and class B universities, reveal more similarities than differences. These findings should of course be considered under the due caveat that our results are based only on data collected for social sciences.

Acknowledgments

Financial support from the National Research Council (grant number PN-II-ID-PCE-2011-3-0746) is gratefully acknowledged by Gabriel Vîiu and Adrian Miroiu.

References

Billaut, J.-C., Bouyssou, D. & Vincke, P. (2010). Should you believe in the Shanghai ranking? An MCDM view. *Scientometrics*, 84, 237–263

Buela-Casal, G., Gutiérrez-Martínez, O., Bermúdez-Sánchez , M.P. & Vadillo-Muñoz, O. (2007). Comparative study of international academic rankings of universities, *Scientometrics*. 71, 349–365

Dill, D. & Soo, M. (2005). Academic quality, league tables, and public policy: A crossnational analysis of university ranking systems. *Higher Education*, 49, 495–533

Egghe, L. (2006). Theory and practise of the G-index. *Scientometrics*, 69, 131–152

Geuna, A. (2001). The changing rationale for European university research funding: Are there negative unintended consequences? *Journal of Economic Issues*, 35, 607–632

Harzing, A.W. (2007). Publish or Perish, available from <http://www.harzing.com/pop.htm>

Hazelkorn, E. (2013). How rankings are reshaping higher education. In Climent, V., Michavila, F. & Ripolles, M. (Eds.), *Los rankings univeritarios: Mitos y realidades*. Ed. Tecnos

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41, 251–261

Kaiser, F., Faber, M. & Jongbloed, B. (2012). U-Map, university activity profiles in practice. In Curaj, A., Scott, P., Vlăsceanu, L., Wilson, L. (Eds.), *European Higher education at the Crossroads: Between the Bologna Process and National Reforms* (pp. 887–903). Dordrecht: Springer

Longden, B. (2011). Ranking indicators and weights. In Shin, J.C., Toutkoushian, R.K. & Teichler, U. (Eds.), *University Rankings. Theoretical Basis, Methodology and Impacts on Global Higher Education* (pp. 73–104). New York: Springer

McCormick, A. (2008). The complex interplay between classification and ranking of colleges and universities: Should the Berlin principles apply equally to classification? *Higher Education in Europe*, 33, 209–218

Miroiu, A. & Andreescu, L. (2010). Goals and instruments of diversification in higher education. *Quality Assurance Review for Higher Education*, 2, 89–101

Păunescu, M., Florian, B. & Hâncean, M.-G. (2012). Internalizing quality assurance in higher education: Challenges of transition in enhancing the institutional responsibility for quality. In Curaj, A., Scott, P., Vlăsceanu, L., Wilson, L. (Eds.), *European Higher education at the Crossroads: Between the Bologna Process and National Reforms* (pp. 317–338). Dordrecht: Springer.

Rauhvargers, A. (2011). *Global University Rankings and Their Impact*. Brussels: European University Association

Sadlak, J. & Liu, N.C. (2007). *The World-class University and Ranking: Aiming beyond Status*. Bucharest: UNESCO-CEPES

Salmi, J. (2009). *The Challenge of Establishing World-class Universities*. Washington DC: World Bank

Salmi, J. & Saroyan, A. (2007). League tables as policy instruments: uses and misuses. *Higher Education Management and Policy*, 19, 31–68

Shin, J.C. (2009). Classifying higher education institutions in Korea: A performance-based approach. *Higher Education*, 57, 247–266

Shin, J.C. & Kehm, B. (Eds.). (2013). *Institutionalization of World-class University in Global Competition*. Dordrecht: Springer

Shulman, L.S. (2005). Classification's complexities. *The Chronicle of Higher Education* (November 11, 2005), 52, p. B20

Stensaker, B. & Gornitzka, A. (2009). The ingredients of trust in European higher education. In Kehm, B.M., Huisman, J. and Stensaker, B. (Eds.), *The European Higher Education Area: Perspectives on a Moving Target* (pp. 125–139). Rotterdam: Sense Publishers

Usher, A. & Medow, J. (2009). A global survey of university rankings and league tables. In Kehm, B.M. and Stensaker, B. (Eds.), *University Rankings, Diversity, and the New Landscape of Higher Education* (pp. 3–18). Rotterdam: Sense Publishers

van der Wende, M. (2008). Rankings and classifications in higher education: A European perspective. In Smart, J. C. (Ed.) *Higher Education: Handbook of Theory and Research* (pp. 49–72), Vol. XXIII, Springer.

van Raan, A. F. J. (2005). Fatal attraction: conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62, 133–143

van Vught, F. (2009). Diversity and differentiation in higher education. In van Vught, F. (Ed.) *Mapping the Higher Education Landscape. Towards a European Classification of Higher Education* (pp. 1–16). Dordrecht: Springer

van Vught, F. & Ziegele, F. (Eds.).(2011). *Design and Testing the Feasibility of a Multidimensional Global University Ranking. Final Report*. Consortium for Higher Education and Research Performance Assessment, CHERPA-Network

Vîiu, G.-A., Vlăsceanu, M., & Miroiu, A. (2012). Ranking political science departments: the case of Romania. *Quality Assurance Review for Higher Education*, 4, 79-97

Vlăsceanu, L., Miroiu, A., Păunescu, M. & Hâncean, M.-G. (Eds.). (2011). *Barometrul calității 2010. Starea calității în învățământul superior din România.* Brașov: Editura Universității Transilvania din Brașov.

Appendix 1. Tests of difference for g-index across academic titles and university classes.

*1.Model summary for ANOVA of g-index of **assistant staff** with regard to university class*				
	Sum of Squares	df	Mean Square	F
Between Groups	20.68	2	10.341	13.29
Within Groups	203.82	262	0.778	Sig.
Total	224.50	264		0.000
Tukey HSD values for ANOVA of g-index of assistant staff with regard to university class				
Comparison	Difference	Lower bound	Upper bound	*p*-value
Class A – Class B	0.212	-0.090	0.515	0.224
Class C – Class A	0.684	0.369	1.000	0.000
Class C – Class B	0.472	0.144	0.799	0.002
*2.Model summary for ANOVA of g-index of **lecturers** with regard to university class*				
	Sum of Squares	df	Mean Square	F
Between Groups	73.7	2	36.85	25.39
Within Groups	754.8	520	1.45	Sig.
Total	828.5			0.000
Tukey HSD values for ANOVA of g-index of lecturers with regard to university class				
Comparison	Difference	Lower bound	Upper bound	*p*-value
Class A – Class B	0.195	-0.085	0.475	0.232
Class C – Class A	1.062	0.710	1.413	0.000
Class C – Class B	0.867	0.487	1.246	0.000
*3.Model summary for ANOVA of g-index of **associate professors** with regard to university class*				
	Sum of Squares	df	Mean Square	F
Between Groups	204.8	2	102.40	24.44
Within Groups	1219.2	291	4.19	Sig.
Total	1424			0.000
Tukey HSD values for ANOVA of g-index of associate professors with regard to university class				
Comparison	Difference	Lower bound	Upper bound	*p*-value
Class A – Class B	0.166	-0.475	0.808	0.813
Class C – Class A	2.107	1.367	2.847	0.000
Class C – Class B	1.941	1.157	2.725	0.000
*4.Model summary for ANOVA of g-index of **full professors** with regard to university class*				
	Sum of Squares	df	Mean Square	F
Between Groups	914	2	457.0	34.83
Within Groups	3936	300	13.1	Sig.
Total	4850			0.000
Tukey HSD values for ANOVA of g-index of full professors with regard to university class				
Comparison	Difference	Lower bound	Upper bound	*p*-value
Class A – Class B	1.053	-0.108	2.215	0.084
Class C – Class A	4.212	3.005	5.420	0.000
Class C – Class B	3.159	1.884	4.433	0.000

Looking beyond the Italian VQR 2004-2010: Improving the Bibliometric Evaluation of Research

Alberto Anfossi[1,2], Alberto Ciolfi[1], Filippo Costa[1,3]

[1]albertofrancesco.anfossi@anvur.it, [1] alberto.ciolfi@anvur.it, [1,3]filippo.costa@anvur.it
[1]ANVUR, Via Ippolito Nievo 35, 00153 Roma (Italy)
[2] Compagnia di San Paolo Sistema Torino, Piazza Bernini 5, Torino (Italy)
[3]Dipartimento Ingegneria dell'Informazione, Università di Pisa, Pisa (Italy)

Abstract
In the recent Italian Evaluation of Research Quality exercise for the period 2004-2010 (VQR), promoted by the Italian Ministry of Education and carried by the National Agency for Research Evaluation (ANVUR), metrics were massively employed. The use of Impact Factor or article citations (or both) are usually considered a powerful tool for supporting the peer review process but the replacement of the latter with an automatic evaluation tool has been always considered risky. Here we propose a possible prescription to overcome some limitations of the bibliometric evaluation carried out within the context of the VQR, while, at the same time, keeping the main distinctive features of the evaluation approach unchanged, namely, a simple evaluation tool based on the combined use of the CIT and IF variables While maintaining the basic elements of the previous algorithm unchanged and keeping the method simple and feasible on a large scale, we argue that the main flaws and limitations can be overcome.

Conference Topic
University Policy and Institutional Rankings

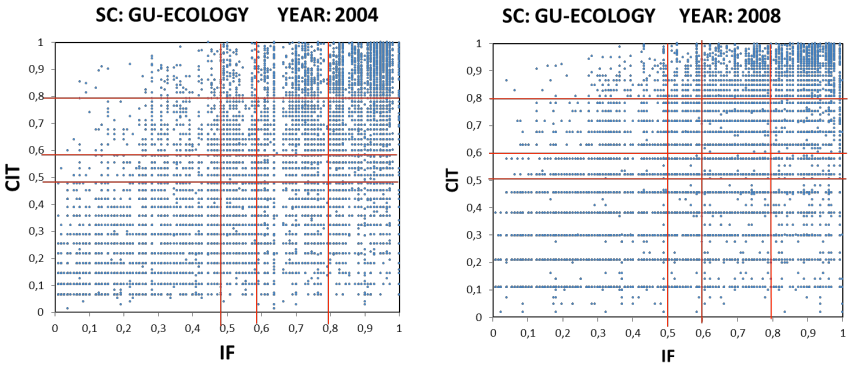
Introduction

The most popular European national research evaluation is the Research Assessment Exercise (RAE) in Great Britain, which started in 1986 and has been replaced, in 2013, by a new exercise - Research Excellence Framework (REF) - where citation-based metrics were employed to inform and supplement Peer Review (PR) evaluation.
In Italy, the first evaluation exercise was carried out in 2005 by the CIVR with reference to the period 2001-2003 (VTR). The VTR was fully based on the PR evaluation method, each submitted research product being assessed by a pool of experts (Minelli et al., 2008). Some studies (Reale et al. 2007; Abramo et al., 2009; Franceschet et al., 2011) analysed the outputs of the VTR comparing peer quality opinions on papers with metrics based on the Impact Factor of the journals publishing the papers, concluding that the two evaluation methods significantly overlap. However, comparison of PR and bibliometric evaluation methodologies has been largely debated in the literature (Barker, 2007; Moed, 2006; Harnad, 2009; Norris et al. 2003, Butler et al., 2003; Bence et al., 2009, Asknes, et al. 2004) with not always concordant outcomes. The use of Impact Factor or article citations or both are usually considered a powerful tool for supporting the PR process but the replacement of the latter with an automatic evaluation tool has been always considered risky.
In the recent Italian Evaluation of Research Quality exercise for the period 2004-2010 (VQR), promoted by the Italian Ministry of Education and carried by the National Agency for Research Evaluation (ANVUR), metrics were massively employed. Around 200.000 research outputs, mainly journal articles or reviews (both called 'paper' in the following), were evaluated, of which 46,5% by use of a bibliometric algorithm (Ancaiani et al., 2014).

Bibliometric Evaluation in the VQR 2004-2010

According to the Ministerial Decree number 17 of July 15th, 2011 promoting the VQR, each paper submitted for evaluation is classified in one of four possible classes of merit, defined as follows: "Excellent" (E): when the paper falls in the top 20% of the world production in a given Subject Category (SC) and in a given year; "Good" (G): when the paper falls in the following 20%; "Acceptable" (A): when a paper falls in the following 10%; "Limited" (L): when a paper falls in the bottom 50%.

In bibliometric areas, the strategy to assign a paper to a given class was based on the combined use of two variables: (i) CIT: number of citations collected by the paper up to December 31st, 2011 and (ii) IF: Impact Factor (or equivalent indexes) of the Journal in the year of publication of the paper. Each paper was submitted by the Organization (i.e. universities or public research bodies) and then uniquely assigned to a thematic evaluation panel (called Group of Experts for Evaluation, GEV) and to a Subject Category (SC), or All Journal Science Category (ASJC) as defined by ISI Web of Knowledge® or Scopus databases, respectively. In each SC/ASJC and for each year it is possible to construct the cumulative distribution function of the two variables[1], thus assigning to each paper its CIT and IF percentile. In the VQR three thresholds for both IF and CIT were defined to distinguish among the four classes specified in the Ministerial Decree. In the space spanned by IF (x-axis) and CIT (y-axis) it is therefore possible to focus on the region Q = [0,1]x[0,1] and plot the publications distribution defined on the basis of their CIT and IF percentile (Fig. 1, where each dot represents a paper denoted by its CIT and IF percentile). Each GEV had the freedom to assign the "off-diagonal" sub-squares (blocks) of the whole region Q, identified by the intersection of the "threshold segments", to a class of merit, thus completing the automatic phase of the evaluation process. Indeed, the diagonal blocks were quite naturally assigned to the four classes: the intersection of "top 20% for CIT" with "top 20% for IF" was straightforwardly associated to the "Excellent" class of merit, and so on. The choice to assign an off-diagonal block to a class was performed according to basically two drivers: first and foremost, the qualitative insight of the GEV on the scientific field and its publication practices (e. g. lag in citations, etc.) and second, the attempt to keep the final assignment as close as possible to the world distribution D specified in the Ministerial Decree.



(a) (b)

Figure 1. Papers distribution in a given SC and in two different years.

Such an approach showed some limitations that we summarize schematically:

[1] CIT: by ordering the total number of paper published in that SC and in that year in decreasing order from the highest to the lowest cited; IF: by ordering the Journals belonging to that SC in that year in decreasing order from the highest impact factor to the lowest.

Absence of "micro calibration": all the GEVs except for GEV 02 (Physical sciences) chose a single assignment (typically, one for years 2004-2008 and one for years 2009-2010), i.e., association of blocks to classes of merit, and did not went through a micro calibration at the level of the single SC and single year. Considering that: (i) for each GEV the number of relevant SC[2] was typically of the order of 50 and (ii) the distribution of the papers in Q was totally not uniform and invariant, rather, it varied significantly from one SC to another and form one year to another (see for instance Fig. 4). The absence of a micro calibration affected the possibility to comply with the distribution D punctually (and not only on average).
Structure of the blocks: (i) as showed in Figure 1 the threshold segments are parallel to the x/y axis. This is not convenient given the discrete nature of the two variables under consideration. (i) It can be easily noted in the plot that the points (corresponding to papers) are distributed in rows, according for instance to the limited number of journals present in a SC. As a consequence, the evaluation may not be robust enough, in the sense that a slight perturbation in the thresholds can modify the final class allocation for whole set of papers. (ii) It is quite hard, if not impossible, to comply with the distribution D by leveraging on the sole degrees of freedom given by the possibility to assign the off-diagonal blocks to a final class of merit. In other words, the constraint of assigning to a single class an entire block is too binding and tends to move too many paper from one class of merit to an another. (iii) The degrees of freedom are even reduced by the need to avoid that two non-adjacent classes of merit (say, "Good" and "Limited") can be adjacent in Q, as shown in Fig. 2.

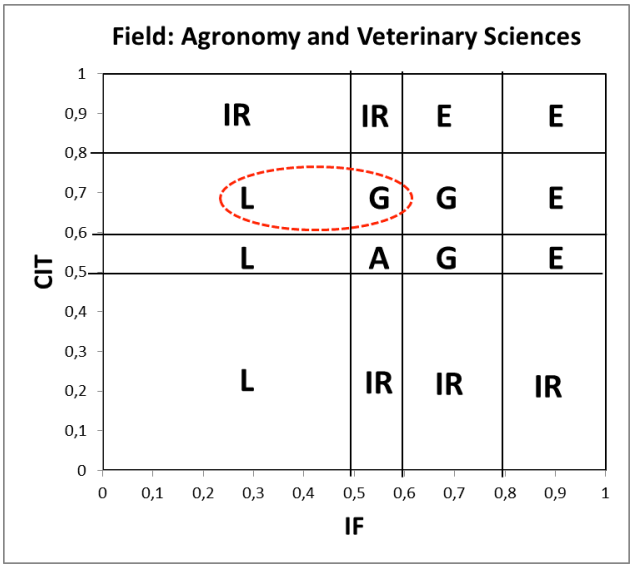


Figure 2. Algorithm used to evaluate research products in the Agronomy and Veterinary Science field: two non-adjacent classes of merit are adjacent in Q (red circle). "IR" indicates products that are lefd undecided by the algorithm and are eventually evaluated by peer review.

The new proposed approach

In the following we discuss a possible prescription to overcome these limitations while, at the same time, keeping the main distinctive features of the evaluation approach unchanged, namely, a simple evaluation tool based on the combined use of the CIT and IF variables. This can be done through the use of three diagonal segments with generic slope (Fig. 3).

[2] By relevant we mean that a great number (more than one hundred) of papers to be evaluated fell under that SC.

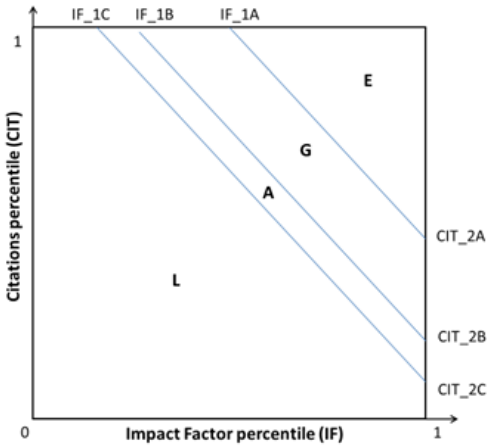


Figure 3. New prescription for combining the CIT and IF variables.

Such a new prescription builds upon three main pillars:

1. The segments identifying the thresholds are now drawn as a linear combination of the CIT/IF thresholds, thus being diagonal and no more parallel to the axes;
2. CIT/IF thresholds do not have to separately satisfy the 20-20-10-50 distribution;
3. The calibration, i.e. where to position the diagonal segments in Q in order to comply with the distribution D, is now performed at the micro level of each SC, for each year and for each GEV (according to general guidelines provided by the GEV itself and based on GEV's proficiency in the specific scientific field);

This would in turn guarantee the effectiveness and the simplicity of the whole process. In Figure 4 we apply this method to some SCs.

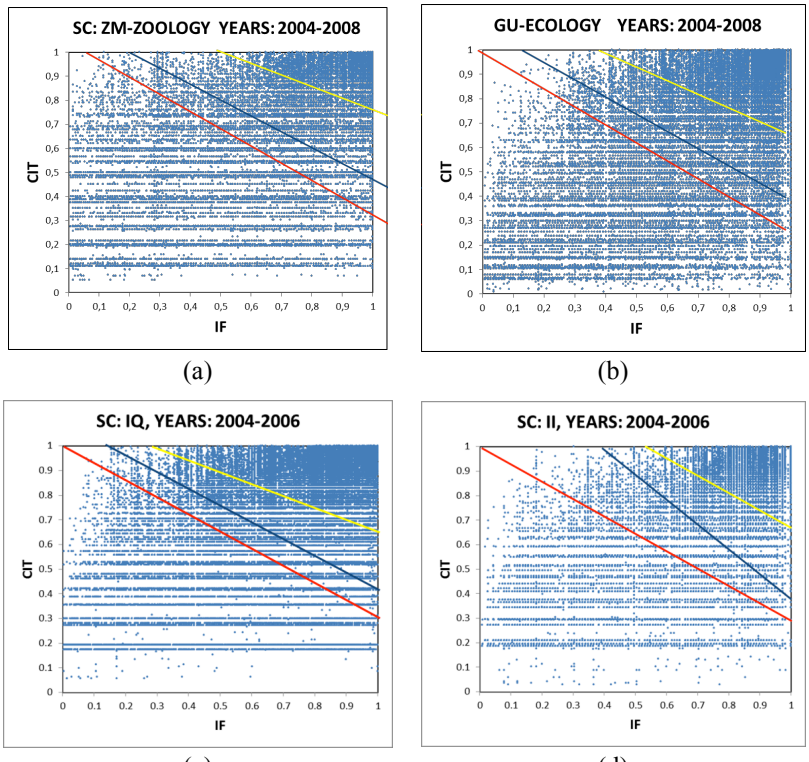
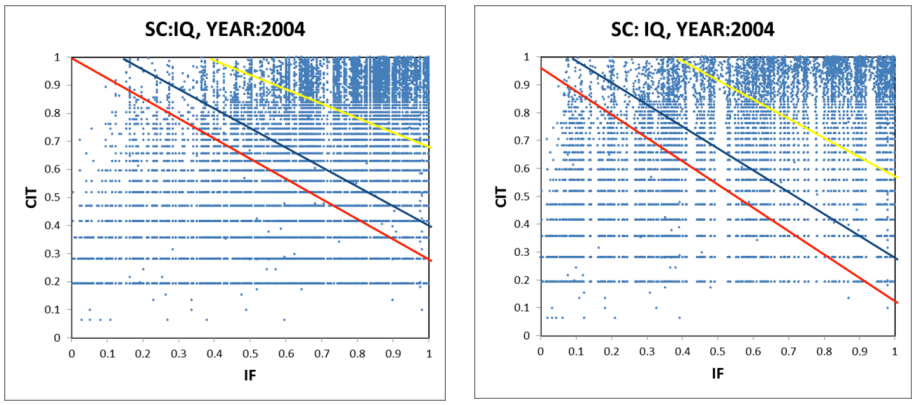


Figure 4. The application of the new algorithm in various SC and years. IQ stands for Electrical and Electronic Engineering, II stands for Engineering Chemical. The straight lines indicate the thresholds for the four classes of merit.

Comments and future developments

This new approach is characterized by a rather marked level of freedom in the choice of the position of the diagonal segments (or, equivalently, of the CIT/IF thresholds). Indeed, there is typically more than one choice that satisfies the distribution D. On the other hand one could impose additional constraints, such as for instance the parallelism between segments, based on additional empiric work and on scientific validation of the procedure (eg. by a PR comparison of the evaluation outcomes). Furthermore, such a freedom might be exploited to accommodate GEV's requirements. For instance, it would be possible to give more relevance to one of the two dimensions (IF, CIT) depending on, say, the year of publication or the citation praxes of specific disciplines (Mathematics vs Medicine being a paradigmatic example).

A significant possibility to further improve the accuracy of the method we discussed comes from a different definition of the cumulative distribution function for the IF variable. Instead of considering the number of journals belonging to a SC, one could consider the number of items (papers) published in the SC (in a given year). Actually, it is common that some journals host few thousands of items per year while other few tens or units. This induces a possible distortion that is quite evident in the plots shown below. As an example, In Figure 5 we analyze the distribution of the SC Electrical and Electronic Engineering in 2004. The distribution of the papers according to the IF and CIT percentile are depicted both considering only the number of journals in the calculation of the IF percentile and by considering also the number of item for each journal. The distributions are subdivided with different lines in order to obtain the target percentages D. It is evident that the equation of the lines is substantially different to guarantee the same final result. It is worth underlining that the lines used to subdivide the distribution reported in Figure 5(a) would result in very different percentages if applied to the distribution in Figure 5(b).



(a) (b)

Figure 5. Distribution of the papers according to the number of journals and papers. (a) IF percentile calculated based on the number of journals (b) the IF percentile is calculated considering the number of items. The distributions are subdivided with lines in order to obtain the target percentage D.

Finally, it would be possible to improve also the CIT dimension by overcoming the concept of SC as "reference set" and move on to clustering strategies based on semantic or on citation networks. This would be more rigorous and meaningful considering the existence of a great number of journals that publish very different subjects, but it would come with a significant enhancement of the complexity of the evaluation procedure, probably not feasible for the numbers implied by a national formal evaluation, at the moment.

Results obtained so far are already highly informative about the existing strength and weakness of the Italian University research system, and provide reliable input for policy interventions. Our proposal is intended to further improve the mix of peer review and bibliometric methods through a more precise calibration of the biblio(metrics) used.
The output turns out to be rather general, thus being applicable to other national assessments based on bibliometric analysis.

References

Ancaiani et al. (2014). Evaluating Scientific Research in Italy: the 2004-2010 Research Evaluation Exercise. *Research Evaluation*, *submitted*

Minelli, E., Rebora, G., Turri, M. (2008). The structure and significance of the Italian research assessment exercise (VTR). *European Universities in Transition*. Edward Elgar Publishing.

Reale, E., Barbara, A., & Costantini, A. (2007). Peer review for the evaluation of academic research: lessons from the Italian experience. *Research Evaluation*, *16*(3), 216-228.

Franceschet, M., & Costantini, A. (2011). The first Italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics*, *5*(2), 275-291.

Abramo, G., D'Angelo, C. A., & Pugini, F. (2008). The measurement of Italian universities' research productivity by a non parametric-bibliometric methodology. *Scientometrics*, *76*(2), 225-244.

Barker, K. (2007). The UK Research Assessment Exercise: the evolution of a national research evaluation system. *Research Evaluation*, *16*(1), 3-12.

Moed, H. F. (2006). Citation analysis in research evaluation (Vol. 9). Springer.

Harnad, S. (2009). Open access scientometrics and the UK Research Assessment Exercise, *Scientometrics*, *79*(1), 147-156

Norris, M., & Oppenheim, C. (2003). Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. *Journal of Documentation*, *59*(6), 709-730.

Butler, L., & McAllister, I. (2009). Metrics or peer review? Evaluating the 2001 UK Research assessment exercise in political science. *Political Studies Review*, *7*(1), 3-17.

Bence, V., & Oppenheim, C. (2004). The influence of peer review on the research assessment exercise. *Journal of Information Science*, *30*(4), 347-368.

Asknes, D. W., Taxt, R. E. (2004). Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. *Research evaluation*, *13*(1), 33–41.

High Fluctuations of THES-Ranking Results in Lower Scoring Universities

Johannes Sorz[1], Martin Fieder[2], Bernard Wallner[2] and Horst Seidler[2]

johannes.sorz@univie.ac.at
[1]University of Vienna, Office of the Rectorate, Universitätsring 1, A-1010 Vienna (Austria)

martin.fieder@univie.ac.at, wallner@univie.ac.at, horst.seidler@univie.ac.at
[2]University of Vienna, Department for Anthropology, Althanstrasse 14, A-1090 Vienna (Austria)

Abstract
A regression analysis of results from the Times Higher Education World University Rankings (THES-Ranking) from 2010-2014 shows high fluctuations in the rank and score for lower scoring universities (below position 50) which lead to inconsistent "up and downs" in the total results. We conclude that these fluctuations do not correspond to actual university performance. They create the impression of the THES-Ranking as a "gamble" for universities below rank 50. We suggest that THE alters its ranking procedure insofar as universities below position 50 should be ranked summarized only in groups of 25 or 50. Additionally, we argue for introducing a standardization process for THES-Ranking data by using common suitable reference data to create calibration curves represented by non-linearity or linearity.

Conference Topic
University Policy and Institutional Rankings

Introduction
Global higher education rankings have received much attention recently and, as can be witnessed by the growing number of rankings being published every year, this attention is not likely to subside. Besides the arguable use of results from global rankings as an instrument for rational university management, they remain influential for stakeholders inside and outside academia. A plethora of regional and national rankings exist, and 10 global higher education rankings are currently attempting to rank academic institutions worldwide. Numerous studies have analyzed and criticized higher education rankings and their methodologies (van Raan, 2005; Buela-Casal et al., 2007; Ioannides et al., 2007; Hazelkorn, 2007; Aguillo et al., 2010; Benito and Romera, 2011; Hazelkorn, 2011; Rauhvargers, 2011; Tofallis, 2011; Saisana et al. 2011; Safon, 2013; Rauhvargers, 2013). This casts justified doubt on a sensible comparison of universities hailing from different higher education systems and varying in size, mission and endowment based on mono-dimensional rankings and league tables (Hazelkorn, 2014). Several studies have demonstrated that data used to calculate ranking scores can be inconsistent. Thus, bibliometric data from international databases (Web of Science, Scopus), used in most global rankings to calculate research output indicators, favor universities from English-speaking countries and institutions with a narrow focus on highly-cited fields, which are well covered in

these databases. This puts universities from non-English-speaking countries, with a focus on the arts, humanities and social sciences, at a disadvantage when being compared in global rankings (Calero-Medina et al., 2008; van Raan et al., 2011; Waltman et al., 2012). Data submitted by universities to ranking agencies (e.g. personnel data, student numbers) can be problematic to compare due to different standards. These incompatibilities are being amplified because university managers have become increasingly aware of global rankings and try to boost their performance by "tweaking" the data they submit to the ranking agencies (Spiegel Online, 2014). Beyond all the data issues, there is the effect that universities with lower positions in the rankings often encounter volatile ups and downs in their consecutive year-to-year ranks. This creates the sensation of contending in a "gamble" in which results are calculated at random by ranking agencies. Such effects make global university rankings in many cases an inappropriate tool for university managers: the ranking results simply do not reflect the universities' actual performance or their management strategies. Volatile jumps are also difficult to explain to the media, which often engage in sensationalism when covering rankings by interpreting subtle changes of scores, even within the margins of statistical deviations, as substantial shifts in performance. Bookstein et al. (2010) found unacceptably high year-to-year variances in the score of lower ranked universities caused by statistical noise in the Times Higher Education World University Ranking (THES), one of the currently most popular global rankings. We again observed puzzling variances in the THES-Ranking 2014-2015, published in October 2014. Accordingly, we here analyze the fluctuations in score and rank of the THES-Ranking by calculating a regression analysis for consecutive years for 2010-2014 to determine the random component of these fluctuations. The methodology of the THES-Ranking was revised several times in varying scale, before and after the split with Quacquarelli Symonds (QS) in 2010 and the new partnership with Thompson Reuters. Times Higher Education (THE) calculates 13 performance indicators, grouped into the five areas Teaching (30%), Research (30%), Citations (30%), Industry income (2.5%) and International outlook (7.5%). However, THE does not publish the scores of individual indicators, only those of all five areas combined. Since 2010, the research output indicators are calculated based on Web of Science data. Most of the weight in the overall score is made up by the normalized average citations per published paper (30%), and by the results of an academic reputation survey (33%) assessing teaching and research reputation and influencing the scores of both areas (Rauhvargers, 2013; THE, 2014). In the past, criticism has been levied against this survey. Academic peers can choose universities in their field from a preselected list of institutions and, although universities can be added to the list, those present on the original list are more likely to be nominated. This leads to a distribution skewed in favor of the institutions at the top of the rankings (Rauhvargers, 2011; Rauhvargers, 2013). THE allegedly addressed this issue by adding an exponential component to increase differentiation between institutions, yet no information is available on its mode of calculation (Baty, 2011; Baty, 2012).

Methods

We used the publicly available data on scores and ranks from the THES-Ranking for the years 2010, 2011, 2012, 2013 and 2014, including only those universities ranked from 1 to 200. We performed the following analysis: i) we regressed the scores of the ranking of the year t-1 on the scores of the year t; ii) we regressed the ranks of the ranking of the year t-1 on the ranks of the year t; iii) we plotted the scores in descending order and iv) we determined the random component of the fluctuations in the ranks from year to year.

Results

Regression of the scores and ranks of two consecutive years

The regression of the scores—particularly of the ranking 2010-2011 regressing on the scores of the ranking of 2011-2012—shows a very high fluctuation/noise (Figure 1a), especially for the lower ranked universities. Moreover, the noise among the lower ranked universities seems to be higher compared to the already very noisy THES-Ranking performed by QS before 2010 (Bookstein et al., 2010, Figure 1). Note that in the rankings in the years following 2010-2011, the noise in the THES-Ranking did improve (Figure 1b-d).

Association between Scores and Ranks

Nonetheless, a general problem of the THES-Ranking remains: the difference in the scores among the 50 highest scoring universities is considerably higher compared to the difference among the lower scoring universities. This clearly suggests a non-linear relationship between scores and ranks (Figure 2 a-e). The consequence is that the ranks of the high scoring universities are much more robust to deviations in the scores from year to year. In the lower ranking universities, however, even very small, more or less random deviations (around 0.5%) lead to unexpected "high jumps" in the ranks from year to year (Figure 1e-h).

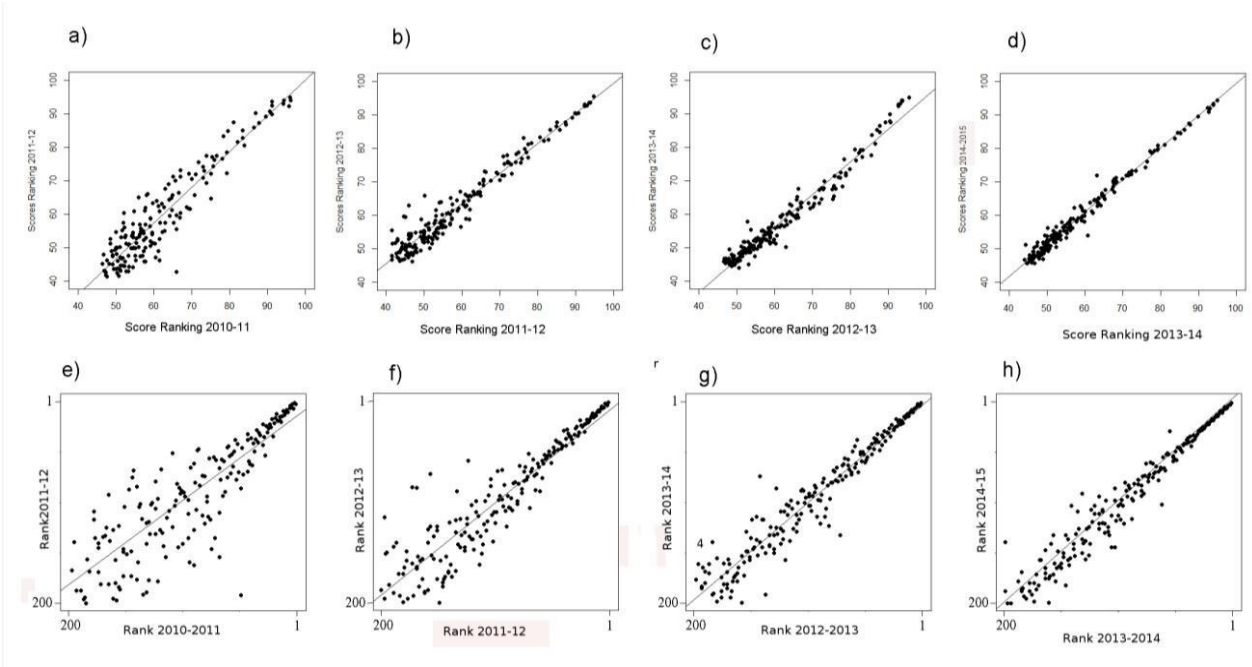


Figure 1a-1d) Scores of the year t-1 regressing on the score of the year t from the ranking 2010-11 on. Figure 1e-1h) Ranks of the year t-1 regressing on the ranks of the year t from the ranking 2010-11 on. Linear regression line indicates perfect association, e.g. no changes in ranks and scores between two consecutive rankings.

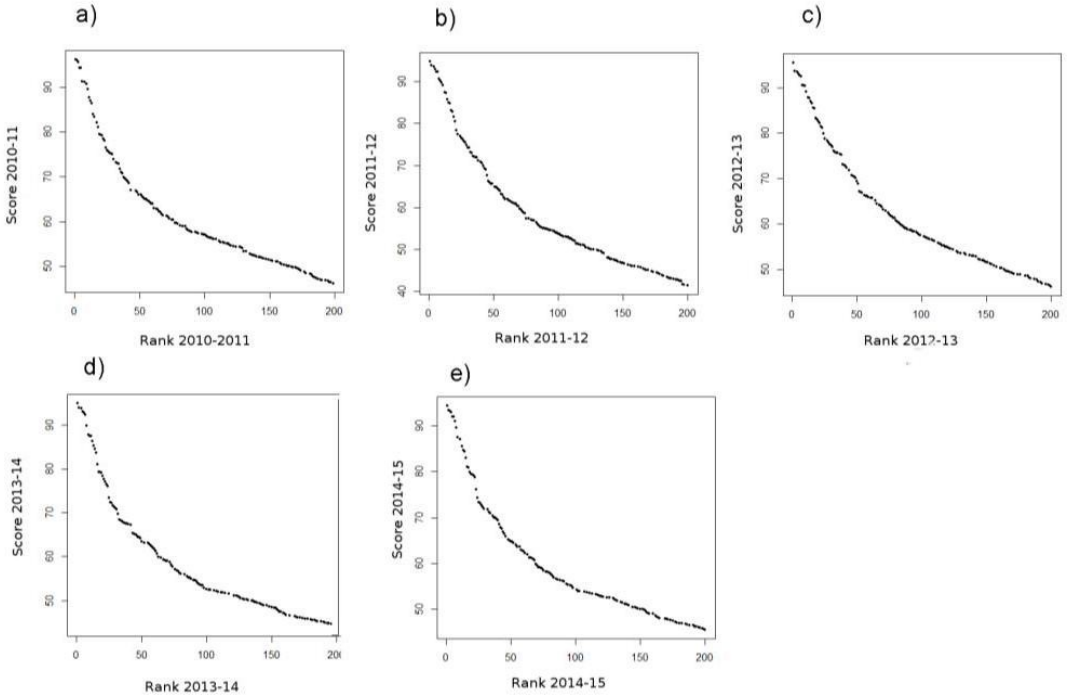


Figure 2 a-e). Ranks plotted against scores for the THES-Ranking a) 2010-11; b) 2011-12; c) 2012-13; d) 2013-2014; e) 2014-15

Discussion and Outlook

High ranking positions achieved by a small group of universities are often self-perpetuating, especially due to the intensive use of peer review indicators, which improve chances of maintaining a high position for universities already near the top (Bowman & Bastedo, 2011; Rauhvargers, 2011). This phenomenon also corresponds to the Matthew effect, which was coined by Merton (1968) to describe how eminent scientists will often get more credit than a comparatively unknown researcher, even if their work is similar: credit will usually be given to researchers who are already famous. The intensive and exaggerated discussion in the media of the "up and downs" of universities in the THES-Ranking is particularly misleading for the lower scoring universities (below approximately a score of 65% and a rank of 50; above scores of 65%, the relationship between ranks and scores is steeper, and it flattens for scores below 65%). This is because the ranking positions suggest substantial shifts in university performance despite only very subtle changes in score. In fact, merely random deviations must be assumed. One reason lies in the weighing of indicators by THE, with the emphasis on citations and peer review (totaling more than 65% of the total score). For lower ranked universities, a few highly cited publications, or the lack thereof, or few points asserted by peers in the reputation survey, probably make a significant difference in total score and position. In a follow up study that is currently under review we compared the results from THES with the results of the ARWU-Ranking (aka Shanghai-Ranking). Although the ARWU-Ranking seems to be more robust than the THES-Ranking (less year-to-year fluctuations probably due to the omittance of peer review indicators), we also found fluctuations below rank 50 and patterns of non-linearity between ranks and scores. Furthermore we found out that year-to-year results do not correspond in THES- and ARWU-Rankings for universities below that rank.

Ranking results have a major influence on the public image of universities and even impact their claim to resources (Espeland & Saunder, 2007; Hazelkorn, 2011). Accordingly, such fluctuations in the THES-Rankings can have serious implications for universities, especially when the media or stakeholders interpret them as direct results of more or less successful

university management. Our initial data in combination with the data from the literature strongly suggests that universities as well as policy makers and stakeholders should avoid to use rankings, especially league-tables, for management purposes or for strategic planning.

More specifically, the THES-Rankings in their current form have very limited value for the management of universities ranked below 50. This is because the described fluctuations in rank and score probably do not reflect actual performance, whereby the results cannot be used to assess the impact of long-term strategies. Thus, results from the THES (and to some extent also the ARWU) should be used only with great discretion. The low correlation between the ranks of the THES and the ARWU ranking, particularly for the universities ranked below 50 in both rankings, creates another serious doubt if rankings should be used for any management purposes at all. Maybe a "meta-analysis" of rankings could be reasonable to derivate consistent and reliable results from rankings. If done, such a meta-analysis should include as many rankings as possible to reduce random perturbations.

Multidimensional rankings, like the U-Multirank (http://www.u-multirank.eu), seem to offer a more versatile picture that reflects both the diversity of higher education institutions and the variety of dimensions of university excellence, allowing university managers to compare institutions on various levels. Although multidimensional rankings do get less public attention than league-tables and they can be prone for errors for the same reasons as monodimensional rankings (e.g., incompatibility of data provided by the universities), from the perspective of a university manager, they offer a more diverse toolset to gauge an institutions strength and weaknesses and to benchmark comparable universities.

"Rankings are here to stay, and it is therefore worth the time and effort to get them right," warns Gilbert (2007). That is especially true for monodimensional rankings, like the THES, that spark a lot of media attention. What could be done to address the fluctuations in the THES-Rankings for universities below rank 50 and to avoid the impression of a "gamble" in which THE "rolls a dice" to determine scores and ranks? THE has already addressed fluctuations to some extent by ranking universities only down to position 200, followed by groups of 25 from 201-300 and groups of 50 from 300 to 400. Nonetheless, based on our data we believe that this is not going far enough and suggest that universities should be summarized in groups of 25 or 50 below the position of 50.

The analyzed curves of scores vs. ranking positions in Figure 2 do have analogous characteristics for example to semi-logarithmic curves produced in analytic biochemistry. The accuracy of such curves is limited to the steepest slope of the curve, whereas asymptote areas deliver higher fuzziness (Chan, 1992). Thus, a further suggestion to avoid the blurring dilemma is the methodological approach of introducing a standardization process for THES-Ranking data. This would involve using common suitable reference data to create calibration curves represented by non-linearity or linearity. However, more research in this area is necessary.

The results presented in this paper are only the starting point and we plan to do more in-depth analyses of the variations in the various indicators in the future. We already have extended our analysis to include the ARWU-Ranking (paper currently in review) and we plan to analyze and compare other major higher education rankings (e.g. the QS-Ranking) in future publications to assess their usability for university management purposes.

References

Aguillo, I.F., Bar-Ilan, J., Levene, M. & Ortega, L.J. (2010). Comparing university rankings. *Scientometrics 85*, 243–256.

Benito, M. & Romera, N. (2011). Improving quality assessment of composite indicators in university rankings: a case study of French and German universities of excellence. *Scientometrics 89*, 153–176.

Baty, P. (2011, October 6) THE Global Rankings: Change for the better, *Times Higher Education*, http://www.timeshighereducation.co.uk/world-universityrankings/2011-12/world-ranking/methodology

Baty, P. (2012, October 4). The essential elements in our world-leading formula, *Times Higher Education*, http://www.timeshighereducation.co.uk/worlduniversity-rankings/2012-13/world-ranking/methodology

Bowman, A.M. & Bastedo, N.M. (2011). Anchoring effects in world university rankings: exploring biases in reputation scores. *Higher Education 61*, 431–444

Bookstein, F.L, Seidler, H., Fieder, M., & Winckler, G. (2010). Too much noise in the Times Higher Education rankings. *Scientometrics 85*, 295–299.

Buela-Casal, G., Gutierrez-Martinez, O., Bermudes-Sanchez, M., & Vadillo-Munoz, O. (2007) Comparative study of international academic rankings of universities. *Scientometrics 71*, 349-365.

Calero-Medina, C., López-Illescas, C., Visser, M.J and Moed, H.F. (2008). Important factors when interpreting bibliometric rankings of world universities: an example from oncology. *Research Evaluation, 17* (1), 71–81.

Chan, D.W. (ed) (1992) *Immunoassay Automation: a Practical Guide*. San Diego, CA: Academic Press.

Espeland, W.N & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds, *American Journal of Sociology, 113* (1), 1–40.

Gilbert, A. (2007). Academics strike back at spurious rankings. *Nature, 447*, 514–515.

Hazelkorn, E., (2007) Impact and influence of league tables and ranking systems on higher education decision-making. *Higher Education Management and Policy, 19* (2), 87–110.

Hazelkorn, E. (2011) *Rankings and the Reshaping of Higher Education: the Battle for World Class Excellence*. Basingstoke: Palgrave-MacMillan.

Hazelkorn, E. (2014) Reflections on a decade of global rankings: what we've learned and outstanding issues, *European Journal of Education, 49* (1), 12–28.

Ioannidis, J. P. A., Patsopoulos, N. A., Kavvoura, F. K., Tatsioni, A., Evangelou, E., Kouri, I., et al. (2007). International ranking system for universities and institutions: A critical appraisal. *BMC Medicine 5*, 30.

Merton, R.K. (1968). The Matthew effect in science. *Science, 159*, 56-63.

Rauhvargers, A. (2011) EUA Report on Global Rankings and their Impact – Report I (European University Association). http://www.eua.be/pubs/Global_University_Rankings_and_Their_Impact.pdf

Rauhvargers, A. (2013). EUA Report on global rankings and their Impact – Report II (European University Association). http://www.eua.be/Libraries/Publications_homepage_list/EUA_Global_University_Rankings_and_Their_Impact_-_Report_II.sflb.ashx.

Safon, V. (2013). What do global university rankings really measure? The search for the X factor and the X entity. *Scientometrics, 97*, 223–244.

Saisana, M., d'Hombres, B., & Saltelli X. (2011). A Rickety numbers: Volatility of university rankings and policy implications. *Research Policy, 40*, 165–177.

Spiegel Online (2014). Deutsche Unis im "THE"-Ranking: Das Wunder von Tübingen. 02.10.2014. http://www.spiegel.de/unispiegel/studium/uni-ranking-hochschulenim- the-ranking-a-994684.html

Times Higher Education. (2014). World University Rankings 2014-2015 methodology http://www.timeshighereducation.co.uk/world-university-rankings/2014-15/worldranking/methodology

Tofallis, C. (2012). A different approach to university rankings. *Higher Education* 63, 1-18.

van Raan, T. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62* (1), 133–143.

van Raan, T., Leeuwen, T., & Visser, M. (2011). Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. *Scientometrics, 88*, 495–498.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, N.C., Tijssen, R.J., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. CWTS Working Paper Series. http://arxiv.org/abs/1202.3941

The Vicious Circle of Evaluation Transparency – An Ignition Paper

Miloš Jovanović[1]

[1]*milos.jovanovic@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis, Appelsgarten 2, 53879 Euskirchen (Germany)

Introduction

The present paper introduces a model, which describes different phases that typically occur in situations, in which a researching subject (e. g. an author, an institution, a country etc.) needs to be evaluated and in which some kind of reward (e. g. monetary in the form of a bonus or funding) is based on this evaluation. This model, the present author calls it the "vicious circle of evaluation transparency", will be underlined by giving examples for each of its phases. In order to be able to observe a process that is described by this model, there first needs to be something that is to be evaluated, for example a research group at a university. Such a need normally comes up, when money is to be divided among different groups or focused on one. The problem of evaluation and rewarding is at the core of the model (see Figure 1).

Figure 1. The "vicious circle of evaluation transparency"-model.

Phase I – Evaluation and rewarding by subjective and intransparent criteria

The first question that might come up in such a situation is the question of how to evaluate a research group. In hierarchically organized universities the leader of a department will decide whether or not and how this group is evaluated. Very often, this person is also the one that conducts the evaluation and, based on this, determines the type and amount of a reward or funding (or some kind of penalty, if the evaluation is negative). In today's world of vast amounts of digital data, it might be hard for only one person to do such an evaluation. Naturally, having one person alone evaluate a group's performance and decide on rewards will lead to a number of persons feeling unfairly evaluated, because the evaluator might not know about their achievements or their work in detail. This criticism might be alleviated in part by expanding the number of evaluators, for example by having a board of evaluators. Another possibility is to improve the transparency of the evaluation by documenting and publishing certain evaluation criteria by which the evaluated subjects can read about the evaluations and try to strive to get a better evaluation. These evaluation criteria are a first step towards phase II of the model.

Phase II – Introduction of "objective" and transparent criteria

These evaluation criteria might be subjective. For example "Quality of work" can be a criterion that is evaluated differently by different people. In order to make evaluation criteria comparable and independent of the evaluating person, "objective" criteria are often introduced. The reason why the word is put into quotation marks is due to the fact that very often these "objective" criteria are not objective at all. The introduction of "objective" and transparent criteria is a simplification of reality, an attempt to put parts of reality into some kind of a score in order to compare them with each other. Bibliometric indicators are one example of such a simplification. In many countries, different kinds of "objective" and subjective evaluation criteria have been introduced, for example in Italy (Abbott, 2009). Normally, these "objective" evaluation criteria (often in the form of different kinds of indicators) are communicated transparently. And while transparency is an important factor for these evaluations, it also leads to one problem in this phase: the fact that the evaluated subjects, in our example researchers at universities, react to the evaluation by starting to change their behavior, in order to maximize their scores in the evaluation. Of course, one reason behind evaluation is to positively influence the behavior of the evaluated researchers. But in Germany, for example, this has led to authors aiming to publish more in internationally known journals that have a US publisher and which are more general in their scope (Michels & Schmoch, 2013). This underlines the fact that authors do not base the decision in which journal they wish to publish in on scientific reasons

alone and constitutes a negative change of behavior. Also, some of the evaluated subjects might complain that the evaluation criteria do not reflect their work adequately and need to be refined. This leads to the next phase.

Phase III – Adaptation and enrichment of "objective" criteria

The need to fairly represent and evaluate researchers' work in the evaluation criteria and to adapt these in order to not allure unwanted change of behaviour leads to reforms in the evaluation system, e.g. new or a mix of indicators are proposed. The current discussion on alternative metrics is an example for phase III (e.g. in Haustein et al., 2014). The problem here is, that phase III is actually reintroducing parts of the simplification of reality, which was conducted in phase II. The evaluation criteria become more complicated again. A country example for this phase is the Czech Republic, which introduced performance-based research funding (phase II). A study by Vanacek (2014) found that the number of publications increased very quickly. He shows that in comparison to the quickly growing number of publications the quality seems to have stagnated and recommends reworking the procedure of evaluation and performance-based funding in order to increase not only the number of publications but also their quality (phase III). But for some research communities, the adaptation and enrichment of the "objective" criteria is no option. Instead, these criteria are rejected. For example, there is an ongoing discussion in the mathematical community. Authors note that bibliometric data lose "crucial information that is essential for the assessment of research". It is pointed out that bibliometric indicators can be manipulated and lead to undesirable publishing practices (Adler, Ewing, & Taylor, 2009). The authors also dismiss reputation, as determined by surveys as a possible way of measuring the quality of a journal. The evaluation of journal editorial processes is not seen as a good way of ranking journals either. Instead, the authors recommend an "honest, careful rating of journals based on the judgment of expert mathematicians", which is the point, where phase IV starts.

Phase IV – Removal of "objective" criteria and return to phase I

Concretely, the IMU recommends that a rating committee of 16-24 experienced and respected mathematicians should be appointed. Without going into too much detail, this committee (via various panels) is then supposed to rate the different journals and assign them to tiers (ranging from tier 1 = high quality journal to tier 4 = low-class journal) (Journal Working Group, 2011). This system is similar to the peer review process. Introducing evaluation by a committee of experts, either by rejecting "objective" evaluation criteria or because the evaluation system has become too complicated, brings the model full circle. The evaluation has reached phase I again. One should note that in phase II of this new cycle, the criteria probably will not be the same as in the first cycle. Newly developed and more sophisticated criteria will take their place.

Conclusion

It is this author's personal opinion that the above described model of evaluation transparency not only describes a typical process in which bibliometric indicators are involved but rather evaluation processes in general. If this is the case, one may discuss possibilities to change this, since a cycle like this is not an optimal solution. An option might be the introduction of diametrically opposed evaluation criteria so that an evaluated subject could not be good in all criteria. Another idea that might serve to fan the discussion on this topic would be the introduction of a changing system of criteria, akin to the disciplines at Olympic Games. The criteria could be published a year before the evaluation takes place and would change each year. This would be a transparent system, while the evaluated researchers would not need to change their behavior in a negative way because the next year the criteria would be different. Whatever changes might be introduced, it is this author's opinion that the vicious circle has to be stopped and replaced by a different system that leads to the desired goal: a fair evaluation of research.

References

Abbott, A. (2009). Italy introduces performance-related funding. *Nature, 460*, 559.

Adler, R., Ewing, J. & Taylor, P. (2009). Citation Statistics: A Report from the IMU in Cooperation with the ICIAM and the IMS. *Statistical Science,* 1-14.

Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., &Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics, 101*, 1145-1163.

Journal Working Group (2011). *Report of the IMU/ICIAM Working Group on Journal Ranking.* Retrieved March 10, 2015 from: http://www.mathunion.org/fileadmin/IMU/Report/WG_JRP_Report_01.pdf

Michels, C. & Schmoch, U. (2013). Impact of bibliometric studies on the publication behaviour of authors. *Scientometrics, 98*, 369-385.

Vanacek, J. (2014). The effect of performance-based research funding on output of R&D results in the Czech Republic. *Scientometrics, 98*, 657-681.

Influence of the Research-oriented President's Competency on Research Performance in University of China -Based on the Results of Empirical Research

Li Gu[1], Liqiang Ren[1], Kun Ding[1], Wei Hu[2]

[1] guli@dlut.edu.cn
WISE Lab, School of Public Administration and Law, Dalian University of Technology, Dalian, 116024 (China)

[2] huwei@moe.edu.cn
Dept. of Personnel, Ministry of Education of China, No.37 Damucang Hutong, Xidan, Beijing, 100816 (China)

Introduction

With the gradual promotion and implementation of China's national innovation-oriented strategy, research universities are playing an irreplaceable role in leading scientific development and technological innovation. Scientific research is one of the basic functions of a research university, which cultivates high-quality innovations and supports research universities in serving their societies (Rhoads, 2014). While high-level research universities need presidents with outstanding quality and ability. Research-oriented presidents, as the scientific research managers and experts, play a very important role in constructing and developing their universities, and they also focus on talent cultivation to realize social missions.

Therefore, the research on the influence of the research-oriented president's competency on research performance has profound connotations and value, which can provide references to guide and explore the systems for selecting, cultivating and assessing research-oriented university presidents.

Method

Research-oriented presidents, as senior managers of research universities, are responsible for teaching university management and for the direct leadership of scientific research. This special position determines the universality and complexity of the factors related to empirical studies on competence characteristics (Angeles, 2014; Sydney & Frances, 2013; Liu & Xu, 2013; Snyder, 2012).

Based on the theoretical analysis of competence characteristics and in combination with the vocational characteristics and main responsibilities of research-oriented presidents, we first constructed a theoretical framework of research-oriented presidents' competence characteristics (Figure 1). Then, we designed a questionnaire system to collect data and data were analyzed using SMRT PLS2.0 software (one of the leading software tools for partial least squares structural equation modeling). The verification results show that the scale's convergent validity was high, and it also had good discriminant validity. Finally, we used the R^2 statistic to analyze the structural model and received good explanation.

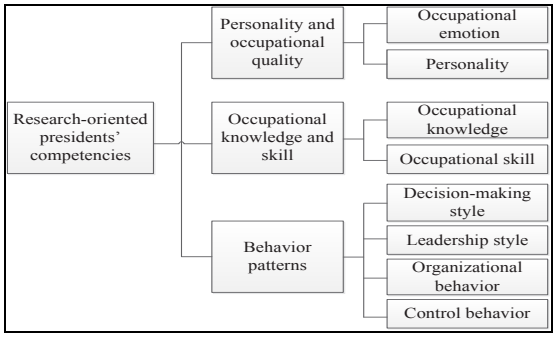


Figure 1. Theoretical Framework of Research-oriented Presidents' Competence Characteristics.

Data

This study selected research-oriented presidents of research universities as its subjects. Therefore, thirty-nine of 985 universities under China's Ministry of Education were selected for the study, and to ensure the comprehensiveness of our investigation, the selected samples included research-oriented presidents, middle management, scientific research management, professors, associate professors, lecturers, assistants, and other research personnel. The descriptive statistics (Table 1) on the study subjects were obtained via statistical data analysis.

Results

Through statistically analysing the sample data, the influence of occupational emotion, personality, occupational knowledge, occupational skill, decision-making style, leadership style, organizational behaviour and control behavior on scientific research performance was respectively checked. The results indicate that the performance had good validity. However, if organizational characteristics are used as an intervening variable, the competence characteristics of research-oriented

presidents have significant positive influences on scientific research performance.

Table 1. Descriptive Statistics on Respondents.

Measurement items		Sample size (N)	Proportion (%)
Gender	Male	292	70.4
	Female	123	29.6
Age	30 and below	37	8.9
	31–35	132	31.8
	36–40	93	22.4
	41–45	63	15.2
	46–50	41	9.9
	51–55	31	7.5
	56 and above	18	4.3
Education	College	3	0.7
	Bachelor's	31	7.5
	Master's	103	24.8
	Doctorate	276	66.5
	Others	2	0.5
Title	Assistant	98	23.9
	Lecture	92	21.9
	Associate Prof.	15	3.6
	Full Prof.	210	50.6
	Academician	0	0
	Others	0	0

Conclusion

Based on the above research results, we constructed a model of research-oriented university presidents' competence characteristics, shown in Figure 2.

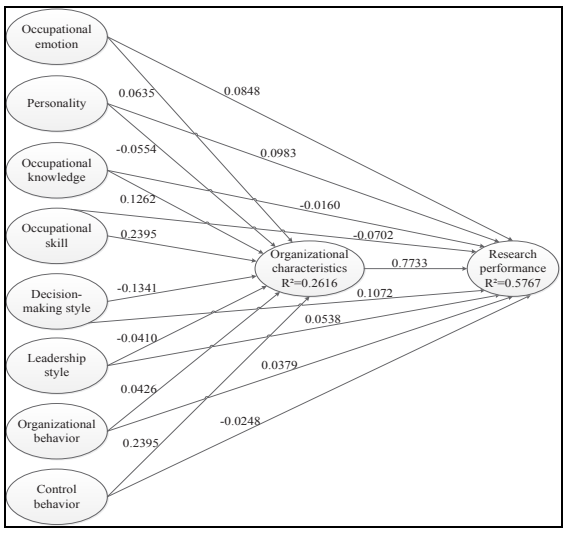


Figure 2. Relational Model of Research-oriented Presidents' Competence Characteristics and Their Universities' Research Performance.

The following conclusions can be drawn by analysing the model of research-oriented presidents' competence characteristics:

(1) From the direct effect perspective: 1) research-oriented presidents' professional emotion, personality traits, decision-making and leadership styles and organizational behavior have significant positive influences on scientific research performance. 2) Presidents' professional knowledge, professional skills and control behavior have significant negative influences on research performance, but further inspection of the analysis results reveals that the negative influence is not absolute.

(2) From the mediating effect perspective, professional emotion, professional skills, organizational behavior and control behavior have significant positive influences on organizational characteristics, whereas personality traits, professional knowledge, and decision-making and leadership styles have significant negative influences on organizational characteristics. However, organizational characteristics as intervening variables between research-oriented presidents' competence characteristics and their universities' scientific research performance can maximize the effects of the presidents' competence characteristics and have significant positive influence on research performance.

Acknowledgments

This work was supported by the project of "Specialized Research Fund for the Doctoral Program of Higher Education" (20130041120049), the project of "Fundamental Research Funds for the Central Universities" (DUT13RW409) and the project of "the Soft Science Research Project of State Intellectual Property Office – Study on Competency and Promotion Polices of Patent Attorneys from the Angle of Patent Application Quality".

References

Angeles, M.M. (2014). Learning for a Sustainable Economy: Teaching of Green Competencies in the University. *Sustainability*, *6*, 2974-2992.

Liu, X.J. & Xu, F. (2013). A Study on the Competency Model for Industry-based Characteristic University Presidents. *Journal of National Academy of Education Administration*, *10*, 60-64.

Rhoads, R.A., Shi, X.G. & Wang, X.Y. (2014). Reform of China's Research Universities: A New Era of Global Ambition. *Education and Society*, *32*, 5-28.

Sydney, F.J. & Frances, K.K. (2013). University Presidents' Perspectives of the Knowledge and Competencies Needed in 21st Century Higher Education Leadership. *Journal of Educational Leadership in Action*, *1*, 1-2.

Snyder. (2012). Higher Education Leadership Competencies: Quantitatively Refining a Qualitative Mode. *American psychologist, 27*, 5-8.

Medical Literature Imprinting by Pharma Ghost Writing: A Scientometric Evaluation

Philippe Gorry[1]

[1] *philippe.gorry@u-bordeaux.fr*
GREThA UMR CNRS 5113, University of Bordeaux, Av. Leon Duguit, 33608, Pessac (France)

Introduction

Misappropriation of authorship, honorary or ghost authorship, undermines academic publishing with a substantial proportion of peer-reviewed medical journals targeted (Flanagin, 1998). Pharmaceutical companies pay professional writers or medical communication companies to produce papers whilst paying other scientists or physicians to attach their names to these papers before they are published in medical or scientific journals. This ghost management is meant to support the marketing of drug products (Sismondo, 2007). Companies use this strategy to communicate competitive message, promote unproven off-label uses, and mitigate perceived drug risks (Fugh-Berman, 2010). Publication planning strategy with fraudulent practices were revealed through internal company communications in the course of the well-known Neurontin® litigation case (Vedula, 2012). Even though ghostwriting realized by pharmaceutical companies has been reported, it remains necessary to measure to what extent ghostwritten articles have impacted medical literature. Healy and Catell (2003) started to answer this question with a sample of 16 ghostwritten articles about a peculiar antidepressant. This pioneering analysis should be extended to a larger collection of ghostwritten articles as well as studied for a longer period of time.

Method

Pharma ghostwriting has been documented initially through 3 original papers: first, D. Healy and D. Cattell reported 16 ghostwritten articles in 2003, later on, A.J. Fugh-Berman (2010) reported 23 new cases, finally in 2012, Vedula and colleagues identified 13 more ghost written publications. Based on legal documents, from US district court following class action and lawsuit against pharmaceutical companies concerning several molecules: estrogen (Prempo®/Premarin®, Wyet), sertraline (Zoloft®, Pfizer), gabapentin (Neurotin®, Pfizer), and paroxetine (Paxil®, GSK), 40 more ghostwritten publications were identified. Therefore, a corpus of 92 publications were retrieved from Pubmed, Scopus or Web of Science databases, and subsequently analyzed for main bibliometric indicators. Descriptive statistics were done using Excel.

Result

A corpus of 92 ghostwritten articles was assembled, covering a period between 1997 and 2008. Two third of theses cases were published between 1998 and 2000. 79 different authors have been identified. While the vast majority of them were co-author of only one ghostwriting paper, 10 authors published two ghost papers and one signed three ghost papers (data shown on the poster). 82% of the identified authors were US academics. However, authors of 10 different countries were identified as representing the main drug pharma market with the noticeable exception of Germany and Japan. Among the different affiliation of the authors, only one pharmaceutical company was identified. Most of the institutions were university with affiliated medical school (data shown on the poster).

Ghostwritten articles were published by average productive author (h-index at the time of ghost publication date: mean=15.84), with some exceptions: Bondareff W, University of Southern California, (h-index=92), Seddon JM, Tufts Medical Center, (h-index=53), Freedman MA, Medical College of Georgia & Jermain DM, Pfizer (h-index= 2). Along the 10 years observation period, there is no noticeable variation in the productivity of the authors (data shown on the poster). Indeed average author h-index reach 29.13 in year 2013.

The corpus covers a large spectrum of medical specialties. However, it is interesting to point out that more than a third of ghostwritten papers concern psychiatry and mental illness (Figure 1).

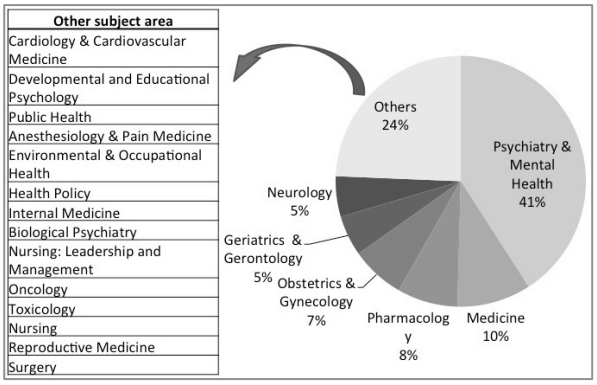


Figure 1. Distribution of ghost written articles by medical specialties.

Publication of ghost articles were scattered throughout 51 different journals. Among these source titles, there are four psychiatric journals, with various impact factor (IF), accounting for a third of the ghostwritten articles (Figure 2 and Table 1).

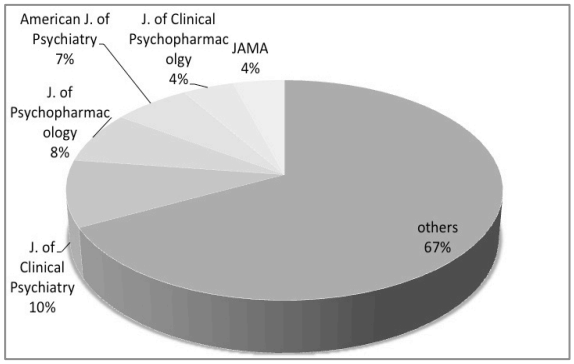


Figure 2. Distribution of ghost written articles by journals.

Table 1. List of the main journal publishing ghost written articles with their impact factor.

Journal	Ghost pub.	SJR impact factor at publication date
Journal of Clinical Psychiatry	9	1,787307692
Journal of Psychopharmacology	7	1,142571429
American Journal of Psychiatry	6	3,599
Journal of Clinical Psychopharmacolgy	4	1,6045
Journal of the American Medical Association	4	3,82875

The average IF of journals where ghostwritten articles are published is in the low-medium range (mean IF=2.51, median IF=1.81). Sometime, there are published in very low IF journal (ex: Climacteric IF=0.091).

Finally, the last evaluation concerns the number of year during which a ghostwritten article can be cited since the date of publication. (Figure 3; no ghostwritten article have been published in 2007). Year after year, ghostwritten articles have on an average 84% chance to be cited.

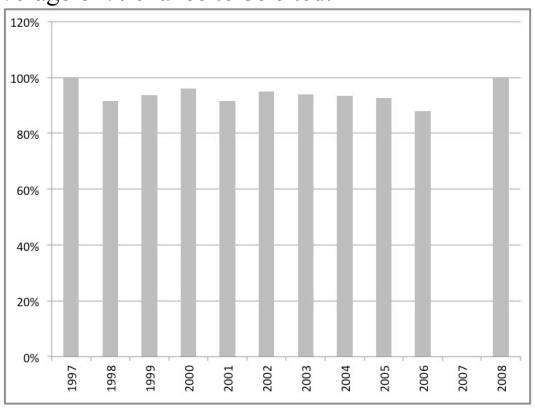


Figure 3. Probability of a ghost written articles to be cited once year since the publication.

On long range, the average ghostwritten article IF is much higher than the average journal IF. Indeed a ghostwritten article is about 10 times more cited than any article published in the same journal (Table 2).`

Table 2. Statistics difference between ghost written & journal article impact factors.

	Ghost Article impact factor	Journal impact factor
mean	7,24	2,68
max	68,13	8,73
min	0,31	0,09

Discussion

With this study, we have been able to conduct a bibliometric analysis on a large number of ghost articles, over a long period of time. Overall, ghostwritten articles are published by average productive author, in low IF journals; they are cited during a long period of time and therefore have a high number of citations (Table 3). Thus, ghostwritten articles might influence the medical community and its practice, which subsequently raises public health concerns.

Table 3. Main bibliometric indicators of ghost written articles.

	Journal impact factor	Ghost written article citations	Last year citation	Total number year citations	author h index
Mean	2,697	84,951	2013	13	29,731
Max	6,984	351	2014	16	68
Min	0,091	4	2008	8	4

Despite numerous declarations by medical journal editors and the conduct of ethics declared by professional medical writers, we would like to underline that none of these ghostwritten articles involved in lawsuit case have been retracted whilst companies have been sentenced by Justice.

Moreover the efficiency of ghostwriting publication strategy could be questioned since only a third of articles have an impact superior to what would be expected. Therefore the return on investment for the pharmaceutical industry might be very low, especially regarding the risk of litigation and the disclosure of such fraudulent marketing practices.

References

Flanagin, A., *et al.* (1998). Prevalence of articles with honorary authors and ghost authors in peer- reviewed medical journals. *Journal of the American Medical Association, 280*, 222-224.

Fugh-Berman, A.J. (2010). The Haunting of Medical Journals: How Ghostwriting Sold ‘‘HRT’’ *PLoS Medicine,* 7, e1000335.

Healy, D. & Cattell, D. (2003). Interface between authorship, industry and science in the domain of therapeutics. *British J. of Psychiatry*, *183*, 22-27.

Sismondo, S. (2007). Ghost management: How much of the medical literature is shaped behind the scenes by the pharmaceutical industry? *PLoS Medicine, 4*, 1429-1433.

Vedula, S.S. et al., (2012). Implementation of a publication strategy in the context of reporting biases. A case study based on new documents from Neurontin litigation. *Trials*, *13*, 136.

Are scientists really publishing more?

Daniele Fanelli[1], and Vincent Larivière[2]

[1] *dfanelli@stanford.edu*
METRICS - Meta-Research Innovation Center at Stanford, Stanford University, 1070 Arastradero Road, Palo Alto, CA 94304

[2] *vincent.lariviere@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, H3C 3J7 Montreal, Qc. (Canada)
Université du Québec à Montréal, Centre Interuniversitaire de Recherche sur la Science et la Technologie (CIRST), Observatoire des Sciences et des Technologies (OST), CP 8888, Succ. Centre-Ville, H3C 3P8 Montreal, Qc. (Canada)

Introduction

The success of researchers and research institutions is increasingly determined by measurable aspects of their performance, in particular the quantity and citation-impact of their publications. The effects that these growing "pressures to publish" might have on publication and research practices are a matter of growing concern and increasing academic interest (de Winter & Dodou, 2014; Fanelli, 2010, 2012, 2013; Tijdink, Vergouwen, & Smulders, 2013; van Dalen & Henkens, 2012).

Much criticisms and concern has been expressed, in particular, for the risk of overemphasising the quantity of a scientist's publication record at the expense of its quality. In order to show a longer lists of publications in their CVs, it is commonly hypothesised, scientists might increasingly resort to questionable practices such as inappropriately subdividing ("salami slicing") their results, publishing trivial and incomplete studies, conducting research hastily and sloppily, selecting out of their findings those that are least "publishable", or even resorting to outright scientific misconduct in the form of duplicate publication, plagiarism and data fabrication (e.g. Angell, 1986; Hayer et al., 2013).

Performance-evaluation policies of institutions in various countries have responded to these concerns by formally removing any quantitative consideration from their performance assessments (e.g. VSNU, 2015). However, there is little evidence to support these policies. No study, in particular, has ever verified whether scientists are have actually responded to growing pressures by churning out more papers. We present preliminary results of a project aimed at filling this gap in the literature.

Methods

We identified individual researchers who published in the Web of Science across the 20th century by selecting all authors identified by three initials (first name and two middle names, plus surname, e.g. Vleminckx-SGE), which reduces the likelihood that these researchers have homonyms. From this initial sample we selected authors who had at least two publications, and from these we then selected authors whose publications spanned a period of at least 15 years. For each of these authors we then counted the total number of papers published in the first 15 years of activity – the period were pressures to publish are hypothesised to be stronger – and we also measured the average number of co-authors.

Results

The raw number of papers published by individual authors has grown very rapidly across the century (Fig. 1). Fractional productivity, however, as measured by dividing the author's total number of papers by the average number of co-authors, shows a net decline (Fig. 2).

Discussion

Although still preliminary, these results suggest that our beliefs about the effects of pressures to publish might be partially incorrect. Authors might have responded to growing performance expectations not, as commonly believed, by subdividing or trivializing their results or by multiplying their effort at the expense of other activities, but by enlarging their network of collaborations in order to make ever smaller contributions to a growing number of projects. Since neither publication nor citation metrics are counted fractionally, this strategy allows scientist to increase their measurable publication rate without necessarily increasing their total research effort.

If scientists' net effort devoted to research is not increasing, then concerns for growing "salami slicing" and other questionable practices might be unjustified. Explanations for recent evidence that retraction and correction rates are growing (Fang & Casadevall, 2011), that publication bias is growing (Fanelli, 2012) and that research bias might be higher in scientifically productive countries (Fanelli, 2010) might need revising. And policies that are currently de-emphasizing "quantity" in

favour of "quality" (e.g. VSNU, 2015) might not have a solid basis in evidence, and could therefore be ineffective or even damaging.

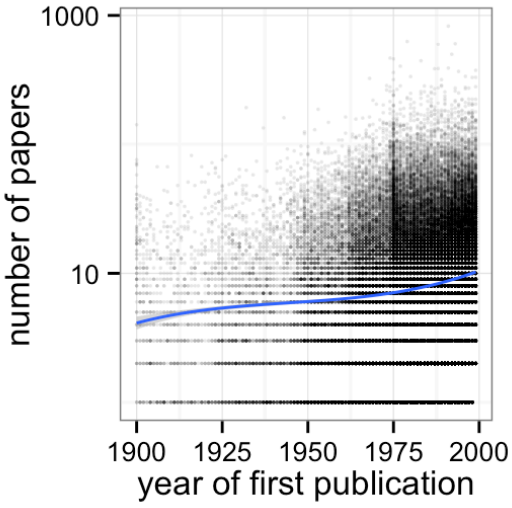


Figure 1. Total number of papers published during the first 15 years of career (N= 70,310). Blue line: cubic polynomial regression fit, with grey areas representing 95%CI.

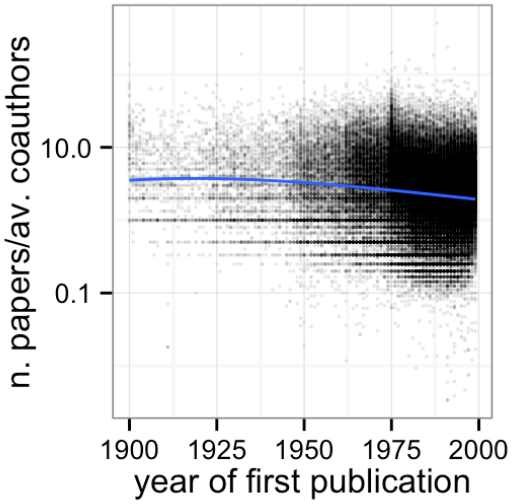


Figure 2. Ratio of total number of papers to average number of co-authors during the first 15 years of career (N= 70,310). Blue line: cubic polynomial regression fit, with grey areas representing 95%CI.

Several limitations to these results, however, remain to be addressed. First, since the likelihood of having two middle names is very unequally distributed amongst countries, our sample might not be sufficiently representative of the corpus of literature in the Web of Science. Second, our method might not be sufficiently robust against disambiguation errors for names from South-East Asian countries, a problem which might have skewed our results. Third, the Web of Science database does not cover a significant proportion of the literature, and its coverage varies by discipline and across the years. Future work will aim at adjusting for these factors, in order to verify whether scientists are actually publishing more or just collaborating more extensively.

Acknowledgments

The authors acknowledge funding from the Canada Research Chairs program as well as from the Social Sciences and Humanities Research Council of Canada.

References

Angell, M. (1986). Publish or Perish - A proposal. *Annals of Internal Medicine, 104*(2), 261-262.

de Winter, J., & Dodou, D. (2014). A surge of p-values between 0.040 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ, PrePrints* (2), e447v443

Fanelli, D. (2010). Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data. *Plos One, 5*(4). doi: 10.1371/journal.pone.0010271

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891-904. doi: DOI 10.1007/s11192-011-0494-7

Fanelli, D. (2013). Why Growing Retractions Are (Mostly) a Good Sign. *PLoS Med, 10*(12), e1001563. doi: 10.1371/journal.pmed.1001563

Fang, F. C., & Casadevall, A. (2011). Retracted Science and the Retraction Index. *Infection and Immunity, 79*(10), 3855-3859. doi: 10.1128/iai.05661-11

Hayer, C.-A., Kaemingk, M., Breeggemann, J. J., Dembkowski, D., Deslauriers, D., & Rapp, T. (2013). Pressures to Publish: Catalysts for the Loss of Scientific Writing Integrity? *Fisheries, 38*(8), 352-355. doi: 10.1080/03632415.2013.813845

Tijdink, J. K., Vergouwen, A. C. M., & Smulders, Y. M. (2013). Publication Pressure and Burn Out among Dutch Medical Professors: A Nationwide Survey. *PLoS ONE, 8*(9), 6. doi: 10.1371/journal.pone.0073381

van Dalen, H. P., & Henkens, K. (2012). Intended and Unintended Consequences of a Publish-or-Perish Culture: A Worldwide Survey. *Journal of the American Society for Information Science and Technology, 63*(7), 1282-1293. doi: 10.1002/asi.22636

VSNU (2015). Protocol for Research Assessments in the Netherlands (2015).

ISSI

2015

PATENT ANALYSIS

COUNTRY LEVEL STUDIES

Tapping into Scientific Knowledge Flows via Semantic Links

Saeed-Ul Hassan[1] and Peter Haddawy[2]

[1] *saeed-ul-hassan@itu.edu.pk*
Information Technology University, 346-B, Ferozepur Road, Lahore (Pakistan)

[2] *peter.had@mahidol.ac.th*
Faculty of ICT, Mahidol University, 999 Phuttamonthon 4 Rd, Salaya, Nakhonpathom 73170 (Thailand)

Abstract
We present a new technique to semantically analyze knowledge flows between countries by using bibliometric data. Using a new approach to keyword-based clustering, the technique identifies the main topics of the research output of a country, as well as the main topics of the citing research of other countries. In this way it provides insight into how research produced by one country is used by others. We present a case study to illustrate the use of our proposed technique in the subject area of Renewable Energy during 2005-2010 using data from the Scopus database. We compare the Japanese and Chinese papers that cite the scientific literature produced by researchers from the United States in order to show the difference in the use of same knowledge. While the Japanese researchers focus on research areas such as efficient use of Photovoltaics and Superconductors, Chinese researchers focus in areas related to Power Systems, Power Management and Hydrogen Production. Such analyses may be helpful in establishing more effective multi-national research collaboration.

Conference Topics
Methods and techniques; Country-level studies

Introduction
The research collaboration facilitated by the Internet and the greatly increased global mobility of researchers have resulted in a new highly dynamic global marketplace for ideas. The possession of knowledge, the value of which depreciates at an increasingly rapid rate, is no longer as valuable as the ability to participate in the knowledge flows associated with these marketplaces. As observed by Hagel et al. (2009) in the context of business competitiveness, "Knowledge flows – which occur in any social, fluid environment where learning and collaboration can take place – are quickly becoming one of the most crucial sources of value creation". Similarly in Science, understanding a research landscape increasingly requires understanding the dynamics of the relevant knowledge flows.

International scientific leadership and influence are commonly viewed as important measures of a country's scientific intellectual strength. This has traditionally been measured in terms of international scientific collaboration and the ability of a country to attract strong researchers and graduate students from abroad. But a further, more direct measure is the extent to which results generated by a country's researchers are influencing and being utilized by researchers abroad, particularly researchers who are not yet directly collaborating with that country's researchers.

In this paper we present a new technique to measure and semantically analyze knowledge flows between countries by using publication and citation data. We select a set of papers authored by the researches of a given source country. Further, we identify the papers cited by the papers only authored by researchers from outside the source country. We cluster these internationally cited papers to identify the main topics. Then, we procure the sets of papers (authored by researchers outside the given country) citing each of the topic clusters. Finally, we in turn cluster each set of citing papers to again identify main topics in order to identify how the knowledge from the topics in the cited papers is being used.

Related Work

In bibliometrics there have been efforts to measure knowledge flows using scientific literature at different levels of detail, namely: among scientists, among journals, among subject categories, among institutions and among countries.

Zhuge (2006) argues that ideas in a scientific article inspire new ideas, which will be recorded and published as new articles after peer review. Therefore, citations between scientific articles imply a knowledge flow from the authors of the article being cited to the authors of the articles that cite it. Zhou and Leydesdorff (2007) use journal-journal citation analysis to investigate international visibility of journals. Zhou et al. (2010) also use journal-journal citation analysis to study the specialization of a research community within a discipline. Johannes and Guenter (2001) measure knowledge export and international visibility of journals by determining the unique subject fields to which the citing journals have been assigned and the unique countries to which the citing authors belong, respectively.

Rowlands (2002) proposes a method to measure the spread of scientific knowledge that is published in a journal. He focuses on journals as units of spread and introduces an indicator to measure the spread of knowledge by looking at the number of different journals that cite the papers published in the primary journal, as shown in Equation 1.

$$RDI = \frac{U}{Cit}, \tag{1}$$

where U stands for the number journals that cite the papers published in the primary journal in a given time window (say T). Cit is the total number of citations received by the articles in the primary journal in T time window and the notion RDI is for Rowlands Diffusion Index. Naturally, diffusion can only increase in an absolute sense, however, empirical results show that the diffusion index proposed by Rowlands is negatively correlated with the total number of citations received (Rowlands, 2002). This leads Frandsen (2004) to provide a different diffusion index, as shown in Equation 2.

$$FDI = \frac{U}{Pub}, \tag{2}$$

where Pub stands for total number of publications in the primary journal, U is the same as above and FDI stands for Frandsen Diffusion Index. Note that Cit is replaced by Pub (i.e. publications). When publications do not change, the Frandsen Diffusion Index cannot decrease, and thus, the Frandsen Diffusion Index is positively correlated with the total number of citations.

Burrell (1991, 1992, 2005 and 2006) shows that the Leimkuhler Curve can provide an intuitive visual representation for the Gini Coefficient Index in giving graphical and numerical summaries of the concentration of bibliometric distributions. Guan and Ma (2007) illustrate the use of the Leimkuhler Curve to reveal the impacts of research outputs of countries. Using the Gini index, Liu and Rousseau (2010) study knowledge diffusion through publications and citations, as shown in Equation 3.

$$G = \frac{2q - 1}{N}, \text{where}$$

$$q = \sum_{i=1}^{N} i \frac{x_i}{M}$$

$$M = \sum_{i=1}^{N} x_i \tag{3}$$

N denotes the number of subject categories, and x_i denotes number of citations in journals mapped with a given subject category i. Note that the Gini index (Burrell, 1992, 2005) can be equally computed using Equation 4.

$$G = 1 - \frac{\sum_{j=1}^{\infty} (r(j))^2}{N.M}, \quad (4)$$

where M and N are the same as in Equation 3, $r(j)$ stands for the number of subject areas with at least j citations and the sum is finite as there is always a subject category with the largest number of citations. Note that Gini based indexes can only characterize the knowledge diffusion and do not quantify the volume of knowledge flow.

Ingwersen et al. (2000) present international citations as an indicator to measure export of knowledge produced by institutions. They measure knowledge export of institutes by calculating the proportion of citations received by a given institute from other countries (outside the host country where the institute is located) relative to total citations received by the institute. Using citation exchange among the scientific articles, we introduce a notion of International Scholarly Impact of Scientific Research (ISISR) to measure international knowledge flows among countries and institutions (Hassan & Haddawy, 2013). However, the measure of ISISR only quantifies knowledge flows and does not elucidate the contents of knowledge that flows across the countries.

The above survey discusses the salient research to quantitatively measure knowledge flows using bibliometric data. However, we believe that apart from the quantitative measures it is extremely important to analyze the contents of the knowledge flows. The scientific work of Zhuge (2009, 2010, 2011 & 2012) sets the theoretical base of semantic analysis in order to extract knowledge from large scale corpus.

Methodology

This section presents analytical techniques used to semantically analyze the knowledge flow from a given source country. We consider a set of papers P` authored by the researchers of a given source country in a given subject area in a given time window. Among the selected papers, we identify the papers P cited by the papers only authored by researchers from outside the source country. We cluster the papers from P to identify the main topics. We procure the sets of papers (authored by researchers outside the given country) citing each of the topic clusters. Next, we in turn cluster each set of citing papers to again identify main topics in order to identify how the knowledge from the topics in the cited papers is being used. The research topics are identified using our proposed Topic with Distance Matrix (TDM) model, an extension of the Latent Dirichlet Allocation (LDA) model proposed by Blei et al (2003).

A number of approaches to model scientific paper content have been proposed (Blei et al., 2003; Hofmann, 1999). These approaches are based upon the idea that the probability distribution over words in a paper can be expressed as a mixture of topics, where each topic is a probability distribution over words. We utilize one such popular model, LDA, proposed by Blei et al. (2003). In LDA, the generation of a paper collection is modeled as a three step process. First, for a given paper, a distribution over topics is sampled from a Dirichlet distribution. Then, for each word in the paper, a single topic is selected according to this distribution. At Last, each word is sampled from a multinomial distribution over words specific to the sampled topic.

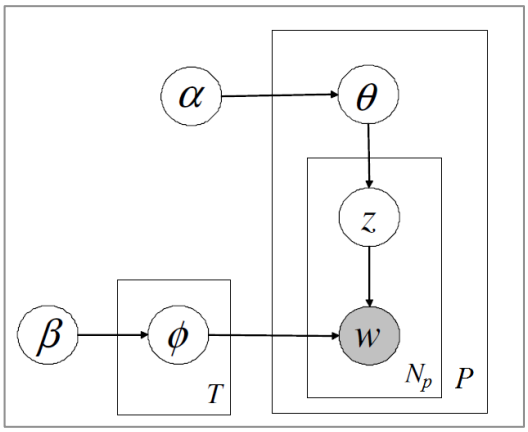


Figure 1. Latent Dirichlet Allocation (LDA) Model.

Using plate notation, the generative process corresponding to the hierarchical Bayesian model is shown in Figure 1. In this model, Φ stands for the matrix of topic distributions for each of T topics being selected independently from a symmetric Dirichlet prior (β). Θ is the matrix of paper specific mixture weights for these T topics, each being drawn independently from a symmetric Dirichlet prior (α). For each word, z denotes the topic responsible for generating that word, drawn from the Θ distribution for that paper, and w is the word itself, drawn from the topic distribution Φ corresponding to z. A paper p is a vector of N_p words, w_d, where each w_{id} is chosen from a vocabulary of size V and P is a collection of papers.

Estimating Θ and Φ provides information about the topics that participate in a publication corpus and the weights of those topics in each paper respectively. A variety of algorithms have been used to estimate these parameters, including variational inference (Blei et al., 2003), expectation propagation (Minka & Lafferty, 2002), and Gibbs sampling (Griffiths & Steyvers, 2004). To induce the probability distribution of Θ and Φ, LDA uses Gibbs Sampling which starts from randomly selected initial states and then revises distributions by changing topics to find correct distributions. Finally, the model provides topic-word relationship by the vector formed probabilistic representations.

Using the LDA, we obtain topic vectors where each value in the vector is associated with a given word that shows the probability of the word occurring under the given topic. For instance, vector T_1 ($word_1$: 0.3, $word_2$: 0.1, $word_3$: 0.2, …, $word_n$: 0.8) shows the probability distribution of all n words for the given topic t_1. Using this information, we represent each paper (from the set P) in the form of a vector where each value in the vector represents the probability distribution of a given word from vocabulary V in the paper for the topic under consideration (say t_1). For instance, P_1 ($word_1$: 0.4, $word_2$: 0.2, $word_3$: 0.0, …, $word_n$: 0.7) shows the probability distribution of words in the paper p_1 for the topic t_1. Note that if a word from V does not appear in p_1 then we assign default zero probability for that word.

Using the Minkowski distance between a given paper-vector P and topic-vector T, we choose papers in order to classify them as belonging to a specific topic (see Equation 5).

$$D = \sqrt{\sum_{i=1}^{n} | a_i - t_i |^2}, \tag{5}$$

where a_i denotes the probability of the term i in paper p_1 for the given topic T, and t_i denotes the probability of term i for the topic T. In order to obtain a set of papers relevant to topic T, a threshold TH is applied with the given percentage of the distance between the minimum and the maximum distance of paper vectors from T. Our experimental results show that the highest F-measure is achieved with TH = 25%. The size of a topic is determined by the

number of papers associated with it. The numbers of topics are determined by computing inter and intra topic similarity. We minimize inter topic similarity and maximize intra topic similarity to obtain the optimal number of topics. To compute the inter similarity between two topic, we use the Jaccard distance index (Jaccard, 1901).

Case Study: Semantic Analysis of Knowledge Flows across Countries in the Field of Renewable Energy

Dataset

We present a case study to illustrate the use of our technique in the subject area Renewable Energy. Using All Science Journal Classification (ASJC), we procured 46,518 publications (journal articles, reviews and conference papers) classified as Renewable Energy, a subarea of Energy(all) from the Scopus database during the time period 2005-2010

We procure 8,590 papers ($P`$) (journal articles, reviews and conference papers) published by researchers from the United States. Among the selected set of papers $P`$, we select 4,362 papers (P) which are cited by papers authored only by researchers from other countries. Further, we select candidate terms to represent each paper. In order to procure such terms, we use author defined keywords from the selected papers. In addition, we extract noun terms from the abstracts and titles of the papers using SharpNLP (http://www.codeplex.com/sharpnlp). We then identify synonyms of the selected noun terms using WordNet 3.0 (http://wordnet.princeton.edu/) and include them as candidate terms as well. Next, we apply the Porter Stemming algorithm (http://tartarus.org/martin/PorterStemmer/) to stem all the selected candidate terms. Finally, we feed this data to our TDM model.

Research Topics Cited by Researchers from Outside the United States in the Field of Renewable Energy

Figure 2 shows four research topics in the field of Renewable Energy cited by researchers from outside the United States. Using Wordle.Net (http://www.wordle.net/), we visualize the contents in each topic. Here, each topic is represented with the most frequently occurring author defined keywords collected from the papers in a given topic. The number of papers belonging to a specific research topic and the size of each research topic are written next to its respective topic. The research topics 1 and 4 are the largest topics cited by researchers from outside the United States. The topic#1 is the largest topic, containing 44% of the 4,362 papers. This topic covers research work related to Solar Cells, Solid Oxide Fuel Cells (SOFC) and Proton Exchange Membrane Fuel Cells (PEMFC). The topic#2 is related to Hydrogen Production. This topic also covers research related to Steam Reforming, a method for producing hydrogen, carbon monoxide, or other useful products from hydrocarbon fuels such as natural gas. Finally, the topic#3 is about Li-ion batteries. Li-ion batteries are an important type of rechargeable battery, particularly used in mobile devices. Finally, the topic#4 covers research related to Sustainable Management. Next we explore how the researcher from different countries cites the knowledge produced by the United States.

Research Topics of the Publications Produced by Chinese and Japanese Researchers that Cite Papers Authored by Researchers from the United States

To understand the difference in the use of the same knowledge, we further analyse that how the scientific knowledge diffuses into other research topics used by different research communities. We compare publications of the researchers from China and Japan that cite the same knowledge produced by the researchers from the United States. We select topic#1 from Figure 2 (the largest topic cited by the researchers from outside the United States in the field of Renewable Energy during 2005-2010). This topic covers research topics related to Solar

Cells (including Thin Film Solar Cells, Solid Oxide Fuel Cells and Proton Exchange Membrane Fuel Cells). Furthermore, we procure all the papers (journal articles, reviews and conference papers) authored by researchers from China and Japan that cite papers in the selected topic. We then identify research topics of the selected Chinese authored and Japanese authored papers.

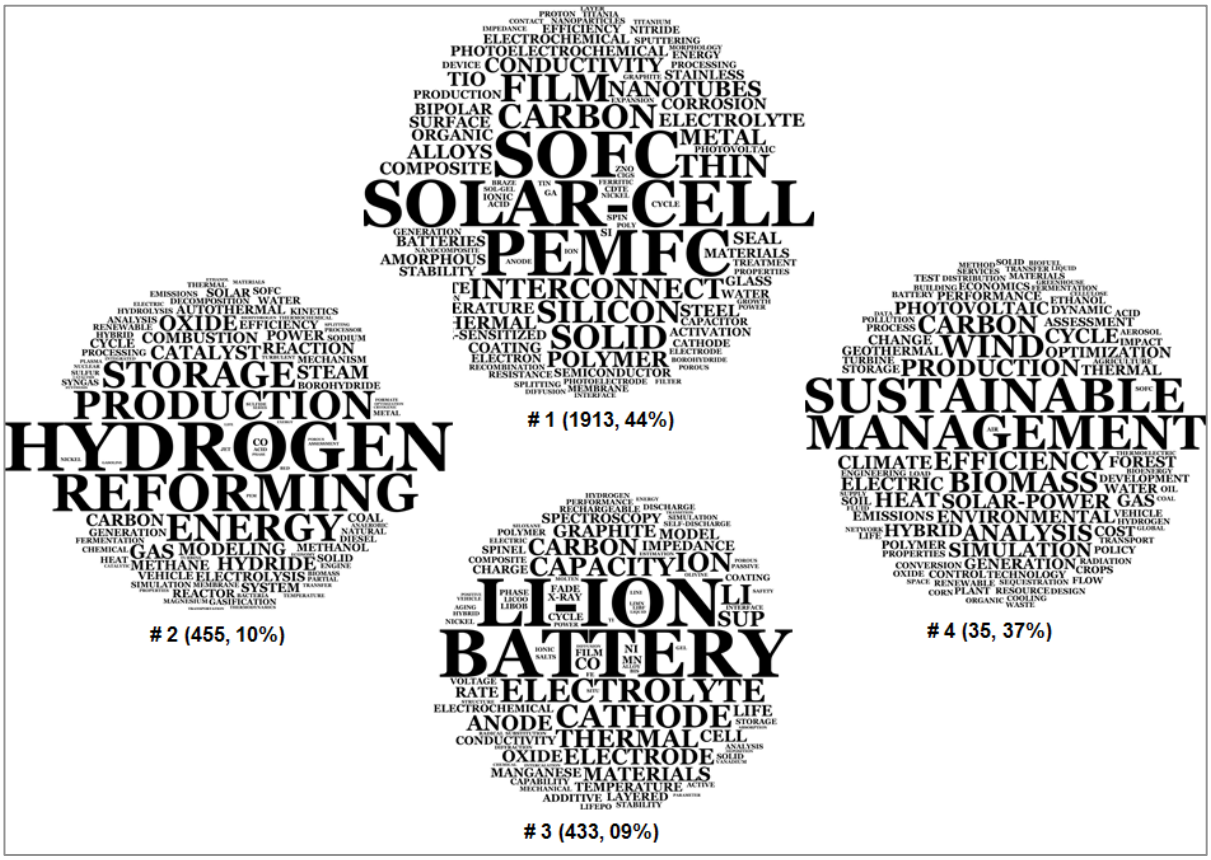


Figure 2. Research Topics Cited by Outside the United States in the Field of Renewable Energy during 2005-2010.

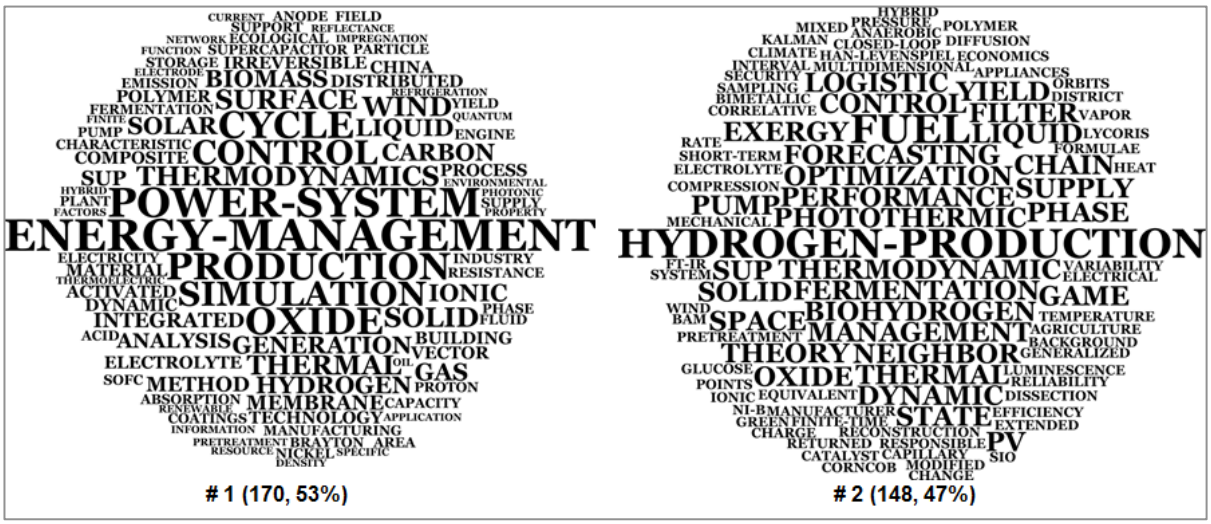


Figure 3. Research Topics of the Scientific Knowledge Produced by the Chinese Researchers (during 2005-2010) that cite the topic#1 in Figure 2.

Figure 3 shows research topics of the scientific knowledge produced by the Chinese researchers during 2005-2010 that cite topic#1 in Figure 2. In Figure 3, topic#1 mainly covers research related to Power Systems, Energy Management and Production. This topic is the largest topic which contains 53% papers out of 318. The topic#2 which contains 47% of the papers mainly focuses on Hydrogen Production.

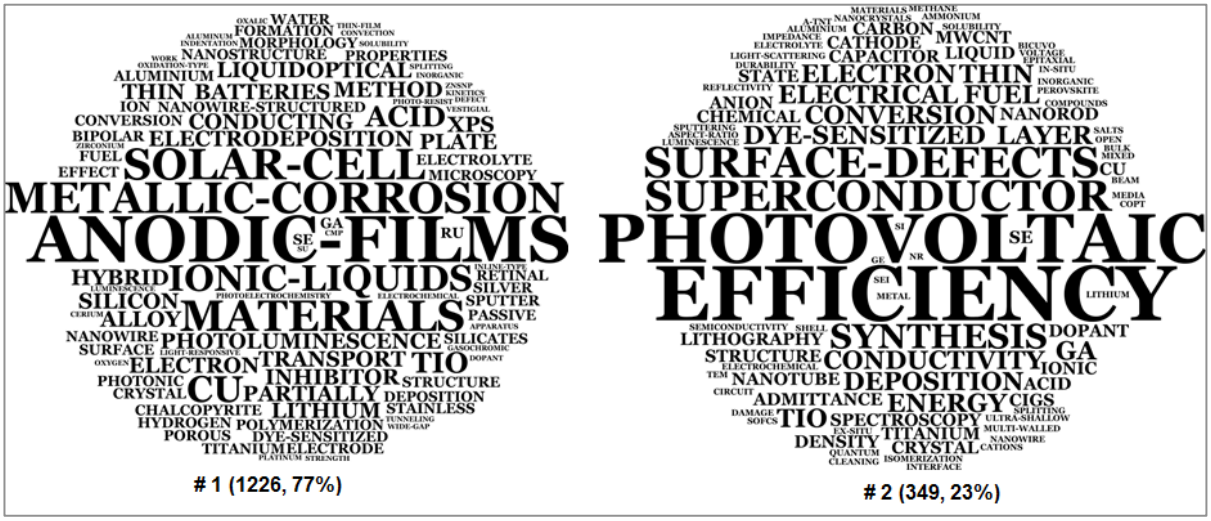


Figure 4. Research Topics of the Scientific Knowledge Produced by the Japanese Researchers (during 2005-2010) that cite topic#1 in Figure 2.

Figure 4 shows research topics of the scientific knowledge produced by the Japanese researchers during 2005-2010 that cite topic#1 in Figure 2. In contrast with China, the Japanese research community utilizes the same knowledge (produced by the United States) in rather different research themes. The Japanese researchers focus on topics related to Metallic Corrosion and Anodic Oxide Films (see topic#1 in Figure 4). Interestingly, we also find another topic (topic#2: 55 papers) describing the efficient use of Photovoltaics, Dye-sensitized Solar Cells and Superconductors. Note that Superconductors play a vital role in providing low-cost renewable energy.

Concluding Remarks

In this paper we have presented a new topic model with distance matrix, called TDM, to semantically analyze knowledge flows across countries by using publication and citation data. We have also presented a case study to illustrate the use of our proposed techniques in the subject area of Renewable Energy during 2005-2010 using data from the Scopus database. We have compared the Japanese and Chinese papers that cite the same scientific literature produced by the researchers from the United States in order to show the difference in the use of same knowledge. The study has shown that Japanese researchers focus in research areas such as efficient use of Photovoltaics, and Superconductors (to produce low-cost renewable energy). In contrast with the Japanese researchers, Chinese researchers focus in the areas of Power Systems, Power Management and Hydrogen Production.

The method of semantic analysis presented in this paper provides an understanding of the internationality of research not provided by studies of researcher mobility and co-authorship patterns. Our case study highlights the diversity in the ways that research produced by a country may be used in different international contexts, even within a relatively narrow research area. Such analyses may be helpful in establishing more effective multi-national research collaboration and in aligning collaboration with national priorities.

References

Blei, M., Ng, A. & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Burrell, Q. L. (1991). The Bradford distribution and the Gini index. *Scientometrics, 21*, 181–194.

Burrell, Q.L. (1992). The Gini index and the Leimkuhler curve for bibliometric processes. *Information Processing and Management, 28*(1), 19-33.

Burrell, Q.L. (2005). Measuring similarity of concentration between different informetric distributions: Two new approaches. *Journal of the American Society for Information Science and Technology, 56*(7), 704-714.

Burrell, Q.L. (2006). Measuring concentration within and co-concentration between informetric distributions: An empirical study. *Scientometrics, 68*(3), 441-456.

Frandsen, T. (2004) Journal diffusion factors: A measure of diffusion?. *Aslib Proceedings, 56*(1), 5-11.

Griffiths, T. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(1), 5228–5235.

Guan, J. & Ma, N. (2007). A bibliometric study of China's semiconductor literature compared with some other major Asian countries, *Scientometrics, 70*(1), 107-124.

Hagel, J., Brown, J. & Davison, L. (2009). *Measuring the forces of long-term change: The 2009 shift index*. Deloitte Development LLC.

Hassan, S. & Haddawy, P. (2013). Measuring international knowledge flows and scholarly impact of scientific research, *Scientometrics*, *94*(1), 163–179.

Ingwersen, P., Larsen, B. &. Wormell, I. (2000). Applying diachronic citation analysis to ongoing research program evaluations. *In B. Cronin & H.B. Atkins (Ed.), The Web of Knowledge (pp. 373-387)*. Medford, N.J.: Information Today, Inc. & American Society for Information Science.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547–579.

Johannes, S. & Guenter, G. (2001). Citation rates, knowledge export and international visibility of dermatology journals listed and not listed in the Journal Citation Reports. *Scientometrics, 50*(3), 483-502.

Liu, Y. & Rousseau, R. (2010). Knowledge diffusion through publications and citations: A case study using eSI-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology, 61*(2), 340-351.

Minka, T., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. *Proceedings of the Eighteenth Conf. on Uncertainty in Artificial Intelligence*, 352–359.

Rowlands, I. (2002). Journal diffusion factor: A new approach to measuring research influence. *Aslib Proceedings, 54*(2), 77–84.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic Author-Topic Models for Information Discovery. *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington.

Zhou, P. & Leydesdorff, L. (2007). A comparison between the China scientific and technical papers and citations database and the Science Citation Index in terms of journal hierarchies and inter-journal citation relations. *Journal of the American Society for Information Science and Technology, 58*(2), 223-236.

Zhou, P., Su, X., & Leydesdorff, L. (2010). A comparative study on communication structures of Chinese journals in the social sciences. *Journal of the American Society for Information Science and Technology, 61*(7), 1360-1376.

Zhou, P. & Leydesdorff, L. (2007). A comparison between the China scientific and technical papers and citations database and the Science Citation Index in terms of journal hierarchies and inter-journal citation relations. *Journal of the American Society for Information Science and Technology, 58*(2), 223-236.

Zhou, P., Su, X. & Leydesdorff, L. (2010). A comparative study on communication structures of Chinese journals in the social sciences. *Journal of the American Society for Information Science and Technology, 61*(7), 1360-1376.

Zhuge, H. (2006). Discovery of knowledge flow in science. *Communications of the ACM, 89*(5), 101-107.

Zhuge, H. (2009). Communities and Emerging Semantics in Semantic Link Network: Discovery and Learning, *IEEE Transactions on Knowledge and Data Engineering*, *21*(6), 785-799.

Zhuge, H. (2010). Interactive Semantics, *Artificial Intelligence*, *174*, 190-204.

Zhuge, H. (2011). Semantic linking through spaces for cyber-physical-socio intelligence: A methodology, *Artificial Intelligence*, *175*, 988-1019.

Zhuge, H. (2012). Knowledge Flow, Chapter 5 in The Knowledge Grid - Toward Cyber-Physical Society, 2nd Edition, *World Scientific Publishing Co.*, Singapore.

Causal Connections between Scientometric Indicators: Which Ones Best Explain High-Technology Manufacturing Outputs?

R. D. Shelton[1], T. R. Fadel[2], P. Foland[3]

[1] *shelton@wtec.org*
WTEC, 1653 Lititz Pike #417, Lancaster, PA 17601 (USA)

[2] *tarek.r.fadel@gmail.com*

[3] *pfoland14@gmail.com*
ITRI, 518 S. Camp Meade Rd., Baltimore, MD 21090 (USA)

Abstract
Scientometric models can connect indicators via cross-country correlations, but these are not enough to assert causality. Sometimes a causal connection can be argued from the physical process. In other cases the causality or its direction is not clear, and the Granger test is often used to clarify the connection. Here it was shown that gross expenditures on R&D (GERD) Granger caused scientific papers in the U.S., EU, and some others, which has policy implications. Granger causality also reinforces earlier findings on why the EU passed the U.S. in papers in the mid-1990s. Downstream, it is difficult to prove the connection between research and gross domestic product (GDP), since the contributions of science are diluted by other factors. New data allows a focus on a sector that is more closely associated with science: high technology (HT) manufacturing outputs. This value-added data permits more accurate models for today's international supply chains. Correlations show that business expenditures on R&D (BERD) and scientific indicators like patents are closely connected with HT manufacturing outputs. However for BERD, either direction of causality is plausible, and enough countries had significant results to show that causality can indeed be in either direction. The connections between papers and patents with HT manufacturing were also investigated; in several countries patents could be said to have Granger caused HT manufacturing.

Conference Topic
Country-level studies

Introduction
Correlation does not imply causality, unless it can be augmented with other evidence. Many researchers have found strong cross-country correlations between national R&D funding and intermediate indicators like papers and patents. These findings bolster the policy argument that researchers deserve more funding, but may sound self-serving. Here however, there is a convincing physical argument that there is philosophical causality. Everyone knows that it take resources to do research. In some "big science" fields like ITER and CERN, it takes international consortia to provide the necessary big funding. Even the lonely bibliometrician needs a computer, data and Internet access, time to do the work, and travel funds to present the results in some pleasant clime.

Downstream in the innovation process, many researchers have also tried to connect those papers and patents to outputs like gross domestic product (GDP), with mixed success. Here the physical connection is not so clear, because science is only one of many factors that are involved. For example, several Asian nations became export powerhouses with skyrocketing GDPs, based initially on imported technologies, which were not reflected in their national papers and patents. Instead, the "New Economic Geography" developed by Paul Krugman (1991) identifies the most significant factors for location of manufacturing, and location of R&D is not high on the list. (He won the 2008 Nobel Prize for this work.) Once prosperous, these nations did invest in indigenous innovation.

In these more difficult cases, analysts rely on statistical tests to provide some evidence of causality. The most common test was devised by Clive Granger (1969). (He also won the Nobel Prize, in 2003.) It is applied to two time series, which the analyst suspects may be related. In simplified terms, a time series x can be said to "Granger cause" a second time series y if the additional knowledge of x allows a significantly better prediction of y than simply the past history of y. The Granger test function is available in several statistical programs; the open source R software was used here (R Core Team, 2014). In the R version, the model order k is the same for both x and y. The null hypothesis that x does not Granger-cause y is not rejected, if and only if no lagged values of x are retained in the regression.
Let y and x be stationary time series. To test the null hypothesis that x does not Granger-cause y, one first finds the proper lagged values of y to include in an autoregression of y:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_k y_{t-k} + \text{residual}_t$$

Next, the autoregression is augmented by adding lagged values of x:

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_k y_{t-k} + b_1 x_{t-1} + \ldots + b_k x_{t-k} + \text{residual}_t$$

One retains in this regression all lagged values of x that are individually significant according to their t-statistics, provided that they collectively add explanatory power to the regression according to an F-test; adapted from Seth (2007). Here the smallest model order that produces significant results is preferred.
Granger testing is not a panacea. It requires that both series be stationary, and scientometric series usually fail the standard Augmented Dickey Freeman (ADF) test. This is often because they have trends such as inflation, population growth, or just more journals in the Science Citation Index (SCI). One normally has to de-trend series, usually by differencing them one or more times. Even when both series are stationary, the Granger test often fails, or worse, shows bi-directional causality, raising more questions than it answers. Furthermore, Granger causality is based on a postulate that cause must precede effect, but is this always true? In the stock market, the prospect of future events, like increased earnings, can influence present stock prices. Thus, one cannot prove true philosophical causality with Granger tests, but may be able to show that one series is a leading indicator for another. True causality has perplexed philosophers for millennia, so we are will not settle the question here. Instead we will just present the most interesting results from many Granger tests for scientometric indicators.

Background

Scientometric models are similar to econometric ones. A nation's innovation establishment can be considered to be an economic system that needs inputs of resources like labor and capital to produce outputs such as products and exports. System inputs and outputs can be measured using indicators. Figure 1 shows the relations between the system model and these indicators. This is a simplified linear model of a more complex situation. In reality there are feedback loops--e.g., an overall one that shows that sales of products can provide resources for investments in R&D.
Previous cross-country analysis showed that there is a strong correlation between inputs and intermediate indicators like papers. Leydesdorff (1990) regressed world share of publications in the SCI as output on GERD as an input. Shelton (2006) identified national inputs most important in encouraging papers. His model suggested that changes in the GERD share have been the driver of national changes in paper share, which can account for the rise of China since 2001 (Jin & Rousseau, 2005; Shelton & Foland, 2010). Later, the models were refined

using components of GERD as explanatory variables (Leydesdorff & Wagner, 2009). Similar models showed that government investments in R&D and higher education spending on R&D (HERD) were especially effective, helping to explain Europe's passing the U.S. in papers during the 1990s (Foland & Shelton, 2010). Conversely, the industrial component of GERD was shown to be more effective in encouraging patents (Shelton & Leydesdorff, 2012). Here these methods are applied to high-technology (HT) outputs as an overall measure of the success of a national innovation enterprise. The preliminary cross-country analysis (Shelton & Fadel, 2014) raised questions about the direction of causality, so a longitudinal approach for time series for individual countries has now been added, using the Granger test.

Such analysis is becoming more common in scientometrics, but sometimes with limited results. After considerable effort, Vinkler (2008) found no significant link between economic performance and research. Peng (2010) found some causality between R&D expenditure and GDP in China, but it is not clear that his series had the required stationarity. LC Lee, Lin & YW Lee (2011) used Granger testing of whether research papers can be said to cause GDP output—aggregated by regions. One result was that there is mutual causality between research and economic growth in Asia, but the results are not so clear in the West. Inglesi, Chang & Gupta (2013) tried Granger testing between research papers and economic growth in Brazil, Russia, India, China, and South Africa (the "BRICS"), which mostly failed to demonstrate causality, except for some positive results for India. Inglesi, Balcilar & Gupta (2014) got more positive results for the connections between U.S. paper output and GDP.

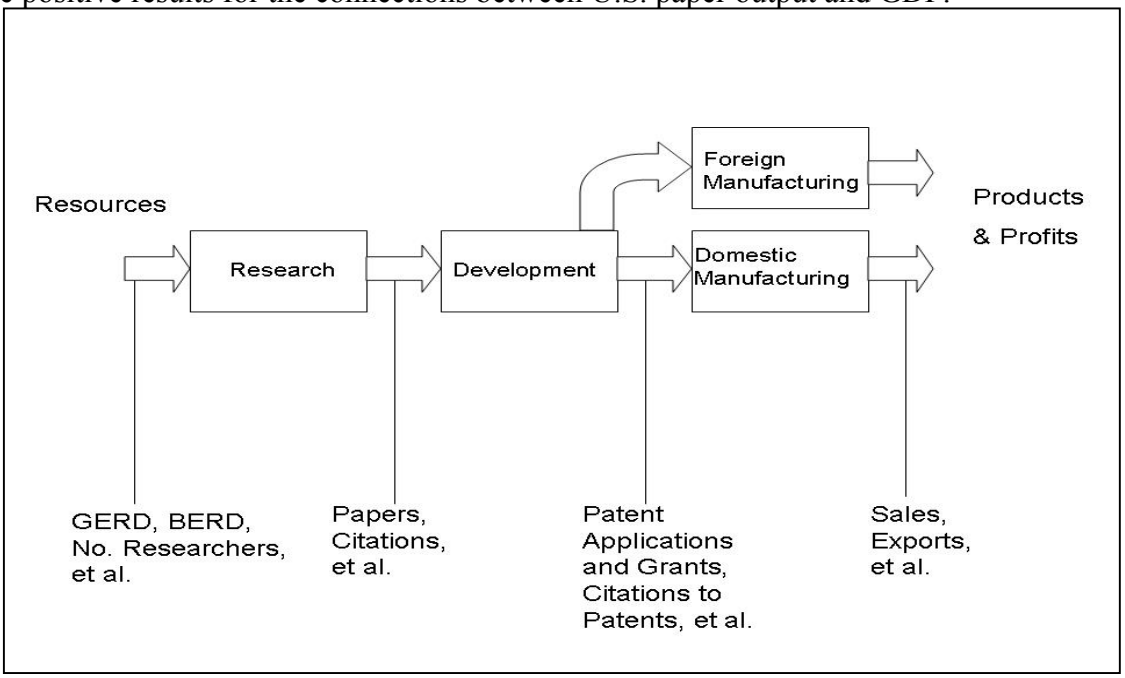


Figure 1. Linear model of an innovation enterprise with some indicators.

While there are some economic papers on factors that best explain *overall* international trade, there are relatively few that focus on the high-technology sector. One economic analysis of whether a country's high-tech exports (as a share of its overall exports) could be explained by R&D investment and country size was done by Braunerhjelm and Thulin (2008). They used the OECD data for 19 countries during 1981-1999. From their economic model, they concluded that overall R&D investment was significant.

Tebaldi (2011) used panel data to analyze factors that are most explanatory of high-technology trade. This approach adds data from more than one year to the usual cross-country analysis. Human capital, inflows of foreign direct investment, and openness to international trade were found to be the most significant of the factors he analyzed.

Data

Indicators like counts of papers and patents come from familiar sources like the SCI (Thomson Reuters 2015), (NSB 2014), and (OECD 2015). They provide insight into the success of national innovation enterprises. However, they are distant proxies for some of the quantities that the public cares most about: jobs, strength of their national economy, and survival of national industries. One scientometric measure of innovation that comes closer to these concerns is the performance of high-technology (HT) industries. Data on HT exports have been complied on a cash basis for decades by the OECD (2015) in its Main Science and Technology Indicators series. However, this measure of industrial output does not capture the nuances of where manufacturing really takes place. For example, the Apple iPad is assembled in China, but most of its components come from Japan, the U.S. and elsewhere (Xing, 2012). Recently a new dataset has been jointly developed by the OECD and the World Trade Organization for manufacturing output on a value-added basis, which avoids double-counting of imported components. This more accurate data, as summarized in (NSB, 2014), allows development of much-improved models that tie these key outputs to inputs like R&D investment. Figure 2 shows some national time series for this measure of HT manufacturing output. Forecasts show that China will soon take the world lead as the U.S. and Japan move final assembly of HT products to China. (Similar graphs for HT exports on a cash basis showed China taking the world lead in 2005.) The Europeans, especially the Germans, seem to have done less of this "off-shoring." There have obviously been big changes in the last decade, and scientometric models might provide insight on why, and what governments might do to respond.

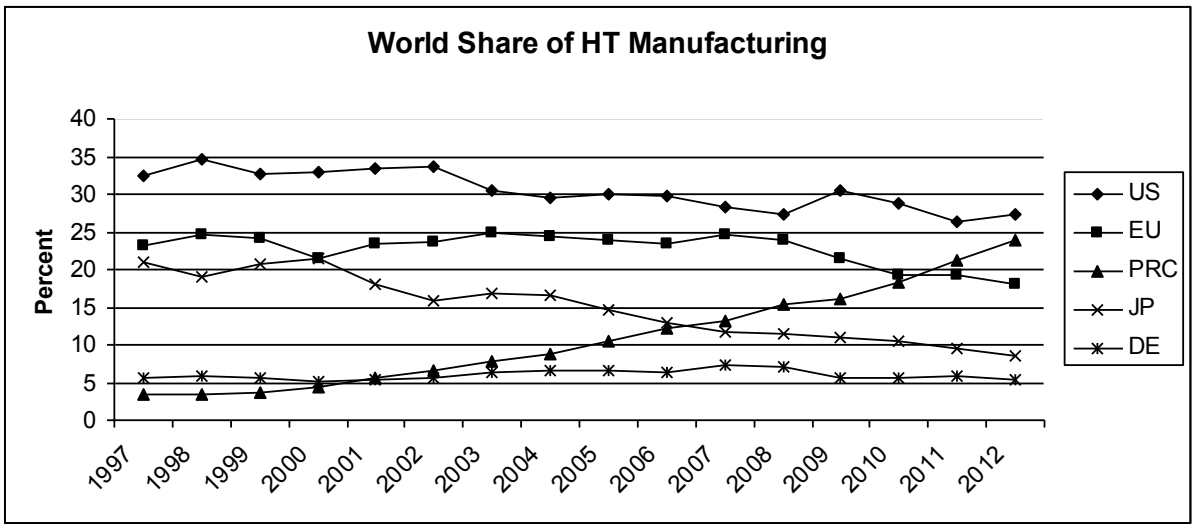


Figure 2. World share of manufacturing of high-technology products, on a value-added basis, for the United States, European Union (28), People's Republic of China, Japan, and Germany.

Causality Methods

Cross-country correlations over the countries in the OECD database are well known. Granger testing can be illustrated by revisiting the key results from Foland & Shelton (2010). That paper provided evidence that the EU passed the U.S. in papers in the mid-1990s because of a U.S. shift in research funding from government to industry, which was less effective in producing papers. At the time, this argument was based on cross-country correlations, and visual inspection of the U.S. and EU15 paper curves, which were very similar to their government GERD (GG) curves, just lagged by a couple of years. Granger testing can now add some quantitative evidence to this conclusion. First the series passed the ADF tests on the

data from 1988 to 2002, once second differences were calculated. The resulting Granger significance probabilities are in Table 1; bold entries are significant ($p < 0.1$).

Table 1. Significance probabilities for Granger tests of Government GERD component (USGGFF) causing papers (USPFF) on the NSI CD--or the reverse. FF means second difference. The "→" symbol means "Granger causes."

Model Order (k)	*USGGFF →USPFF*	*USPFF →USGGFF*
1	**p = 0.095**	p = 0.52
2	**p = 0.041**	p = 0.19
3	**p = 0.092**	p = 0.73

Thus the government GERD indicator can be said to "Granger cause" papers in the U.S. in this time interval. The most significant result was for a model order of two years, and there was no significant reverse causality. This provides additional evidence that relative changes in the Government GERD component led to the EU becoming much more efficient than the U.S. in producing papers, and led to its passing the U.S. in the mid-1990s to become the world leader in this indicator.

The Granger test has low power, that is, it often does not find significant results, particularly when the sample size is small. The sample size for Table 1 is only N = 15, preventing the use of higher model orders, so it is fortunate that some definitive results were obtained. To seek more definitive results, longer series were extracted for US, EU15, Japan, Netherlands, and Turkey from the Web of Science and the OECD for 1980 – 2012 where possible. After the second differences necessary for stationarity, this resulted in N = 30 samples for 1982-2012.

One experiment investigated whether total GERD (using constant $ and PPP weights) could be said to cause papers in the WoS (articles, letters, and reviews), with whole counts from the SCI-E and SSCI indexes. The results showed that U.S., EU15, and Japanese papers were indeed Granger caused by their national GERD with the significance probabilities in Table 2. None showed reverse causality. It did take a much higher model order to demonstrate Japanese causality. It was not possible to demonstrate significant results for the Netherlands or Turkey.

With these longer series, there is also the possibility that structural changes may take place over years. Sometimes a sliding window is used to examine shorter intervals within a longer one (Inglesi, Balcilar & Gupta, 2014). Here an auxiliary analysis simply examined the most recent years 2000 – 2012 (N = 13). The U.S. still exhibited Granger causality with the best result of p = 0.012 for a model order of k = 2. However, the other four country results for this shorter interval were not significant.

Table 2. Significance probabilities for Granger tests of GERD (G) causing papers (P) in the WoS (or the reverse) for 1983-2012. All used second differences.

Order (k)	*USG→USP*	*USP →USG*	*EUG →EUP*	*EUP→EUG*	*JPG→JPP*	*JPP→JPG*
1	**0.0067**	0.53	0.41	0.67	0.30	0.65
2	**0.0024**	0.89	0.53	0.94	0.53	0.53
3	**0.013**	0.90	0.76	0.82	0.54	0.54
4	**0.034**	0.91	**0.085**	0.92	0.54	0.79
5	0.10	0.86	0.14	0.96	**0.064**	0.80
6					**0.0029**	
7					**0.0090**	
8					**0.011**	

A similar test for Patent Cooperation Treaty (PCT) applications (OECD, 2014) in the U.S. was not so conclusive. Only for a model order of k = 6, could it be said that GERD Granger caused PCT patents, with p = 0.09. There was no reverse causality, however.

Another experiment tried to confirm a finding from Foland & Shelton (2010), that higher education spending on R&D (HERD) was closely associated with more papers. The dataset again included the U.S., EU15, the Netherlands, Japan, and Turkey, for the data range 1988-2002. Significant results were obtained only for the last two countries (Table 3). It was necessary to use fairly large model orders for Japan. The series passed the ADF tests with second differences, and there was no reverse causality for these model orders. Thus it can be said that, in Japan and Turkey at least, HERD Granger caused papers in these years. This might be useful for professors in those countries to mention in their battles for more funding.

Table 3. Does higher education spending Granger cause scientific papers?

Model Order (k)	*Japan HERD →Japan Papers*	*Turkey HERD →Turkey Papers*
1	p = 0.56	p = 0.24
2	p = 0.82	**p = 0.049**
3	p = 0.37	p = 0.12
4	**p = 0.090**	p = 0.21
5	**p = 0.016**	p = 0.30

Correlations for the Value-Added HT Manufacturing Indicator

Simple correlation over the 40 or so countries in the database of input resources in (OECD, 2014) can provide insight into which investments might be most productive in encouraging HT exports and manufacture. However, since many indicators simply increase with the size of the country, it is necessary to find explanatory variables whose correlations are much greater than those for measures like population or GDP. Furthermore, the U.S. and China are outliers; it is necessary to either omit them, or use log measures, if the contributions of smaller countries are to affect the results.

Table 4 from Shelton & Fadel (2014) shows the coefficients of determination (R^2) for two measures of performance of national HT industries with a number of explanatory or independent variables. For both measures, business expenditure on R&D (BERD) is best, with gross expenditure on R&D (GERD) not far behind. The correlations are far better for the new value-added data for HT manufacturing in the last column, than for the earlier exports on a cash basis. Indeed a quite accurate regression model can be constructed for this case (Equation 1), where NM9 is HT manufactures and BN9 is BERD, both in current dollars in 2009. Figure 3 shows the scattergram for this model.

$$\log NM9 = 0.385 + 0.944 \log BN9 \quad (R^2 = 84.1\%) \tag{1}$$

One would expect that there would be a delay between R&D investments and downstream benefits. For some indicators like patent grants, models that incorporate these delays can be more accurate (Shelton & Monbo, 2012). Here, correlations do not change much with lags, thus they did not improve the models enough to warrant the increased complexity. To see if a multiple linear regression would improve the model, a step-wise regression on HT manufacturing in 2009 was performed using the nine independent variables in Table 4. None of the other variables was significant in a multiple regression, once BERD was included as an explanatory variable, making a simple univariate regression without lags reasonable.

Table 4. Coefficients of determination (R^2 in %) of HT exports and overall HT manufacturing with explanatory variables in 2009. Uses log scales. More recent data downloads produce somewhat different correlations, and the values are sensitive to missing data points.

	Exports (Cash Basis)	*Overall Output (Value-Added)*
Papers SCI	41.7	71.0
Patents Triadic	48.8	69.9
Patent PCT Apps	34.3	61.5
GERD	44.8	79.8
BERD	49.0	84.5
Researchers	26.2	61.4
Business Researchers	29.3	71.6
Size GDP	27.3	56.9
Size Population	13.1	34.3

Despite the precision of the regression model in Equation (1), however, there is an alternate explanation for the trends of HT manufactures in the last decade. Could it be that HT manufacturing causes R&D investment, instead of the reverse? Indeed, it is the income from these sales that does provide some of those resources. OECD states that it picked the sectors for inclusion in the HT set precisely because these industries invest an extraordinary fraction of their income in R&D. And these correlations are too good to be true for BERD solely causing HT manufacturing--there are simply too many other factors that must also contribute. There have been frequent news accounts of Western and Japanese firms moving manufacturing to China and other low wage countries to increase their profits. China was also favored because its vast market offered potential for huge growth in HT sales.

This alternate explanation brings into question the efficacy of a nation increasing its HT manufacturing by encouraging greater business investment in R&D. It is possible that the results might be disappointing if the executives of the HT companies still prefer to locate the manufacturing abroad, the top path in Figure 1, so that some other nation reaps the benefits of the sales of HT goods. A policy remedy that addresses both explanations would be more likely to succeed. R&D investment policies could be coupled with trade policies that encourage location of manufacturing where the investments were made.

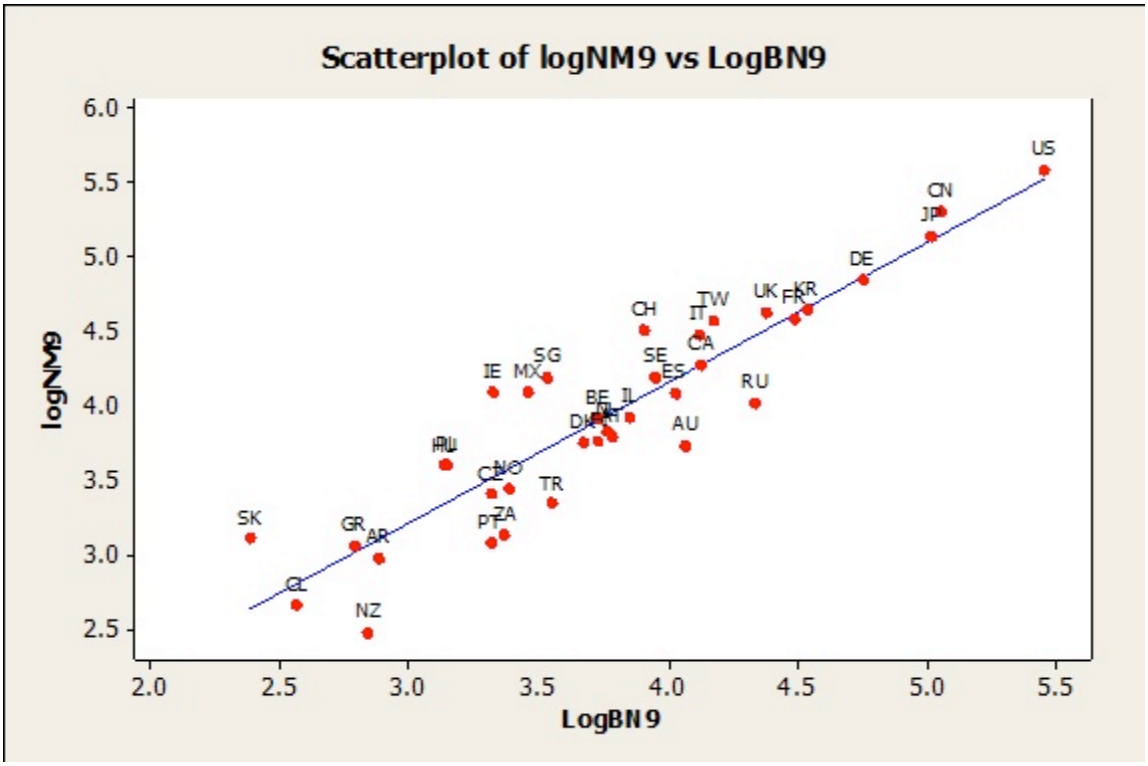


Figure 3. Scattergram of overall high-technology manufacturing vs. business expenditure on R&D in 2009. The cluster in the center contains BE, DK, IL, FI, and NL. HU and PL also overlap.

Further both manufacturing and BERD could be the results of an exogenous variable, some underlying third series. For example many of them seem to be closely tied to recent perturbations of the business cycle over the 1998 - 2011 data range available.

Causality Results for Value-Added HT Manufacturing

Table 1 shows that BERD has the highest correlation with HT manufacturing, so it will be analyzed first. Overall results for the sum of all countries in the OECD database were not significant. Findings for those individual countries with significant results are in Table 5. All are for model order k = 1, but orders up to k = 3 do not add countries to the list. Both series use current dollar values, and BERD used PPP weighting. The data ranges from 1999-2012.

Table 5. Does BERD Granger cause HT manufacturing (Mfg), or the reverse? Entries are significance probabilities; p < 0.1 is significant (bold type).

Country	*Mfg →BERD*	*BERD→Mfg*
Korea	0.21	**0.097**
Hungary	0.16	**0.0013**
Romania	0.57	**0.023**
PRC	**0.025**	0.32
Canada	**0.019**	0.43
Germany	**0.016**	0.19
Russia	**0.060**	0.54
Finland	**0.0014**	**0.010**

Of the some 24 countries with complete OECD data, 15 passed both ADF tests using second differences. The entries in bold type are the only ones that were significant from the Granger tests. While these results do not settle the question, they do show that (Granger) causality can indeed run in either direction for these indicators. Policymakers in Korea, Hungary, and Romania could benefit from knowing their country's business R&D investment did Granger cause its HT manufacturing output in these years, and may want to encourage more of this virtuous cycle. (Taiwan also showed this direction of causality for its available data from 2000-2012, using model order k = 2.) Chinese, Canadian, German, and Russian policymakers might be pleased to find that their country's HT manufacturing output Granger caused more BERD investment. Those in Finland would probably not find bi-directional causality very useful.

The second highest correlation in Table 1 was with overall GERD. As expected, these results were not as conclusive as those for the BERD component. Of the some 40 countries in the OECD Group, 30 had complete data. Of these 13 passed the ADF test for stationarity for both time series, using second differences. Using k = 1, only Hungary and Korea showed positive results (p = 0.0029 and p = 0.069, respectively). In the reverse direction of Mfg causing GERD, only Canada and Germany showed significant results (p = 0.026 and p = 0.0075 respectively. The Slovak Republic showed bi-directional causality with p = 0.091 for GERD causing Mfg and p = 0.025 in the reverse direction. These results seem to show that the higher correlation of BERD with manufacturing is necessary to get more definitive results.

BERD and GERD are not always thought of as scientometric indicators, though. What can be said about causality of HT manufacturing for traditional intermediate scientometics indicators like papers and patents? Only a couple of countries had significant results for papers, but the PCT patent applications were more interesting (Table 6). Using second differences, the ADF tests showed that 29 countries of the 37 countries with data had both series stationary, and 10 countries, plus the EU as a whole, showed Granger causality. The results are for order k =1, except for Denmark and the Czech Republic where k = 2. Two countries had bidirectional causality: Germany and the Netherlands.

Table 6. Do PCT international patent applications Granger cause HT manufacturing, or the reverse? Entries are significance probabilities; p < 0.1 is significant (bold type).

Country	*Patents→Mfg*	*Mfg→Patents*
EU28	**0.060**	0.14
Austria	**0.036**	0.24
Belgium	**0.046**	0.24
Canada	**0.060**	0.15
Czech Republic	**0.055 (k = 2)**	0.54
Denmark	**0.012 (k = 2)**	0.78
Korea	**0.063**	0.92
New Zealand	**0.0064**	0.11
Switzerland	**0.014**	0.40
Germany	**0.0014**	**0.0047**
Netherlands	**0.050**	**0.055**

So, there are quite a few countries where it can be said that their patenting activity Granger causes HT manufacturing output. This connection was suggested by the correlation results in Table 1, of course. There are good physical reasons that make this causality plausible, but the results do *not* imply that a national initiative to file more PCT applications would necessarily

result in more manufacturing. The Granger tests do add quantitative evidence that investments in science and technology indeed bear fruit in outputs that the public cares about.

Conclusions

For further work, statistical testing for causality can enrich study of the connections between scientometric indicators, and there are many others. However, the Granger test often fails, even when strong cross-country correlations exist and there are good physical reasons to suspect causality. There are other tests, like Toda & Yamomoto (1995), which can be employed. And more sophisticated data analysis might also help: other methods of de-trending, sliding windows for long series, panel data, et al. As always, one needs to be cautious of spurious results from data mining; running many tests is likely to turn up some positive results by chance.

The results here show that GERD did Granger cause papers and patents for the U.S., which is probably true for some others as well. This quantitative evidence bolsters the case that R&D funding is important for the success of a nation's science. In particular, the U.S. has a goal of maintaining its science leadership, but is rapidly falling behind in the funding race with China. In a rare good year, the U.S. increases its GERD by a real 3%; Chinese GERD has been increasing by more than 15% annually for decades.

New data on value-added manufacturing outputs provides quantitative insight on which inputs can be most effective in encouraging high-technology industries. Not surprisingly, there is a strong connection between such success and investments in R&D, particularly by the business sector. In countries where this can be demonstrated to be a cause of these successes, governments might wish to adopt policies, such as tax incentives, which can encourage such investment. Intermediate indicators like patents can also be good explanatory variables, showing quantitatively that traditional scientometric measures indeed provide useful information about outputs that directly affect a nation's prosperity.

Of course there are many other benefits of science and technology beyond the manufacture and sale of the HT products considered here. Science can lead to better healthcare, cleaner air and water, solutions of problems like global warming, improved communications that allow more extensive cooperation and collaboration, and many others. Most of these benefits can accrue to everyone, regardless of their nationality. Even in the competitive analysis of national market share of HT manufactures considered here, one should not lose sight of the overall performance of the sector. Worldwide sales have almost doubled over the last decade with only a slight pause during the Great Recession, reaching over $1.5 trillion in 2012. This growth has created millions of new jobs and a cornucopia of wonderful new products most people can enjoy--the ubiquitous cell phone has provided the first rapid communications in some of the poorest countries.

Acknowledgments

This work was partly funded by NSF cooperative agreement ENG-0844639. These findings do not necessarily reflect the views of NSF.

References

Braunerhjelm, P. & Thulin, P. (2008). Can countries create comparative advantages? R&D expenditures, high-tech exports, and country size in 19 OECD countries 1981-1999. *International Economic Journal*, *22*, 95-111.

Foland, P. & Shelton R.D. (2010). Why is Europe so efficient at producing scientific papers, and does this explain the European Paradox? *11th International Conference on S&T Indicators*, Leiden.

Granger, C.W.J., (1969). Investigating causal relations by econometric models and cross spectral methods. *Econometrica, 37*(3), 424-438.

Inglesi-Lotz, R., Balcilar, M, & Gupta, R. (2014). Time-varying causality between research output and economic growth in US. *Scientometrics, 100*, 203-216.

Inglesi-Lotz, R., Chang, T. & Gupta, R. (2013). Causality between research output and economic growth in BRICS. *Quality & Quantity.*

Jin, B. & Rousseau, R. (2005). China's quantitative expansion phase: Exponential growth, but low impact. *Proceedings of the 10th International Conference on Scientometrics and Informetrics,* Stockholm.

Krugman, P. (1991). *Geography and Trade*, Cambridge: MIT Press.

Lee, L.C., Lin, P.H., Chuang, Y.W., & Lee, Y.Y. (2011). Research output and economic output: a Granger causality test. *Scientometrics, 89*, 465-478.

Leydesdorff, L. (1990). The prediction of science indicators using information theory. *Scientometrics, 19*, 297-324.

Leydesdorff, L., & Wagner, CS. (2009). Macro-level indicators of the relations between research funding and research output. *Journal of Informetrics, 3*(4), 353-362.

NSB. (2014), *Science and Engineering Indicators 2014*. Arlington: National Science Board.

OECD. (2015). *Main Science and Technology Indicators*. Retrieved January 1, 2015 from: http://stats.oecd.org/Index.aspx?DataSetCode=MSTI_PUB

OECD. (2014). *Trade in Value-Added*. Retrieved February 23, 2014 from: http://www.oecd.org/sti/ind/49894138.pdf

Peng, L. (2010) Study on the relationship between R&D expenditures and economic growth of China. *Proceedings of the 7th International Conference on Innovation and Management*. 1725-1728.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved on Dec. 1, 2014 from http://www.R-project.org/

Seth, A (2007) Granger causality. Retrieved April 17, 2015 from http://www.scholarpedia.org/article/Granger_causality

Shelton, R.D. & Fadel, T.R. (2014). Which scientometric indicators best explain national performance of high-tech outputs? *15th Collnet Conference*, Ilmenau, Germany.

Shelton, R.D. & Monbo, S.D. (2012). Input-output modelling and simulation of scientometric indicators: A focus on patents. *Proceedings of the 17th International Conference on S&T Indicators*, pp. 756-767. Montreal.

Shelton, R.D. & Leydesdorff, L. (2012). Publish or Patent: Bibliometric Evidence for Empirical Trade-offs in National Funding Strategies. *Journal of the American Society for Information Science and Technology*, *63*(3), 498-511.

Shelton, R.D. (2006). Relations between national research investment and publication output: Application to an American paradox. *Scientometrics*, *74*(2), 191-205.

Shelton, R.D. & Foland, P. (2010). The race for world leadership of science and_technology: Status and forecasts. *Proceedings of the 12th International Conference on Scientometrics and Informetrics*, pp. 369-380.

Thomson Reuters (2015) Web of Knowledge. Retrieved on January 1, 2015 from http://wokinfo.com/

Tebaldi, E. (2011). The determinants of high-technology exports: A panel data analysis. *Atlantic Economic Journal,* 39:343-353.

Toda, H.Y. & Yamamoto, T. (1995) Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics, 66*, 225-250.

Vinkler, P. (2008). Correlation between the structure of scientific research, scientometrics, and GDP in EU and non-EU countries. *Scientometrics, 74*, 237-254.

Xing, Y. (2012). *The PRC's High-Tech Exports: Myth and Reality*, ADBInstitute Working Paper #357. Retrieved January 13, 2014 from: http://www.adbi.org/working-paper/2012/04/25/5055.prc.high.tech.exports.myth.reality/

Scientific Production in Brazilian Research Institutes: Do Institutional Context, Background Characteristics and Academic Tasks Contribute to Gender Differences?

Gilda Olinto[1] & Jacqueline Leta[2]

[1]*gilda@ibict.br*
Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Rua Lauro Muller, 455 - 4° andar, CEP 22290 – 160, Rio de Janeiro (–Brazil)

[2]*jleta@bioqmed.ufrj.br*
Universidade Federal do Rio de Janeiro (UFRJ), Av. Brigadeiro Trompowisky s/ nº, Prédio do CCS, Bloco B – sala 39, CEP 21941-590, Rio de Janeiro (Brazil)

Abstract
Despite the recent changes that occurred in the Brazilian science, this field is still strongly anchored on male figures, as it happened at the beginning of its institutionalization. This paper detaches the contribution of Brazilian Research Institutes for the development of Brazilian science and the importance of contextual, background and academic tasks involvement in scientific production in those institutes, giving special attention to gender differences. Data from government graduate programs evaluation forms were obtained for the analyses presented here which take into account all professor-researchers - 890 women and 1,470 men - affiliated to 72 graduate programs under the responsibility of 31 Brazilian Research Institutes (BRI), the majority of which supported by the Federal Government. The main findings include: women are a minority in those institutes, are concentrated in the health and biological sciences, show higher scientific production than their male colleagues, especially in journal articles and among those involved in highly evaluated graduate programs. We believe the set of results presented in this paper may contribute to a better understanding of women's participation not only in BRI, which are dedicated to specific scientific areas, but also in Brazilian science in general and so contribute to gender governmental policy.

Conference Topic
Country level studies

Introduction
The process of science institutionalization in Brazil started about a century ago, when in Europe and in the USA this activity was already structured, both in science academies and in research institutions. One of the first steps contributing to this process in Brazil was the creation, in 1900, of the Federal Serotherapy Institute at Manguinhos, in Rio de Janeiro (which was afterwards named Instituto Oswaldo Cruz), considered the first Brazilian Research Institute to win international recognition (Weltman, 2002). In the following decades, the first public universities were created, as the University of Brazil (later renamed Universidade Federal do Rio de Janeiro), founded in 1920, and the University of São Paulo, in 1934. However, only in the nineteen fifties, with the creation of the first agencies for the promotion of scientific development in the country, this process advanced significantly: CAPES assumed the responsibility of structuring and monitoring graduate programs (Masters and Doctorate), throughout the country, while the other agency, the CNPq assumed the task of promoting scholarships and research projects.

Considering the above mentioned initiatives, it is possible to say that, in the second half of the twentieth century, one witnesses a strong governmental effort towards structuring scientific institutions, and also an induced and spontaneous expansion of graduate programs. In 2010, three decades later, the country already counted with an extensive system of S&T, including: 83,170 doctors-researchers, 64,588 students enrolled in doctorate courses, 2,840 graduate programs, 27,523 research groups, and 452 research institutes and universities throughout the

country (MCTI, 2014). The effort to train and qualify S&T human resources, build up and modernize the infrastructure of research institutions and, more recently, create legal tools to allow the increase and maintenance of science funding, resulted in an outstanding growth of scientific output in the years 2000, especially output in journals indexed by international bibliographic databases (Regalado, 2010; Leta et al., 2013).

It is important to point out that such growth is also result of a combination of factors, besides the previously mentioned ones. Among these factors, the following could be mentioned: (1) the inclusion of Brazilian journals in databases, which resulted in an expressive growth of Brazilian production in international bases in the last few years (Leta, 2012); and (2) the creation of evaluation mechanisms of graduate programs, which stimulate and reward output in journals, mainly in international journals (Mugnaini & Sales, 2011). About this last aspect, it is important to highlight that graduate programs - which cover all areas of knowledge and a great part of the institutions of higher education and research, especially those of the public sector - became the leading stronghold of Brazilian science. Thus, policies and evaluation mechanisms directed to these programs are reflected in Brazilian scientific outputs and outcomes.

The institutionalization, growth and international recognition of Brazilian science have not promoted significant changes in aspects of scientific stratification, more specifically an equalitarian representation of men and women in scientific activities. Although the last decades have witnessed a significant growth in the number of women in the country's academic and scientific fields – in higher education, in graduate programs and as professors and/or researchers at universities and research institutions (INEP, 2007) – they are still a minority in several areas, in higher academic levels and in administrative functions of higher prestige (Olinto, 2011; Gauche, Verdinelli & Silveira, 2013). This scenario, although not exclusive of Brazilian scientific field, calls attention to the fact that, in face of the many recent changes that occurred in the country's science, this field is still strongly anchored on male figures.

Many factors support the maintenance of this scenario in Brazil and in the world, where women are excluded of certain areas, a phenomenon known as horizontal gender segregation, and they do not advance in their careers, a phenomenon known as the vertical gender segregation (Shienbinger, 2001). In a previous study (Leta et al., 2013), considering the symbolic value of different academic tasks that are part of the academic career, the hypothesis posed was that female Brazilian scientists would be involved in tasks of lesser prestige and, consequently, would be less productive and advance less in their careers than their male peers. We inquired into this issue examining productivity and involvement in academic tasks of the population of over 52,000 professor-researchers who participated in Brazilian graduate programs (our unit of analysis was each professor-researcher linked to a Brazilian graduate program, and whose academic characteristics and performance are yearly included in evaluation forms provided by the federal government). This study revealed a higher participation of men in articles published in annals of events, but major differences between male and female professors-researchers were not observed. Even though it may be considered positive the fact that both sexes have an equal share of academic-scientific tasks, the population analyzed in the mentioned study was very heterogeneous. Subtle differences were found, however, when the analysis considered the area of graduate work in which the professor-researcher was linked to. The health area was the closest one to our hypothesis: women tend to get more involved in activities of lesser prestige, like teaching graduate courses, and less involved in activities of higher prestige, like publishing in journals. Academic area and the nature of the institution are some aspects, among others, that may have an impact in the characteristics and the amount of scientific output of both men and women. In order to reduce diversity, in the present study, the focus turned to the participants of

graduate programs who are affiliated to Brazilian Research Institutes. The central question of this study is: how do gender differences in scientific performance are related to the characteristics of the academic and institutional context, as well as the involvement in several academic tasks of professor-researchers in graduate programs of Brazilian Research Institutes?

Research Institutes and Women

The largest part of the Brazilian Research Institutes belongs to the public sector and is linked to the Ministry of Science, Technology and Innovation (MCTI). Among the oldest is the National Observatory, founded in 1827, in the city of Rio de Janeiro. Presently there are thirteen other Research Institutes linked to the MCTI, the majority directed towards research in exact sciences and engineering. Other ministries also maintain Research Institutes, as the Ministry of Agriculture, responsible for Embrapa, created in 1973 with the purpose of developing research in agriculture; the Ministry of Health is responsible for the Brazilian National Cancer Institute (INCA), founded in 1961, and for the Oswaldo Cruz Foundation (at present – Fiocruz), created in 1900.

Until recently, women's presence and contribution at Research Institutes was poorly explored as a research topic in studies about gender and science. Among a few recent studies, the one by Brito Ribeiro (2011) inquired into the distribution of male and female researchers at Research Institutes linked to the MCTI in two career functions: researcher and technologist. This author points out to the small proportion of women in those institutes: about 30% in both types of careers. Nevertheless, that fraction still decreases substantially when the research areas of these institutions are considered. In the Brazilian Center of Research in Physics, for instance, there are only 17% of women in those two careers. The author also presents data about the distribution of men and women in higher prestige posts at these institutions, like presidency and boards of directors: out of 362 senior administrators, only 36 (10%) were occupied by women in 2010, a clear indication of vertical gender segregation. A more thorough analysis was done recently taking into account 571 researchers, with doctor degrees, affiliated to Fiocruz (Rodrigues, 2014), an institution that plays a central role in health research in the country. This author points out that male researchers have a *per capita* output quite superior to that of female ones. A different situation is found in Fiocruz, however, when the analysis focuses on administrative positions. Differently from other Research Institutes, especially those oriented towards exact sciences and engineering, Fiocruz is concerned with gender equity, and thus started a Pro-Equity Gender Program in 2009. This initiative might explain the large number of women in administrative positions in this institution. In 2013, out of 768 administrators with salary bonus, 382 (49.7%) were women, which is close to parity. However, women are still an absolute minority occupying the highest prestige posts, as president and directors.

The scenario previously described is shared by Research Institutes of other countries. One of the most prominent Research Institutes in the world, the Massachusetts Institute of Technology, has recently published a study on gender equity in the institution. Compared with previous studies (1999 and 2002), it showed major advances in two Schools. In the School of Science and School of Engineering, particularly, "the number of women in faculty increased significantly (from 30 to 52 in science and 32 to 60 in engineering) and in both schools women now hold several senior administrative positions" (Gillooly, 2011). However, despite these advances, women are still a minority, especially among those that occupy positions of higher prestige and salary, as tenured faculty members, of which women represent only 15% and 12% in the two schools, respectively. At the Centre National de la Recherche Scientifique (CNRS), the largest Research Institute in France, a country with a solid tradition in science and a pioneer in actions and policies that benefit women, Hermann

& Cyrot-Lackmann (2002) observed that women represent from 22% to 38% of the total CNRS's researchers and, what seems to be more significant, 31% of the research directors are in the highest prestige positions. Yet, as seen in the MCTI Institutes in Brazil, at the CNRS in France, this representation also varies according to the area of study: in Physical & Mathematical Sciences and Engineering Sciences only 12% and 9%, respectively, are women; and in Life Sciences, 28% of the research directors are women.

Different theories and models are considered by the literature to explain the phenomenon of female segregation in science and they include personal, biological, cultural, social and institutional aspects; and empirical studies based on these theories and models usually point out to gender imbalances favoring men (Barrios, 2013; Epstein, 2007; European Commission, 2009; Fox, 2005; Long, 1992; Meulders et al., 2010; Prpic, 2002).

The present focus on gender differences in institutional contexts suggests that male researchers would show better performance in different academic tasks and also present greater scientific production, like publishing in prestigious journals. Rewards for better performance would include the occupation of prestigious posts. Such arguments allow one to bring about the concept of scientific capital, proposed by Bourdieu (2003): a kind of symbolic or tacit capital, which opens opportunities and promotes recognition and which would tend to help perpetuate gender differences in science. Researchers with higher rates in publications and with high involvement in prestigious academic-scientific tasks accumulate scientific capital and, in a "snow ball" feedback effect, would tend to keep to themselves positions of higher academic prominence. In an opposite movement, researchers with less involvement in the more valued activities accumulate less scientific capital and would tend to be less involved in the more valued tasks, as well as to have a greater burden of less valued tasks, as, for instance, teaching assignments. Considering this model, the present study intends to investigate the relation between gender, academic background, institutional context, including the involvement in academic tasks, and scientific output of professor-researchers affiliated to the BRI.

Data collection and method

This study uses the documental analysis technique applied to information retrieved from three pre-established PDF forms with information used in the 2009 national evaluation of graduate programs (CAPES, 2013). Information provided includes aspects of academic and scientific performance as well as personal and academic characteristics of 52,294 professor-researchers affiliated to 2,247 graduate programs. Since a key characteristic, the professor-researcher's gender, was not included in CAPES' forms, a series of strategies was developed to allow for this classification (Leta et al., 2013).

For the present study, we have selected a subset of the 2009 original population and took into account information about all professor-researchers affiliated to 72 graduate programs under the auspices of 31 Brazilian Research Institutes (BRI), which were classified by us in three main groups: (1) supported by funds from the Federal government (Public/Federal), (2) supported by funds from State governments (Public/States) and (3) supported by the private sector (Private).[1]

[1] First group: Brazilian Center of Research in Physics (CBPF), Centre of Nuclear Technology Development (CNEN/CDTN), Institute of Nuclear Engineering (CNEN/IEN), Institute of Radio Protection and Dosimetry (CNEN/IRD), Oswaldo Cruz Foundation (Fiocruz), Research Centre (FIOCRUZ/ CPqGM), René Rachou Research Centre (FIOCRUZ/CPqRR), Institute of Military Engineering (IME), Institute of Pure and Applied Mathematics (IMPA), Brazilian National Cancer Institute (INCA), National Institute of Metrology, Quality and Technology (INMETRO), National Institute of Research in the Amazon (INPA), National Institute for Space Research (INPE), National Institute of Industrial Property (INPI), Technological Institute of Aeronautics (ITA), Botanical Garden Foundation of Rio de Janeiro (JBRJ), National Laboratory for Scientific Computing (LNCC)

It is important to mention that not all BRI are included in this study since a few of them do not have a graduate program under their responsibility. Examples are Embrapa and IBICT, major research institutes in the areas of agricultural sciences/biology and information science, respectively. These Institutes do have graduate programs but they are organized in collaboration with public universities.

Once the BRI were identified and data cleaned, all information was exported to a matrix of SPSS (Statistical Package for the Social Sciences), version 12. The population of the study represented in this matrix, and focus of the analyses presented here, can be so defined: BRI professor-researchers who participated in graduate programs in Brazil in 2009 (N=2,362). Among the variables that characterize each professor-researcher are: (a) personal and academic characteristics of the professor-researcher (gender, S&T area and year of doctoral title), (b) characteristics of institution of affiliation/ graduate programs (economic sector, area and evaluation grade); (c) academic roles performed by each professor-researcher (graduate courses, graduate advising, banking participation, project leadership) and (d) publication output (journal articles, articles in Annals and other types of publications). For the classification of S&T area of the graduate programs, we utilized the categories considered by CNPq (2013).

Results

The analyses are presented in two main sections: (a) characteristics of the institutional context in which professor-researchers participate and aspects of his academic background and (b) academic tasks and the scientific output of the professor-researchers, with emphasis given to gender differences.

Characteristics of the Institutions and of professor-researchers background

Table 1 shows the distribution of the 2,362 professor-researchers according to three macro-characteristics of the graduate programs of the BRI to which these professionals are linked: the economic sector, the area of knowledge and the performance grade.

Considering the economic sector, data show that the greatest part of professor-researchers are linked to the institutions maintained by the Federal Government and very few of these professionals are active in programs belonging to private institutions: only 3%. These results are different from those obtained for Brazilian graduate programs considered as a whole, which showed that 55% of the institutions belonged to the federal government, 30% states government and 15% to the private sector (CAPES, 2014).

The distribution of professor-researchers according to the academic areas of the BRI graduate programs (which represent the areas of expertise of these professionals) is, however, more homogeneous, although it is clear that a massive number of professors are concentrated in two major groups: Engineering and Exact Sciences, in one hand, and in Health and Biological Sciences, in the other hand. These areas together absorb 80.3% of the professor-researchers in the BRI.

---

and National Observatory (ON). The second group: Nuclear and Energy Research Institute (CNEN/IPEN), Institute of Medical Assistance to the State Civil Servants (IAMSPE), São Paulo Institute of Biology (IBSP), São Paulo Institute of Botanic (IBT), São Paulo Institute of Fishery (IP), Institute of Ecological Research (IPÊ), São Paulo Institute of Technological Research (IPT), Pernambuco Institute of Technology (ITEP) and Institute of Zoology (IZ / APTA). Third group: Recife Centre of Studies and Advanced Systems (CESAR), Brasilia Institute of Public Law (IDP), Latin American Institute of Research and Education in Odontology (ILAPEO) and Institute of Technology for the Development (LACTEC).

Table 1. Number and % of professor-researchers according to the economic sector, areas and grades of Graduate Programs from Brazilian Research Institutes – 2009.

ECONOMIC SECTOR	*N*	*%*
Public / Federal	1,933	81.8
Public / States	357	15.1
Private	72	3.0
Total	2,362	100
AREAS		
Engineering	489	20.7
Exact Sciences	476	20.2
Health Sciences	601	25.4
Biological Sciences	331	14.0
Human Sciences	71	3.0
Social Applied Sciences	14	0.6
Agrarian	31	1.3
Other/interdisciplinary	349	14.8
Total	2,362	100
CAPES EVALUATION		
Grade 2	38	1.6
Grade 3	356	15.1
Grade 4	623	26.4
Grade 5	693	29.3
Grade 6	489	20.7
Grade 7	163	6.9
Total	2,362	100

Table 2. Distribution (%) of professor-researchers from Brazilian Research Institutes according to academic areas and other characteristics by gender – 2009.

[1]

Contextual aspect	Percentage[1]	
	Women	**Men**
Professor-researchers [2]	37.7 (n= 890)	62.3 (n=1,470)
ACADEMIC AREAS	%	%
Engineering	8.5	28.1
Exact Sciences	10.8	25.9
Health Sciences	38.1	17.8
Biological Sciences	20.9	9.9
Other areas/interdisciplinary	21.7	18.4
TOTAL	100	100[3]
OTHER CHARACTERISTICS	% yes	% yes
Public / Federal	83.7	80.8
PHD before 2000	58.1	66.1
PHD abroad	16.4	30.0
Program with grade 2 to 3	14.5	17.9
Program with grade 5 – 7	59.0	55.8
Program with grade 6 to 7	20.6	31.9

Percentages calculated within each gender category. [2] We were not able to attribute the sex of two professor-researchers. [3] Partial and total percentages provided by SPSS.

The final contextual aspect, presented in table 1, refers to the performance grade of the graduate programs issued by CAPES. These grades are recorded in a scale from 2 to 7, and the meaning of these assessments is: from grade 5 the program is considered to be at a good

level, able to participate in institutional programs etc. Grades 6 and 7 are assigned to programs of high performance, and some aspects that contribute to the assignment of these grades, besides scientific productivity, are institutional agreements as well as institutional exchange of researchers, professors and students. In table 1, it is also possible to observe that the great majority of professor-researchers participate in programs that received grades from 5 to 7.

The following Table 2 aims to identify gender differences in institutional affiliation and aspects of personal background of the professors/researchers in BRI.

It is possible to note that women represent less than 40% of this population (N=890), a fraction similar to the one obtained in a previous study which focused on professor-researchers of all graduate programs in the country (Leta et al., 2013). Data also show that women are predominant in the areas of Biological and Health Sciences, whereas men form a great majority in Engineering and Exact Sciences, which points to the phenomenon of horizontal segregation of gender, a characteristic also observed in Brazilian graduate programs in general (Leta et al., 2013).

Table 2 also presents other relevant information related to gender, calling attention to gender differences favoring men: a higher proportion of men show longer careers than women (which in fact might reflect the recent increase in women's entrance in scientific careers), relatively earn more degrees abroad and participate more in graduate programs of higher prestige.

Gender and scientific production of professor-researchers of Brazilian Research Institutes

Table 3 shows the distribution of men and women according to the number and the kind of published work in 2009 - articles in journals, complete works in annals of events and abstracts in annals of events.

Table 3. Distribution (%) of professor -researchers from Brazilian Research Institutes by sex and number of journal articles, annals full article and annals abstract – 2009.

Publication	Journal Article		Annals full Article		Annals Abstract	
	Women	**Men**	**Women**	**Men**	**Women**	**Men**
0	30.6	38.7	76.7	66.7	68.9	80.3
1-2	33.9	31.7	14.7	15.6	15.7	10.9
3+	35.5	29.6	8.5	17.7	15.4	8.8
Total	**890**	**1,470**	**890**	**1,470**	**890**	**1,470**

These results call attention to the high percentage of both men and women without any work published in 2009, particularly those with zero annals full article and annals abstract. This table also stresses the higher women's performance as far as journal articles are considered: a lower proportion of women are included among those with zero contribution to this kind of publication and a higher proportion of this gender group are among those contributing with one or two journal articles, and especially among those considered more productive: three or more articles. It is important to keep in mind that this is the kind of published work that has more value in the scientific field in general, and is also the kind of publication that contributes the most to the grades attributed to the graduate programs by Brazilian Agencies. In Annals, a type of publication that is highly valued in technological fields, as Engineering, it is possible to see an alternate pattern between men and women: men with better performance in annals full articles and women in annals abstracts.

Scientific production is influenced by a large number of factors, including the academic area, years of academic experience (Bonaccorsi & Daraio, 2003), education abroad (Velema, 2012), etc. Table 4 presents the publication mean of the different types of publications of the BRI professor-researchers by gender, as well as by gender controlled by the above-mentioned factors – area, experience and education abroad –, and also the CAPES grade of the program, a particular aspect in the Brazilian scientific area.

Taking into account the general mean performance and gender, table 4 also shows, as in table 3, that women outperformed men in BRI in 2009 in mean number of journal articles (women published a 2.51 and men 2.12 articles, mean results with similar standard deviation) and the mean number of annals abstract (W=1.14 and M=0.75), while men attained higher means of annals full articles (W=0.74 and M=1.48). With these results, and considering the higher academic value attributed to publication in journal articles, one can say that women of the BRA show higher performance in relation to men.

Focusing on differences between academic fields, in Table 4, as expected, mean number of journal articles is higher in biological, health sciences and in exact sciences than in engineering. This difference could partially account for the women's higher general performance in the BRI, previously mentioned. But even considering journal publication in this specific group, it can also be observed that women in the biological and health areas publish, in average, more journal articles than men. Men, on the other hand, show higher performance in journal articles in exact sciences and engineering. These gender tendencies are not clear in the other two types of publication.

Table 4. Mean of types of publications of professor-researchers from Brazilian Research Institutes by sex considering academic area, Graduate Program evaluation and PHD period and PHD country – 2009.

	Publication Means					
	Journal Article		**Annals Full Article**		**Annals Abstract**	
	Women	**Men**	**Women**	**Men**	**Women**	**Men**
GENERAL MEAN PERFORMANCE	2.51	2.12	0.74	1.48	1.14	0.75
ACADEMIC AREA						
Engineering	0.99	**1.11**	2.66	2.96	0.45	0.32
Exact Sciences	2.24	**2.71**	1.88	1.42	0.86	0.65
Health Sciences	**2.99**	2.90	0.28	0.23	1.26	1.51
Biological Sciences	**3.27**	3.19	0.09	0.07	1.25	1.26
GRADUATE PROGRAMS						
Low evaluated (2 and 3)	**1.12**	0.90	0.99	2.07	0.98	0.30
High evaluated (6 and 7)	**3.66**	2.52	1.23	2.26	0.47	0.45
PHD period						
Before 2000	2.97	**2.40**	0.72	1.60	1.07	0.76
2000 and After	1.88	**1.57**	0.77	1.25	1.23	0.74
PHD country						
Brazil	2.59	2.08	0.72	**1.27**	1.25	0.87
Abroad	2.19	2.25	0.88	**2.07**	0.59	0.49

Table 4 also shows that belonging to programs with higher grades seems to have a positive impact in the output of men and women in journal articles and annals full articles. However, what stands out in the comparison of the two types of program (low and high performance) is

that women's mean number of journal articles is much higher than men's in high performance programs, where men are predominant (Table 2).

Data also suggest that professional experience, estimated through the time elapsed since PHD conclusion, contributes positively, for both women and men, to a greater output in journal publishing. On the other hand, both gender groups with more recent PHD degrees tend to publish more annals full articles. The other factor considered - PHD country- suggests that being educated abroad is more relevant to male output: men educated abroad show a much higher performance than women in this category. Regarding this last result, it could be pointed out that full articles in annals is the type of output that appears more often in the technological areas, like engineering, where 20% of the professor-researchers of the BRI are institutionally related (Table 2). It is also possible to consider that this kind of publication, which is associated to the participation in events, especially international events, may contribute to the development of professional contacts, favored by the period of experience abroad. If this is the case, women are not profiting, as much as their male colleagues, of their experience abroad.

Professor-researchers have several assignments besides publishing results based on their research projects. These assignments comprise, among others, graduate teaching, dissertation advising, banking participation and tasks involved in project leadership. How the involvement with these assignments is related with their publication output, and how gender might interfere in this process is explored in table 5.

Table 5. Mean number of involvement in academic tasks of professor-researchers from Brazilian Research Institutes by publication level and gender – 2009.

Academic Task	Professor-researchers			
	with no journal article		with 3 or more journal articles	
	Mean		Mean	
	Woman	**Man**	**Woman**	**Man**
Graduate Teaching	0.90	1.10	1.17	1.08
MS Advisor	0.59	0.70	0.83	0.98
PHD Advisor	0.37	0.63	**0.80**	**0.98**
Banking participation	0.94	1.42	1.00	1.18
Project Leader	0.87	0.82	**1.64**	**1.37**

Table 5 show that, in average, those BRI professor-researchers who have not published in 2009 – those with zero articles – tend to have less involvement with the different academic tasks considered, notably involvement with doctoral degree advising and project leadership. Besides, the comparison between men and women shows that men, independently of publication quantity, tend to be more involved in academic tasks, except in graduate teaching and project leadership, in which women show higher performance, but only a small positive difference. Women higher involvement in this specific task - project leadership -, especially among the more productive ones, might contribute to explain their higher performance in journal articles as previously shown in tables 3 and 4.

Concluding remarks

This work focused on gender differences in scientific production of professor-researchers attached to in BRI, aiming at identifying how institutional and background aspects may be related do their production, as well as how the diverse academic tasks performance by these men and women might interfere with their scientific production.

Considering institutional and background aspects, the results show that these professor-

researchers are allocated in the public sector, are concentrated in four academic areas, and the majority in programs that received high grades from government evaluation process (Table 1). Results also show that women are a minority in those institutes and are concentrated in the health and biological science, whereas men are concentrated in engineering and exact sciences (Table 2). Women also show higher scientific production, especially in journal articles, the most valued type of academic publication (Tables 3 and 4). Women's performance is especially outstanding when they are involved in highly evaluated graduate programs. Female professor-researchers only show lower production output in relation to their male colleagues in journal articles of traditionally masculine areas: exact sciences and engineering. But male predominance in these areas is not consistently maintained when the other types of scientific productions are considered. The last results highlighted here refer to the involvement in academic tasks by level of production. Data show that the involvement of both men and women in those tasks seems to be positively related to their productive levels, especially PHD advising and project leadership. Men, however, tend to be more involved in most academic tasks, regardless of their productive levels, with the exception of project leadership, in which women are more involved, notably the highly productive ones (Table 5).

The originality of the data presented in this study is the inclusion of different types of scientific production in the analyses of gender differences in science, as well as the examination of associations of these different types of productions with contextual and academic background, as well as with involvement in academic tasks. The originality of this study is also in the selection of a particular study field: the research institutes that have an outstanding place in the development of modern science, as institutions created with the specific purpose of scientific development. Despite their relevance for the scientific field, only few studies about gender and science focus on these institutions. In Brazil, the great majority of BRI are supported by the Federal Government, are dedicated to specific scientific areas and the graduate programs under their responsibility are well recognized by the scientific community and, as data analyses shown here, tend to receive high grade marks from the national graduate programs evaluation. These indicators of excellence make it valuable the analysis of gender differences in those institutions aiming at contributing to better understand women's participation in Brazilian science and also contribute to gender governmental policy.

Intended further analyses with the BRI data will make use of statistical multivariate models trying to evaluate the relative contribution of the different contextual, background and academic tasks involvement, as well as gender in scientific production of professor-researchers. These analyses will help to indicate the importance of institutional and gender cultures, and patterns of academic practices in scientific production.

Acknowledgment

The authors are grateful to CNPq for financial support.

References

Barrios, M., Villarroya, A., Ollé, C., & Ortega, L. (2013). Gender inequality in scientific production. *Proceedings of ISSI 2013*, *1*, 811-818.

Bonaccorsi, A. & Daraio, C. (2003). Age effects in scientific productivity. The case of the Italian National Research Council (CNR). *Scientometrics*, *58*(1), 49-90.

Bourdieu, P. (2003). *Os usos sociais da ciência*. São Paulo: UNESP.

Brito Ribeiro, L.M.B. (2011). Gênero e Ciência: A Presença Feminina em Institutos Públicos de Pesquisa. Anais ANPAD - XXXV Encontro da Associação Nacional de Pós-Graduação em Administração –, Rio de Janeiro.

CAPES, Cadernos de Avaliação. Received December, 2014 from http://conteudoweb.capes.gov.br/conteudoweb/CadernoAvaliacaoServlet.

CAPES, GeoCapes. Distribuição de docentes, ano 2009. Received December, 2014 from http://geocapes.capes.gov.br/geocapes2/.

CNPq, Areas do Conhecimento. Available at: http://www.memoria.cnpq.br/areasconhecimento/index.htm Accessed in December 2013.

Epstein, C. (2007). Great divides: the cultural, cognitive, and social bases of the global subordination of women. *American Sociological Review*, *12*(1), 1-25.

European Commission. *She figures 2009*: *statistics and indicators on gender equality in science*. http://ec.europa.eu/research/sciencesociety/document_library/pdf_06/she_figures_2009_en.pdf.

Fox, M. F. (2001). Women, science and academia: graduate education and career. *Gender and society*, *15*(5), 654-666.

Fox, M.F. (2010). Women and men faculty in academic science and engineering social-organizational indicators and implications. *American Behavioral Scientist*, *53*(7), 997-1012.

Gauche, S., Verdinelli, M.A., & Silveira, A. (2013). Composição das equipes de gestão nas universidades públicas brasileiras: segregação de gênero horizontal e/ou vertical e presença de homosociabilidade. Anais do IV Encontro de Gestão de Pessoas e Relações de Trabalho. Brasília, DF.

Gillooly, P. (2011). MIT News. New report details status of women in science and engineering at MIT. Available at: http://newsoffice.mit.edu/2011/women-mit-report-0321

Hermann, C. & Cyrot-Lackmann, F. (2002). Women in Science in France. *Science in Context, 15*(4), 529–556.

INEP. (2007). A mulher na educação superior brasileira: 1991-2005 / Organizadores: Dilvo Ristoff ... [et al.]. – Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. 292 p. ISBN 85-86260-82-7.

Leta, J. (2012). Brazilian growth in the mainstream science: The role of human resources and national journals. *Journal of Scientometric Research*, *1*, p. 44-52.

Leta, J., Olinto, G., Batista, P.D., & Borges, E.P. (2013). Gender and academic roles in graduate programs: analyses of Brazilian government data. *Proceedings of ISSI 2013*, *1*, 796-810.

Leta, J., Thijs, B., & Glänzel, W. (2013). A macro-level study of science in Brazil: seven years later. *Encontros Bibli*, *18*, 51-66.

Long, J. S. (1992). Measures of sex differences in scientific productivity. *Social Forces*, *71*(1), 159-178.

MCTI. Indicadores Ciência, Tecnologia e Inovação de 2014. Tables 3.1.2, 3.5.2, 3.5.5 and 3.6.1. Received from http://www.mct.gov.br/index.php/content/view/740.html?execview=

Meulders, D., Plasman, R., Rigo, A., & O'Dorchai, S. (2010). Horizontal and vertical segregation. Meta-analysis of gender and science research – Topic report. *7th RTD Framework Programme of the European Union*.

Mugnaini, R. & Sales, D.P. (2011). Mapeamento do uso de índices de citação e indicadores bibliométricos na avaliação da produção científica brasileira. Anais ENANCIB - Encontro Nacional de Pesquisa em Ciência da Informação. *Brasília: Thesaurus*, *12*, 2361-2372.

Olinto, G. (2011). A inclusão das mulheres nas carreiras de ciência e tecnologia no Brasil. *Inc. Soc*., *5*(1), 68-77.

Prpic, K. (2002). Gender and productivity differentials in science. *Scientometrics*, *55*(1), 27-58.

Regalado A. (2010). Brazilian Science: Riding a Gusher. *Science*, *330*(6009), 1306–1312.

Rodrigues, J.G. (2014). A trajetória feminina na pesquisa na Fundação Oswaldo Cruz: um estudo exploratório. Tese (Doutorado em Informação, Comunicação em Saúde) – Fundação Oswaldo Cruz, Instituto de Informação Científica e tecnológica em Saúde, Rio de Janeiro, 2014.

Shienbinger, L. (2001). O feminismo mudou a ciência? Bauru, SP: EDUSC. p. 384.

Velema, T. (2012). The contingent nature of brain gain and brain circulation: their foreign context and the impact of return scientists on the scientific community in their country of origin. *Scientometrics*, *93*(3), 893-913.

Weltman, W. L. (2002). A produção científica publicada pelo Instituto Oswaldo Cruz no período 1900-17: um estudo exploratório. *História, Ciências, Saúde Manguinhos*, *9*(1), 159-86.

Comparing the Disciplinary Profiles of National and Regional Research Systems by Extensive and Intensive Measures

Irene Bongioanni[1], Cinzia Daraio[2], Henk F. Moed[2] and Giancarlo Ruocco[1,3]

irene.bongioanni@gmail.com, Giancarlo.Ruocco@roma1.infn.it
[1]Sapienza University of Rome, Department of Physics, Rome, (Italy)

daraio@dis.uniroma1.it, henk.moed@uniroma1.it
[2]Sapienza University of Rome, Department of Computer Control and Management Engineering Antonio Ruberti, Via Ariosto, 25, Rome, 00185 (Italy)
[3]Center for Life NanoScience@LaSapienza, IIT, Sapienza University of Rome, Viale Regina Elena 295, Rome (Italy)

Abstract
In this paper, by modeling national and regional research systems as complex systems, we compare the dynamics of their disciplinary profiles using extensive (size dependent) indicators as well as intensive (size independent) average productivity indicators of scientific production. Our preliminary findings show that the differences between the disciplinary profiles among countries in the world is of the same order of magnitude of the differences among European countries, that in turn, is of the same order of magnitude of the dynamics among regions within a country. While additional research (that is in progress) is needed to confirm these findings, we describe the main advantages (features) of our approach and outline its usefulness to support evidence-based policy making.

Conference Topics
Methods and techniques; Citation and co-citation analysis; Indicators; Science policy and research assessment; Country-level studies

Introduction, scope and structure of this paper
The dynamics of national or regional research systems is one of the most important topics in quantitative science and technology research. Interestingly, a lot of studies have analyzed the disciplinary specialization of countries (see e.g. Glanzel, 2000; Glanzel & Schlemmer, 2007; Glanzel et al., 2006, 2008; Hu & Rousseau, 2009; Tian et al., 2008; Wong, 2013; Wong et al., 2012; Yang et al., 2012; Horlings & Van den Besselaar, 2013; Radosevic & Yoruk, 2014) or have investigated the disciplinary specialization of regions within a particular country, or have conducted case studies on individual regions and/or on a few number of selected disciplines (see e.g. Zhu et al., 2009; Glanzel, Tang & Shapira, 2011).
Much less studied are the disciplinary profiles of European countries at the regional level. To the best of our knowledge there are not empirical analyses at European level, investigating the evolution of the disciplinary composition (i.e. the 27 Scopus Subject categories) of regions. Moreover, none of the existing studies have analyzed in a comparative way, the range of variability (briefly: the dynamics) of national and regional research systems which is the aim of our paper. We investigate here this dynamics in terms of both extensive measures of scientific production (i.e. total number of scientific publications, citations and so on) and in terms of intensive average scientific productivity (i.e. number of publications per author).
In particular, the investigation of the dynamics of intensive measures of scientific production has an important policy relevance. According to the macroeconomic theory, we have growth convergence when smaller (poorer) countries, in terms of output per capita (e.g. GDP per capita), grow faster than larger (richer) countries. In the context of research systems, we can say that there is a convergence if smaller scientific systems, in terms of scientific output per capita, grow faster than larger one. This is an important question, related to the policy

decision of supporting catching up countries depending on whether there is convergence or not. This question is extremely important also at the regional level, for which there is an increasing interest in the smart specialization of regions, defined in terms of technological specialization, linked to the degree of innovativeness of the regions, to develop effective policies of cohesion (McCann & Ortega-Argilés, 2013; Camagni & Capello, 2013). Despite the fact that scientific specialization is commonly considered as a relevant factor for the technological specialization of regions, there is not available evidence on the scientific specialization of regions and their dynamics. Even more scant is the empirical evidence aiming at analyzing the dynamics of the scientific profiles of regions together with those at the national level, to derive informative policies to support research at national and regional level, able to take into account the complementarity/substitution relationship between national and regional research systems. We try to fill this gap, providing an investigation of the dynamics of the disciplinary profiles at the national and regional level using extensive and intensive measures.[1]

Bongioanni, Daraio, Moed and Ruocco (2014) provided a first exploration at the world country level. In the current paper, the analyses are extended systematically in the following three manners.

a) The paper analyzes a series of both extensive (size dependent) and intensive (size independent) bibliometric indicators of research productivity, impact and collaboration. Table 2 gives a list of all indicators included in the study. Data was extracted from the Scopus database and relate to the scientific production of world countries and 27 Scopus subject categories from 1996 to 2012.
b) The analyses do not only relate to *national* research systems, but also to *regions* within European countries. In terms of the Nomenclature of Territorial Units for Statistics, NUTS-2 units were analyzed.
c) We describe the main features and advantages of our approach to investigate the scientific convergence of national and regional research systems.

The model

A spin glass is a disordered assembly of spins (e.g. dipole magnets) that are not aligned in a regular pattern. The term “glass” comes from an analogy between the “magnetic” disorder in a spin glass and the positional disorder of a conventional, chemical glass, e.g., a window glass. In window glass or any amorphous solid the atomic bond structure is highly irregular; in contrast, a crystal has a uniform pattern of atomic bonds. In ferromagnetic solid, magnetic spins all align in the same direction; this would be analogous to a crystal. The individual interactions in a spin glass are a mixture of roughly equal numbers of ferromagnetic bonds (where neighbors prefer to have the same orientation) and antiferromagnetic bonds (where neighbors tend to orientate in the opposite directions). These patterns of aligned and misaligned magnets create what are known as frustrated interactions - distortions in the geometry of atomic bonds compared to what would be seen in a regular, fully aligned solid. They may also create situations where more than one arrangement of spins is stable.

In the physics of complex systems, a mathematical framework is developed to analyze spin glass systems. This paper uses certain elements of this framework. National or regional research systems are conceived as analoga of spins and their complex interactions give rise to disordered, spin glass like, systems. Their orientation is described in terms of the distribution of a research system’s publication output or related bibliometric measures over the various research disciplines. A research system’s disciplinary orientation is described as a vector the

[1] This is the first step of our analysis. Further research will be subsequently devoted to the exploration and investigation of the link between scientific and technological profiles of regional and national research systems.

elements of which contain the percentage of publications in the various disciplines. The rationale for using the spin glass model lies in the ability to analyze the dynamical interactions among research units in a wider system analogously to the analysis of spin orientations in spin glasses.
The following Table 1 summarizes the analogy between the main physical notions of a spin glass model and the corresponding notions in the research system model (see also the Appendix of Bongioanni, Daraio & Ruocco, 2014).

Table 1. Spin glass model: main physical notions and their corresponding notions for research system.

Notion in the physical system	Notion in the Research system
Spin	Country/region
Spin components	Scientific disciplines
J couplings	Country-to-country or region-to-region interactions
Energy (it has to be minimized to find stable solutions)	Generalized cost function (to be minimized)
Overlap	Similarity measure

Within the framework of this model, Bongioanni, Daraio & Ruocco (2014) proposed to compare the disciplinary patterns of research systems, by computing the 'overlaps' quantities, that are similarity measures between disciplinary patterns, borrowed from the physics of complex systems. The main variables analysed here are the Pa(i) i.e. the shares of articles published in a subject category i for a given country (or region) a over the sum of publications made during 1996-2012. Similar variables are based on the number of citations received, or the number of internationally co-authored papers. Table 2 gives an overview of all indicators used in this study. The measure of the overlap between the pattern of disciplinary profiles of two countries a and b, $P_a(i)$ and $P_b(i)$ respectively, that is the measure of similarity between systems, is defined as:

$$q_{ab} = \frac{1}{D}\sum_{i=1}^{D} \sigma_a(i)\sigma_b(i),$$

where

$$\sigma_a(i) = \frac{P_a(i) - \langle P_a \rangle}{\sqrt{\langle {P_a}^2 \rangle - \langle P_a \rangle^2}},$$

in which <A> stands for average of A, $\sigma_a(i)$ and $\sigma_b(i)$ represent the normalised shares of the indicator considered, for country (or region) a and b, respectively; and D is the number of subjects or disciplines analysed, which in this study amounts to 27 and are derived from Scopus. We note that if we use as variables $\pi_a(i) = P_a(i) - \langle P_a \rangle$ instead of $P_a(i)$, q_{ab} coincides with the Salton's cosine (calculated with the variables π).
The overlap measure or similarity of profiles between two countries a and b, q_{ab}, ranges from −1, meaning precisely the opposite profile, to 1, meaning precisely the same profile, with 0 representing independence and intermediate values indicating in-between levels of similarity or dissimilarity. Moreover, the overlap can be calculated with respect to another country, with respect to an average or standard value or with respect to a given distribution.

Interpreting the distribution of the overlaps to shed lights on the dynamics of the overall system.

An interesting property of the computed overlap measures between two countries (or regions)' profiles relates to their distribution. The distribution of the overlap reveals whether there is a *converge*nce in the overall system towards a unique disciplinary profile or whether there is a divergence of the system towards different disciplinary configurations. In particular, according to Bongioanni, Daraio and Ruocco (2014) the interpretation of the distribution of the overlap values is as follows: one pick on one shows a convergence towards the *same* disciplinary profile for all countries, while two picks point to two *different* configurations of disciplinary profiles.

We point out that this is one of the main advantages of our approach compared to currently bibliometric approaches used for comparing disciplinary profiles. Although a systematic comparison of our approach with other existing methods is in progress, we think that our approach offers an easy way, based on the investigation of the distribution of the overlap, to check whether there is convergence or not without having to adopt one of the alternative methods developed in the theory of growth to measure convergence. The most applied method to assess convergence in this context, adopted also in the context of scientific convergence (see e.g. Horlings & van den Besselaar, 2013), is based on *regressions*. Within this framework (see e.g. Barro & Sala-i-Martin, 1992), it is said that there is *beta-convergence* (where beta is the coefficient of the initial level of per capita output in the growth regression) when poor economies tend to grow faster than rich economies (and hence the beta coefficient is lower than zero, implying that the higher initial level of output per capita negatively affects the growth rate). Another related concept is that of *sigma-convergence*, which happens when the dispersion of the output per capita decreases over time. The sigma-convergence is often measured by analyzing the variation of the standard deviation (or the coefficient of variation or the concentration) of the output per capita over time. However, this regression based approach has been questioned in the growth literature (see e.g. Durlauf, 2000) and other studies of convergence have applied different methods, including a test on the distribution of the output and how it evolves over time, reaching often very different results (see e.g. Durlauf, Kourtellos, & Tan, 2005). Our approach, offers an interesting alternative to estimate the convergence, by analyzing the distribution of the overlaps and their dispersion.

Another interesting property of our approach is related to the exploitation of the *ultrametric* structure of the overlap values to obtain "automatically" clusters of the national or regional research systems analysed, without having to carry out a specific clustering exercise.[2]

Note that the indicators reported in bold in Table 2 are average productivity indicators, that is intensive (size independent) indicators of the scientific production, while the others are extensive (size dependent) indicators of scientific production.

In this paper the following overlaps were computed:

- Of each main country in the world against all other countries, using a set of 41 countries, including all member states of the European Union and major countries from the rest of the world.
- Of each 27 European country against all other European countries, to provide an aggregate benchmark for the regional analysis.
- Of each NUTS-2 region against all other regions, using a set of 266 NUTS-2 regions in member states of the European Union.

[2] Research on this point is in progress.

Table 2. Indicators applied in the study

Indicator	Description
PUB	Number of articles (integer count).
PUBf	Number of articles (fractional counts based on authors affiliations).
C	Total citations (4 years window, i.e., for articles in 2006; citations are from 2006-2009).
CPP	Total citations per paper (4 years window, i.e., for articles in 2006; citations from 2006-2009).
HCPUB	Number of articles in top 10 per cent of most highly cited articles in a discipline.
PUBINT	Number of internationally co-authored papers.
PUBNAT	Number of nationally (but not internationally) co-authored papers.
PUBINST	Number of papers co-authored by members of different institutions within a country.
PUBSA	Number of non-collaborative (single address) papers.
NA	Number of publishing authors in a particular year, by discipline.
APUB	Number of articles (integer count) divided by NA
APUBf	Number of articles (fractional counts based on authors affiliations) divided by NA
AC	Total citations (4 years window, i.e., for articles in 2006; citations are from 2006-2009) divided by NA
ACPP	Total citations per paper (4 years window, i.e., for articles in 2006; citations from 2006-2009) divided by NA
AHCPUB	Number of articles in top 10 per cent of most highly cited articles in a discipline divided by NA
APUBINT	Number of internationally co-authored papers divided by NA
APUBNAT	Number of nationally (but not internationally) co-authored papers divided by NA
APUBINST	Number of papers co-authored by members of different institutions within a country divided by NA
APUBSA	Number of non-collaborative (single address) papers divided by NA

Legend to Table 2: Data was extracted from the Scopus database and relate to the scientific production of world countries and NUTS2 European regions for 27 Scopus subject categories from 1996 to 2012.

Results are presented in two sections. The first part explains the base notion of a disciplinary profile, compares pair-wise profiles of countries and NUTS2 regions, and analyzes the structure within the set of profiles. It focuses on one single indicator: the number of articles (PUBf) published in 2012. The second part analyzes also average productivity indicators (APUBf) and dynamical aspects.

Disciplinary profiles of countries and regions

Figure 1 shows large differences in the distribution of research articles among subject fields between USA and China. The first country has a strong focus on medical sciences and biomedical research, including biochemistry, genetics and molecular biology, neurosciences, and on social sciences and humanities. The latter shows a large publication activity in physical sciences and engineering: chemistry, materials science, physics, and engineering and computer science.

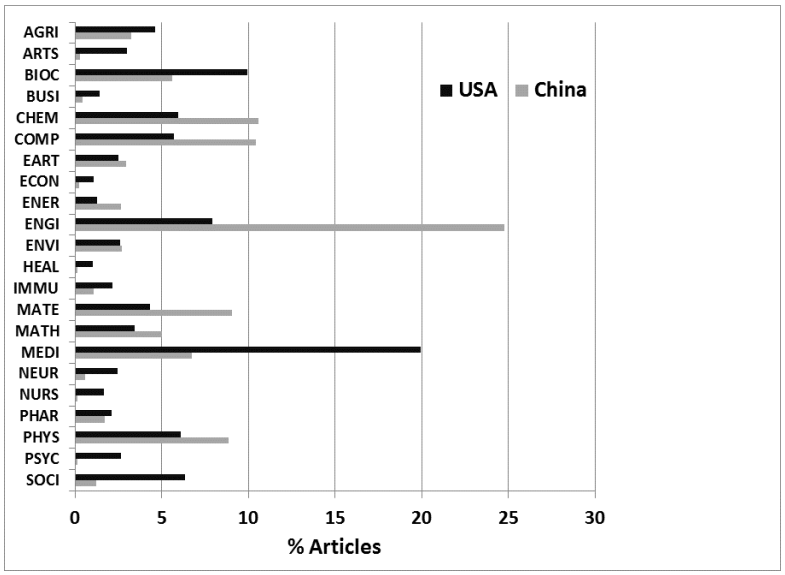


Figure 1. Disciplinary profiles of two countries large countries: China vs. USA. Data relate to the year 2012, and are extracted from Scopus.3 In this figure, four small disciplines have been left out: Dentistry, Decision Sciences, General, and Veterinary Sciences. Chemical Engineering is merged with Chemistry.

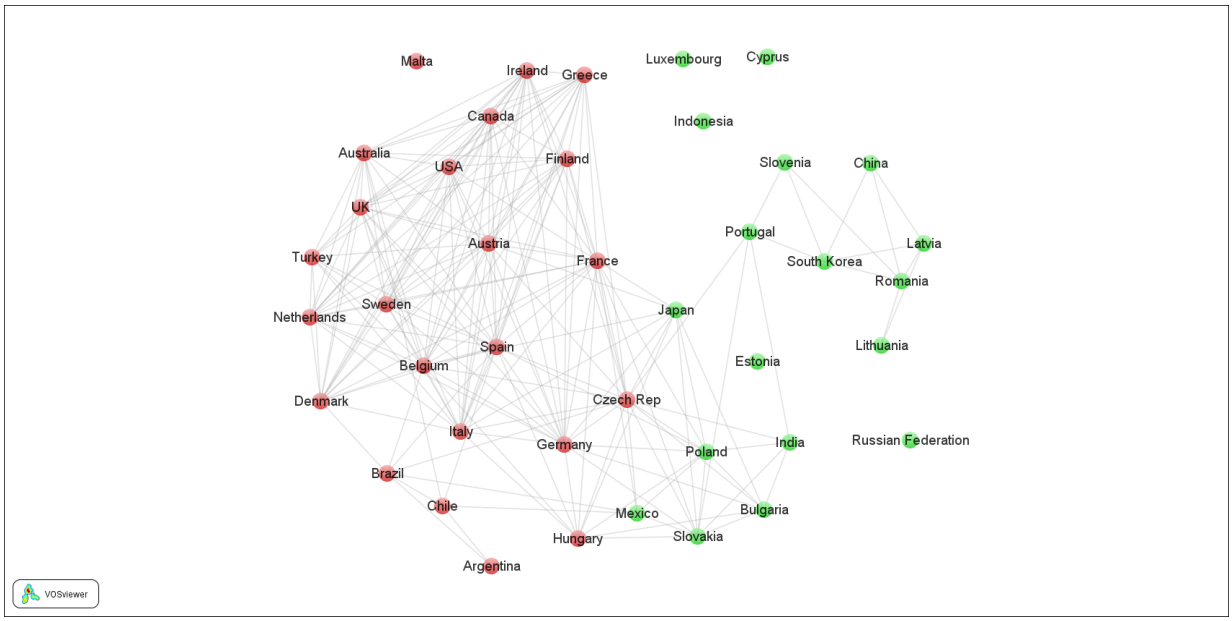


Figure 2. VoS-Viewer Map of the de degree of overlap of disciplinary profiles among 41 countries. For more details about VoS viewer, the reader is referred to www.vosviewer.com

Figure 2 shows a map of a set of 41 countries, including all member states of the European Community, and major countries from the rest of the world. Interestingly, the cluster module in the VoS Viewer identified two clusters of countries. These clusters correspond to the

3 The labels of the disciplines are the following: AGRI: Agricultural and Biological Sciences; ARTS: Arts and Humanities; BIOC: Biochemistry, Genet, Mol Biol; BUSI: Business, Managmnt, Accounting; CHEM: Chemistry; COMP: Computer Science; DECI: Decision Sciences; DENT: Dentistry; EART: Earth and Planetary Sciences; ECON: Economics, Econometrics and Finance; ENER: Energy; ENGI: Engineering; ENVI: Environmental Science; GENE: General; HEAL: Health Professions; IMMU: Immunology and Microbiology; MATE: Materials Science; MATH: Mathematics; MEDI: Medicine; NEUR Neuroscience; NURS: Nursing; PHAR: Pharmacology, Toxicology and Pharmaceutics; PHYS: Physics and Astronomy; PSYC: Psychology; SOCI: Social Sciences; VETE: Veterinary Sci.

different profiles illustrated in Figure 1. The countries indicated with red circles, located at the left hand side of the plot, tend to have a biomedical disciplinary profile, similar to USA and the Netherlands. At the right hand side a group of countries indicated by green circles tends to have a physical-sciences profile, like China, and Russia. Many Central and Eastern-European countries belong to this group: apart from South Korea, also India, Indonesia, Mexico, Portugal, and the small countries Luxembourg and Cyprus.

Several studies in the past have found differences in disciplinary profiles between countries. But to the best of our knowledge, no study has systematically analyzed geographical regions within countries. Figures 3 and 4 show results for the so called NUTS-2 regions. In total, 266 NUTS2 regions were identified. Table 3 presents the quantiles of the distribution of the number of published articles (year 2012) among regions. The distribution is highly skewed. The top 25 per cent of regions has published more than 4,146 articles in 2012. 5 per cent has published more than 11,612 articles. The bottom 25 per cent has published less than 496, and the bottom 10 per cent less than 89. Figure 3 shows disciplinary profiles of two pairs of NUTS2 regions: Inner London and the German city Stuttgart. The figure reveals the same main profiles as Figure 1 did at the level of countries: a biomedical profile in Inner London, and a physical sciences profile in Stuttgart.

Table 3. Quantiles of the distribution of number of publications among NUTS2 regions

Level	***Score***
Number of NUTS2 regions	266
Average articles/region	3,326
Level	*Quantile*
100% Max	46,451
90%	8,247
75% Q3	4,146
50% Median	1,815
25% Q1	496
10%	89
0% Min	1

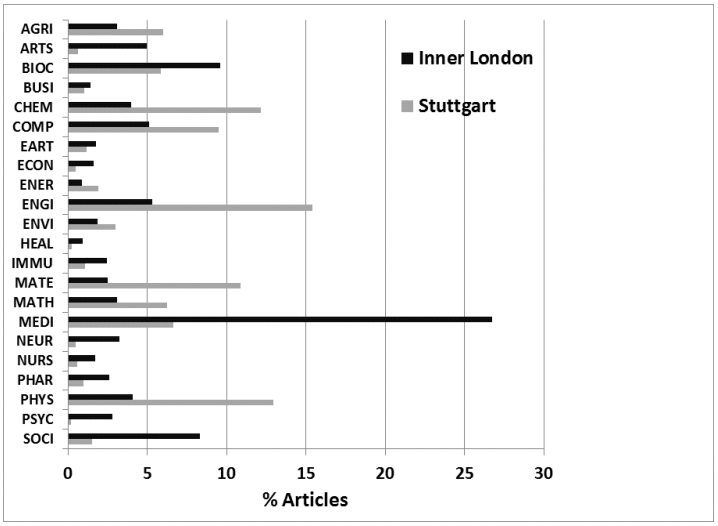


Figure 3. Disciplinary profiles of Inner London (UK) vs. Stuttgart (Germany)

different profiles illustrated in Figure 1. The countries indicated with red circles, located at the left hand side of the plot, tend to have a biomedical disciplinary profile, similar to USA and the Netherlands. At the right hand side a group of countries indicated by green circles tends to have a physical-sciences profile, like China, and Russia. Many Central and Eastern-European countries belong to this group: apart from South Korea, also India, Indonesia, Mexico, Portugal, and the small countries Luxembourg and Cyprus.

Several studies in the past have found differences in disciplinary profiles between countries. But to the best of our knowledge, no study has systematically analyzed geographical regions within countries. Figures 3 and 4 show results for the so called NUTS-2 regions. In total, 266 NUTS2 regions were identified. Table 3 presents the quantiles of the distribution of the number of published articles (year 2012) among regions. The distribution is highly skewed. The top 25 per cent of regions has published more than 4,146 articles in 2012. 5 per cent has published more than 11,612 articles. The bottom 25 per cent has published less than 496, and the bottom 10 per cent less than 89. Figure 3 shows disciplinary profiles of two pairs of NUTS2 regions: Inner London and the German city Stuttgart. The figure reveals the same main profiles as Figure 1 did at the level of countries: a biomedical profile in Inner London, and a physical sciences profile in Stuttgart.

Table 3. Quantiles of the distribution of number of publications among NUTS2 regions

Level	***Score***
Number of NUTS2 regions	266
Average articles/region	3,326
Level	*Quantile*
100% Max	46,451
90%	8,247
75% Q3	4,146
50% Median	1,815
25% Q1	496
10%	89
0% Min	1

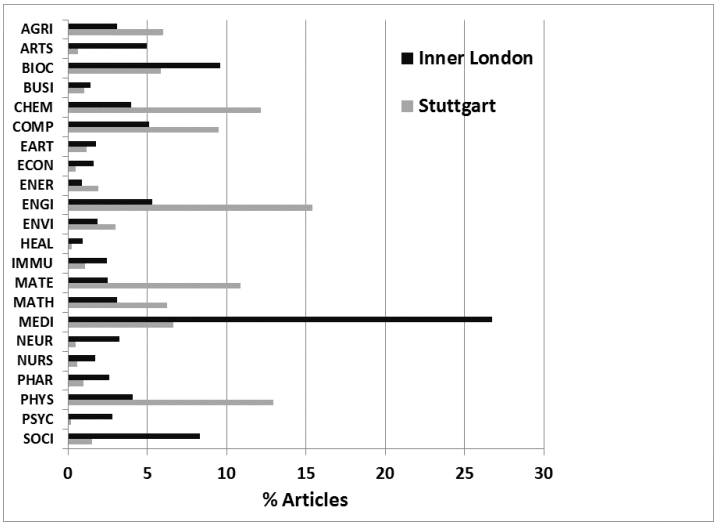


Figure 3. Disciplinary profiles of Inner London (UK) vs. Stuttgart (Germany)

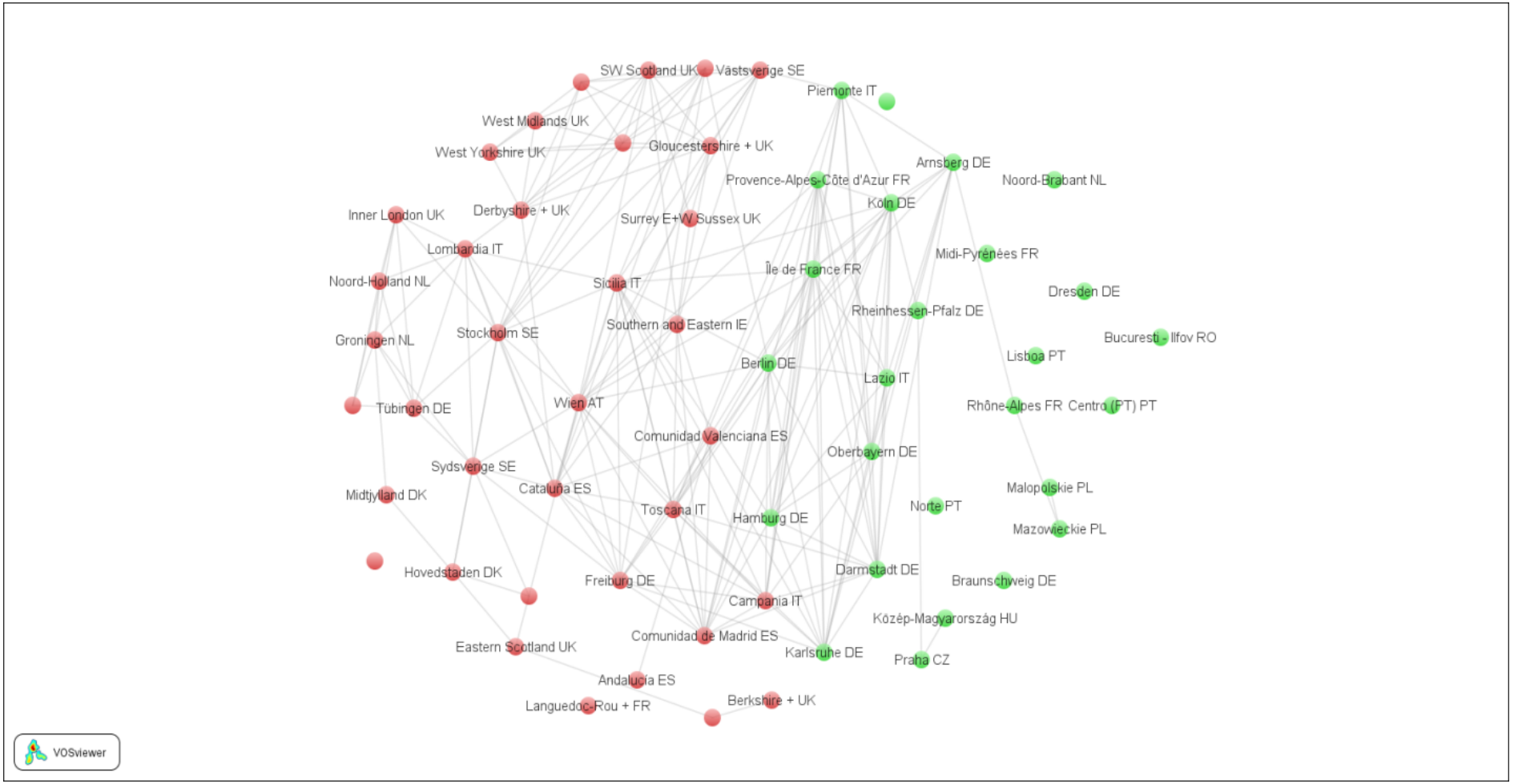


Figure 4. VoS-Viewer Map of the de degree of overlap of disciplinary profiles among 62 NUTS 2 regions. Results are based on an analysis of 62 NUTS2 regions. Due to inconsistencies in the primary data, regions from Belgium and Finland are missing in this graph. Not all circles have labels.

among countries

among regions

Analysis of distributions of overlap values

convergence towards a unique disciplinary profile

less dispersed

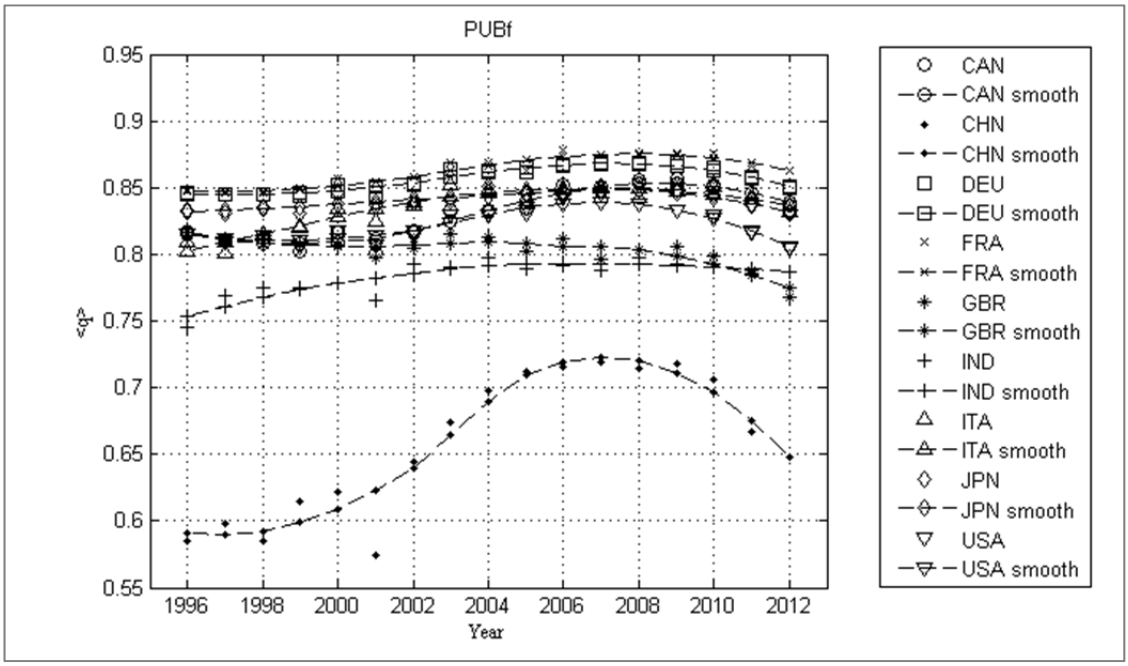
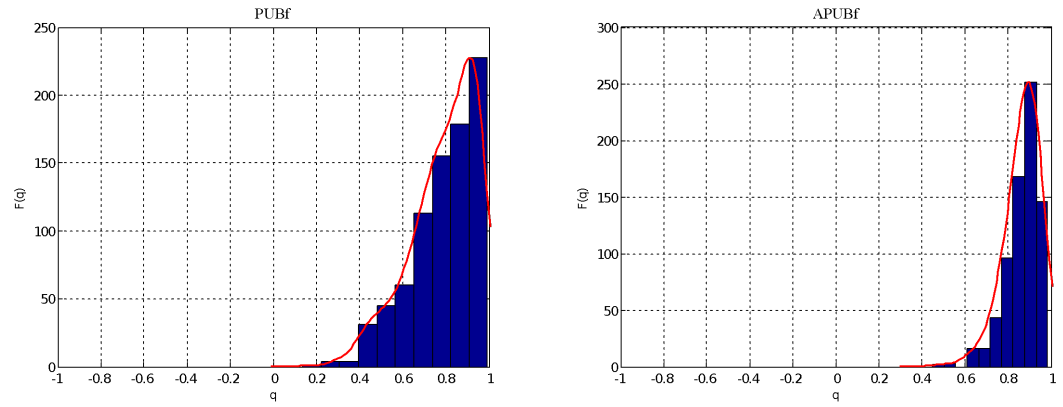
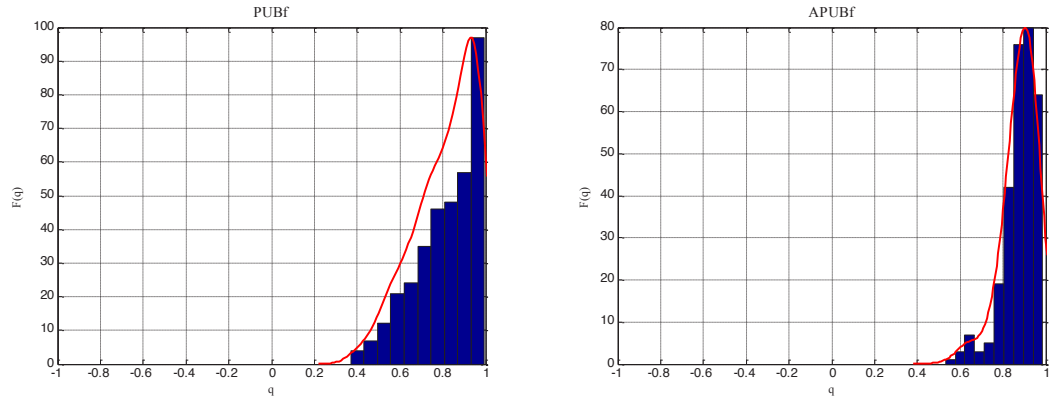


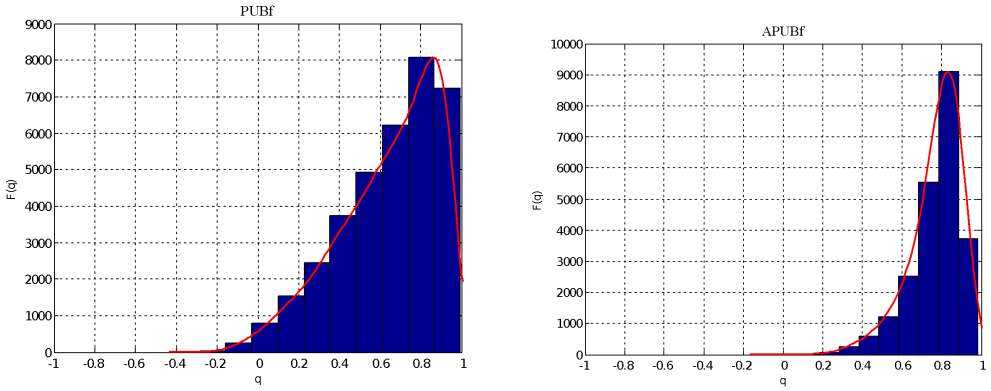
Figure 5. Dynamics of overlaps between 9 leading nations and all other countries for the fractional number of publications (PUBf).



TOP PANEL. World Distribution of the overlaps calculated on each country against all other countries in the world for the *extensive* (size dependent) indicator of scientific production PUBf (top-left panel) and the *intensive* average productivity indicator APUBf (top -right panel).



MIDDLE PANEL. European Distribution of the overlaps calculated on each European country against all other European countries for the *extensive* indicator of scientific production PUBf (middle-left panel) and the *intensive* average productivity indicator APUBf (middle-right panel).



BOTTOM PANEL. European Regions (NUTS2 units) Distribution of the overlaps calculated on each European region against all other European regions for the *extensive* indicator of scientific production PUBF (bottom-left panel) and the *intensive* average productivity indicator APUBf (bottom-right panel).

Figure 6. Distributions of the overlaps calculated at World, European and Regional level for extensive (PUBf) and intensive (APUBf) indicators.

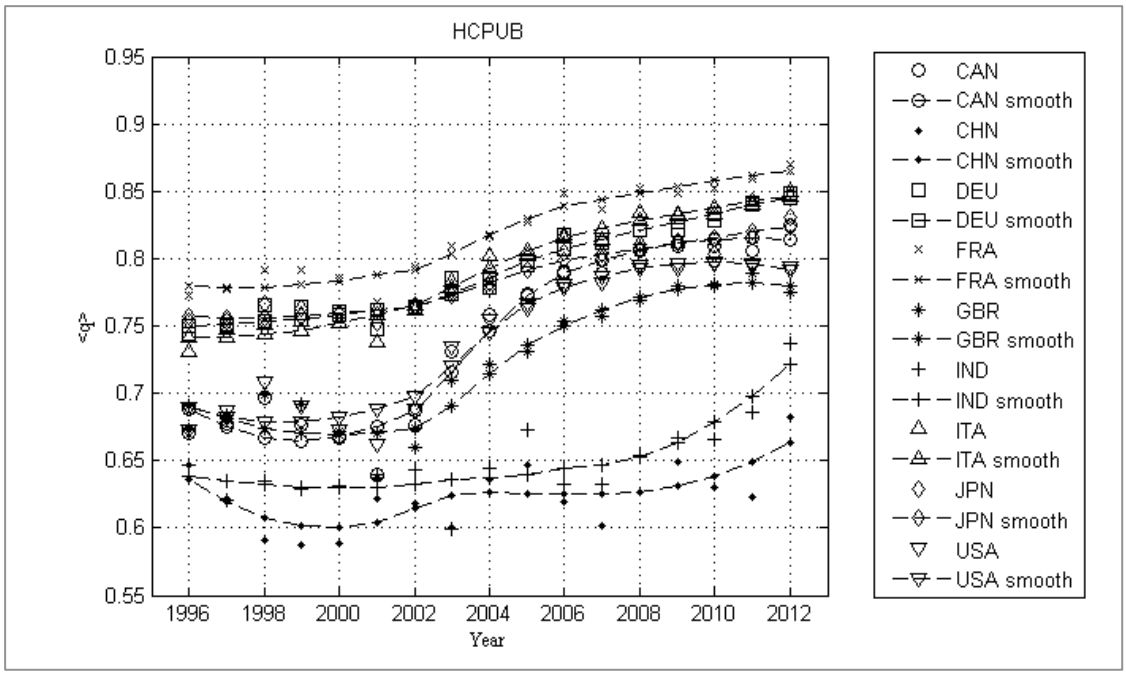


Figure 7. Dynamics of overlaps between 9 leading nations and all other countries for the number of highly cited publications (HCPUB)

An important aspect is the *dynamics* of the overlap values: how do the overlap distributions develop over time, and how does the position of specific countries evolve. Figures 5 and 7 present for 9 leading nations the development over time of the average overlap with all other countries, for the fractional number of publications (PUBf) and the number of highly cited publications (HCPUB), respectively. Although Figure 6 shows during the last 4 years a slight decline in overlap for most countries, Figure 7 reveals a trend towards convergence, especially for India and China. Perhaps the latter two countries increased their contribution to the international research front, but they maintained to some extent their own disciplinary profiles.

Conclusions

A tentative conclusion that should be substantiated in future empirical research is that the *variability of disciplinary profiles among countries is of the same order of magnitude of the variability among regions within a country and that the same happens for their convergence* rates, as shown by the distributions of the overlap calculated and displayed in this paper. The same dynamics observed for the extensive measures of scientific production is observed for the intensive average productivity, which appears to have a more concentrated distribution for all the level of the analysis carried out. Further research is in progress to support these preliminary findings and to illustrate the advantages of our approach, including the application of the ultrametric property of the overlap values to determine "automatic" clustering of the investigated national and regional systems of research. The step further will be then to link the scientific structure of national and regional systems with their technological structure to evaluate their dynamics at national and regional level.

Acknowledgments

This work was supported by Elsevier that provided the data within the Elsevier Bibliometric Research Programme (EBRP) for the Project "Assessing the Scientific Performance of Regions and Countries at Disciplinary Level by Means of Robust Nonparametric Methods: New Indicators to Measure Regional and National Scientific Competitiveness".

References

Barro, R. J. & Sala-i-Martin, X. (1992), Convergence, *Journal of Political Economy, 100*(2), 223–251.

Bongioanni, I., Daraio, C., & Ruocco, G. (2014), A Quantitative Measure to Compare the Disciplinary Profiles of Research Systems and their evolution over time, *Journal of Informetrics*, *8*, 710-727.

Bongioanni, I., Daraio, C., Moed, H.F., & Ruocco, G. (2014), Disciplinary Profiles and Performance of Research Systems: a World Comparison at the Country level, *Proceedings of the Science and Technology Indicators Conference 2014 "Context Counts: Pathways to Master Big and Little Data",* 3-5 September 2014, edited by E. Noyons, published by Universiteit Leiden CWTS 2014, pp. 50-63, ISBN 978-90-817527-1-8.

Camagni, R., & Capello, R. (2013). Regional innovation patterns and the EU regional policy reform: Toward smart innovation policies. *Growth and Change*, 44(2), 355-389.

Durlauf, S. (2000), "Econometric Analysis and the Study of Economic Growth: A Skeptical Perspective," in *Macroeconomics and the Real World*, R. Backhouse and A. Salanti, eds., Oxford: Oxford University Press.

Durlauf, S. N., Kourtellos, A., & Tan, C. M. (2005). Empirics of growth and development. *International Handbook of Development Economics*, 1.

Glanzel, W., Debackere, K., & Meyer, M. (2008). 'Triad' or 'tetrad'? On global changes in a dynamic world. *Scientometrics*, *74*, 71-88.

Glänzel, W., Leta, J., & Thijs, B. (2006). Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics*, *67*(1), 67-86.

Glanzel, W. (2000). Science in Scandinavia: A bibliometric approach. *Scientometrics 48*, 121-150.

Glanzel, W. & Schlemmer, B. (2007). National research profiles in a changing Europe (1983–2003): An exploratory study of sectoral characteristics in the triple helix. *Scientometrics*, *70*(2), 267–275.

Hu, X. J., & Rousseau, R. (2009). Comparative study of the difference in research performance in biomedical fields among selected Western and Asian countries. *Scientometrics*, *81*(2), 475–491.

Horlings, E. & van den Besselaar, P. (2013), Convergence in science growth and structure of worldwide scientific output, 1993-2008, Rathenau Instituut, Working Paper 1301.

McCann, P. & Ortega-Argilés, R. (2013). Smart specialization, regional growth and applications to European Union cohesion policy. *Regional Studies*, 1-12.

Radosevic, S., & Yoruk, E. (2014). Are there global shifts in the world science base? Analysing the catching up and falling behind of world regions. *Scientometrics*, *101*(3), 1897-1924.

Tang, L. & Shapira, P. (2011). Regional development and interregional collaboration in the growth of nanotechnology research in China. *Scientometrics*, *86*, 299–315.

Tian Y., Wen, C., & Hong, S. (2008). Global scientific production on GIS research by bibliometric analysis from 1997 to 2006. *Journal of Informetrics,* 2, 65-74.

Wong, C. Y. (2013). On a path to creative destruction: science, technology and science-based technological trajectories of Japan and South Korea. *Scientometrics*, *96*, 323–336.

Wong, C. Y. & Goh, K. L. (2012). The pathway of development: science and technology of NIEs and selected Asian emerging economies. *Scientometrics*, *92*, 523–548.

Yang, L. Y., Yue, T., Ding, J. L., & Han, T. (2012). A comparison of disciplinary structure in science between the G7 and the BRIC countries by bibliometric methods. *Scientometrics*, *93*, 497–516.

Zhou, P., Thijs, B., & Glänzel, W. (2009), Regional analysis on Chinese scientific output, *Scientometrics*, *81*(3), 839-857.

New Research Performance Evaluation Development and Journal Level Indices at Meso Level

Muzammil Tahira[1], Rose Alinda Alias[1], Aryati Bakri[1] and A. Abrizah[2]

[1]*mufals@yahoo.com* , [1]*alinda@utm,com,* [1]*aryatib@utm.com*
[1]Department of Information System, Faculty of Computing,
Universiti Teknologi Malaysia (UTM), Skudai 81310, Johor Bahru, (Malaysia)

[2]*abrizah@um.edu.my*
[2]Department of Library & Information Science,
Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, (Malaysia)

Abstract
This study applies scientometric approach to meso level data. The objective was to evaluate Institutional level h-index's (IHI) reliability with respect to other Journal Related Indices (JRI). Most of the studies in the literature considered journal's h-index as contrasted measure. Nevertheless, there has been no study that explores the relation between IHI and institutional level JRI. To get further evidence, we have explored the inter-correlation of IHI with a set of JRI. For this purpose data from Web of Science, Journal Citation Report and time cited features were used. Our unit of analysis was Malaysian engineering research with a wider time span of 10 year's data (2001-2010) and a larger set of journals (1381 journals). Previous studies are are used for comparative analysis. This paper puts forward a better understanding to considering new impact indices at meso level for evaluation purpose.

Conference Topic
University policy and institutional rankings, Science policy and research assessment

Introduction
Journal Impact Factor (JIF) was introduced by the Institute of Scientific Information (ISI) via Journal Citations Report (JCR) about 30 years ago. It has a long tradition as an Impact Factor (IF) indicator for scholarly research output. Alike, h-index and many of its variants have been introduced and displayed on JCR site (www.webofknoweldge.com). IF can be used as a measure of research quality/impact of journals (Braun, Glanzel & Schubert, 2006). In general research performance evaluation (RPE) practices, it has become a "chief quantitative measure of the quality of researcher, and even the institution" but, it cannot be used as a direct measure of quality (Amin & Mabe, 2003; Bornmann et al., 2011). JIF remains the primary criterion when it comes to assessing the quality of journals and authors (Raj & Zainab, 2012). IF should not be used as a sole measure of a journal rank (Bornmann, et al., 2011).

To overcome the limitations, of IF, researchers suggested that it should be used with new alternative tools (Braun, Glanzel, & Schubert, 2006; Prathap, 2011; Bornmann et al., 2011; Yang Yin, 2011) or as a measure of research quality / impact of journals (Braun, Glanzel & Schubert, 2006). An interesting debate was started by Braun, Glanzel, and Schubert, (2006) who suggested that the h-index can be used as a measure of research quality or impact of a journal. The notion of Journal h-index was introduced by (Braun, Glanzel, & Schubert, 2005). Who found it a promising measure for the journal (Braun, Glanzel, & Schubert, 2006). After the introduction of h-index, a number of studies made a comparative analysis of both measures and their variants. Both impact indices (h and IF) are easily comprehensible (Leydesdorff, 2009) and have received worldwide recognition. However, prior studies, as reviewed in the subsequent paragraphs were concerned with the evaluation of journal's h-index to JRI.

Mingers, Macri and Petrovici (2012) examined Journal level h-index against Impact Factor 2year (JIF), Impact Factor 5 year (IF5y) and peer judgment for management journals. They preferred journal h-index to IF because of the former's selective time frame and the formulaic problem. Another study in the field of management was carried out by Moussa and Touzani (2010) using Google-Scholar (GS) as source data. They used a variant of the h-index, the hg-index along with two and five years IF. There was a substantial agreement found (>0.85) between JIF 5y and the hg -index ranking. They suggested hg-index as an alternative to the GS based journals. Soutar and Murphy (2009) studied 40 marketing journals and ranked them according to IF and h-index, and compared their list with Australian journal ranking. They suggested these indices as the basis for moving some journals up and other journals down. Their study supported the use of GS as an alternative way to measure citations in marketing. Harzing and Van der Wal compared h-index calculated from GS with the impact factors computed from the Web of Science (WoS^{TM}) and with peer reviewed journal ranking (2009) by undertaking a larger-scale investigation of over 800 business and management journals.

A comparative analysis of IF and h-index was carried out by Bador and Lafouge (2010) on pharmacology and psychiatry journals from JCR with two-year publications. The journals correlation coefficient between IF and h-index was high. They inferred that IF and h-index can be totally corresponding when analyzing journals of the similar scientific subject. Bornmann, Mutz and Daniel (2009) studied the journal's h-index of twenty organic chemistry journals from WoS^{TM} database for two years time span. They analyzed a number of impact indicators including the IF, and journal's h-index and its variants *g* index, h^2 index, A, and *R* index. They found "a high degree of correlation between the various measures" (Bornmann, Mutz & Daniel, 2009).

Yang Yin (2011) analyzed 20 top journals in the field of science and engineering using data from WoS^{TM}. The researcher hypothesized "that the combination of complementary journal indicators could provide a simple, flexible and practical alternative approach for evaluating scientific journals" (p.2). Yang Yin considered the journal h-index with another JRI e.g. EigenFactor score. There is a positive correlation although not strong among these indices. They suggested getting published research work in high Eigenfactor scores journals. These indices can also be combined to complement each other.

Research Objectives

The objective of past studies was to evaluate a journal's h-index validity and reliability with respect to other JRI. Most of these studies considered journal's h-index as contrasted measure with JIF, JIF (5Y), and EigenFactor Score (EF). These studies are meaningful to understand the properties of newly introduced indices and potential use of journal's h-index as a complement aid with IF and its variants (Bador & Lafouge, 2010; Bornmann et al., 2012; Yang Yin, 2011) or, as a supplement (Braun, Glanzel, & Schubert, 2006).

Nevertheless, there has been no study to explore the relation of IHI with JRI. To have further evidence of validity of h-index at the institutional level, we hypothesized that IHI is a potential index for RPE that can be used to complement or as a supplement along with JRI for RPE at the institution level.

Methods and Materials

The empirical part of this study focuses on one non-Western country, Malaysia, which has a developed and well-defined scholarly publishing industry based in its universities. Research productivity, citations record, and institutional journal data of twelve Malaysian universities are retrieved from WoS^{TM} and JCR'2011 from the Web of Science. Only those universities that have at least fifty publications during the past ten years were selected. "The statistical methodology of EFA can be used to examine for latent associations present in a set of

observed variables, and reduce the dimensionality of the data to a few representative factors" (Schreiber et al., 2012, p.349). It is mainly used to identify a smaller set of salient variables from a larger set and to explore the underlying dimensions or factors that explain the correlations among a set of variables (Conway and Huffcutt, 2003). Initially, we used eleven indices for the present study. These are Total publications (TP), Total Citations (TC) Citation Per Publications (CPP), Institutional H-Index (IHI), JIF, Cumulative Journal Impact Factor (CIF), Journal Impact Factor 5y (JIF5y), Cumulative Journal Impact Factor 5y (CJIF5y), Average Impact Factor (AIF), Median Impact Factor (MIF), Immediacy-index (Imm-index) and EigenFactor Score (EF).The definitions and the acronym used are described in Table 1.

Table 1. Definitions of indices used at Meso level.

Indicators	*Definition*
1. Total Publications (TP)	Total publications of a university over the set criteria
2. Total Citations (TC)	Total citations of a university over the set criteria
3. Institutional H-Index (IHI)	An institution has index h if h of institutional publication has at least h citation each and other publication have fewer than or equal to h citations each.
4. Journal Impact Factor (JIF)	The average number of times articles from the journal published in the past two years has been cited in the JCR year (Thomson- Reuters 2015).
5. Cumulative Journal Impact Factor (CIF)	This is the cumulative value of Journal Impact Factor of each university.
6. Impact Factor five Years (IF5y)	The average number of times articles from the journal published in the past five years have been cited in the JCR year (Thomson-Reuters 2015).
7. Cumulative Impact Factor Five Years (CIF5y).	This is the cumulative value of five years Journal Impact Factor of each university.
8. Average Impact Factor (AIF)	This is the average of the Impact Factor of each university.
9. Median Impact Factor (MIF)	This is the median of the Impact Factor of each university.
10. Immediacy-index (Imm-index)	This is calculated by dividing the number of citations to articles published in a given year by the number of articles published in that year Thomson-Reuters 2015).
11. EigenFactor Score(EF)	**"**Eigenfactor score is calculated by the ratio of the total number of citations for the JCR year to the total number of articles published in the last 5 years". Thomson-Reuters 2015).

Data Processing

To get a meaningful evaluation, we used a wider set of WoS^{TM} engineering journals (1381 journals) considered by our sample (12 Malaysian universities) institutions with a wider horizon of ten years (2001-2010) under specified nine categories. Our research term was "Malaysia" in "Address", limited to document type research article and reviews only and

refined by nine engineering research categories. These engineering categories are engineering electrical, electronic, engineering manufacturing, engineering biomedical, engineering industrial, engineering civil, engineering chemical, engineering mechanical, engineering environmental and engineering multidisciplinary.

Data were suffered from affiliation problem, change of journal title and abbreviation of a journal name. All the data were checked manually for publications, citations, institutional affiliation, and journal name change or emergence cases. The selected twelve universities got their articles published in 1381 journals. According to JCR'2011, almost all journals in our data set were IF. There were only 22 journal articles published in six journals, and ten proceedings had no impact factor. It is assumed that the said journals/proceedings may have IF prior to 2011. These records were included in the journal list for analysis purpose. Firstly, all the records were retrieved in a spreadsheet file, and IBM SPSS version'19 was used for statistical analysis purpose.

Table 2 provides the university-wise total journal records. The publication share of research university (RU) status was 66 % (908) while; the non-RU status universities shared 34 % (473) of the total journals.

Table 2. Distribution of journals (N=1381).

No	*University*	*Total journals and proceedings*	*University Status*	*Contribution%*
1	University of Malaya (UM)	191		
2	Universiti Sains Malaysia (USM)	188		
3	Universiti Putra Malaysia (UPM)	187	Research	
4	Universiti Teknologi Malaysia (UTM)	184	Universities= 908 journals	66
5	Universiti Kebangsaan Malaysia (UKM)	158		
6	Universiti Teknologi Mara (UiTM)	87		
7	University of Multimedia (MMU)	81	Non-Research Universities=473 Journals	34
8	Universiti Teknologi PETRONAS (UTP)	78		
9	International Islamic Universiti Malaysia (IIUM)	77		
10	University of Nottingham Malaysia Campus (UNMC)	61		
11	MONASH Universiti Sunway Campus (MONASH)	51		
12	Universiti Tenaga Nasional (UNITEN)	38		
	Total	1381		100

The RU universities are more bound to published in IF journals to get more research funding. These universities receive a big amount of budget for R&D purposes and have to face pressure and make policies accordingly (http://www.hir.um.edu.my), and this is especially prevalent in Asian countries (Leydesdorff, 2009). The first five public universities (RU) published in 150-200 journals. Comparatively the private universities had fewer publications and published in 50 to 100 journals. The average number of journals for RU and non-RU universities is 182 and 68 respectively.

Analysis and Findings

Exploratory Factor Analysis (EFA)

In a tie with the problem, this section proceeds accordingly with descriptive statistics, data normality and EFA for our set of indices as presented in Table 3.

Descriptive Statistics and Normality Analysis of Complete Dataset

Descriptive statistics along with Skewness and Kurtosis are presented in Table 4. The results of the normality test based on raw data (excluding outliers) are reported in Table 5. The Skewness and Kurtosis are valid tests to find the normality of data. Their values show a normal distribution of data adequately normal. Keeping in view the requirement of EFA statistical application we used two other options as well. We also examined the relation between the raw, logarithmically transformed shifted (ln(x + 1) and square root transformation.

Table 5 shows a better Kaiser-Meyer-Olkin (KMO) results and a slight better-explained variance for log data. For this reason, we found the logarithmic transformed data more adequate for EFA. Bornmann, Mutz and Daniel (2008; 2009) used a cut-off threshold >0.6 for extraction loading factors while Schreiber, Malesios and Psarakis (2012) fixed it at > 0.685 for Varimax rotation.

Schreiber *et al*. (2012) argued that small sample size for EFA can produce reliable results. Quite a few factors and high communalities are in favour of small sample sizes (Preacher and MacCallum, 2002). Further, to measure a sampling adequacy, a specific test Kaiser-Meyer-Olkin (KMO) of value >5 is acceptable (Kaiser, 1974). KMO value (Table 6) of the present data sample is >0.5 with high communalities (>0.85) (Table 7). Based on KMO values and variance explained (Table 6 and 7), we finally utilized logarithmically transformed data. We identified two unknown factors through Eigen values (>1) via variance explained.

This is evident that EFA can be used and is appropriate for our formulated problem and dataset. Initially, we considered eleven indices, TP, TC, IHI and 8 of JRI (JIF, CIF, IF avg, MIF, CIF, CIF5Y, Imm-Index, and EF). This set of indices produced inadequate results for EFA. After omitting the TP, we applied EFA to TC, IHI, and 8 JRI (IF, CIF, IFavg, MIF, CIF, CIF5Y, Imm- Index, and EF).

Table 3. Analysis of Complete dataset for institutional level indices applied

University	*TP*	*TC*	*IHI*	*JIF*	*CIF*	*AIF*	*MIF*	*IF(5Y)*	*CIF(5Y)*	*Imm-Index*	*EF*
USM	724	4027	26	311.36	1609.71	2.229	1.35	331.43	1705.82	49.752	2.506
UPM	551	2309	20	255.12	879.04	1.600	1.12	262.86	886.18	40.100	2.070
UM	495	2388	23	337.45	948.07	1.950	1.50	318.54	871.69	52.598	2.481
UTM	475	2259	23	262.16	883.14	1.883	1.12	280.76	910.61	39.835	2.277
UKM	386	1490	17	233.65	624.13	1.634	1.25	246.65	629.14	36.081	1.975
UiTM	139	359	9	144.85	239.58	1.815	1.39	154.08	248.73	21.922	1.318
IIUM	138	251	7	100.01	174.87	1.270	1.02	103.96	177.20	14.640	0.960
MMU	532	2231	19	120.22	583.83	1.099	1.17	128.66	576.70	18.130	0.874
UNMCC	126	616	13	102.82	248.58	1.973	1.55	100.34	241.58	15.450	0.776
UTP	142	329	9	122.97	263.12	1.853	1.31	134.24	287.38	19.896	1.179
MONASH	76	302	10	87.87	131.94	1.713	1.59	94.86	140.93	13.533	0.887
UNITEN	71	139	6	50.86	91.77	1.293	1.22	55.65	100.24	7.460	0.351

Analysis of EFA

Table 6 reports the results of KMO values of the transformed data for the appropriateness of factor analysis. The next table 7 reveals the results of communalities for 3 EFA models that are the "variance in observed variables accounted for by a common factor" (Hatcher, 1994).

Table 4. Descriptive statistics

Indices	*Descriptive Statistics*						
	Mean	*St.dev*	*Median*	*Min*	*Max*	*Skewness*	*Kurtosis*
TP	321.25	229.079	264.00	71	724	0.364	-1.47
TC	1391.67	1246.835	1053.0	139	4027	0.776	-0.17
IHI	15.17	7.004	15.00	6.00	26.0	0.151	-1.60
IF	177.44	96.683	133.90	50.85	337.45	0.452	-1.34
CIF	556.48	457.445	423.47	91.77	1609.7	1.115	1.02
MIF	1.30	0.182	1.28	1.02	1.59	0.239	-1.01
AIF	1.69	0.332	1.76	1.10	2.23	-0.427	-0.44
IF(5Y)	184.34	97.047	144.15	55.65	351.43	0.351	-1.58
CIF(5Y)	564.68	471.04	432.04	100.24	1705.8	1.317	1.87
Imm-index	27.45	15.356	20.91	7.46	52.60	0.471	-1.32
EF	1.47	0.748	1.249	0.35	2.51	0.179	-1.56

Overview of Statistical Procedure for EFA

Table 5. Test for normality of data

	Kolmogorov-Smirnov[a]			*Shapiro-Wilk*		
	Statistic	*Df*	*Sig.*	*Statistic*	*df*	*Sig.*
TP	.283	12	.009	.863	12	.053
TC	.233	12	.071	.852	12	.038
IHI	.186	12	.200*	.918	12	.267
IF	.208	12	.158	.881	12	.090
CIF	.235	12	.067	.856	12	.043
AIF	.183	12	.200*	.929	12	.369
MIF	.114	12	.200*	.960	12	.782
IF(5Y)	.228	12	.085	.876	12	.078
CIF(5Y)	.212	12	.143	.829	12	.020
Imm-index	.228	12	.086	.904	12	.178
Eigen Factor	.180	12	.200*	.937	12	.458

*At a 5% Significance Level

Table 6. Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy

	X	*√x*	*ln(x + 1)*
KMO	0.564	0.540	0.695
Sig.	0.00	0.00	0.00

Table 8 provides Initial Eigenvalues >1 and indicates that the total variance explained by first two factors is 75%, and 17 % of cumulative variance explained by both factors are 91%.
Component matrix (Table 8) illustrates that the set of indices clearly loads on two extracted factors. Rotated Component Matrix Table (9) for EFA model shows that the indices have

substantial loading on two established factors. It indicates the loading of two institutional 'impact of the productive core indices' (TC and IHI) and six others JRI have high loading (> 0.90) and a slight less for EF (>0.891).

Table 7. Communalities for 3 EFA models

	X		*√x*		*ln(x + 1)*	
Indices	*Initial*	*Extraction*	*Initial*	*Extraction*	*Initial*	*Extraction*
TC	1	0.893	1	0.9	1	0.896
IHI	1	0.883	1	0.877	1	0.866
IF	1	0.94	1	0.951	1	0.953
CIF	1	0.934	1	0.958	1	0.962
IF(avg)	1	0.854	1	0.865	1	0.841
MIF	1	0.869	1	0.844	1	0.87
IF(5Y)	1	0.954	1	0.963	1	0.967
CIF(5Y)	1	0.879	1	0.925	1	0.950
Imm- Index	1	0.918	1	0.943	1	0.955
EF	1	0.869	1	0.861	1	0.870

AIF and MIF both have substantially high loading on the second factor>0.9. MIF is more accurate measure than the average value, due to the impact factor's skewed distribution (Costas & Bordons, 2007). IF and CIF and IF5y and CIF5y require two years and five years time span with different strengths of productivity. EF is another index based on 5-year data excluding journal self-citation to rate the total importance of journal. Journals generating higher impact on the field have larger Eigenfactor scores (Bergstrom, 2007). "EF improves upon JIF and somewhat robust indicators of quality and prestige of the journal due the inclusion of 5 year's data, exclusion of journal self-citations" (YangYin, 2010, p.3). Rather a high journal EF depicts producing of high-impact scientific findings in a specific area (YangYin, 2010; Saad, 2006). IF (5y) indicates the speed with which citations to a specific journal appear in the published literature. Immediacy index that is based on one-year data shows the same value as CIF on the first factor. They both require a different strength of data. Surprisingly they all loaded on the same factor along with IHI.

Table 8: Total variance explained for 3 EFA models.

Data type		*Initial Eigenvalues*			*Extraction Sums of Squared Loadings*			*Rotation Sums of Squared Loadings*		
		Total	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
Raw indices	1	7.401	74.006	74.006	7.401	74.006	74.006	7.269	72.687	72.687
	2	1.594	15.940	89.946	1.594	15.940	89.946	1.726	17.259	89.946
√x	1	7.432	74.325	74.325	7.432	74.325	74.325	7.314	73.142	73.142
	2	1.655	16.547	90.872	1.655	16.547	90.872	1.773	17.730	90.872
ln(x+1)	1	7.457	74.569	74.569	7.457	74.569	74.569	7.343	73.427	73.427
	2	1.672	16.720	91.290	1.672	16.720	91.290	1.786	17.862	91.290

Table 9. Rotated component matrix

Indices	*Components*	
	1	*2*
C	**.945**	-.055
IHI	**.929**	.059
IF	**.965**	.147
CIF	**.978**	-.074
AIF	-.133	**.907**
MIF	.309	**.880**
IF(5Y)	**.970**	.159
CIF(5Y)	**.974**	-.038
Imm-index	**.950**	.230
EF	**.891**	.275
Eigenvalues	**7.401**	1.595
Variance explained	**75%**	17%

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Values >.5 are bold.

Conclusions

The caveats of h-index, JIF, and traditional metrics have been discussed in the abundant literature. Previous studies are meaningful to understand the properties of newly introduced indices and potential use of Institutional's h-index as a complement aid with IF and its variants. (Bador & Lafouge, 2010; Bornmann et al., 2012; Yang Yin, 2011) or, as a supplement (Braun, Glanzel & Schubert, 2006).

The present study describes the case of Malaysian engineering research applying the scientometric approach, method and techniques for RPE. Based on the ten years data analysis from WoS^{TM}, we applied a set of comparatively new indices. To achieve the research objectives, empirical analyses were carried out, and hypotheses were examined statistically.

The major findings of the study demonstrate that there seems to be increasing the trend to get published in IF journals. A steady increase of IF publications is observed from 2001 in the Malaysians scientific productivity of all studied disciplines including engineering. The ambition to publish in IF WoS^{TM} recognized publications is reinforced by the Malaysian Research Assessment (MyRA) exercise, which requires institutions to publish papers that are indexed in the citation database. This is due to the Malaysian Ministry of Education policies towards research and publications during two five years plans (2001-2005; 2006-2010). RU status universities (shared 68% and 74% publications and citations). These universities have published in 66% of total journals. Overall, the RU universities lead in positioning order with the application of indices. USM is an exceptional case and remained in position one with respect to almost all indicators. While others showed a noteworthy change in their positioning order. IHI has stronger functional relation with institutional citation data followed by publication record. Institutional citation data is the best predictor of IHI. Often used metric C (as total impact indicator) and the EF (as prestige indicator) have a high association with IHI. This establishes the property of h- index as prestige impact measure of scientific productivity. This index appears a useful yardstick, because of good functional relationship with C and P and has shown some discriminatory power for ranking purpose. The EFA suggests the same distinguishing behaviour of IHI like P and C. The findings put forward a better understanding

about the consideration of new impact metric for RPE at the meso level. Malaysian engineering institutional case indicates that h-index and others metric have not only strong association for total institutional citation data but also with institutional cumulative journal indices. However, the total variance explained for two components yields about 75% for its first component and 16% for the second component. Therefore, findings are based within the limitations of the statistical analysis.

Publishing in high-quality IF journals is important if a country is to realize its ambition to have its universities amongst the top rated universities in the world. This is not peculiar to Malaysia. The Ministry of Education Malaysia is targeting two research universities in the country to be in the top world 100 best universities by 2020. Other countries also place a high emphasis on publishing in IF journals and would want to be ranked as top world universities, even if they are not always explicit in saying so. Given the significant number of papers that have now been published by Malaysian institutions (56, 571 in Web of Science, Essential Science Indicators, Web of Science 2015), there is an opportunity to carry out further analysis. It would be interesting, for example, to provide analysis at a discipline level to get a feeling for the strengths of the institution at a lower level. It would also be informative to consider other normalization measures to ascertain if they provide a better correlation with the MyRA ranking.

Acknowledgement

The work of Muzammil Tahira and A. Abrizah was supported by the Ministry of Higher Education Malaysia (HIR-MOHE) UM.C/HIR/MOHE/FCSIT/11.

References

Amin, M. & Mabe, M. (2000). Impact factor: Use and abuse. *Perspectives in Publishing, 1*, 1-6.

Bador, P. & Lafouge, T. (2010). Comparative analysis between impact factor and h-index for pharmacology and psychiatry journals. *Scientometrics*, *84(1),* 65-67.

Bornmann, L. Marx, W., Gasparyan, A.Y., & Kitas, G. D. (2012). Diversity value and limitations of the journal impact factor and alternative metrics. *Rheumatol International*, *32*, 1861-1867.

Bornmann, L., Mutz, R., Hug, S. E., & Daniel, H.D. (2011). A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. *Journal of Informetrics, 5*(3), 346-359.

Bornmann, L., Mutz, R., & Daniel, H.D. (2009). Do we need the h-index and its variants in addition to standard bibliometric measures? *Journal of the American Society for Information Science and Technology, 60*(6), 1286-1289.

Bornmann, L., Mutz, R., & Daniel, H.D. (2008). Are there better indices for evaluation purposes than the h-index? A Comparison of nine different variants of the h-index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, *59*(5), 830–837.

Braun, T., Glanzel, W., & Schubert, A. (2006). A Hirsch-Type index for journals. *Scientometrics*, *69*(1), 169-173.

Conway, J.M. & Huffcutt, A. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods, 6(*2), 147-168.

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology, 59*(2), 278-287.

Mingers, J., Macri, F., & Petrovici, D. (2012). Using the h-index to measure the quality of journals in the field of business and management. *Information Processing and Management*, *48,*234-241.

Moussa, S. & Touzani, M. (2010). Ranking marketing journals using the Google Scholar-based hg-index. *Journal of Informetrics, 4*, 107–117.

Prathap, G. (2011). Correlation between h-Index, Eigenfactor™ and Article Influence™ of chemical engineering journals (Letter). *Current Science*, *100*(9), 1276.

Raj, R.G. & Zainab, A.N. (2012). Relative measure index: A metric to measure quality. *Scientometrics, 93*(2), 305-317.

Schreiber. Malesios, C. C. & Psarakis, S (2012). Exploratory factor analysis for the Hirsch Index, 17 h-type variants, and some traditional bibliometric indicators. *Journal of Informetrics*, *6,* 347– 358.

Schreiber, M., Malesios, C. C. & Psarakis, S. (2011). Categorizing Hirsch Index variants. *Research Evaluation*, *20*(5), 397–409.

Thomson-Reuters. (2015).Web of Science ™ Retrieved on March 20, 2015 from http://admin-apps.webofknowledge.com/JCR/help/h_impfact.htm)

Yang Yin, C. (2011). Do impact factor, h-Index and Eigenfactor™ of chemical engineering journals correlate well with each other and indicate the journals' influence and prestige? *Current Science*, *100* (5), 648-653.

Factors Influencing Research Collaboration in LIS Schools in South Africa

Jan Resenga Maluleka[1], Omwoyo Bosire Onyancha[2], Isola Ajiferuke[3]

[1] *malulrj@unisa.ac.za* [2] *onyanob@unisa.ac.za*
University of South Africa, Dept of Information Science, PO Box 329 Unisa, 0003

[3] *iajiferu@uwo.ca*
University of Western Ontario London, Ontario, Canada, N6A 5B7

Abstract
The study sought to explore the underlying factors that influence research collaboration in Library and Information Science (LIS) schools in South Africa. The population for the study consisted of 85 academic teaching staff employed by LIS schools in South African universities. A survey design was used to obtain data for the study, through a questionnaire containing open- and close-ended questions. A total of 85 teaching staff in 10 LIS schools in South Africa were alerted, through email, to the location of the Web-based questionnaires, developed using the Stellarsurvey software. A total of 51 questionnaires were completed and returned for analysis. The findings suggest that factors such as networking, sharing of resources, enhancing productivity, educating students, overcoming intellectual isolation, and accomplishments of projects in a short time as well as learning from peers influenced research collaboration in LIS in South Africa. Factors that are likely to hinder effective collaboration in LIS research include bureaucracy, lack of funding, lack of time, as well as physical distance between researchers. The findings further suggest that even though there are drawbacks to collaboration, majority of LIS researchers thought that collaboration is beneficial and should be encouraged.

Conference Topic
County-level studies

Introduction
In today's global economy, there is an increasing importance of collaborative relationships between individuals, organisations, and even countries. Collaboration, defined as a "process where two or more individuals or organizations deal collectively with issues that they cannot solve individually" (Ocholla, 2008:468) and "the working together of researchers to achieve the common goal of producing new scientific knowledge" (Katz & Martin, 1997), can be found in all the spheres of human life, for example in politics, economics or even in religion. Katz & Martin (1997) are of the opinion that research collaboration has significant benefits such as intellectual championship, joint development of skills, effective transfer of knowledge and the improvement of potential visibility of researchers. For example, collaboration can build partnerships and help empower researchers to accomplish projects that were never going to be easy to do individually. Collaboration brings together experiences, skills, knowledge and the know-how of different researchers into one particular project. By way of research collaboration, researchers from different countries (both developed and developing countries) come together for different purposes, among which are sharing of information, knowledge and technological transfer as well as finding solutions to specific problems (Onyancha, 2009). Researchers collaborate in order to accomplish tasks that cannot be accomplished as isolated individuals. Onyancha & Ocholla (2007), too, note that securing research grants is to a large extent becoming increasingly pegged on whether the intended research would be conducted collaboratively. Collaboration can be important especially in developing countries where there might be a lack of scientists and resources in certain fields. The few available researchers in developing countries can collaborate with those in developed countries for the former to be active in research as well as flourish as scientists.

According to Katz and Martin (1997), collaboration among scholars in both natural and social sciences has been steadily increasing for decades, covering different disciplines, development categories, institutions, geographic regions and countries. The increasing attention on research collaboration in LIS has also been pointed out by Onyancha and Maluleka (2011). Sugimoto (2011) argues that research in the field of LIS has followed similar patterns of increased collaboration as in other fields. According to Ocholla (2008), collaboration and partnerships could be forged amongst LIS institutions in a country and internationally or regionally in areas such as teaching, research, student and staff exchange, conferences and workshops, curriculum development, publications, research supervision and examination and distance teaching/research.

Rationale for the study

An examination of the published literature reveals that several studies have been conducted to examine research collaboration in different fields or disciplines including LIS. The focus of these studies includes identifying the collaborating authors, institutions, and/or countries (e.g. Sun, 2006; Onyancha & Ocholla, 2007), measuring the strengths of research collaboration (e.g. Yamashita & Okubu, 2006) and examining the nature of collaboration (e.g. Katz & Martin, 1997; Smith & Katz 2000). Several other studies have majorly focused on answering the question 'who' or 'what' of collaboration. In other words, studies that have been conducted previously on collaborative research have largely focused on the frequency of collaboration between the authors, the nature of collaboration and the strength of collaboration across disciplines. To the best of the researchers' knowledge, little has been done to answer the question 'why?' The current study therefore aims to investigate those factors that may influence collaboration in LIS schools in South Africa. The main objective of this study is to find out the underlying reasons and/or factors that influence collaboration, a situation that may explain the quantitative results (e.g. trends, patterns, and type of research collaboration) reported in previously published works.

Research Questions

The following research questions were posed in order to fulfil the study's main objective;

- What factors hinder and/or would hinder effective research collaboration in LIS schools in South Africa?
- What factors do and/or are likely to foster effective research collaboration in South African LIS schools?
- To what extent do the enhancers and inhibitors of collaboration influence research collaboration in LIS schools in South Africa?

Methodology and Materials

The study adopted a survey design to seek for the LIS academics' views on factors that influence research collaboration in LIS research in South Africa. Neuman (2007:273) argues that survey research is developed within the positivist approach and it is the mostly and widely used design in the social sciences. Similarly, Leedy and Ormrod (2010:187) argue that survey research *involves acquiring information about one or more groups of people – perhaps about their characteristics, opinions, attitudes, or previous experiences by asking them questions and tabulating their answers.*

In this study, the survey involved all academic teaching staff employed by LIS schools in South African universities. They include teaching assistants, junior lecturers, lecturers, senior lecturers, associate professors, and professors. Honorary professors, research fellows, extraordinary professors, or any other scholars who are linked to a particular department but without being fulltime were excluded as they appeared to have more than one institutional

affiliation. With only ten LIS schools offering LIS education in South Africa, there was no sampling conducted as all schools were included in the study. The total number of the teaching staff was also small, leading us to include all academics in the target population for this study. Table 1 shows the number of staff in the LIS departments by the parent University.

Table 1. LIS Schools in South Africa

School name	Acronym	Number of teaching staff
University of South Africa	UNISA	19
University of Pretoria	UP	24
University of KwaZulu-Natal	UKZN	6
University of Zululand	UZ	7
University of Fort Hare	UFH	4
University of Cape Town	UCT	8
University of the Western Cape	UWC	6
Durban University of Technology	DUT	5
University of Limpopo	UL	4
Walter Sisulu University	WSU	2
TOTAL		85[1]

The instrument of data collection for the study was a questionnaire, which was deemed to be the most appropriate. The questionnaire contained both closed-ended and open-ended questions, the former being the majority. There were a total of 20 questions focusing on specific items that were linked to the research questions. We used the "*Stellarsurvey*" online survey software as a platform for the questionnaires.[2] We then sent emails to all the identified LIS researchers in South African LIS schools. The emails contained a link directing them to the website which invited them to participate in the study. Respondents were given three weeks to complete the questionnaire online. After three weeks a reminder was sent to participants again reminding those who had not responded to do so.

Results and discussion

Profile of the respondents

Out of the 85 teaching staff members that were approached to participate in the study, only 51 completed the questionnaires, leading to a response rate of 64.6%. It was found that 43% (i.e. 22) of the respondents were male while 29 (57%) were female. All respondents had a university qualification ranging from a bachelor's degree to doctoral degree. The majority of the respondents (i.e. 21 or 41%) had a master's degree as their highest qualification, followed by those with a doctoral degree (i.e. 19 or 37%) and then those with honours (11 or 22%). The majority of the respondents were employed as lecturers (27 or 54%), followed by junior lecturers (9 or 18%) and full professors (5 or 10%) while senior lecturers and associate professors stood at 3 (3%) each. The results shows that the majority of the respondents are actively involved in research either as masters and doctoral students or as supervisors and mentors for these students.

[1] The number of the teaching staff was retrieved from the LIS departments' websites.
[2] The software is available at: http://stellarsurvey.com/.

The status of collaboration in LIS research

It was found that 43 (84%) of the respondents collaborated in the conduct of research while only 8 (16%) indicated that they never collaborated before. The results in Figure 1 (a) reveal that 45 (88%) respondents believe and agree that collaboration in research is important while 2 (4%) were neutral with only 4 (8%) saying collaboration in research is not important.

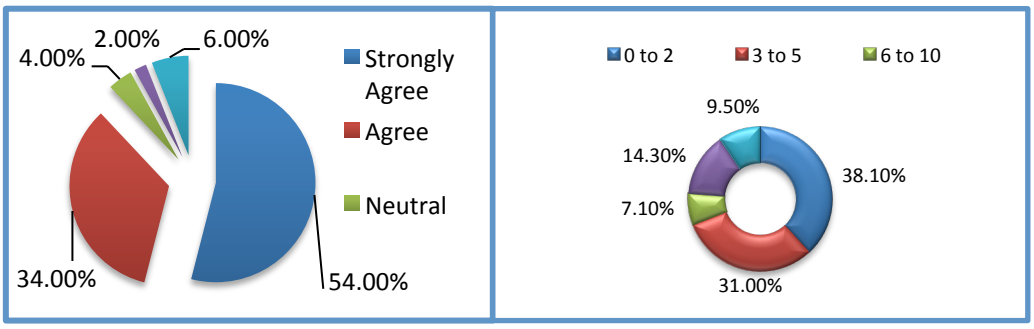


Figure 1. (a) Importance of collaboration (N=51) (b) The number of collaborated projects that are already published.

It is strange to note that while 84% of the respondents indicated that they collaborated, there was a sizable number, who may have included the ones who reported that they collaborated, who might have felt that collaboration is not important. This group could include researchers who are forced, by circumstances (e.g. institutional policies on co-supervision of students or mentorship of junior colleagues). When we looked at collaborative projects already completed (Figure 2 (b)), 32 (62%) respondents had already completed three or more projects collaboratively while only 19 (38%) had completed between 1 and 2 projects collaboratively.

It was worth noting that the current generation of researchers are actively engaged in collaborative research. Results tend to imply that the researchers prefer sharing and working together as compared to the past where the degree of collaboration among researchers has been reported to be low.

It has been shown that research collaboration in South Africa has increased tremendously in the previous decade (i.e. 2001-2009) (Sooryamoorthy, 2009). There are a number of reasons that may have influenced this pattern on collaborative research. Universities in South Africa have realised that they are losing their most experienced researchers who were approaching retirement age before the young developing researchers were fully equipped in the area of research. In some universities such as UNISA, huge funds have been invested into the development of young researchers through initiatives such as the mentorship programmes. This is done in view of Liebowitz's (2009) suggestion that formal mentoring programmes are popular techniques used for knowledge sharing, knowledge retention, knowledge transfer, and also to enhance worker skills. In this programmes, senior researchers are assigned mentees who learn from them on a daily basis for a specific period of time. Research funding organisations such as the National Research Foundation (NRF) of South Africa are also making funds available for collaborative and multidisciplinary research. Doctoral students are also funded to conduct post-doctoral research in collaboration with their mentors. The responses from the questionnaire also suggest that other universities have made it compulsory for supervisors to publish at least one article collaboratively with their students from the latter's theses and dissertations. The above is evident from the feedback from the respondents and it may be the reason why the majority of the respondents in the survey indicated that they are engaged in collaborative research, although some of them also indicated that collaboration is not important.

Looking at the group of people that the respondents mostly collaborated, it was noted that the researchers in LIS schools in South Africa largely collaborate with fellow researchers when taking the occasional, often and most often times of collaboration into account; the three account for 80% (see Table 2). This suggests that LIS researchers prefer collaborating with fellow researchers, preferably in their own field of interest. The main reason may be that working on a project with someone who understands one's subject area and the methodologies involved may result in the project being completed at a faster pace than if the opposite had to happen.

Another point worth highlighting is the results on collaboration with international researchers which was very low, with over 70% of the respondents indicating that they never collaborated at this level. This pattern is contrary to previous studies' findings, which revealed that most research in Africa is published in collaboration with international researchers (see Narvaez-Berthelemot, Russell, Arvanitis, Waast, & Gaillard, 2001). It is therefore unfortunate to find that researchers in LIS schools largely collaborate locally as opposed to engaging in international collaboration as researchers collaborating at the international arena have a competitive advantage over their peers because they have a chance of using resources from both institutions to which they are affiliated. The other notable advantage worth mentioning about international collaboration is the fact that it allows researchers a chance to publish in international journals, share international experiences which will allow them an opportunity to gain international visibility. Narvaez-Berthelemot, Russell, Arvanitis, Waast, & Gaillard (2001) note that researchers in developing countries would also benefit from their peers in developed countries in terms of publication of their research in international journals. The authors opine that "*the less productive the developing country, the greater the dependence on international co-authorship for mainstream publication*". Katz and Martin (1997) observe that most governments have been keen to increase the level of international collaboration engaged in by the researchers whom they support in the belief that this will bring about cost-saving or other benefits. The main reason given by respondents for not collaborating at this level was distance and logistical problems that exist when working with someone from another country. The other reason worth noting is the fact that researchers from bigger institutions or developed countries may undermine the contribution of the other researchers from poorer countries or smaller institutions. The opposite may also happen where researchers from smaller institutions may lack self-belief, contribute less and end up not playing an equal role in the whole collaborative venture.

Table 2. Group of persons that respondents collaborated with

	Never	Rarely	Occasionally	Often	Most often
Students	33.3%	7.7%	25.6%	23.1%	10.3%
Mentor	24.3%	18.9%	13.5%	16.2%	27.0%
Mentees (other than students)	50.0%	14.7%	20.6%	11.8%	2.9%
Fellow Researchers	5.0%	15.0%	30.0%	45.0%	5.0%
Senior Researchers	28.2%	15.4%	15.4%	20.5%	20.5%
International Researchers	45.9%	24.3%	10.8%	13.5%	5.4%

It seems like there is need for institutions to initiate programmes geared towards supporting the researchers in overcoming problems faced during international collaboration. The researchers also need to take advantage of the latest technologies that can easily allow them to work together without having to travel between countries. For LIS researchers in South Africa to remain at par with their international counterparts, they need to engage with them and work with them collaboratively so that they don't work in isolation.

Table 3. Groups likely to collaborate with in the future

	Never	Rarely	Occasionally	Often	Most often
Students	2.6%	7.9%	23.7%	39.5%	26.3%
Mentor	24.3%	16.2%	16.2%	21.6%	21.6%
Mentees(other than students)	25.7%	14.3%	34.3%	20.0%	5.7%
Fellow Researchers	0.0%	12.2%	22.0%	43.9%	22.0%
Senior Researchers	12.5%	20.0%	10.0%	35.0%	22.5%
International Researchers	12.5%	20.0%	30.0%	20.0%	17.5%

Enhancers and Impact of collaboration

Merlin (2000), Katz and Martin (1997), Bozeman and Corley (2004) give a summary of the following factors that are likely to foster effective collaboration in research:

- Collaborative research allows young researchers, access to expertise /experts with specialised knowledge and expertise in a particular area and learns directly from them.
- These partnerships gives researchers an opportunity to share resources where researchers from smaller institutions will get access to resources from big institutions and again institutions to supplement each other
- Multidisciplinary research allows a cross pollination of ideas and collaborative research allows partners to learn from one another
- There are more chances of getting funds if a collaborative initiative is submitted to funding organisation. Secondly a project can get funds from both organisations with will make it possible to carry out
- Working alone in a particular project can make one feel lonely and isolated. Working in a team helps one to overcome that intellectual isolation.

For this study, respondents were asked to indicate the extent to which factors such as networking, sharing of resources, enhancing productivity, educating students, overcoming intellectual isolation, accomplishment of projects in a short time, learning from peers, and incentives influence them (researchers) to engage in collaborative research.

The results indicated that over 44 (86%) respondents engage in collaborative research to strengthen their networks with other scholars. The respondents reported that networking helps to bring these scholars who happen to have common interests together and create partnerships that often last for longer. Researchers usually work alone on their projects which leaves them isolated. Networking or coming together with fellow researchers to work on a project together may help overcome that isolation. The importance of networking was also highlighted by 37 (73%) respondents who indicated that they collaborate in research to overcome intellectual isolation. Another patch of respondents numbering 38 (75%) also agreed to be collaborating with an aim of sharing resources. This can be very significant to researchers from smaller institutions and underdeveloped countries with little resources. Such partnerships can allow them to take advantage of the available resources in both institutions, some of which may not be available in their smaller institutions.

Learning from peers was also one of the most common factors among respondents on why they collaborate in research. The results show that 43 (84 %) respondents collaborate in research to learn from their peers. This usually happens where two or more scholars with different expertise come together to solve a research problem. Each researcher brings a special skill that may not be known by the others and that brings an opportunity for all to learn from one another. There were mixed feelings among respondents when it came to having to collaborate to get incentives. In South Africa, a number of institutions usually attach incentives to publications published in selected peer reviewed journals, book chapters, peer reviewed conference proceedings and books that earn subsidy from the Department of Higher

Education and Technology (DoHET). Only 24 (47%) respondents indicated that incentives may influence them to collaborate with 21 (41%) saying incentives have very little influence on them when it comes to collaborating. It has been informally noted by researchers at some forums of discussion that some researchers at times choose not to collaborate so that they don't share incentives made available and opt to work alone. This can have serious implications because those who are skilled enough will work alone and continue getting incentives while they are not leaving anyone to take over from them when they retire which will create a knowledge gap. Having incentives for research in an academic setting is motivating and encouraging for researchers but it has negative implications for the future.

Reasons for collaborating

Respondents were requested to give specific reasons that are likely to foster collaborative initiatives with particular groups such as, students; mentors; mentees (other than students); colleagues in the same department; fellow researchers; and international researchers.

Reasons for collaborating with students and mentees (other than students)

The responses received for this question were not that surprising considering the population for this study. Respondents indicated that they collaborate with students to impart knowledge and help the latter to obtain their qualifications. Some respondents indicated that collaborating with students is part of their jobs. A number of promoters feel that it takes a lot of time to do postgraduate supervision and as a result, they make sure that they get an article out of the whole project so that their efforts do not go to waste. It was also interesting and encouraging to note that some supervisors feel that students bring fresh perspectives on themes and ideas that they may be having at the time. This means that such supervisors give students a platform and opportunity to participate in the whole project while taking their ideas into consideration. Furthermore, respondents indicated that they would like to share their experiences on a particular subject and help capacitate their mentees while strengthening their relationships with their students at the same time exploring areas outside their subject specialisation.

Reasons for collaborating with mentors and managers

There was a general consensus among those respondents, who are being mentored by senior colleagues, that it is important to tap into the mentor's experience and knowledge in order to develop skills and research avenues. Mentorship of young researchers where the latter learns from the senior and experienced colleagues is again at the centre stage. Field (2001:270) is of the opinion that a mentor should play an important role in the career development of mentees, by providing them with background information and support for individual growth, as well as making them aware of opportunities available.

The other important thing about having a mentor is the creation of an opportunity to connect with the mentor's professional networks. This allows the mentee to grow and expand his/her professional boundaries. Mentorship can either be formal or informal. The best example of a formal mentorship is that of a supervisor working with a post graduate student. Informal mentoring may happen between the experienced and the less experienced through a personal connection. One respondent mentioned that mentors know their mentees best, and it is advantageous to work with someone who knows and understands his/her mentee well. Having worked with someone before gives the mentee an advantage of knowing how the mentor does things and what the latter expects of him/her. This is important during collaboration where responsibilities are shared because it will be helpful in deciding which role should be played by whom. Other respondents indicated that a natural consequence of being a young researcher and wanting to learn definitely motivated them in the conduct of collaborative research with their mentors.

Reasons for collaborating with colleagues in the same department

Being in the same department will most likely mean that one knows and understands each other's strengths and weaknesses. Respondents indicated that they collaborate with colleagues with the aim of producing high quality papers in a short space of time to enhance their productivity. Some respondents mentioned a desire to pursue niche areas in their departments as a reason for collaborating with fellow researchers. They indicated that such collaborative research has the potential to generate income for them and increase their research output. Some respondents indicated that they work on departmental joint projects and they have no choice or can't avoid them as they are in the same department. This group may not yield desired results because collaboration is not conducted between willing partners who are committed to seeing the project through to the end.

Other respondents mentioned that co-supervision of students' work automatically gets them to work together and eventually they publish together with the students. In view of the fact that some LIS schools in South Africa have closed down or changed focus to non-LIS disciplines, the onus is left to the few available LIS schools to ensure the survival of the profession. The closing down of LIS schools has put too much pressure on the few academics left in LIS as they are expected to service the increasing student numbers and also conduct research so they stay relevant. This situation encourages collaboration where researchers will share responsibilities and reduce the time and effort required to complete a task.

Reasons for collaborating with colleagues from other departments

The respondents indicated that collaborating with someone from another department in the conduct of research widens their horizons. The respondents further mentioned that such collaboration is very important because it helps with the establishment of interdisciplinary networks and exposure to a wide variety of research methods. The other notable reason mentioned by the respondents is the cross-pollination of ideas that will result from collaborating with someone from a different department or discipline.

Reasons for collaboration with International Researchers

This type of collaboration as discussed in the sections above enables researchers to share international experiences, foster international networks, and can help researchers do comparative studies with peers from other countries. Respondents who indicated that they have collaborated at the international level believe that global perspective is key to providing comprehensive research studies. Researchers can never work in isolation and the same should happen in LIS. International collaboration according to some respondents can increase researchers' chances of accessing funds and publications as well as get international visibility.

Barriers to collaboration

This section explores the issues that LIS scholars perceive to hinder effective research collaboration in LIS schools in South Africa. Katz and Martin (1997) gave a summary of the following barriers to collaboration:

- Financial implications in the form of travel costs , moving of equipment's and so forth
- Increased administration resulting from more people/institutions involved,
- Lack of time from some collaborators, or additional time required as different parts of the research will be done in different locations
- Different management cultures, financial systems and rules on intellectual property rights

Table 4. Barriers to collaboration

	To a great extent	Somewhat	Very little	Not at all
Bureaucracy	42.2%	33.3%	22.2%	2.2%
Lack of funding	43.5%	28.3%	19.6%	8.7%
Intellectual property rights	9.1%	29.5%	36.4%	25.0%
Lack of time	43.5%	28.3%	15.2%	13.0%
Clash of values	9.1%	31.8%	34.1%	25.0%
Ethics	15.9%	18.2%	27.3%	38.6%
Distance between researchers	15.2%	19.6%	23.9%	41.3%

For this study, respondents were first asked to indicate the extent to which barriers such as bureaucracy, lack of funding, intellectual property rights, lack of time, clash of values, ethics, and distance between researchers may have prevented them or are likely to prevent them from engaging in collaborative research. Secondly respondents were requested to indicate the extent to which a number of personal traits and characteristics may be a barrier/s to research collaboration. Table 4 provides the extent to which some factors act as barriers to effective collaboration.

Table 5. Personal traits or characteristics that may be a barrier to research collaboration

	To a great extent	*Somewhat*	*Very little*	*Not at all*
Gender	6.7%	15.6%	20.0%	57.8%
Level of education	31.1%	44.4%	20.0%	4.40%
Competencies	70.5%	29.5%	0.0%	0.0%
Honesty	72.7%	13.6%	6.8%	6.8%
Respect	80.0%	11.1%	6.7%	2.2%
Self-discipline	72.1%	23.3%	4.7%	0.0%
Work Ethic	75%	20.50%	4.5%	0.0%
Mutual Intent	75%	20.50%	4.5%	0.0%
Attitude	70.5%	25.0%	4.5%	0.0%
Interpersonal skills	47.7%	45.5%	2.3%	4.5%
Reliability	74.4%	23.3%	0.0%	2.3%
Nationality	4.7%	2.3%	20.9%	72.1%

A good majority of respondents (i.e. 39 or 76%) indicated that bureaucracy may be a barrier to collaboration. We believe that academics work under tight deadlines and the pressure to deliver is high and therefore too much red tape may sometimes delay their progress. Again over 36 (71%) respondents indicated that lack of funding maybe a barrier to collaboration. It should be noted that many institutions make funds available for research but if access to those funds is a problem then little research will be done. If a project does not receive funds then it will never get off the ground. It was interesting and surprising to note that 34 (66%) respondents indicated that ethics has very little impact on whether they collaborate or not. We opine that ethics is very important in research and perhaps that is why institutions around the world have adopted specific ethical principles when it comes to research. Only 17 (34%) respondents indicated that ethics may be a great barrier and influence their decision to collaborate. The distance between researchers also seem not to be a problem among respondents with 33 (65%) respondents indicating that it will not stop them from collaborating. The latest computer technologies such as Skype make it possible to work with someone who is in another country as if one were in the same room, so the issue of distance is increasingly becoming a thing of the past.

The majority of the respondents (i.e. 29 or 57.8%) did not see gender as barrier to collaboration. However someone's level of education was considered very important by the respondents. Over 38 (75%) respondents indicated that someone's level of education may be a barrier to collaboration. This may be influenced by the fact that researchers collaborate to accomplish goals that they cannot accomplish on their own; as a result, someone who is not academically capable may not be a good partner to have especially when one is under pressure to deliver. This was supported by the fact that all respondents suggested that somebody's inadequate competencies is definitely a barrier to collaboration. Personal characteristics such as honesty, respect, self-discipline, as well as attitude had over 46 (90%) respondents strongly indicating that the attributes will definitely block them from collaborating. Everybody wants to be associated with a well-mannered and respected person as well as someone who is not troublesome.

Reasons for not collaborating

Just like in the study by Katz and Martin (1997), this study investigated those underlying reasons that may hinder collaboration in LIS in South Africa. Respondents were asked to provide reasons that best describe why they may not collaborate with the following groups: students, mentors, Mentees other than students, colleagues in the same department, fellow researchers, seniors or managers and international researchers. The following were results as obtained from the survey.

Reasons for not collaborating with students and mentees

There was a general feeling amongst respondents that they will never work with students who are lazy and not prepared to work. This factor cannot be overemphasized as respondents mentioned issues like, lack of competencies, poor work ethic, and not following instructions on the students' side as main reasons they may not collaborate with students. Students who are repeating the same mistakes or not considering any advice or guidance given to them may be left without mentors. The respondents feel that such students may delay them at times as they do not stick to deadlines and agreements. Senior researchers may want to share their knowledge and skills but if the partner is not willing to learn then it defeats the whole purpose. Senior researchers are rated and evaluated according to their output and therefore wasting time on someone who does not want to learn or not willing to learn may be costly for them. Other responses included lack of mutual understanding, lack of commitment, time constraints as well as if the two parties do not share common research goals.

Reasons for not collaborating with mentors and managers

There were no surprises when it came to reasons why researchers will not collaborate with their seniors or managers in the conduct of research. A number of respondents were concerned about the fact that their mentors or seniors make them do all the work but equally share the credit which is somehow discouraging to them. Even though this is obviously unethical, it is common knowledge that some mentors abuse their positions and take advantage of their mentees. Young researchers will be expected to do all the work with little contribution from their more senior collaborating partner. Respondents further mentioned that mentors always demonstrate authority, lack empathy and never listen to their suggestions. Ignoring the contribution made by the more junior researchers may be demoralising and may result in the young researchers losing interest in conducting research because of the lack of self believe. Managers or mentors have an obligation to build as any form of advice or feedback is supposed to build as opposed to being too harsh. Many masters and doctoral students never complete their studies as some mentors give poor feedback or criticism that is aimed at breaking the students. Some of the respondents mentioned a lack of work ethic, lack of time, and not getting valuable advice or input from their mentors as other reasons for not collaborating with their mentors. Mentors normally have a lot of commitments, and a

collaborative project with a student may not be a priority to them, while the student's development and growth will be depending on it. This can therefore discourage students from wanting to collaborate with mentors.

Reasons for not collaborating with colleagues in the same department

This was a very interesting question and some of the responses given were somehow unexpected. Respondents mentioned that some colleagues have drawn their own conclusions about others which affect or influence their decision to collaborate. This is again a question of underestimating others and having one's own biased perceptions of others before they get to know them. That is a personal problem and has to do with everybody's personality and can only be solved over time, even though it poses challenges. Other respondents indicated that they will never collaborate with colleagues in their department because some colleagues never give their ideas a chance. This is a problem everywhere; colleagues who are mostly quiet may keep their ideas to themselves in such partnerships. Others are not good in expressing themselves and will mostly keep to themselves. This may result in ideas that end up being used although they are not the best, just because they came from the most vocal participants. One respondent indicated that in some instances, the most vocal colleagues may have a good command of the English language, while their ideas lack substance. Some of the other reasons raised include selfish colleagues, clash of ideas, competencies, attitude; lack of work ethic, and professional jealousy which was really unexpected. Some colleagues may feel that involving others in projects and working together may improve their profile and maybe become a threat to them in the work environment. Such colleagues end up being selfish and holding on to information and blocking their fellow colleagues. Others indicated they are so busy to an extent that they do not have time to do any other extra work, including collaborative research. Issues relating to office politics and intellectual property rights were also highlighted as possible reasons why some respondents do not enter into collaborative initiatives with fellow colleagues in the same department.

Reasons for not collaborating with fellow researchers

This question aimed to get responses on why LIS researchers are not collaborating or may not collaborate with fellow researchers in other departments as well as those in other universities. Many responses given were similar to the ones given in the immediate question above. However the issue of different research interests came out ahead of others. Even though many universities encourage multi-disciplinary research, researchers seem to prefer working with scholars who understand their area of interest and methodologies involved in the research, just to name but a few. Other reasons included unethical behaviour, time and distance between researchers, and different agendas among collaborating researchers.

Reasons for not collaborating with international researchers

Most of the barriers already indicated in the preceding questions were also mentioned here. Other reasons which were given by respondents regarding this question and are worth mentioning include distance and logistical problems, lack of communication, and topical issues, just to list a few. There is a general feeling from many local researchers that it is really not easy to work with someone who is very far especially in another country, even though the technologies available today make this possible and better than before.

Conclusions

The study by Sooryamoorthy (2009) revealed that collaboration in research in South Africa has been growing steadily over the years. This implies that, even though there are difficulties and drawbacks associated with collaboration in research, LIS researchers are mainly focusing in all the benefits that come with such partnerships and therefore engaging in collaborative research. It is important to mention that, even though the benefits of collaboration are evident,

the drawbacks cannot be ignored. A re-look at the enhancers and inhibitors of research collaboration suggests that the distance between researchers, past relationships and the institution of affiliation most influenced who collaborated with whom. The results imply that LIS researchers prefer partnering with colleagues who are nearer, mainly from the same institution. The collaboration networks suggest that issues discussed above have had a major impact on the current status of collaboration in LIS research in South Africa.

Collaboration links between supervisors and students are very much evident and seem to be the most influencing factor on research collaboration among LIS researchers in South Africa. It is also very encouraging to see some partnerships between senior researchers from different schools which is crucial for the growth and development of research in the field. Ocholla (2008) has observed that collaboration of LIS schools is weak and largely informal. This was very evident in the current study, too. Collaboration mainly happened between individuals while departments rarely collaborate hence there is no evidence of students from a particular university collaborating with their peers from other universities. This finding concurs with the views of Ocholla & Bothma (2007) who indicated that collaboration among LIS schools and researchers in such areas as "teaching, research, student and staff exchange, conferences, workshops, curriculum development, publications, research supervision, examination is very important yet very minimal"**.**

Acknowledgments

This paper reports part of the findings of Mr Jan Malulaka's Masters dissertation (University of South Africa, 2014), supervised by Prof Omwoyo Bosire Onyancha, Chair and Professor, Department of Information Science and Prof Isola Ajiferuke, University of Western Ontario.

References

Bozeman, B. & Corley, E. (2004). Scientists collaboration strategies: Implications for scientific and technical human capital. *Research Policy*, *33*(4), 599-616

Field, J. (2001). Mentoring: a natural act for information professionals? *New Library World*, *102*(7/8), 269-273.

Katz, J.S. & Martin, B.R. (1997). What is research collaboration? Retrieved February 20, 2010 from: http://www.sussex.ac.uk/Users/sylvank/pubs/Res_col9.pdf.

Leedy, P.D. & Omrod, J.E. (2010). *Practical Research: Planning and Design 9th Ed*. New Jersey: Pearson education, Inc.

Liebowitz, B. (2009). What's inside the suitcases? An investigation into the powerful resources students and lecturers bring to teaching and learning. *Higher Education Research and Development, 28*(3), 261-274.

Merlin, G. (2000). Pragmatism and Self-Organization Research Collaboration on the individual level, *Research Policy, 29*(1), 31-40.

Narvaez-Berthelemot, N., Russell, J. M., Arvanitis, R., Waast, R., & Gaillard, J. (2001). Science in Africa: An overview of mainstream scientific output. In: M. Davis & C. S. Wilson (eds.). *Proc. 8th International Conference on Scientometrics and Informetrics,* Sydney, July, 16-20, 2, 469-476.

Neuman, W.L. (2006). *Social Research Methods: qualitative and quantitative approaches*. (6th ed.) Boston: Allyn and Bacon.

Ocholla, D N. (2008). The current status and challenges of collaboration in library and information studies (LIS) education and training in Africa". *New Library World*, *109*(9/10), 466 – 479.

Ocholla, D.N. & Bothma T.D.J. (2007). Library and information education and training in South Africa. *New Library World, 108*(1/2), 55-78.

Onyancha, O.B. (2009). Towards Global Partnerships in Research in sub-Sahara Africa: an informetric study of the national, regional and international country collaboration in HIV/AIDS literature in eastern and southern Africa. *South African Journal of Libraries and Information Science, 75*(1), 86-99.

Onyancha, O.B. & Maluleka, J.R. (2011). Knowledge production through collaborative research in sub-Saharan Africa: how much do countries contribute to each other's knowledge output and citation impact? *Scientometrics, 87*, 315–336.

Onyancha, O.B. & Ocholla, D.N. (2007). Country-wise collaborations in HIV/AIDS research in Kenya and South Africa, 1980–2005. *Libri*, *57*(4), 239-254.

Smith, D. & Katz, S. (2000). Collaborative Approaches to Research. A report to the Higher Education

Funding Council for England. Centre for Policy Studies in Education, University of Leeds.

Sooryamoorthy, R. (2009). Do types of collaboration change citation? Collaboration and citation patterns of South African science publications. *Scientometrics, 81*(1), 177-193.

Sugimoto, C.R. (2010). Collaboration in information and library science doctoral education. *Library & Information Science Research, 33, 3-11.*

Sun, Y. (2006). Bibliometric analysis of scientific research collaboration between Japan and China. *Int. Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting.*

Yamashita, Y. & Okubu, Y. (2006). Patterns of scientific collaboration between Japan and France: Inter-sectoral analysis using Probabilistic Partnership Index (PPI). *Scientometrics*, *68*(2), 303-324.

The Diffusion of Nanotechnology Knowledge in Turkey

Hamid Darvish[1] and Yaşar Tonta[2]

[1] *hssdarvish@gmail.com*
Kastamonu University, Faculty of Arts & Sciences, Information Management Department, Kastamonu (Turkey)

[2] *yasartonta@gmail.com*
Hacettepe University Department of Information Management, 06810 Beytepe, Ankara (Turkey)

Abstract
This paper assesses the diffusion of nanoscience and nanotechnology in Turkey in the last decade using bibliometric and Social Network Analysis (SNA) techniques. We extracted a total of 10,062 articles and reviews from Web of Science (WoS) authored by the Turkish scientists between 2000 and 2011. We divided the data set into two 6-year periods (2000-2005 and 2006-2011). Almost three quarters (7,398) of all papers were published between 2006 and 2011. For each period, we compared the number of nanotechnology papers, the universities' output along with their levels of collaboration with one another, the diffusion and adoption of nanotechnology, the most prolific authors and the nanotechnology research topics studied most often by the Turkish researchers. We found that nanotechnology research and development (R&D) in Turkey is on the rise and its diffusion and adoption has increased tremendously in the second period. This is due primarily to the fact that the government identified nanotechnology as a strategic field a decade ago and decided to provide constant support for nanotechnology R&D. Overlay maps showed that nanotechnology R&D in Turkey concentrated primarily in Materials Sciences, followed by Chemistry, Physics, Clinical Medicine and Biomedical Sciences.

Conference Topics
Country-level studies, Mapping and visualization

Introduction[1]

Nanoscience and Nanotechnology is the study of materials at atomic levels within the 1 to 100 nm range (i.e., at a magnitude of 10^{-9} of a meter) (Mehta, 2002). Although Nanotechnology has been introduced more than half a century ago by Feynman (1960), it took some time for the nanotechnology research to pick up. Many countries have invested heavily in nano-related technologies in the past two decades. The US government, for example, has allocated 1.74 billion US dollars to nano-related technologies in 2011 (Sargent, Jr., 2013). European countries under the 7th Framework Program have also heavily invested in joint projects among its members. Consequently, the number of scholarly publications in nano-related technologies in North America, Europe and Far Eastern countries has increased.

Turkey as a developed country prepared its strategic plan by taking nano-related research and development into account. Nanotechnology including nanophotonics, nanoelectronics, and nanoscale quantum computing is one of the eight strategic fields of research and technology mentioned in Turkey's "Vision 2023 Technology Foresight Study" that was prepared as part of the "National Science and Technology Policies 2003-2023 Strategy Document" by the Supreme Council of Science and Technology (SCST) more than a decade ago (Ulusal, 2004, pp. 19-20). Nanotechnology as a research field has been receiving state support since 2007 in Turkey (about one billion Turkish Lira, or circa 500 million USD). The Turkish Scientific and Technological Research Council (TUBITAK) and the Ministry of Development (MoD) support nanotechnology projects financially. For example, MoD continues for more than a decade to invest to improve the infrastructure of nanotechnology research facilities and

[1] This paper is based on the findings of first author's Ph.D. dissertation entitled "Assessing the diffusion of nanotechnology in Turkey: A Social Network Analysis approach." (Darvish, 2014).

supported the establishment of nanotechnology research centers. In addition, it supports several nanotechnology-related projects carried out by research institutes and universities.

Thanks to state support, nanotechnology has become a major field of research in Turkey. Universities invested heavily in nanotechnology in the last decade. More than 20 nanotechnology research centers were set up mostly in universities. Among them are Bilkent, Middle East Technical, Hacettepe, Sabancı, İstanbul Technical and Boğaziçi Universities. More than 10 universities are offering both undergraduate and graduate degrees in nanotechnology. More than 100 commercial companies and start-ups of various sizes have also invested in nanotechnology (e.g., Normtest, Arçelik, Yaşar Holding, Yeşim Textile and Zorlu Energy) and developed commercial nanotechnology products in a number of sectors including surface coating, textile, chemistry, automotive and construction industries, and polymer and composite materials. Turkey has been among the first three countries in terms of the growth of nanotechnology research with some 2,000 scientists working in this field (Bozkurt, 2015, p. 49; Denkbaş, 2015, p. 84; Özgüz, 2013). The number of nanotechnology related scientific papers published by Turkish researchers and listed in Web of Science (WoS) is ever increasing (more than 2,500 in 2014 alone).[2]

This paper aims to assess the diffusion of nanoscience and nanotechnology in Turkey between 2000 and 2011 using bibliometrics and Social Network Analysis (SNA) techniques. It identifies the total production of nano-related publications by Turkish researchers and the key fields in which nanotechnology is applied in Turkey (e.g., biomedicine, pharmacy, and metallurgy). The adoption of nanotechnology by the most prolific universities and the diffusion of nanotechnology knowledge through collaboration among them is also studied.

Literature Review

Scientists have investigated the diffusion of innovation and knowledge in societies from different perspectives. Rogers (2003, p. 5) defines the diffusion of an innovation as "the process by which an innovation is communicated through certain channels over time among the members of a social system." Social interactions between scientific domains and practitioners are instrumental to the diffusion of innovation and knowledge. According to Rogers, the key elements in the diffusion process are: innovation/knowledge, communication channels, time and social systems (p. 7). An innovation starts with a few people and has a few adopters, but eventually it gains the momentum until it reaches its peak. Rogers likens the diffusion process of an innovation to a mathematically-based bell curve (also known as "Rogers adoption/innovation curve") and categorizes the adopters accordingly (i.e., starting from the left tail of the curve to the right, 2.5% of the adopters are called "innovators", 13.5% "early adopters", 34% "early majority", 34% "late majority", and the remaining 16% on the right tail of the curve as "laggards").

Poire (2011) looks at the timeframe of the adoption of innovations along with the impact of innovations on the economy. He argues that "it takes about 28 years for a new technology to become widely accepted, followed by a period of rapid growth lasting about 56 years. Some 112 years after invention, the innovation reaches maturity and grows in-line with population increases" (Roy, 2005, p. 9). Using these yardsticks, he convincingly charted the adoption processes of textiles, railways, automobiles, computers and nanotechnology. He predicts that nanotechnology, which according to him came into being in 1997, will be more widely adopted by 2025, followed by a 56-year long rapid adoption period (until 2081) during which time nanotechnology products will become an integral part of our everyday life like computers.

[2] Search on WoS was carried out on January 11, 2015.

If an innovation is communicated among the members of a social system, as Rogers indicated, then studying social systems is important because scientists work and collaborate within such systems. Assessing social relations among scientists reveals how collaborative they are. Conventionally, Derek de Solla Price (1965) studied the scholarly communication process between scientists, thereby opening the door to the quantitative study of science.

Social Network Analysis is a paradigm in which relational interaction among members signifies the role of people in a network structure (Wellman & Berkowitz, 1997). The diffusion of knowledge in a network of people can thus be studied by exploring the social structure of the network along with the relations and collaboration (or lack thereof) among network members using SNA concepts such as density and centrality. For example, poorly connected "structural holes" in a densely connected network are crucial for connecting "clusters" (groups of people) in a network structure and for the diffusion of knowledge in the network (Burt, 1992). Newman (2000) referred to clustering as "community structure". The value of a person in a social network is therefore linked to his/her potential to establish connections between clusters that are separated by structural holes.

Scientific discovery comes with a group of specialized people who "attend, read and cite the same body of literature and attend the same conferences" (Chen et al., 2009, p. 192). Bibliometric methods such as co-citation (Crane, 1972) or co-author (Girman & Newman, 2002) analyses were used to study the diffusion of knowledge in the network of scientists as well as to track the level of collaboration among different partners along with the emergence of new research areas. As a collaborative model involving universities (research centers), funders and industries, the Triple Helix was proposed to streamline the diffusion of knowledge (Leydesdorff & Etzkowitz, 1998).

Scientometricians use visualizations in addition to other indicators to track or investigate new scientific developments over time. For example, science overlay maps were introduced as a novel approach to illustrate the bodies of research precisely surrounded by global scientific domains (Rafols, Porter & Leydesdorff, 2010). Science overlay maps can represent different types of data and large data sets such as network of authors, publications and universities succinctly and "help benchmark, explore collaborations, and track temporal changes" (Rafols, Porter & Leydesdorff, 2010, p. 1871).

Nanotechnology has been the subject of several studies in the past and reviewing them is beyond the scope of this paper. However, we should mention Milojević (2009, 2012) who studied the coginitive content of nanoscience and nanotechnology as well as its diffusion using SNA techniques and mapped the evolution and socio-cognitive structure of it. We should also mention one particular study that measured the growth and diffusion of nanotechnology on a global level on the basis of the number of publications produced by countries as well as the most prolific institutions and authors along with the most cited authors, papers and journals (Kostoff, Stump, Johnson, Murday, Lau & Tolls, 2006). China, Far Eastern countries, USA, Germany and France were among the most prolific ones.

As mentioned earlier, Turkey is among the first three countries based on the growth of nanotechnology research. Turkey's contribution to nanotechnology literature was also evident at the global level (Kostoff, Koytcheff & Lau, 2007). Recently, the state of nanotechnology centers and companies carrying out research and manufacturing nano-related technologies in Turkey was studied with a view to compare them quantitatively with their counterparts in China, India and Germany, for example (Aydoğan-Duda & Şener, 2010; Aydoğan-Duda, 2012). The present study attempts for the first time to map the nanotechnology output of Turkish universities and investigate the diffusion of nanoscience and nanotechnology knowledge in Turkey at the micro level by means of Social Network Analysis and bibliometrics. The results can be considered as a stepping stone for comparative studies for future studies.

Method

The aim of this research is to assess the diffusion of nano-related technology by mapping of collaborative social structure of scientists in Turkey between 2000 and 2011. We attempted to address the following issues: (a) the most prolific universities publishing nanotechnology research; (b) the rate of diffusion of nanotechnology knowledge and its adoption within universities between 2000 and 2011; and (c) key areas of nanotechnology research.

In order to answer the research issues, we used a compound textual query on nanotechnology modified from Kostoff's[3] and searched (WoS). We retrieved a total of 10,062 papers (with at least one author of each paper affiliated with a Turkish university or research institute) published between 2000 and 2011. We then divided the data set into two 6-year periods (2000- 2005 and 2006-2011) to further assess the diffusion of nano-related technology in Turkey.

We analyzed co-occurrences among universities to capture collaborations in network structures. VOSviewer was used to implement the method of "associative strength" that clustered bibliometric data based on their similarities and mapped the network structure. A geocoder[4] was used to get the geo-coordinates for each city and Google Maps was used to overlay the relationships among cities on a geographic map. Bibexcel was used to calculate the most frequent collaborators from selected universities in the research. The top ranked universities in each period (2000-2005 and 2006-2011) were selected on the basis of their co-occurrence in terms of scientific collaboration on nanotechnology. Gephi, VOSviewer and GoogleMaps were used to map the network structure.

Findings

The number of Turkey's scientific publications on nanotechnology increased from 215 papers in 2000 to 1,748 in 2011, more than an eight-fold increase (Fig. 1). Almost three quarters (7,398) of all papers (articles and reviews) were published between 2006 and 2011 while the rest (2,664) were between 2000 and 2005. This increase is mainly due to Turkey's making nanotechnology a priority field in its 2003-2023 strategic plan and providing state support to nanotechnology research and development starting from 2007. The number of newly-established universities, hence the number of researchers studying nanotechnology, has also increased tremendously in this period.

There are about 180 universities in Turkey, two-thirds being state-funded. Using the fractional counting method, Figure 2 shows the top ranked universities based on the number of nanotechnology papers they published between 2000 and 2011. The Middle East Technical, Hacettepe, İstanbul Technical, Gazi and Bilkent Universities are the top ranking ones. All but four (Bilkent, Koç, Fatih and Sabancı) universities in Figure 2 are state funded.

[3] Personal communication with Prof. Ronald N. Kostoff (20 April 2012). Search query is available from the authors upon request.

[4] Available from http://www.gpsvisualizer.com/geocoder/.

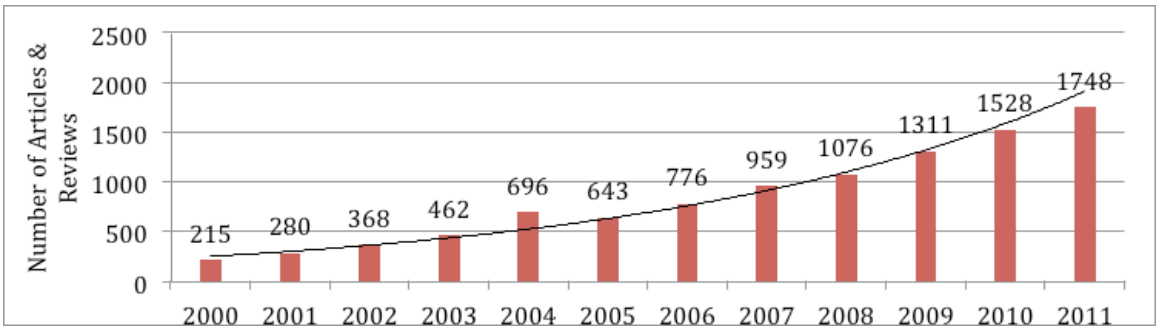


Figure 1. Number of nano-related technologies publications in Turkey: 2000-2011 Source: Thomson's ISI Web of Science as of November 2013.

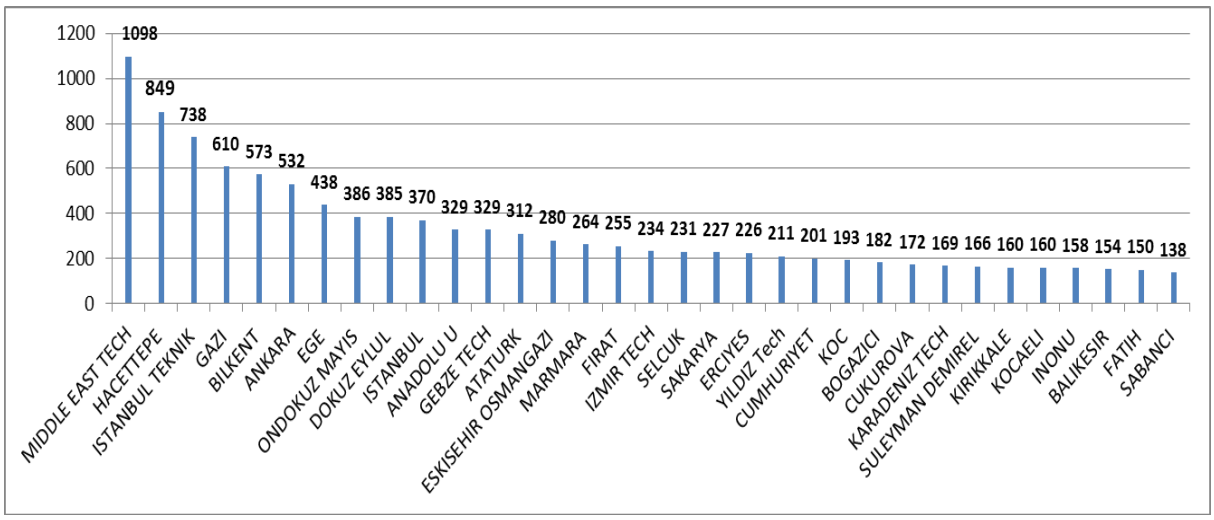


Figure 2. Number of nanotechnology papers of the top Turkish universities between 2000 and 2011 Source: Web of Science as of November 2013.

To assess the level of collaboration and the diffusion of nanotechnology knowledge among universities, we examined the average co-occurrence frequencies of all universities in published papers and created separate networks for the periods of 2000-2005 and 2006-2011 (Fig. 3). The collaboration network was much sparser in the first period with a few universities such as Hacettepe and METU acting as hubs of research on nanotechnology and cooperating with others. The network was much denser in the second period with more universities both acting as hubs of nanotechnology research and collaborating with their counterparts. This is an indication of an increasing level of collaboration among universities in carrying out nanotechnology research within a relatively short period of time.

The diffusion of nanotechnology knowledge in Turkey can be examined from a somewhat different angle by looking at the number of provinces where nanotechnology research took place. Turkey is divided into 81 administrative provinces. The information presented in Figure 4 is less granular than that in Figure 3 due to a few provinces such as İstanbul, Ankara and İzmir having several universities (both old and new). Nevertheless, the number of provinces where nanotechnology research is carried out went up from 40 in the first period (2000-2005) to 72 in the second period (2006-2011). The geographical spread is due to new universities being established in some provinces for the first time and to the government support that enabled researchers both in new and old universities to collaborate further.

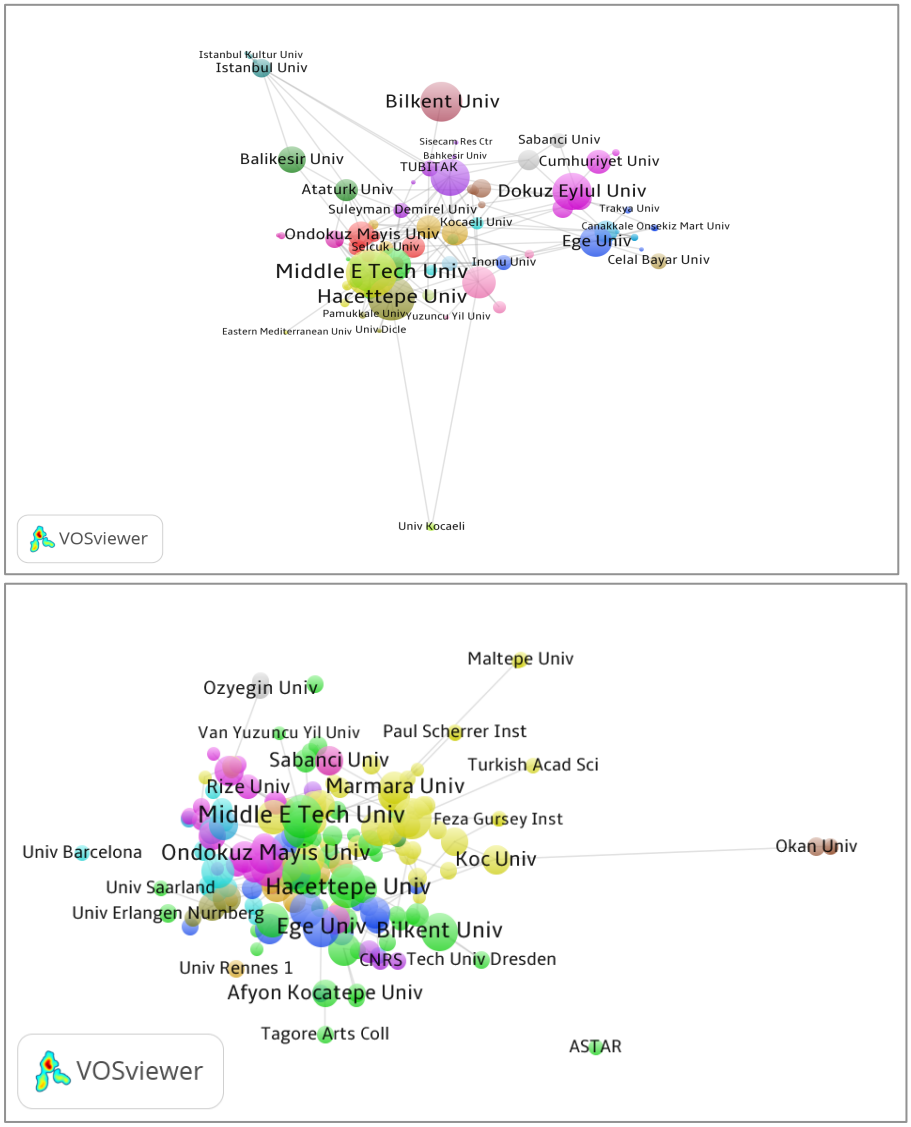


Figure 3. Collaboration of Turkish universities on nanotechnology (top) 2000-2005; (bottom) 2006-2011.

Table 1 shows the top 15 universities with the highest co-occurrence frequencies in both periods. The average co-occurrence frequency for the top 15 universities has almost tripled from 17 in 2000-2005 to 46 in 2006-2011. Note that the top 15 universities in the second period are slightly different from the ones in the first period, as some of the more prolific and more collaborative universities with higher frequencies of co-occurrence replaced the previous ones. We used the fractional counting method and found that the average number of nanotechnology papers published by the top 15 universities in the first period increased from 9 in 2000 to 27 in 2005, and from 35 in 2006 to 77 in 2011 in the second period, indicating more than an eight-fold increase (Table 2).

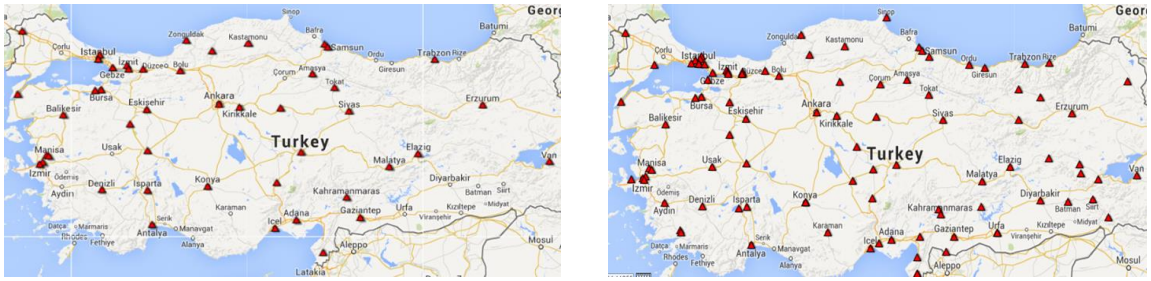


Figure 4. Geographical distribution of nanotechnology research activities in Turkish provinces (l) 2000-2005; (r) 2006-2011.

Table 1. Top 15 Turkish universities with the highest co-occurrence frequencies of collaboration between 2000 and 2011

2000-2005		2006-2011	
University	N	University	N
Hacettepe	30	Hacettepe	63
Middle East Technical	29	Gazi	63
Ankara	21	Middle East Technical	60
Gazi	20	Istanbul Technical	57
Istanbul Technical	18	Ankara	53
Gebze Institute of Technology	17	Gebze Institute of Technology	47
Dokuz Eylül	15	Ondokuz Mayıs	42
Marmara	14	Ege	41
Bilkent	14	Istanbul	41
Abant İzzet Baysal	13	Erciyes	40
Kırıkkale	12	Bilkent	38
Ege	12	Dokuz Eylül	34
Ondokuz Mayıs	11	Anadolu	34
Erciyes	11	Atatürk	33
Kocaeli	11	Fırat	31
Average	**17**	**Average**	**46**

Table 2. Number of papers published by universities with the highest co-occurrence frequencies in the second period (2006-2011)

University	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Hacettepe	79	85	89	97	95	107
Gazi	36	77	95	85	99	98
Middle East Technical	77	93	59	131	143	143
İstanbul Technical	52	64	65	88	91	121
Ankara	40	62	70	49	73	54
Gebze Institute of Technology	20	25	33	45	49	55
Ondokuz Mayıs	37	32	35	55	76	74
Ege	16	39	28	60	95	77
İstanbul	25	28	30	42	57	63
Erciyes	16	12	20	41	32	45
Bilkent	34	41	58	63	61	99
Dokuz Eylül	31	43	35	51	52	58
Anadolu	15	29	39	41	45	55
Atatürk	23	18	37	33	55	53
Fırat	17	19	23	31	45	50
Average	**35**	**44**	**48**	**61**	**71**	**77**

Next, we examined the diffusion of nanotechnology knowledge in Turkey using a more refined approach and identified the new authors collaborating each year in order to find out the adoption rate of nanotechnology research. Regardless of whether they appeared in the same paper or not, each new collaboration between any two authors was counted as one and considered a new adoption. The number of collaborating authors was just 214 at the beginning (2000) whereas it rose to 2,989 in 2011 (Table 3 and Figure 5). The number of new adopters was rather slow in the first period (2000-2005) with an average of 216 collaborations per year but the "tipping point" seems to have been reached in 2006 when the number of new adopters jumped from 282 in 2005 to 1622, an almost six-fold increase. The average number of new adopters in the second period (2006-2011) rose to 1868, more than eight times of what it was in the first period. Altogether, the number of cumulative new adopters soared in 12 years and was 13,692 in 2011. The annual rate of cumulative increase in percentages ranged between 11% (2004) and 54% (2006). Needless to say, the increase in the number of new adopters is primarily due to nanotechnology becoming a major research field in Turkey and nanotechnology research being supported by government funds.

Table 3. Number of new and cumulative adopters between 2000 and 2011

Year	# of new adopters	# of cumulative adopters	Rate of cumulative increase (%)
2000	214	214	0
2001	177	391	45
2002	193	584	33
2003	381	965	39
2004	115	1080	11
2005	282	1362	21
2006	1622	2948	54
2007	1668	4652	37
2008	1907	6559	29
2009	1919	8478	23
2010	2225	10703	21
2011	2989	13692	22

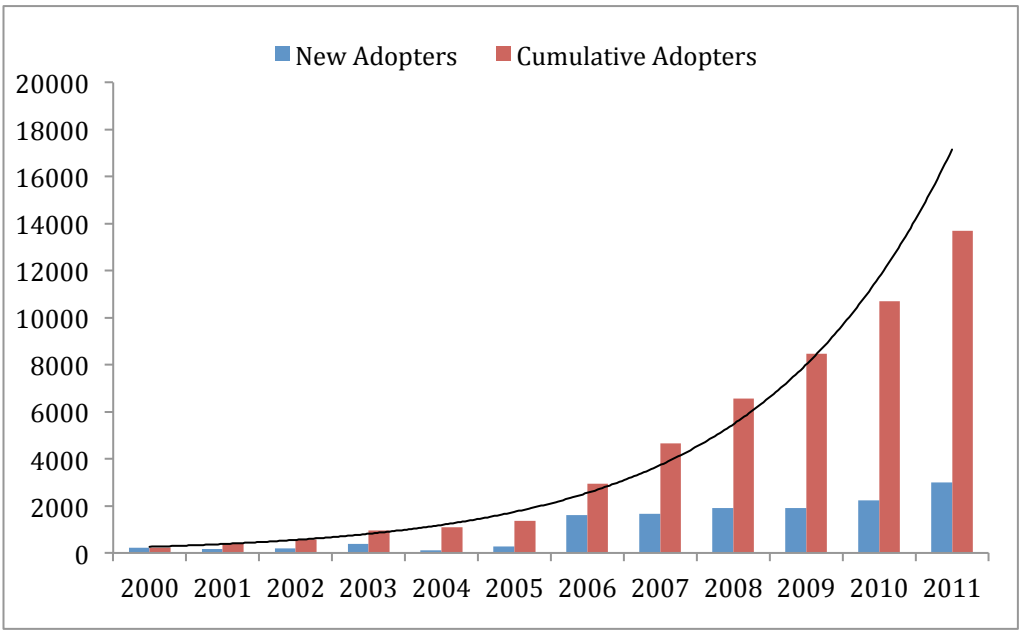


Figure 5. The growth of adoption of nanotechnology knowledge based on the number of collaborating authors (2000-2011).

Next, we identified the most prolific Turkish researchers in nanotechnology between 2000 and 2011 based on the number of papers they authored or co-authored. The fractional counting method was used for co-authored papers. Table 4 shows the top 20 researchers in both periods along with their total number of co-authors. The total number of papers authored or co-authored by the top 20 researchers almost doubled in the second period (from 645 to 1,189). Nine researchers appeared in both periods (italicized in the table) with different ranks. This means that 11 new researchers became more productive than they were in the first period and replaced the less productive ones in the second period or they entered the field anew. O. Buyukgungor of Ondokuz Mayis University, for instance, is at the top of the second period with 149 papers to his credit even though he did not appear in the top 20 of the first period. The top 20 most prolific researchers co-authored more papers with their colleagues in the second period (216 and 315, respectively). The number of co-authors of nine researchers who appeared in both periods increased 42% in the second period, indicating that they were influential in diffusing the nanotechnology knowledge to their colleagues. The same can probably be said for the remaining 11 researchers who appeared in the top 20 list in the second period.

Finally, we identified the research topics in nanotechnology that were studied more often by the Turkish scientists. We created separate overlay maps of research topics for both periods using ISI's 224 Subject Categories listed in WoS records. Both co-authorship networks and overlay maps were shared with five senior and five junior experts in nanoscience whose publications appeared in leading journals and their comments with respect to their places in the network were recorded (not reported here) (Darvish, 2014).

Table 4. The most prolific Turkish scholars in nanotechnology (2000-2011) Source: WoS (as of November 2013)

2000-20005			2006-2011		
N	First author & affiliation	# of co- authors	N	First author & affiliation	# of co-authors
53	Erkoc S (METU)	29	149	Buyukgungor O (Ondokuz Mayıs)	37
49	*Sokmen I (Dokuz Eylül)*	*16*	78	*Yagci Y (ITU)*	*19*
42	*Ciraci S (Bilkent)*	*13*	*75*	*Denizli A(Hacettepe)*	*18*
39	*Denizli A (Hacettepe)*	*12*	72	Yakuphanoglu F (Firat)	28
38	*Yagci Y (ITU)*	*10*	*67*	Ozkar S (METU)	23
37	Celik E (Bilkent)	11	*67*	*Toppare L (METU)*	*15*
37	*Sari H (Bilkent)*	*11*	*64*	*Ozbay E (Bilkent)*	*13*
36	Turker L (METU)	28	62	Yesilel OZ (Osmangazi)	17
30	*Yilmaz VT (Dokuz Eylül)*	*8*	*61*	*Sokmen I (Dokuz Eylül)*	*17*
30	*Toppare L (METU)*	*7*	58	Ozcelik S (Gazi)	12
29	Hascicek YS (Gazi)	8	52	Demir HV (Bilkent)	13
28	Ovecoglu ML (ITU)	7	49	Baykal A (Bilkent)	10
27	Elmali A (Ankara)	8	45	Turan R (METU)	10
26	Elerman Y (Ankara)	8	44	Sahin E (Bilkent)	11
26	Piskin E (Hacettepe)	8	*44*	*Yilmaz VT (Dokuz Eylül)*	*13*
26	*Kasapoglu E (Cumhuriyet)*	*8*	43	Caykara T (Gazi)	15
26	Balkan N (Bilkent)	5	*41*	*Sari H (Ankara)*	*9*
22	Yilmaz F (METU)	6	*40*	*Ciraci S (Bilkent)*	*12*
22	Turan S (Marmara)	8	*39*	*Kasapoglu E (Cumhuriyet)*	*12*
22	*Ozbay E (Bilkent)*	*5*	39	Albayrak C (Ondokuz Mayıs)	11

Each color in the map represents a subject category and the node size is proportional to its co-occurrence frequency with other nodes (Fig. 6). It appears that the nanotechnology papers authored by Turkish researchers in both periods were primarily related with Materials Science

(black) followed by Chemistry (blue), Physics (purple), Clinical Medicine (red), Biomedical Sciences (light green), Environmental Science and Technology (orange), and Computer Science (fuchsia). Subject categories appeared in overlay maps clearly show the priorities of Turkey in nanotechnology research and development and are commensurate with the nanotechnology products developed by commercial companies based in Turkey.

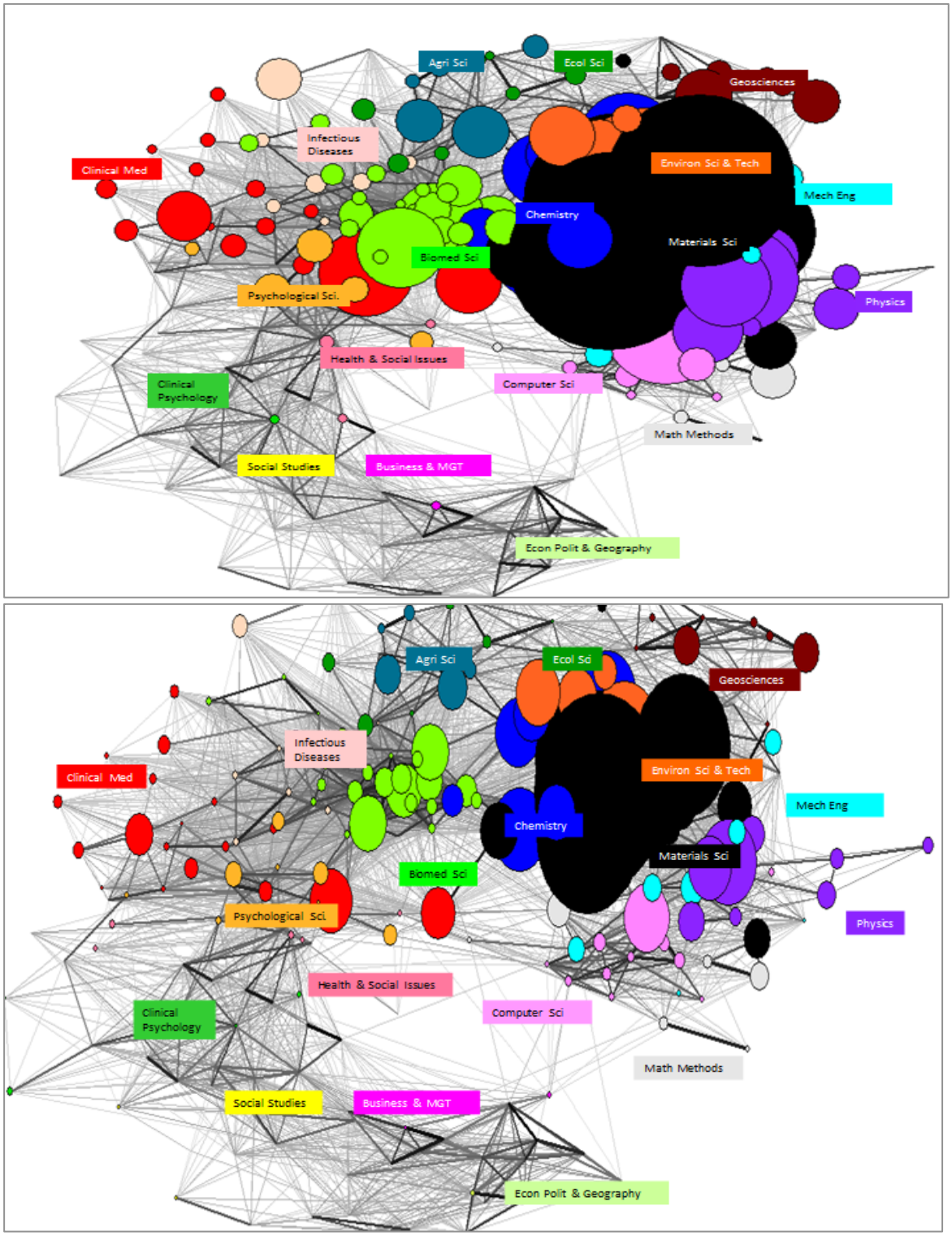


Figure 6. Overlay maps of subject categories of nanotechnology papers authored by Turkish scientists (top) 2000-2005; (bottom) 2006-2011.

Conclusion

Our analysis clearly shows that nanotechnology R&D in Turkey is flourishing. The number of nanotechnology papers published by Turkish scientists has tripled once the Turkish government has identified nanotechnology as one of the eight strategic fields in its national science and technology policy of 2003-2023 and decided to invest in nanotechnology accordingly. This decision has tremendously increased the diffusion and adoption of nanotechnology as a research field. Nanoscientists became more collaborative and more prolific in their research. This is somewhat similar to the experience of India, China, Iran and Latin American countries in that the importance of nanotechnology has increased once they identified it as a promising technology in their national development plans (Aydoğan-Duda, 2012).

The key areas of nanotechnology research and applications in Turkey are primarily in Materials Science, Chemistry, Physics, Clinical Medicine and Biomedical Sciences. All but Clinical Medicine appear in Milojević's list of areas as having the highest number of nanoscience and nanotechnology papers published in the literature (Milojević, 2012). The diversity of nanotechnology research shows that Turkish scientists are well aware of the trans- and interdisciplinary nature of nanotechnology as a discipline, although collaborative nanotechnology research in some areas such as Mathematics, Computer Science and Social Sciences seems to be currently lacking in Turkey.

Nanoscience stimulates scientific research in Physics, Chemistry, Biology and Medicine. Results revealed that notably well-established universities are instrumental in nanoscience research and newer universities are catching up. Turkey recognized the importance of nanotechnology as a strategic field relatively early. Based on Poire's timeframe of innovations becoming the drivers of economy, we can say that the diffusion of nanotechnology and its widespread adoption in Turkey will likely continue to accelerate until early 2030s.

References

Aydoğan-Duda, N. (2012). Nanotechnology: A Descriptive Account. Making it to the Forefront in Aydogan-Duda, N. (Ed). *Nanotechnology: A Developing Country Perspective*. 1, (pp. 1-4). New York: Springer.

Aydoğan-Duda, N., & Şener, I. (2010). Entry Barriers to the Nanotechnology Industry in Turkey in Ekekwe, N. (Ed). *Nanotechnology and Microelectronics: Global Diffusion, Economics and Policy*. (pp. 167-173). Hershey, PA: IGI Global.

Bozkurt, A. (2015). Türkiye, 10 yıldır "en küçük" dünyanın farkında, artık büyük adımlar atması gerekiyor (Turkey is aware of the "smallest" world for 10 years, but it should take big steps). *Bilişim: Aylık Bilişim Kültürü Dergisi*, 43(172), 44-53. http://www.bilisimdergisi.org/

Burt, R.S. (1992). *Structural Holes: The Social Structure of Competition*, Cambridge, MA: Harvard University Press.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an Explanatory and Computational Theory of Scientific Discovery. *Journal of Informetrics*, *3*(3), 191-209.

Crane, D. (1972). *Invisible Colleges: Diffusion of Knowledge in Scientific Communities.* Chicago, IL: University of Chicago Press.

Darvish, H. (2014). *Assessing the Diffusion of Nanotechnology in Turkey: A Social Network Analysis Approach*. Unpublished PhD Dissertation, Hacettepe University, Ankara.

Denkbaş, E.B. (2015). Nanoteknolojiye yapılacak yatırımlar, ülkelerin ekonomik gücünü yansıtabilecek bir parametre olacak (Investments in nanotechnology will become a parameter reflecting economic powers of countries). *Bilişim: Aylık Bilişim Kültürü Dergisi*, *43*(172), 78-87. http://www.bilisimdergisi.org/

Feynman, R.P. (1960). There's Plenty of Room at the Bottom. Caltech Engineering & Science. Retrieved, Feb.14, 2014, from http://calteches.library.caltech.edu/1976/1/1960Bottom.pdf

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks, *PNAS*, 99(12), 7821–7826.

Kostoff, R.N., Koytcheff, R.G., & Lau, C.G.Y. (2007). Global nanotechnology research literature overview, *Technological Forecasting & Social Change, 74*, 1733-1747.

Kostoff, R.N., Stump, J. A., Johnson, D., Murday, J.S., Lau, C.G.Y., & Tolls, W.M. (2006). The structure and infrastructure of global nanotechnology literature. *Journal of Nanoparticles Research*, *8*, 301-321.

Leydesdorff, L. & Etzkowitz, H. (1998). The Triple Helix as a model for innovation studies (Conference Report), Science & Public Policy, *Research Policy 25*(3), 195-203.

Mehta, M., (2002). Nanoscience and nanotechnology: Assessing the nature of innovation in these fields. *Bulletin of Science, Technology and Society*, *22*(4), 269-273.

Milojević, S. (2012). Multidisciplinary cognitive content of nanoscience and nanotechnology. *Journal of Nanoparticle Research*, *14*(1), 1-28.

Milojević, S. (2009). *Big Science, Nano Science? Mapping the Evolution and Socio-Cognitive Structure of Nanoscience/Nanotechnology Using Mixed Methods*. Unpublished PhD Dissertation, University of California, Los Angeles.

Newman, M. E. J. (2000). The structure of scientific collaboration networks. *PNAS*, *98*(2), 404–409.

Özgüz, V. (2013). Nanotechnology Research and Education in Turkey (presentation slides). Retrieved, December 27, 2014, from: http://rp7.ffg.at/upload/medialibrary/12_Oezguez.pdf.

Poire, N.P. (2011). *The great transformation of 2021: How the looming sustainability crisis will revolutionize capitalism, fracture the nation-state, and topple American supremacy*. Lulu.com.

Price, D. J. de Solla. (1965). Networks of scientific papers. *Science*, *49*(3683), 510–515.

Rafols, I., Porter, A. L., & Leydesdorff, L. (2010). Science overlay maps: a new tool for research policy and library management. *Journal of the American Society for Information Science and Technology, 61*(9), 1871–1887.

Rogers, E. M. (2003). *Diffusion of Innovations*. 5th ed. New York: The Free Press.

Roy, J.M.A. (2005). Nanotechnology: Investing in products of the future (Research report). Retrieved, January 12, 2015, from http://www.ianano.org/Hambrecht-report.pdf.

Sargent, Jr., J.F. (2013). *Nanotechnology: A policy primer*. Washington, DC: Congressional Research Service. Retrieved, April 28, 2014, from http://www.fas.org/sgp/crs/misc/RL34511.pdf.

Ulusal Bilim ve Teknoloji Politikaları 2003-2023 Strateji Belgesi. (2004, November 2). (National Science and Technology Policies 2003-2023 Strategy Document). Version 19. Turkish Scientific and Technological Research Council. Retrieved, December 6, 2014, from http://www.tubitak.gov.tr/tubitak_content_files/vizyon2023/Vizyon2023_Strateji_Belgesi.pdf

Wellman, B. & Berkowitz, S. D. (1997). *Social Structures: A Network Approach*. 2nd ed. Cambridge: University of Cambridge.

The Network Structure of Nanotechnology Research Output of Turkey: A Co-authorship and Co-word Analysis Study[1]

Hamid Darvish[1] and Yaşar Tonta[2]

[1]*darvish@gmail.com*
Kastamonu University, Faculty of Arts & Sciences, Department of Information Management, Kastamonu (Turkey)

[2] *yasartonta@gmail.com*
Hacettepe University, Department of Information Management, 06800 Beytepe, Ankara (Turkey)

Abstract
This paper aims to assess the diffusion of nanotechnology knowledge within the Turkish scientific community using co-citation and co-word analysis techniques. We retrieved a total of 10,062 records of nanotechnology papers authored by Turkish researchers between 2000 and 2011 from Web of Science (WoS) and divided the data set into two 6-year periods. We identified the most prolific and collaborative top 15 universities in each period based on their network properties. We then created co-authorship networks of Turkish nanotechnology researchers in each period and identified the most prolific and collaborative top 15 authors on the basis of network centrality coefficients. Finally, we used co-word analysis to identify the major nanotechnology research fields in Turkey on the basis of the co-occurrence of words in the titles of papers. Findings show that nanotechnology research in Turkey continues to increase due to researchers collaborating with their colleagues. Turkish researchers tend to collaborate within their own groups or universities and the overall connectedness of the network is thus low. Their publication and collaboration patterns conform to Lotka's law. They work mainly on nanotechnology applications in Materials Sciences, Chemistry and Physics, among others. This is commensurate, more or less, with the global trends in nanotechnology research and development.

Conference Topic
Country-level studies, Mapping and visualization

Introduction
Nanotechnology is a relatively new field studying materials at atomic levels within the 1 to 100 nanometer (nm) range (one nm is equal to one billionth of a meter, or, 10^{-9}) (Nanotechnology, 2015). It involves physics, chemistry, medicine, and biotechnology, among others, and promises a great deal of innovation for, and benefit to, society as a whole. Turkey identified nanotechnology early on (2003) as one of the eight strategic fields to support and invested considerably in nanotechnology infrastructure and education. It set up several "centers of excellence" in universities for nanotechnology research and development (R&D). Among them are the Research Center for Nanotechnology and Biotechnology of the Middle East Technical University (METU) and the National Nanotechnology Center in Bilkent University. The former is the first such center established with 15M USD government support while the latter is the first largest multi-purpose nanotechnology center established with 70M USD investment. Universities themselves also invested in nanotechnology. Altogether, there are currently more than 20 nanotechnology research centers in Turkey (Bozkurt, 2015;

[1] This paper is based on the findings of first author's PhD dissertation entitled "Assessing the diffusion of nanotechnology in Turkey: A Social Network Analysis approach." (Darvish, 2014).

Denkbaş, 2015; Özgüz, 2013). The private sector has also invested in nanotechnology in Turkey. Currently, more than 100 companies working in this field and they already developed several nanotechnology products and commoditized them.

In parallel with both government's and private sector's financing of nanotechnology research, several universities initiated multidisciplinary nanotechnology degree programs both at undergraduate and graduate levels (MSc and PhD). The undergraduate and graduate programs of Bilkent University's "Material Science and Nanotechnology", METU's "Micro and Nanotechnology" and Hacettepe University's "Nanotechnology and Nanomedicine" are among them.

The substantial interest and investment in nanotechnology triggered nanotechnology research in Turkey. In fact, Turkey is among the top three countries in the world in terms of the growth rate of nanotechnology research. More than 2,000 researchers are active in this field producing some 2,500 papers in 2014 alone[2] (Bozkurt, 2015, p. 49; Denkbaş, 2015, p. 84; Özgüz, 2013). In this paper, we investigate the development of nanotechnology research in Turkey using bibliometric and Social Network Analysis (SNA) techniques to study the network characteristics of more than 10,000 papers authored by Turkish researchers between 2000 and 2011. We compare the diffusion of nanotechnology research between 2000-2005 and 2006-2011 by measuring the network properties such as degree, betweenness and closeness centrality coefficients of the most prolific and collaborative universities and researchers for each period. We also identify the major nanotechnology research strands in Turkey using co-word analysis.

Literature Review

Information scientists have studied the growth of science and communication using bibliometrics and Social Network Analysis (SNA). While the former deals mainly with the effects of scientific productivity using citation analysis, the latter mainly focuses on the pattern of relationships among scientists. The network composed of co-authorship among scientists is a true indication of their cooperation in research activity.

The "small world" effect is a phenomenon that has been studied by scientists in different fields. This phenomenon conjectures that each member (node) in a society is linked to others (edges) through friends. Literally, every node in a small world is connected through an acquaintance. Newman (2000) found out that average distance from one person to the other by an acquaintance is proportional to the logarithm of the size of the community, implying one of the small world properties. Moreover, he found out that traversing between the two randomly selected nodes of a network takes an average of six steps.

In social contexts, Moody (2004) analyzed the structure of a social science collaboration network over a period. He discovered that collaboration between graduate students in a specific topic creates a small world of scientists and removes restrictions between them. Small world networks may manifest themselves in several shapes and models. Therefore, a good understanding of small world models helps us understand the network characteristics, too. For example, according to Watts (2003) a social network can be categorized as active or passive. Granovetter (1974) studied an active social network from the perspective of finding a job while Burt (1992) looked at such a network as social capital preluding the "rich get richer" phenomenon. In this study, the co-authorship network of structure is represented in a passive sense where the nodes and the edges connecting them are treated as actors and their relationships. Small world models are comprised of clusters or components. Clusters embedded in a network structure reveal a property called "clustering coefficient". According to Watts and Strogatz (1998), one can define a clustering coefficient *C*, which is the average

[2] Search on WoS was carried out on January 11, 2015.

fraction of pairs of neighbors of a node which are also neighbors. That is to say, if node *A* neighbors with node *B* and node *B* is a neighbor of node *C*, then there is a probability that node *A* is also a neighbor of node *C*.

According to Otte and Rousseau (2002, p. 443), betweenness, closeness and degree centrality are well known measures used in analyzing networks. Betweenness centrality is defined as the number of shortest paths going through a node. Thus, a node with high betweenness centrality will have a large impact on the diffusion of knowledge in the network (assuming that knowledge diffusion follows the shortest paths). Centrality is the total number of links that a node has. Degree centrality identifies the most influential node in the diffusion of knowledge in the social network. Closeness measures how far a node is from other nodes in the network structure. Closeness centrality is a measure of how long it will take to diffuse knowledge in a network (Centrality, 2015).

Betweenness centrality plays an important role in the structures of social networks. According to Freeman (2004), the discovery of the structural properties of scientific papers is measured by the betweenness centrality. Actors with a high level of betweenness centrality play a pivotal role in connecting different groups within the network. Betweenness centrality characterizes preferential attachments, cliques, or brokers. Preferential attachments play an important role in network development (Barabasi & Albert, 1999, p. 509). In other words, people in social networks tend to work with well-known people that lead to the concept of "strong and weak ties", characterizing a group of people attached to one node with high centrality. This is called the "star network model" (Moody, 2004; Scott, 2000).

Newman (2000) stated that collaboration among scientists in networks is a good example of showing preferential attachment. As mentioned earlier, if two nodes have high degrees of centrality, the probability of being acquainted with a mutual friend gets higher. Only a small percentage of people in a social network are well connected while the rest are loosely connected (Lotka's law). The productivity of authors in a network resembles Lotka's law in that a small number of researchers publish the majority of papers while large numbers of researchers publish one or two papers (Martin, Ball, Karrer & Newman, 2013). Each group of authors creates a community in which a node with a high degree of centrality is the central node. Therefore, collaboration networks consist of separate clusters representing different scientific fields where they may connect through lower degree connectors. Each community comprises several star networks and these clusters may be connected by a node of lesser degree. Newman (2000) referred to clustering as "community structure".

Co-authorship analysis is used by bibliometricians to track temporal and topological diffusion of scientific publications. Co-authorship stimulates the knowledge diffusion in scientific communities (Chen et al., 2009, p. 192). Thus, co-authorship analysis is used quite often to study the diffusion of innovation and knowledge. For example, Özel (2010) assessed the diffusion of knowledge in business management among academia in Turkey from 1928 to 2010 by studying the co-authorship relationships of academics in business management.

Co-word analysis of texts helps map scientific fields and reveals the cognitive structure of the scientific domain (Chen, 2004). Callon, Courtial, Turner, and Bauin (1983) used the co-word analysis to study the literature over time in terms of the frequencies or co-occurrences of words in titles, abstracts, or more generally, in text. PageRank measuring the popularity of web pages is a similar metric (Page & Brin, 1989). For example, the appearance of a certain author in the references of a corpus of articles reflects the prestige of that author in the network structure.

As we mentioned earlier, the growth rate of nanotechnology research in Turkey is quite encouraging and researchers contribute to the global nanotechnology literature (Kostoff et al., 2006; Kostoff, Koytcheff & Lau, 2007). Although the state of the art of nanotechnology centers and companies has been studied quantitatively (Aydoğan-Duda & Şener, 2010;

Aydoğan-Duda, 2012), their research output in terms of scientific papers has yet to be studied in detail. This is the first such study to investigate the diffusion of nanotechnology in Turkey and the level of collaboration among the most prolific universities and researchers using co-authorship and co-word analysis.

Method

This paper aims to depict the development of nanotechnology in Turkey between 2000 and 2011 by identifying the network structure of nanotechnology papers authored by Turkish researchers and finding out the most productive universities and researchers who help diffuse the nanotechnology knowledge by collaborating with their peers. Social network analysis, co-authorship and co-word analysis tools were used to map the nanotechnology network structure and the collaboration patterns. We attempt to answer the following research questions:

1) Which universities and researchers contribute most to the diffusion of nanotechnology research in Turkey by collaboration?
2) Do co-authorship networks in nanotechnology literature exhibit a "small world" network structure?
3) What are the main nanotechnology research interests of Turkish scholars?

To answer these questions, we retrieved a total of 10,062 records of nanotechnology papers (articles and reviews) from Web of Science (WoS) published between 2000 and 2011 by Turkish authors. We divided the data set into two equal periods (2000-2005 and 2006-2011) to better identify the trends. Almost three quarters of papers (7,398 papers or 73.5%) were published in the second period. Elsewhere, we presented the descriptive statistics for each period on the number of nanotechnology papers published by universities and analyzed the diffusion and adoption of nanotechnology in Turkey by means of the output of the most prolific authors (Darvish & Tonta, 2015). In this paper, we investigate the diffusion of nanotechnology in Turkey by studying the network properties of nanotechnology literature. We first identified the top 15 most prolific universities and authors by means of social network analysis tools. We then identified the scientists with the highest coefficients of centrality in the network structure. We used co-authorship, co-word[3] and factor analyses to track the collaboration patterns and research interests of Turkish nanotechnology scholars between the two periods. We used Bibexcel, VOSviewer, Pajek and Gephi to create files and map the bibliometric data, calculate the properties of the social network structure (e.g., the betweenness, closeness, and degree centralities and the PageRank of each node) and depict the network's features visually.

Findings

Table 1 shows the network properties of the top 15 selected universities in each period (2000-2005 and 2006-2011) ranked by the degree centrality coefficients of their nanotechnology papers. Middle East Technical (METU), Bilkent and Hacettepe Universities are at the pinnacle of the list and they contributed to the network with the highest number of nanotechnology papers. İstanbul Technical (İTU), Erciyes and Kocaeli Universities are at the bottom of the list with the lowest degree centrality coefficients in the 2000-2005 period. Nodes with higher degree centralities participate more in the network than that with the lower ones and the network structure adheres to the small world phenomenon.

[3]The co-word analysis was conducted based on software: http://www.leydesdorff.net/software/fulltext/index.htm

Table 1. Centrality coefficients of nanotechnology papers of the top 15 universities between 2000-2005 and 2006-2011

2000-2005					2006-2011				
University	# of papers	Degree centrality	Closeness centrality	Betweenness centrality	University	# of papers	Degree centrality	Closeness centrality	Betweenness centrality
Middle East Technical	353	0.523	0.467	0.113	Bilkent	356	0.620	0.588	0.069
Bilkent	183	0.515	0.495	0.124	Gebze Institute of Technology	227	0.603	0.541	0.068
Hacettepe	283	0.401	0.495	0.072	Hacettepe	552	0.574	0.524	0.022
Ondokuz Mayis	65	0.357	0.359	0.041	Middle East Technical	646	0.562	0.511	0.054
Dokuz Eylül	108	0.333	0.393	0.109	Istanbul Technical	481	0.534	0.468	0.031
Gebze Institute of Technology	71	0.314	0.499	0.110	Anadolu	224	0.470	0.379	0.042
Kirikkale	36	0.288	0.457	0.119	Gazi	490	0.457	0.373	0.070
Ege	84	0.276	0.359	0.126	Ondokuz Mayis	309	0.450	0.415	0.067
Abant İzzet Baysal	11	0.252	0.612	0.184	Istanbul	245	0.445	0.394	0.045
Gazi	127	0.244	0.373	0.156	Ege	315	0.431	0.382	0.035
Marmara	64	0.225	0.336	0.215	Ankara	348	0.418	0.363	0.071
Ankara	181	0.224	0.373	0.072	Dokuz Eylül	270	0.323	0.429	0.060
Kocaeli	21	0.218	0.325	0.425	Firat	185	0.317	0.452	0.051
Erciyes	58	0.162	0.466	0.098	Erciyes	166	0.256	0.452	0.049
Istanbul Technical	214	0.109	0.363	0.151	Atatürk	219	0.230	0.316	0.091
Avg		**0.296**	**0.425**	**0.141**	**Avg**		**0.446**	**0.439**	**0.055**

The average degree centrality for the top 15 universities rose from 0.296 in the first period to 0.466 in the second period, indicating an almost 60% increase. Istanbul Technical University's degree centrality increased five times between the two periods, making it one of the top nodes in the second period. Kırıkkale, Abant İzzet Baysal, Marmara and Kocaeli Universities with relatively fewer number of papers did not make it to the top 15 universities in the 2006-2011 period and were replaced by Anadolu, İstanbul, Fırat and Atatürk Universities.

Bilkent University is at the top of the 2006-2011 list with the highest closeness centrality coefficient (0.588) followed by Gebze Institute of Technology (0.541) (which was in the 6th place in the first period). Their high closeness centrality coefficients indicate that sub-networks within the whole network are almost 60% connected. However, their betweenness centrality coefficients are relatively low, which means that the flow of information among sub-clsuters within the whole network is slow. Hacettepe and Middle East Technical Universities are also at the top of the 2006-2011 list. These four universities form a cohesive network structure in 2006-2011. However, the average closeness centrality coefficient stayed almost the same for both periods (0.425 and 0.439, respectively). In other words, it took equally long to spread nanotechnology knowledge for the top 15 universities in each period.

In general, betweenness centrality coefficients are much lower for all universities. In fact, the average betweenness centrality has decreased from 0.141 to 0.055 in the second period, indicating that sub-clusters in the network structure became less connected in the second period for the top 15 universities. Atatürk, Ankara, Gazi, Bilkent, Gebze Institute of Technology and Ondokuz Mayıs Universities have the highest betweenness centrality coefficients in the second period, an indication of relatively higher flow of information among sub-clusters within the network than the rest. Dokuz Eylül, Hacettepe and Ankara Universities

have the lowest betweenness centrality coefficients in the first period and Hacettepe, İstanbul Technical and Ege Universities in the second period.

Next, we studied the co-authorship network structures in both periods using social network analysis (SNA) techniques (Fig. 1). SNA enabled us to discern the nodes that might be crucial to the diffusion of nanotechnology knowledge. The network consists of 470 nodes and 1,042 edges in 2000-2005 and 945 nodes and 4,915 edges in 2006-2011. The rates of growth for nodes and edges (ties) increased two- and four-folds, respectively, between the two periods. However, the level of collaboration has not changed so much. There is a minimal change in density (from 0.009 to 0.011) between the two periods, but the network is still quite sparse. Nonetheless, the average degree and clustering coefficients show that clusters within the network are somehow connected for both periods. For example, the average clustering coefficient for 2000-2005 is 0.75, indicating that 75% of the nodes were connected. Since the network has grown in the second period, the rate of connectedness has decreased (0.51), indicating that newly formed clusters were not that cohesive yet.

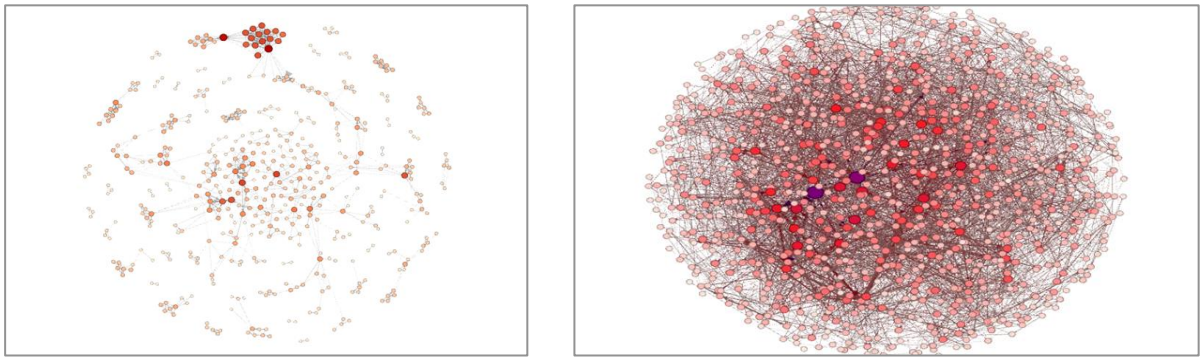


Figure 1. Co-authorship network of scientists working on nanotechnology between: (1) 2005-2011and (r) 2006-2011

The network in the second period adheres to the transitivity relations, indicating that the network at meso level is well connected even though the sub-clusters are not that well connected (especially in the periphery of the network) (Fig. 1). That is to say that there has been some progress in terms of creating new sub-clusters in the co-authorship network, although links among sub-clusters have yet to be formed. In other words, almost all scientists have co-authored with one or more authors in their own cluster but not beyond.

Table 2 shows the top 15 Turkish authors and their affiliations with the highest centrality coefficients (closeness, betweenness, degree, and PageRank) between 2000 and 2005 who contributed to the diffusion of nanotechnology with their scientific papers. Some scientists appear in more than one columns of centrality due to their high collaboration level in the network structure. For example, Yakuphanoğlu F (Fırat University), Yağcı Y and Öveçoğlu MN (İTU), Çelik E (Dokuz Eylül) and Denizli A (Hacettepe) appeared in three columns with high degree (collaborator), betweenness (broker and gatekeeper), and PageRank coefficients (prolific author) while Yılmaz F and Toppare L (METU), Morkoç H (Atatürk), Özdemir I (Dokuz Eylül) and Pişkin E (Hacettepe) appeared at least in two columns out of four (degree, betweenness, closeness and PageRank centralities). They were highly influential in the diffusion of nanotechnology in Turkey between 2000 and 2005.

Table 2. Network properties of the top 15 Turkish authors based on co-authorship degree centralities: 2000-2005.

Rank	Degree centrality	Betweenness centrality	Closeness centrality	PageRank
1	Balkan N (Fatih)	Yilmaz F (METU)	Sarı H (Bilkent)	Ovecoğlu MN (ITU)
2	Teke A (Balıkesir)	Gencer A (Hacettepe)	Sökmen I (Dokuz Eylül)	Çelik E (Dokuz Eylül)
3	Yağci Y (ITU)	Koralay H (Firat)	Kasapoğlu E (Cumhuriyet)	Denizli A (Hacettepe)
4	Yakuphanoğlu F (Firat)	Okur S (Izmir Inst Tech)	Çiraci S (Bilkent)	Hasçiçek YS (Gazi)
5	Ovecoğlu MN (ITU)	Denizli A (Hacettepe)	Aytor O (Bilkent)	Yağci Y (ITU)
6	Çelik E (Dokuz Eylül)	Yavuz H (Hacettepe)	Biyikli N (METU)	Yakuphanoğlu F(Firat)
7	Yilmaz F (METU)	Güneş M (Kirikkale)	Özbay E (Bilkent)	Toppare L (METU)
8	Toppare L (METU)	Yakuphanoğlu F (Firat)	Doğan S (Bilkent)	Yilmaz VT (Ondokuz Mayıs)
9	Doğan S (Bilkent)	Balkan N (Fatih)	Morkoç H (Atatürk)	Pişkin E (Hacettepe)
10	Morkoç H (Atatürk)	Çelik E (Dokuz Eylül)	Sari B (Gazi)	Erkoç Ş (METU)
11	Denizli A (Hacettepe)	Pişkin E (Hacettepe)	Talu M (Gazi)	Kurt A (Koç)
12	Erol A (Istanbul)	Güven K (Erciyes)	Kartaloğlu (Bilkent)	Elmali A (Ankara)
13	Özdemir I (Dokuz Eylül)	Yağci Y (ITU)	Yilgor E (Koç)	Hincal AA (Hacettepe)
14	Turan R (METU)	Ovecoğlu MN (ITU)	Yilgor I (Koç)	Ozdemir I (Dokuz Eylül)
15	Dag O (Bilkent)	Menceloğlu YZ (Sabancı)	Andaç O (Ondokuz Mayıs)	Oral A (Sabancı)

Co-authorship map of the first authors for the first period is shown on the left-hand side of Figure 2. Most of the authors listed in Table 2 are also on the map. Although most authors were from universities with high degree centralities, other authors whose universities did not have high degree centralities were also instrumental in the diffusion of nanotechnology knowledge in the network during the 2000-2005 period (e.g., Yilgor E and Yilgor I from Koç, Koralay H and Yakuphanoğlu E from Fırat, and Kasapoğlu E from Cumhuriyet Universities).

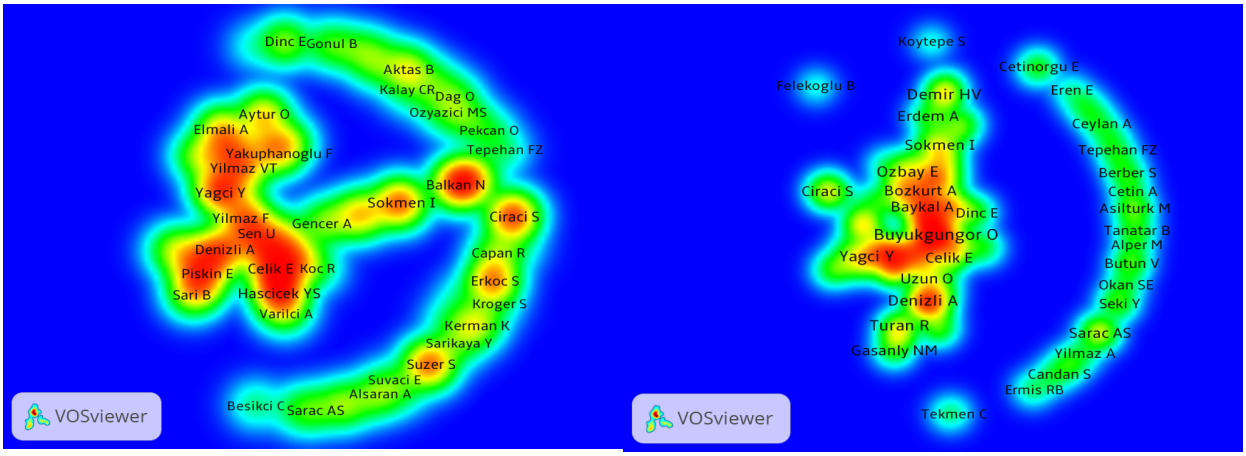


Figure 2. Co-authorship map of Turkish nanotechnology scientists between: (l) 2000-2005 and (r) 2006-2011.

Table 3 shows the top 15 authors who were influential in the diffusion of nanotechnology in Turkey between 2006 and 2011. Interestingly, Büyükgüngör O of Ondokuz Mayıs University has the highest centrality coefficients in all four categories but one (the betweenness centrality) even though he was not in the top 15 authors in the first period. His name appears in the center of the 2006-2011 network of Figure 2 as a prestigious researcher playing an important role in the dissemination of nanotechnology knowledge in the network structure. (His research field is Crystallography.) Similarly, Özçelik S of Gazi University is at the top 15 in all four categories. Six authors appear in at least three columns: Denizli A (Hacettepe), Şahin E (Gazi), Yağcı Y (İTU) and Toppare L (METU) in degree, betweenness and PageRank columns, and Özbay E and Çıracı S (Bilkent) in degree, closeness and PageRank columns. An

additional six authors appear in at least two columns: Yeşilel ÖZ (Osmangazi) and Baykal A (Fatih) in closeness and PageRank columns; Yıldız A (Fatih) and Yılmaz F (METU) in degree and betweenness columns; Çakmak M (Koç) in betweenness and PageRank columns; and Turan R (Ege) in degree and PageRank columns.4 It should be pointed out that even though Fatih and Karadeniz Technical Universities failed to have the highest degree centrality coefficients in neither period, some of their scientists (e.g., Yildiz A and Bacaksız E, respectively) played an important role nonetheless in the diffusion of nanotechnology knowledge in the network.

The centrality coefficients of four authors were high in both periods: Yağcı Y (İTU), Denizli A (Hacettepe), and Toppare L and Yılmaz F (METU). They were highly active in spreading the nanotechnology knowledge in Turkey between 2000 and 2011 as prolific authors, collaborators, brokers and gatekeepers, and diffusers.

Table 3. Network properties of the top 15 authors based on co-authorship degree centralities: 2006-2011.

Rank	Degree centrality	Betweenness centrality	Closeness centrality	Page Rank
1	Büyükgüngör O (Ondokuz Mayis)	Yilmaz F (METU)	Büyükgüngör O (Ondokuz Mayis)	Büyükgüngör O (Ondokuz Mayis)
2	Şahin E (Gazi)	Büyükgüngör O (Ondokuz Mayis)	Yeşilel ÖZ (Osmangazi)	Özbay E (Bilkent)
3	Toppare L (METU)	Özçelik S (Gazi)	Demir HV (Bilkent)	Özçelik S (Gazi)
4	Yilmaz F (METU)	Toppare L (METU)	Nizamoğlu S (Bilkent)	Toppare L (METU)
5	Özçelik S (Gazi)	Yağcı Y (ITU)	Çağlar Y (Anadolu)	Denizli A (Hacettepe)
6	Yağci Y(ITU)	Şahin E (Gazi)	İlican S (Anadolu)	Turan R (Ege)
7	Özbay E (Bilkent)	Yildiz A (Fatih)	Çağlar M (Anadolu)	Şahin E (Gazi)
8	Turan R (Ege)	Çakmak M (Koç)	Özbay (Bilkent)	Çıracı S (Bilkent)
9	Çakmak M (Kirikkale)	Şahin O (Dokuz Eylül)	Özçelik S (Gazi)	Yeşilel ÖZ (Osmangazi)
10	Yerli A (Sakarya)	Yilmaz M (Istanbul)	Baykal A (Fatih)	Yağci Y (ITU)
11	Yildiz A(Fatih)	Turan R (METU)	Köseoğlu Y(Fatih)	Sökmen I (Dokuz Eylül)
12	Çetin K (Ege)	Bacaksiz E (Karadeniz Technical)	Toprak MS (Fatih)	Arslan H (Hacettepe)
13	Çiraci S (Bilkent)	Denizli A (Hacettepe)	Çiraci S (Bilkent)	Oskar S (METU)
14	Denizli A (Hacettepe)	Şen S (Yalova)	Durgun E (Bilkent)	Çakmak M (Koç)
15	Sari H (ITU)	Balkan A (Fatih)	Akgol S (Adnan Menderes)	Baykal A (Fatih)

The collaboration network of Turkish scientists who work on nanotechnology seems to be well connected at the micro level but not so much at the macro level. In other words, researchers tend to collaborate within their own sub-clusters (i.e., groups or universities) more often. The frequencies of the total number of publications that first authors contributed to adhere to Lotka's law:

$$f(y) = .2459 \div y^{1.2881} \quad (1)$$

where *f(y)* denotes the relative number of authors with *y* publications (the K-S DMAX = 0.6323) (Rousseau, 1997), indicating that a small number of well-known scientists have stronger positions in the network. As mentioned earlier, although some scientists from smaller universities with the lower degree centrality coefficients have appeared in the network structure as a turning point, one can call them as non-elite authors. However, their impact on knowledge diffusion is remarkable.

[4] Note that some author names with the same initials are affiliated with two different universities in this period (e.g., Çakmak M at both Koç and Kırıkkale Universities and Turan R at both Ege and Middle East Technical Universities). They may well be the same authors who may have moved from one university to the other during this period.

We also carried out a co-word analysis on the words that appear in the titles of articles extracted from WoS to find out the most frequently used terms between 2000 and 2005, and between 2006 and 2010. The first 75 most frequently occurring words in each period were collected, processed and compiled by the software.5 Non-trivial words were eliminated. In order to analyse the word/document occurrence matrix in terms of its latent structure, SPSS software version 16.0 was used to factor analyse the co-occurrence of words. Factor analysis maps each word to a different component (research strand) with the highest factor loading. SPSS created two factors from the list of the co-words. Table 4 and 5 show the output of factors for the periods of 2000-2005 and 2006-2011 along with the loadings of different words in each factor (not all 75 words listed in the tables). According to eigenvalues, the first factor explains 56% of the variance in the entire data set for the period of 2000-2005 while the second one explains the rest of the variance (44%). For the 2006-2011 period, the first factor explains 35% of the variance in the entire data set while the second and third ones explain 33% and 32% of the variance, respectively.

Table 4. Factor analysis of co-words in the titles of nanotechnology papers (2000 and 2005).

Rotated component matrix[a]

Words	Factor 1	Words	Factor 2
CHEMICAL	.999	PLASMA	.999
QUANTUM	.999	TREATMENT	.999
STEEL	.998	CONDUCTING	.990
HYDROGEN	.997	CERAMIC	.982
COPOLYMER	.992	SOL-GEL	.982
FIELD	.992	LAYER	.945
PROPERTIES	.984	OPTICAL	.945
ELECTRICAL	.973	SURFACE	.945

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 8 iterations.

We then produced a normalized cosine extraction of the words and mapped the network structure of co-word analysis in each period using Kamada & Kawai algorithm embedded in Pajek (Fig. 3). Words that appear in both periods belong mainly to Multidisciplinary Science and Materials Science. Represented fields in both periods are as follows: Surface Materials ("Doped", "Alloy", and "Plasma"); Chemistry and its subfields ("Coating", "Crystal" "Catalyst", and "Sol-Gel"); and Physics ("Quantum", "Dot" and "Nanotube"). It appears that Turkish nanoscientists work primarily in Material Sciences, followed by Physics and, to some extent, Biotechnology.

[5] We used the software available at http://www.leydesdorff.net/software/fulltext/index.htm to create a normalized cosine symmetric co-occurrence matrix of labels.

Table 5. Factor analysis of co-words in titles of nanotechnology papers (2006 and 2011).

Rotated component matrix[a]

Words	Factor 1	Words	Factor 2	Words	Factor 3
COPOLYMER	.766	STEEL	.673	DOT	.687
COMPLEXES	.697	WELL	.655	MORPHOLOGY	.676
CRYSTAL	.674	AQUEOU	.651	ADSORPTION	.654
THERMAL	.653	ZNO	.642	ENERGY	.644
SPECTROSCOPIC	.650	PARTICLE	.626	PREPARED	.641
CHARACTERISTIC	.643	MATERIAL	.625	QUANTUM	.620
COPOLYMER	.766	TEMPERATURE	.620	ELECTRICAL	.619
METAL	.636	CELL	.618	MODIFIED	.610

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Discussion and Conclusion

In this paper, we assessed the network structure of nanotechnology papers authored by Turkish scientists between 2000 and 2011. We used the social network analysis techniques and studied the network properties from different perspectives. We first identified the top 15 universities for each period (2000-2005 and 2006-2011) on the basis of centrality coefficients. They played pivotal roles in the dissemination of nanotechnology knowledge in Turkey. We then created the co-authorship network of nanotechnology scientists and analyzed the network properties (coefficients of degree, betweenness, closeness centralities and PageRank) of the top 15 authors in each period. We also used the co-word analysis to identify the major nanotechnology research fields in Turkey on the basis of the co-occurrence of words in the titles of papers.

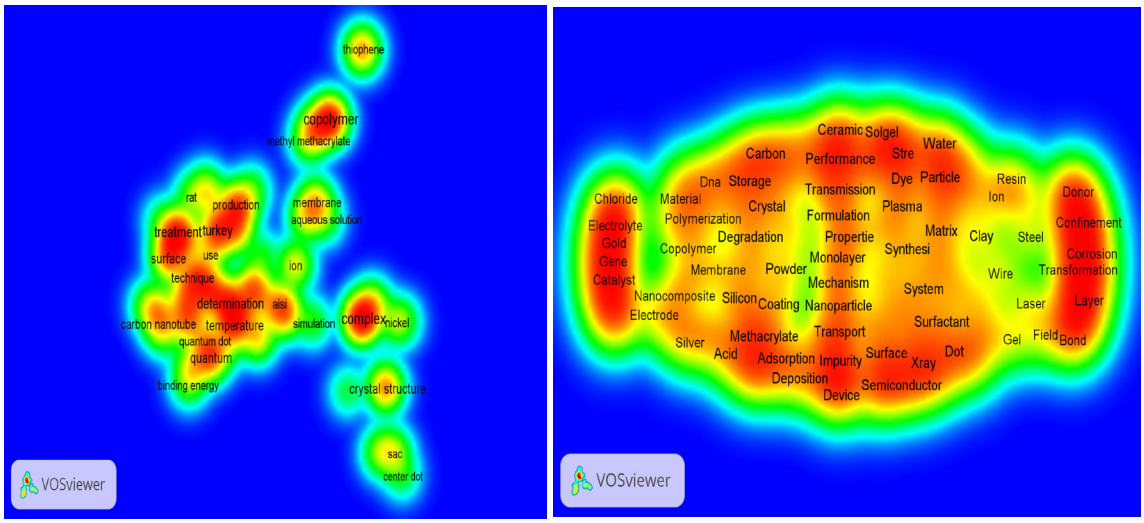


Figure 3. Network of co-word analysis in nanotechnology in Turkey: (l) 2000-2005 and (r) 2006-2011.

Although the number of nodes in the network has increased in the second period (2006-2011), the overall connectedness of the network structures is low. The centrality coefficients of the network structure of the top 15 universities revealed that the social network structure is denser

at the micro level than that at the macro level. While the betweenness centrality remained low and the closeness centrality did not change much, the degree centrality increased almost 60% in the second period, which is an indication of the small world phenomenon in the network structure.

The research output of Turkish nanoscientists and collaboration among them conform to some extent to Lotka's law in that a few researchers tend to publish the bulk of nanotechnology papers while the rest are less prolific. This indicates that Turkish scientists tend to work with prolific authors. The taxonomy identified by the co-word analysis shows that Turkish nanoscientists mainly work in Materials Sciences, Chemistry and Physics. Nanotechnology research continues to flourish due to collaborations at the micro level within the Turkish scientific community and the diffusion of nanotechnology knowledge is accelerating. Bibliometric indicators and network properties reported in this research may help policy-makers to understand the interdisciplinary character of nanoscience and nanotechnology better and develop funding mechanisms accordingly.

References

Aydogan-Duda, N. (2012). Nanotechnology: A descriptive account. Making it to the forefront in Aydogan-Duda, N. (Ed). *Nanotechnology: A Developing Country Perspective*. 1, (pp. 1-4). New York: Springer.

Aydogan-Duda, N., & Şener, I. (2010). Entry barriers to the nanotechnology industry in Turkey in Ekekwe, N. (Ed). *Nanotechnology and Microelectronics: Global Diffusion, Economics and Policy*. (pp. 167-173). Hershey, PA: IGI Global.

Barabasi, AL. & Albert, R. (1999). Emergence of scaling in random networks. *Science Magazine*, *286*(5439), 509-512.

Bozkurt, A. (2015, January). Türkiye, 10 yıldır "en küçük" dünyanın farkında, artık büyük adımlar atması gerekiyor (Turkey is aware of the "smallest" world for 10 years, but it should take big steps). *Bilişim: Aylık Bilişim Kültürü Dergisi*, 43(172), 44-53. Retrieved June 6, 2015 from: http://www.bilisimdergisi.org/s172/pages/s172_web.pdf.

Burt, R.S. (1992). *Structural Holes: The Social Structure of Competition*, Cambridge, MA: Harvard *University Press.*

Callon, M., Courtial, J.P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191-235.

Centrality. (2015). Retrieved, January 20, 2015, from http://en.wikipedia.org/wiki/Centrality.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. *PNAS*, *101*(Suppl. 1), 5303-5310.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, *3*(3), 191-209.

Darvish, H. (2014). *Assessing the diffusion of nanotechnology in Turkey: A social network analysis approach*. Unpublished PhD dissertation. Hacettepe University, Ankara.

Darvish, H. & Tonta, Y. (2015). The diffusion of nanotechnology knowledge in Turkey (submitted).

Denkbaş, E.B. (2015, January). Nanoteknolojiye yapılacak yatırımlar, ülkelerin ekonomik gücünü yansıtabilecek bir parametre olacak (Investments in nanotechnology will become a parameter reflecting economic powers of countries). *Bilişim: Aylık Bilişim Kültürü Dergisi*, 43(172), 78-87. Retrieved June 6, 2015 from: www.bilisimdergisi.org/pdfindir/s172/pdf/78-87.pdf.

Freeman, L.C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. Vancouver: Empirical Press.

Granovetter, M. (1974). *Getting a Job: A Study of Contacts and Careers.* Cambridge, Mass: Harvard University.

Kostoff, R.N., Koytcheff, R.G., & Lau, C.G.Y. (2007). Global nanotechnology research literature overview, *Technological Forecasting & Social Change, 74*, 1733-1747.

Kostoff, R.N., Stump, J. A., Johnson, D., Murday, J.S., Lau, C.G.Y., & Tolls, W.M. (2006). The structure and infrastructure of global nanotechnology literature. *Journal of Nanoparticles Research*, *8*, 301-321.

Martin, T., Ball, B., Karrer, B., Newman M. E. J. (2013). *Coauthorship and citation in scientific publishing*. Retrieved December 27, 2014 from http://arxiv.org/abs/1304.0473.

Moody, J. (2004*)*. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review, 69*(2), 213-238.

Nanotechnology. (2015). Retrieved, January 20, 2015, from http://en.wikipedia.org/wiki/Centrality

Newman, M. E. J. (2000). The structure of scientific collaboration networks. *PNAS*, *98*(2), 404–409.

Otte, E., & Rousseau, R. (2002). Social Network Analysis: A powerful strategy, also for the information sciences. *Journal of information Science*, *28*(6), 443–455.

Özel, B. (2010). *Scientific collaboration networks: Knowledge diffusion and fragmentation in Turkish management academia*, Unpublished PhD dissertation, Bilgi University, Istanbul.

Özgüz, V. (2013). Nanotechnology research and education in Turkey (presentation slides). Retrieved, December 27, 2014, from: http://rp7.ffg.at/upload/medialibrary/12_Oezguez.pdf

Page, L., & Brin, S. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*, 107-117.

Rousseau, R. (1997). *Sitations: an exploratory study*. *Cybermetrics*. Retrieved, February 14, 2014, from http://cybermetrics.cindoc.csic.es/articles/v4i1p4.pdf.

Scott, J. (2000*). Social Network Analysis: A Handbook.* 2nd ed. London: Sage.

Watts, D. (2003). *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton & Company.

Watts, D. J. & Strogatz, S. (1998). Collective Dynamics of "small-world" Networks. *Nature*, *393*(6684), 440-441.

Analysis of the Spatial Dynamics of Intra- v.s. Inter-Research Collaborations across Countries[1]

Lili Wang[1] and Mario Coccia[2]

[1] *wang@merit.unu.edu*
UNU-MERIT, Keizer Karelplein 19, 6211 TC, Maastricht (The Netherlands)

[2] *mario.coccia@ircres.cnr.it*
CNR -- National Research Council of Italy, 10024 Moncalieri, Torino (Italy)

Abstract
The purpose of this paper is to analyse the evolutionary pattern of international research collaborations. Using publication data from 1997 to 2012, this study decomposes international collaborations into two complementary types, intra-collaboration (within the same geographical area) and inter-collaboration (across different geographical areas). Our results show that the geographical concentration of international research collaborations is reducing. The formation of new network structure of international research collaborations is driven by the increase of inter-research collaborations of countries across different geographical areas rather than intra-collaborations of countries within the same geographical area.

Conference Topic
International collaboration

Introduction
Scientific collaborations have been widely acknowledged to be efficient in managing time and labour in research labs (Coccia, 2014; Solla Price & Beaver, 1966), improving research quality (Presser, 1980; Narin et al., 1991; Katz & Hicks, 1997) and spurring the breakthroughs of scientific research for supporting competitiveness (Coccia, 2012). A number of factors have contributed to the continuous increase of international research collaborations and co-authored papers (Beaver & Rosen, 1978; Frame & Carpenter, 1979; Katz & Martin, 1997). Along with the steady rise of international scientific collaborations, a better understanding on the structure of the global research network across geo-economic areas and its evolutionary pattern are needed for scholars and policy makers.

The high heterogeneity across countries – in terms of size, scientific capacity of the national system of innovation, etc. – generates a variety of patterns of the international research collaborations (Melin, 1999; Narin et al., 1991; Ozcan & Islam, 2014). A main issue in economics of science is to determine how and to which extent countries are engaged in international research collaborations so as to understand the behaviour of knowledge flows and to design research policies for improving the scientific research production which will in turn to enhance national competitiveness.

Luukkonen et al. (1992) maintain that the map of collaborative connections between countries corresponds to a geographical map. Frame et al. (1977, p. 502), considering data of 1973, claim that: "the production of mainstream science is more heavily concentrated in the hands of a few countries". Hoekman et al. (2010), using data on co-publications in European countries, show that research collaborations are geographically localized and despite a research heterogeneity in European countries in terms of research collaboration patterns, there

[1] Mario Coccia gratefully acknowledges financial support from United Nations University -The Maastricht Economic and Social Research Institute on Innovation and Technology (Contract ID 606U U-04 76) where this joint research was conducted while he was a visiting researcher.

is “a gradual convergence is taking place toward a more integrated interconnected European science system” (Hoekman et al., 2010, p. 672).

The purpose of this research is to investigate the evolutionary pattern of international research collaborations across countries. Emphasis is placed on two complementary collaboration types, i.e. intra- and inter-collaborations. The former refers to research collaborations conducted by countries within the same geographical area; the latter refers to research collaborations engaged by countries from different geographical areas.[2] Increase of intra-collaborations indicates that cooperation is more and more bounded within certain geographical territories, while increase of inter-collaborations signals the fade of geographical limit.

The main research questions of this paper are:

- How does the distribution of international collaborations across countries evolve over time?
- What type of research collaborations (inter- or intra-) plays a more important role in reshaping the global collaborative scientific network across geo-economic areas?
- How do inter- and intra- connections change in the global collective network?

The analysis of the temporal and spatial evolution of these patterns is of great scientific interest for researchers and policy makers in order to better master knowledge flow and optimize collaborative research output across countries.

Data and methodology

The data of this study are collected from publications in academic journals covered by the Science Citation Index (SCI) and Social Sciences Citation Index (SSCI). In particular, this study refers to dataset by National Science Foundation (2014)-National Center for Science and Engineering Statistics, special tabulations from Thomson Reuters (2013), SCI and SSCI. Collaboration data cover two years 1997 and 2012 and 40 countries (see the list in Appendix A). These 40 countries produce about 97% of the global total articles over 1997-2012. The 40 countries are classified into eight geographical areas: North America, South America, Europe Union, Other Europe, Middle East, Africa, Asia and Australia/Oceania (see Appendix A).

The analysis consists of the following steps:

- Firstly, to analyse the worldwide distribution of international collaborations, this study uses Lorenz curves and Gini coefficient. Lorenz curve is indicated by $L(X)$, then Gini coefficient can be derived as follows:

$$Gini\ coefficient\ (G) = 1 - 2\int_0^1 L(X)dX \quad (1)$$

  G is main indicator of concentration of the distribution of data.
- Secondly, to map the research connections between countries, both absolute collaborative output (number of articles) and collaboration intensity are considered. The former data set demonstrates the major players in the global collaboration research network while the latter puts all countries into one comparable framework. Although the matrix of co-authored papers between countries provides us main information concerning the output co-occurrence, the number of collaborated output might have different meanings for the collaborating country pair due to their different research capacity. For instance, suppose that a research collaborative pair is formed by Country A (of which the number of total publications is 1000) and Country B (of which the number of total publications in 10,000). Collaboration intensity (the ratio of collaborative output to national total publications) presents a stronger collaboration

[2] The under studied geographical areas are: North America, South America, Europe Union, Other Europe, Middle East, Africa, Asia and Australia/Oceania.

link for country A than B. Therefore, extra caution should be exercised when analysing the collaborative connections between research partners.

Based on eight geographical groups, this study disentangles intra-collaborations (between countries located in the same geographical area) from inter-collaborations (between countries of different geographical areas).[3]

Salton and Jaccard indexes are both valuable in measuring relative collaboration intensity (cf. Luukkonen et al., 1993). The collaboration index by Salton's measure (CSI) is

$$CSI = \frac{CO_{ij}}{\sqrt{P_i * P_j}} \quad (2)$$

whereas, the Jaccard's measure (CJI) is given by:

$$CJI = \frac{CO_{ij}}{P_i + P_j - CO_{ij}} \quad (3)$$

Where $\boldsymbol{CO_{ij}}$ is the number of co-authored papers between country *i* and country *j*

$\boldsymbol{P_i}$ is the total publication number by country *i*

$\boldsymbol{P_j}$ is the total publication number by country *j*

In addition, to understand the intra- and inter- collaborations by Salton and Jaccard indices (equations (2) and (3)), the adapted intra- and inter- collaboration intensities are

- $\mathrm{CSI}_{\mathrm{intra}} = \frac{\mathrm{CO_{ij}}}{\sqrt{\mathrm{P_i * P_j}}}$ (*i* & *j* ∈ same geographical area) (4)
- $CSI_{inter} = \frac{CO_{ij}}{\sqrt{P_i * P_j}}$ (*i* & *j* ∈ different geographical areas) (5)
- $CJI_{intra} = \frac{CO_{ij}}{P_i + P_j - CO_{ij}}$ (*i* & *j* ∈ same geographical area) (6)
- $CJI_{inter} = \frac{CO_{ij}}{P_i + P_j - CO_{ij}}$ (*i* & *j* ∈ different geographical areas) (7)

Coefficient of variation is also applied to assess the dispersion of data.

- Thirdly, from a dynamic perspective, this study applies network analysis to explore the structure of international collaborations and its changes from 1997 to 2012. In particular, intra- and inter- scientific ties across countries are distinguished from each other in the networks.

Empirical analysis

Global distribution of scientific research and collaborations

It has been well recognized that research capability and resources are unevenly distributed in the world, and hence scientific research output is concentrated in certain countries which are scientifically strong (Frame et al., 1977). By measuring the statistical dispersion of total publications and international collaborations, Table 1 shows that the Gini coefficient of internationally co-authored papers is lower than that of total publications, which means the former is distributed more evenly across countries than the latter. Most importantly, the Gini coefficients for both types of scientific outputs are decreasing over years. This means that the distributions of total publications and internationally co-authored papers both became less geographically concentrated in the later years.

[3] Refer to Appendix A for detailed group information.

Table 1. Gini Coefficient over years

	1997	*2002*	*2007*	*2012*
Total publications	0.67	0.63	0.61	0.59
Internationally co-authored papers	0.60	0.58	0.56	0.54

Dynamics of international collaborations

Salton and Jaccard measures are considered for estimating the collaboration intensity (Figure B1 and B2, in the Appendix B). The arithmetic mean of Salton measure is as twice as that of Jaccard measure, which is in line with Hamers, et al. (1989). However, the coefficient of variation in Jaccard is somewhat higher than that of Salton (see Fig. B1 and B2), indicating a greater dispersion of collaboration intensities is measured by Jaccard index. As the aim of this study is to analyse collaborative research variability between countries, intensities derived from Jaccard index seem to be more suitable.[4]

At the level of geographical groups, Figure 1 shows the relationship of the intra- and inter-collaboration intensities between 1997 and 2012. Red dots represent the inter-collaboration intensity and green ones represent intra-collaboration intensities. A dot being above diagonal line indicates that the collaboration intensity of this observed unit has increased in 2012 in contrast to that of 1997. Likewise, a dot underneath the diagonal indicates that the international collaboration intensity has decreased in 2012 compared to that of 1997. The fact that all the dots lying above the diagonal line suggests that both intra- and inter- collaboration intensities in all geographical areas have improved over years. On the other hand, by comparing the red and green dots, it is of great interest to observe that inter-collaborations in all geographical areas have increased dramatically while intra-collaborations stay mostly low and close to the diagonal line. The intra-collaboration intensity in the European Union (EU) is the only exception with high level of intra-collaborations in both 1997 and 2012, which is a phenomenon of "Europeanisation" as discussed by Mattsson et al. (2008). In general, this figure shows that intra-collaborations tend to be static while inter-collaborations exhibit high dynamics of growth.

[4] In the rest of the paper, we present only results calculated based on Jaccard measure. Similar results using Salton measure are available upon request.

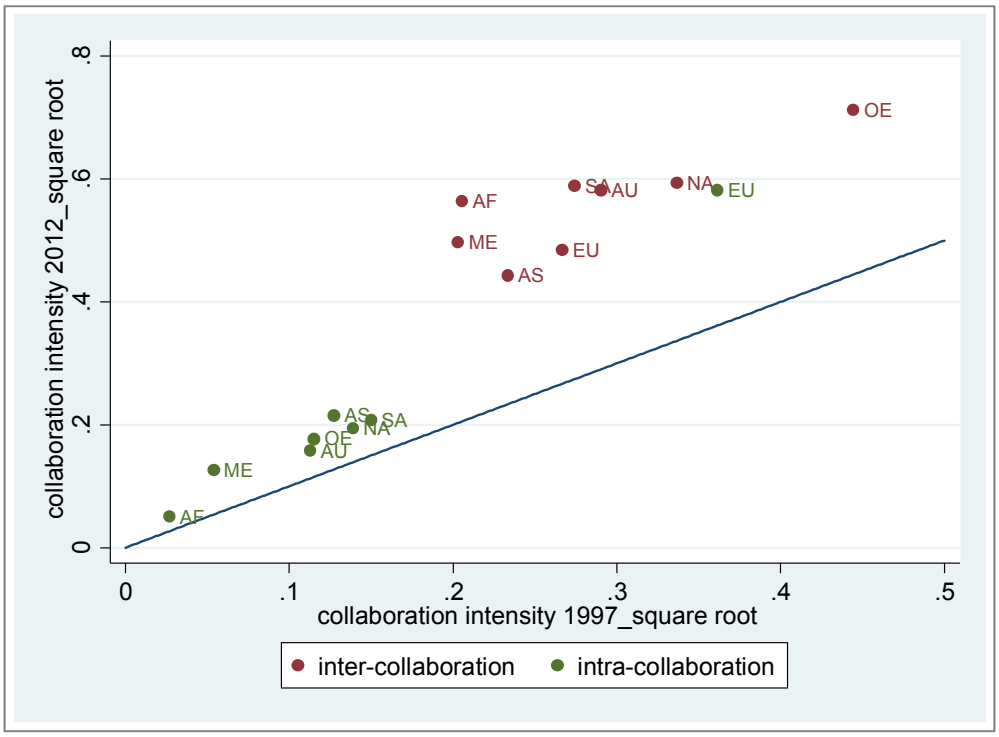


Figure 1. Comparison of international collaboration intensity (inter vs. intra)

Note: 1) The eight geographical areas are: North America (NA), South America (SA), European Union (EU), Other Europe (OE), Middle East (ME), Africa (AF), Asia (AS) and Australia/Oceania (AU). 2) Collaboration intensity is measured by Jaccard index.

To further understand the changes of collaborative performance in individual countries, Figure 2 presents the intra- and inter-collaboration intensity in the 40 under studied countries. Countries in European Union are the only ones showing growth of both intra- and inter-collaborations. This can be the result of European Commission's policy which stimulates cooperation between European countries. In the rest countries, the intra-collaboration performance looks all static, while inter-collaborations have risen obviously. Among all the countries, a group of Asian countries (China, India, Japan, Singapore, and South Korea) show relatively slow growth in inter-collaborations.

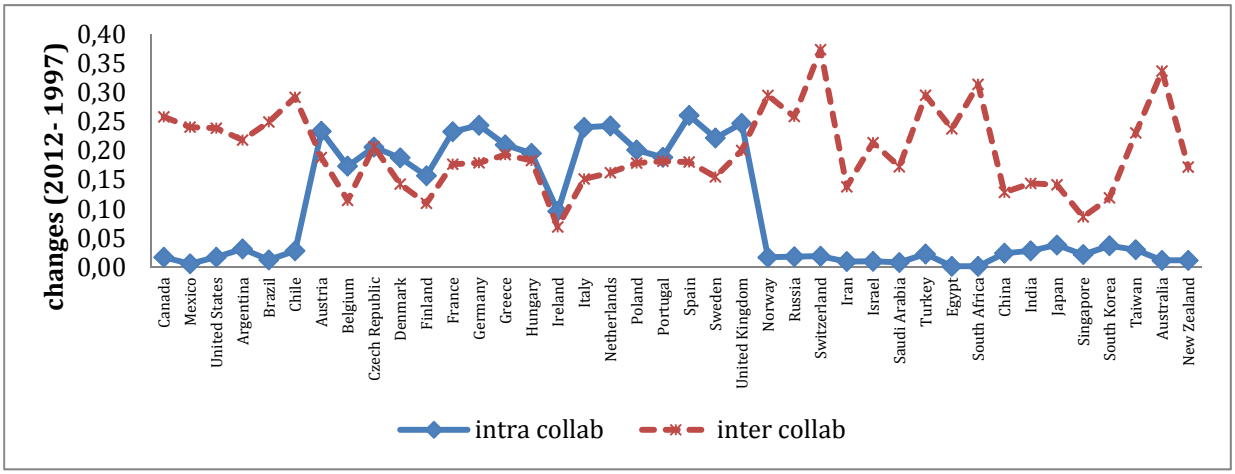


Figure 2. Changes of international collaboration intensity by country (inter vs. intra)

Note: 1) Collaboration intensity is measured by Jaccard index. 2) The value of y-axis is calculated by the collaboration intensity in 2012 minus that in 1997.

Networks of research collaborations

Based on Jaccard collaboration intensity, collaborative networks across 40 countries in 1997 and 2012 are provided in Figure 3 and 4. The thickness of each edge between two nodes reflects the strength of their collaborative relationship. The higher collaboration intensity one country pair has, the thicker their connection line is. In order to distinguish between intra- and inter-collaborations, geographical areas are presented in different colours.[5] Lines connecting nodes in different colours represent inter-collaborations, while those between nodes in same colours represent intra-collaborations. The size of each node embodies its aggregated collaboration intensity (including both intra- and inter- collaborations).

Figure 3 shows that scientific collaboration networks have been, to some degree, formed by geographic ties. Apart from the intensive connections between European countries (intra-collaborations), there are a few geographically biased small clusters are of great interest. The rectangular cluster in Nordic countries (formed by Denmark, Sweden, Norway and Finland) and the triangular cluster in South America (formed by Chile, Brazil and Argentina) both indicate that scientific collaborations are geographically localized. Besides these small clusters, in North America, a strong tie is observed between United States and Canada. In Asia, China is mainly connected with Japan. In Australia/Oceania, New Zealand has a strong connection only with Australia.

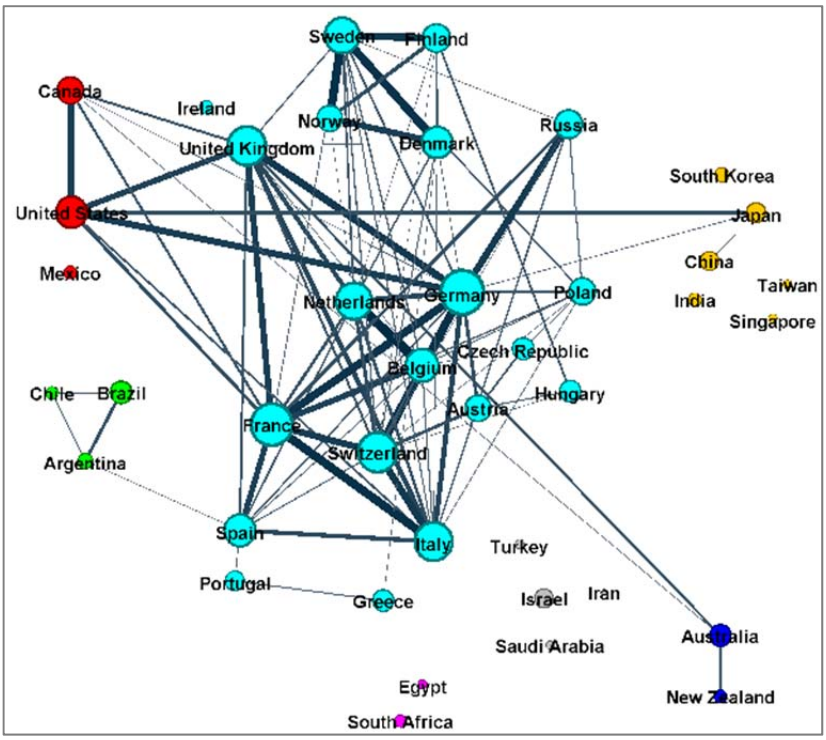


Figure 3. Network of global research connections in 1997.

Note: 1) A filter of 0.0083 is applied in this figure, which means that edges with collaboration intensity less than 0.0083 are omitted. 2) The thickness of each edge between two nodes reflects the strength of their collaborative relationship. 3) The size of each node embodies its aggregated collaboration intensity.

[5] To emphasize the effect of geographical locations, *European Union* and *Other Europe* are regarded as one group in the network figures (Fig. 3 and 4).

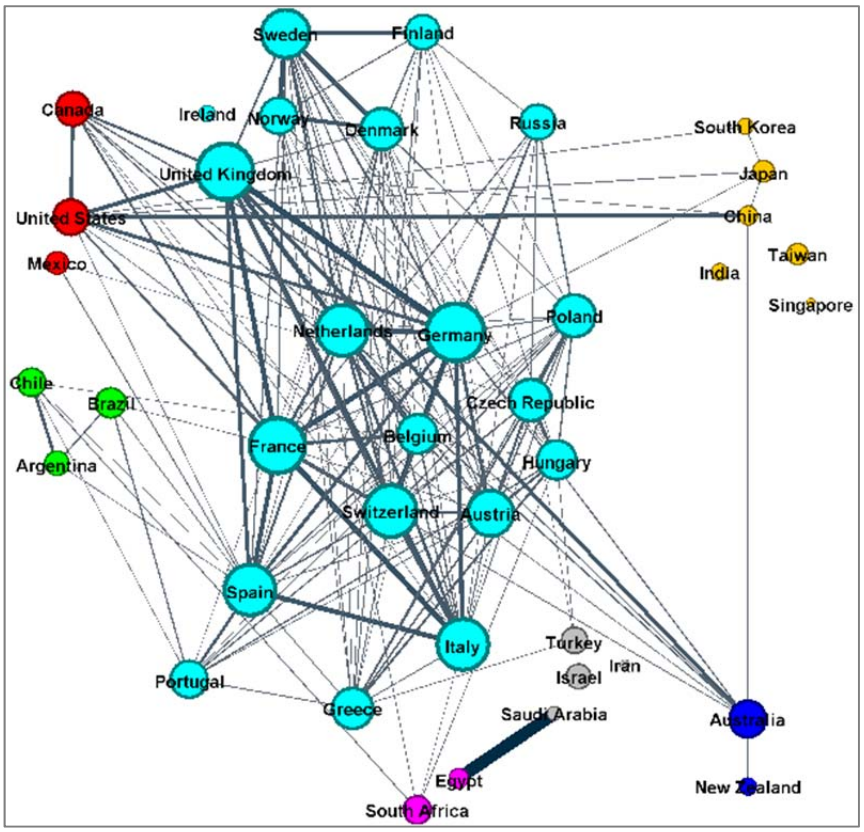


Figure 4. Network of global research connections in 2012.

Note: 1) The network in 2012 is much denser than that of 1997. In order to keep the visualization compact and readable, filter applied in this figure is as twice high as the 1997 figure. Edges with collaboration intensity less than 0.016 are omitted. 2) The thickness of each edge between two nodes reflects the strength of their collaborative relationship. 3) The size of each node embodies its aggregated collaboration intensity.

In order to understand the dynamics of international collaborations, it is necessary to compare the structure of networks in the earlier year 1997 (Fig. 3) with that of the later year 2012 (Fig. 4). In contrast with 1997, the aggregated collaboration intensity (embodied by the circle size of each node) for most countries has increased in 2012. In particular, an important observation is that, the variety of inter-collaborations (lines between different coloured nodes) has grown significantly in 2012, while the connection strength between major intra-collaborative partners (nodes with the same colours) stayed roughly at original level of 1997.

In contrast with the structure in 1997 (Fig. 3), the rectangular Nordic cluster and triangular South American cluster in 2012 have both increased their inter-connections with countries beyond their geographic neighbours (see Fig. 4). The strong tie between Chile and Brazil (i.e. intra-collaboration) has been weakened while both Chile and Brazil developed new inter-collaborative partnerships with countries from other geographical areas. Similarly, the tie between Finland and Denmark became relatively weaker, whereas both of them established more connections with various countries. Due to the effect of "Europeanisation" of this geo-economic area, the new major collaboration partners are still within Europe, but far beyond the old Nordic limit in the later year.

Asian countries, though still with relatively low collaboration intensity, have increased scientific cooperation with the United States (i.e. known as type of inter-collaborations). In particular, China has developed a very strong collaborative tie with the United States and a reasonable partnership with Australia, which are both inter-collaborations. Yet as the second

largest producer of scientific publications, China did not develop any new strong collaborative ties (i.e. intra-collaborations) within its own geographical area.

Located in North America, Mexico seemed to have developed new collaborative research partners only beyond its own geographical area (i.e. inter-collaborations). As one of the most dynamic countries regarding international research collaborations, South Africa seemed to have built inter-collaborative relationships mainly in Europe and South America. Different from the isolated situation in the earlier stage (1997), Egypt and Saudi Arabia developed an extremely strong research partnership in 2012.[6] Their connection with each other was so strong that they hardly had any cooperation with any third countries.

Conclusions

The main lessons learned of this research can synthetized as follows:

1) The Gini coefficients for total publications and collaborations were both smaller in 2012 than 1997, indicating that the distribution among the under studied 40 countries became more and more balanced. Nevertheless, it is worthwhile to note that the distribution of total publications was more divergent than that of internationally co-authored papers.
2) In the process of evolution of international collaborations, evidence shows significant difference between intra- and inter- collaborations. In all geographical areas, except European Union, the intra collaboration performances exhibited a steady-state pattern, whereas inter-collaborations in the global network research structure have risen dramatically.
3) From a dynamic point of view, the comparison of 1997 and 2012 networks shows that inter-collaborations (between countries from different geographical areas) have grown significantly in the later stage, while the connection strength between major intra-collaborative partners stayed mostly unchanged. This finding indicates that recent research network across countries has a higher global inter-connection beyond geographical territorials, which is likely driven by advances of ICT and transportation new technologies and improvement of socio-economic systems.

In short, the increase of research collaborations between countries from different geographical areas has reshaped the global structure of international scientific collaborations. In the modern process of knowledge production, countries seem to be looking for more diverse collaborative partners worldwide.

References

Beaver, de B.D. & Rosen, R. (1978). Studies in scientific collaboration – Part. I. The professional origins of scientific co-authorship, *Scientometrics, 1*, 65-84.

Coccia M. (2010). Democratization is the Driving Force for Technological and Economic Change, *Technological Forecasting & Social Change*, *77*, 248-264.

Coccia, M. (2012). Political economy of R&D to support the modern competitiveness of nations and determinants of economic optimization and inertia, *Technovation*, *32*, 370-379

Coccia M. (2014). Driving forces of technological change: The relation between population growth and technological innovation-Analysis of the optimal interaction across countries, *Technological Forecasting & Social Change, 82*, 52-65

de Solla Price D. & de B. Beaver D. (1966). Collaboration in an invisible college, *American Psychologist*, *21*, 1011-1018.

Frame J . D. & Carpenter M. P. (1979). International research collaboration, *Social Studies of Science*, *9*, 481-497.

Frame, J.D., Narin, F. & Carpenter, M.P. (1977). The distribution of world science, Social *Studies of Science*, *7*, 501-516.

[6] Although Egypt and Saudi Arabia are classified into different groups, they are in geographically adjacent. Therefore their collaborative relationship can be still regarded as a result of geographical localization.

Hamers, L., Hemeryck, Y., Herweyers, G., Janssen, M., Keters, H., Rousseau, R., & Vanhoutte, A. (1989). Similarity measures in scientometric research: the Jaccard index versus Salton's cosine formula. *Information Processing & Management*, *25*(3), 315-318.

Hoekman, J., Frenken, K. & Tijssen, R. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe. *Research Policy*, *39*, 662-673.

Katz, J. S. & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*, 1-18.

Katz, J.S. & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics*, *40*, 541-554.

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration, *Science, Technology & Human Values*, *17*, 101-126.

Luukkonen, T., Tijssen, R.J.W., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics*, *28*, 15-36.

Mattsson, P., Laget, P., Nilsson, A. & Sundberg, C-J. (2008). Intra-EU vs. extra-EU scientific co-publication patterns in EU. *Scientometrics*, *75*, 555-574.

Melin, G. (1999). Impact of national size on research collaboration. *Scientometrics*, *46*, 161-170.

Narin, F., Stevens, K. & Whitlow, E.S. (1991). Scientific co-operation in Europe and the citation of multinationally authored papers, *Scientometrics*, *21*, 313-323.

Ozcan, S. & Islam, N. (2014). Collaborative networks and technology clusters –The case of nanowire. *Technological Forecasting and Social Change*, *82*, 115-31.

Presser, S. (1980). Collaboration and the quality of research. *Social Studies of Science*, *10*, 95-101.

Appendix A. Country/economy of the sample

nr	country	Geo-Economic Area
1	Canada	North America
2	Mexico	
3	United States	
4	Argentina	South America
5	Brazil	
6	Chile	
7	Austria	European Union
8	Belgium	
9	Czech Republic	
10	Denmark	
11	Finland	
12	France	
13	Germany	
14	Greece	
15	Hungary	
16	Ireland	
17	Italy	
18	Netherlands	
19	Poland	
20	Portugal	
21	Spain	
22	Sweden	
23	United Kingdom	
24	Norway	Other Europe
25	Russia	
26	Switzerland	
27	Iran	Middle East
28	Israel	
29	Saudi Arabia	
30	Turkey	
31	Egypt	Africa
32	South Africa	
33	China	Asia
34	India	
35	Japan	
36	Singapore	
37	South Korea	
38	Taiwan	
39	Australia	Australia/Oceania
40	New Zealand	

Appendix B:

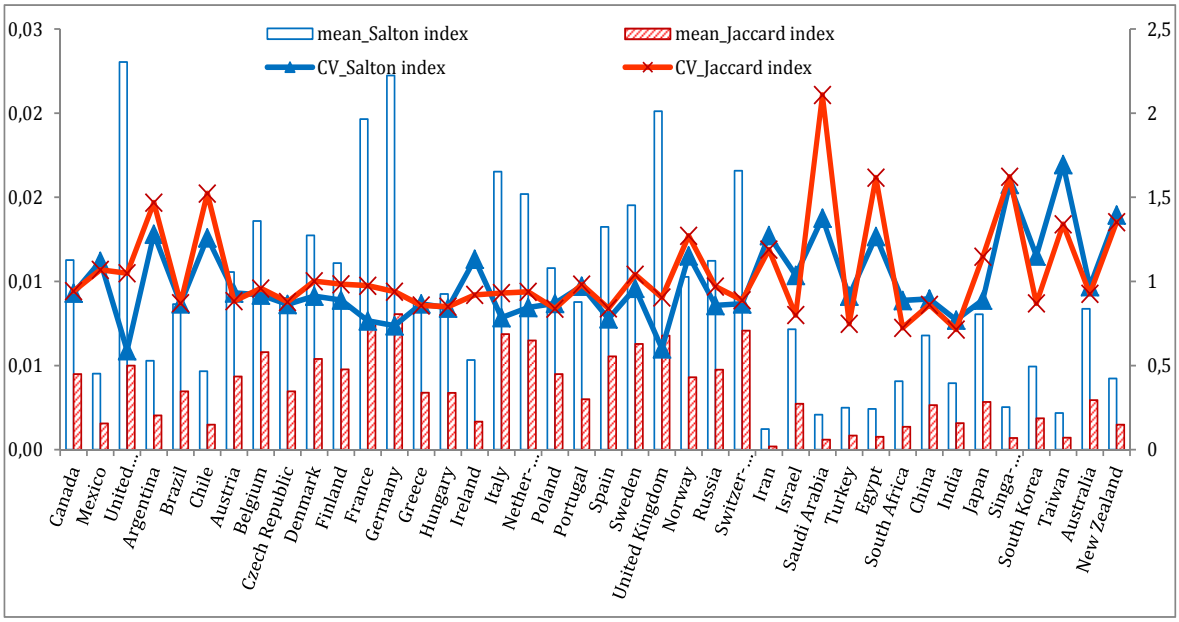


Figure B1. Mean and coefficient variation for collaboration indices (Salton vs. Jaccard) 1997

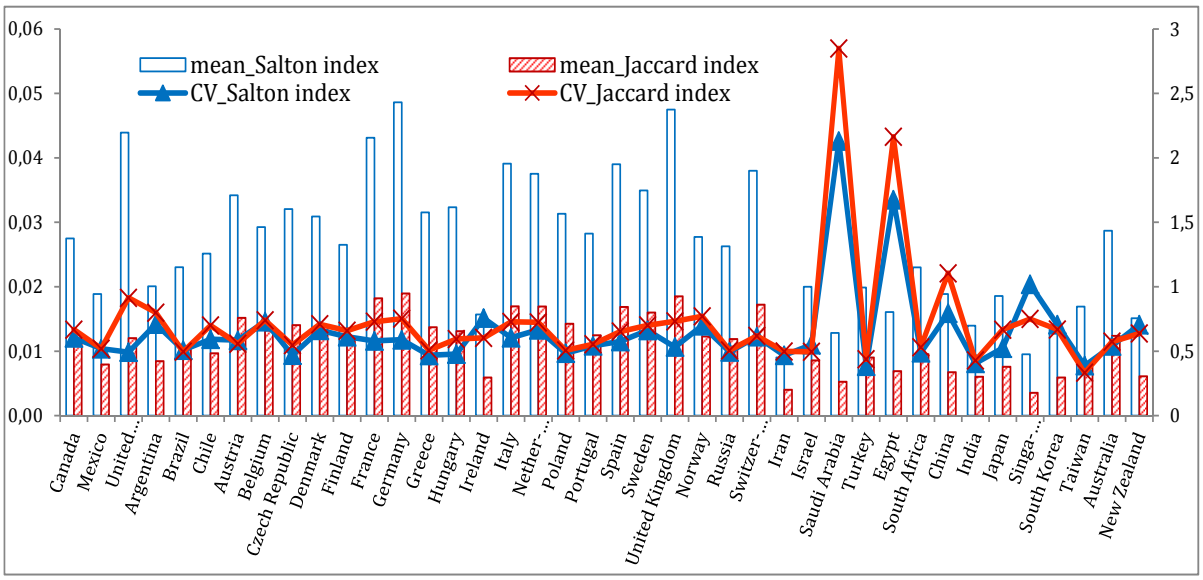


Figure B2. Mean and coefficient variation for collaboration indices (Salton vs. Jaccard) 2012

Nanotechnology Research in Post-Soviet Russia: Science System Path-Dependencies and their Influences

Maria Karaulova[1], Oliver Shackleton[1], Abdullah Gök[1] and Philip Shapira[1,2]
[1] Manchester Institute of Innovation Research, Manchester Business School, University of Manchester (United Kingdom)
[2] School of Public Policy, Georgia Institute of Technology (USA)

Abstract
This paper contributes to the analysis of Russian research dynamics and output in nanotechnology. The paper presents an analysis of Russian nanotechnology research outputs during the period of 1990-2012. By examining general outputs, publication paths and collaboration patterns, the paper identifies a series of quantified factors that help to explain Russia's limited success in leveraging its ambitious national nanotechnology initiative. Attention is given to path-dependent institutionalised practices, such as established publication pathways that are dominated by the Academy of Sciences, the high centralisation of the entire research system, and issues of internal collaborations of actors within the domestic research system.

Conference Topic
Country-level Studies

Introduction
Nanotechnology has been an interest of bibliometric research since the early 2000s after the United States and China adopted large-scale policy and funding programmes to stimulate scientific development by massively investing in this interdisciplinary research area. China has been among the countries with a large increase in research outputs in nanotechnology, and is the emerging economy that is frequently the focus of researchers (Appelbaum et al., 2011; Bhattacharya & Bhati, 2011; Liu et al., 2009).

Other emerging and transitional economies have also invested in nanotechnology development. Russia is a particular case among these countries, because the National Nanotechnology Initiative that was adopted in 2007 was a political as well an economic, scientific and technological project. The Russian government picked up on global trends and invested greatly in development of nanotechnology. On a purchasing power basis, it is suggested that public investment in Russian nanotechnology has rivalled that of the US and China (Schiermeier, 2007). Lux Research (2013) estimates that Russian nanotechnology investment has consistently been the third largest in the world after the US and China: Russia invested over $1 bln in 2010 and 2011 in nanotechnology projects, and just under $1 bln in 2012. However, with lower than anticipated results in nanotechnology, the Russian government has decreased its investment programme and the share of Russia in world nanotechnology funding dropped from 15% to 13% in 2013. It is anticipated to continue decreasing.

Important changes and structural reforms of Russian science (including nanoscience) have been implemented only relatively recently, in the mid- to late-2000s, almost two decades after the dissolution of the Soviet Union in 1991. Until then, Russian science was relatively unchanged from rules and institutional developed during the Soviet era. The Academy of Sciences of Russia maintained its Soviet-style organisation up until 2013 when it was subjected to a radical reform. Universities were reformed in 2008 and 2009 to move them away from mainly teaching and to develop research capabilities and to try to emulate US research clusters. The funding structure for Russian science was tied to four-year umbrella research programmes accompanied by small-scale research foundations until 2013, when decisions were made to reform Russia's Federal Targeted Programmes and Grant

Programmes towards more grant-based system. Importantly, the Russian National Nanotechnology Initiative and the associated surges in interest and investment pioneered the system-wide initiatives that started several years before other large-scale top-down changes.

Existing literature on nanotechnology research and innovation in Russia is less prodigious than for other "Rising Powers" countries, particularly China but also including Brazil and India. Scientometric analyses often examine Russian nanotechnology development as a benchmark for other emerging economies, mainly China and India (Liu et al., 2011, 2009) rather than deeply probing within the Russian system. At the same time, there is an important strand of scientometric work on Russian science and technology (including nanotechnology) produced by the Russian research community itself. In these cases, research is often descriptive or addresses internal debates within Russia (Terekhov, 2012, 2011), and sometimes lacks a critical approach. Additionally, most of these studies remain mostly background reference country reports (and are frequently only available in Russian).

There are, of course, some exceptions. For example, Klochikhin (2012) contextualised Russian nanotechnology policy in terms of post-Soviet path-dependencies and asked whether it was possible to break out from technological inertia to a new development trajectory. There are other studies of Russian nanotechnology that pose similar questions, be it from the industry and market formation perspective (Ananyan, 2005), or regulation (Gokhberg et al., 2012). A recent overview of the Russian Science, Technology and Innovation system (Karaulova et al., 2014) provides background for discussion of persisting path-dependencies. In the present paper, we build on, and extend, this prior work to examine Russia's technology development policies and to reflect on the challenges posed by its persistent and deeply-embedded path-dependent practices.

Data and Methodology

The dataset for our research covers the time period from 1990 to 2012, which includes the transitional period after the breakup of the Soviet Union, the Russian Nanotechnology Initiative (NNI) development (2004 – 2007) and the post-NNI period of nanotechnology research. We first provide an updated profile of nanotechnology research in Russia since the breakup of the Soviet Union until 2012. Second, we investigate the possible emergence of new trends of research of Russian nanotechnology after the adoption of large-scale policy programs. Third, we use self-reported publication data in order to illustrate the path-dependent nature of Russian nanotechnology research.

The bibliometric analysis draws on datasets of nanotechnology publications and patents developed by researchers at Georgia Institute of Technology and the Manchester Institute of Innovation Research. Two data sources are used: the Web of Science (scientific publications) (WoS) and Derwent Innovations (patents). Both data sources are published and made available in the Web of Knowledge by Thomson Reuters. Nanotechnology records in the databases are identified using the two-stage search strategy detailed in Porter et al. (2008), and updated in Arora et al. (2012). A keyword search based on a Boolean query is applied. Unrelated records are then removed by applying exclusion terms.

The defining characteristic that we used to identify Russian publications was that at least one author of each included publication had to have a Russian affiliation address (Soviet Union in 1990-1992; Russia subsequently). The primary language of publications in the dataset is English, but specialised editions that include translated articles originally published in Russian are included as well. In total 33,538 Russian nanotechnology publication records were identified in 1990-2012. We acknowledge that there are limitations in using WoS for capturing the totality of Russian science activity (but see also subsequent discussion in this paper of Russian journal publishing strategies).

A feature of the Soviet Union, carried over into the Russian Federation, is that science was and is developed in parallel – but not always in cooperation – with researchers elsewhere in the world. This influences the choice of terminology used by Russian researchers. For example, it has been observed that there is a rich tradition of nanotechnology research in Russia. Alexander Terekhov traces the technological development of Russian nanotechnology back to 1980s when the understanding of the physical properties of ultra-dispersed states enabled Soviet researchers to construct the first lasers and to conduct experiments at the nano-scale (Terekhov, 2013). But the term nanotechnology was not necessarily used at that time. A simple search strategy would not pick up on many Russian nanotechnology publications, especially in earlier years, which are crucial to understand trends of overall growth and development. We judge that the more complex and nuanced approach we apply is better able to capture the emergence and development of the Russia nanotechnology field.

After the publication data was collected and cleaned from unrelated records, further data cleaning to remove duplicates and consolidate organizational and author names was undertaken using VantagePoint text mining software. Cleaning is a large part of our methodology. One of the biggest problems of country report studies that use bibliometric analysis is the issue of varied affiliation reporting. We have addressed various problems through intense cleaning of the data. One problem of aggregation relates to affiliation (location, funding source, author) categories that the database recognizes as separate, but are actually the same. This is an issue that occurs in the self-reported semi-structured publication data. There are variations in reporting of affiliation data, different ways to spell the name of the organization, abbreviations and others. If left unchallenged, the data may be potentially distorted: the contributions of certain actors may appear as less than it reality, which can be misleading. Another major cleaning issue is disambiguating terms that were lumped together. For example, the process of disambiguation of the "Tech Univ" field and further aggregation of the items highlighted that the original very general field contained mainly records published in three large technical universities, and in a number of smaller ones. Table 1 illustrates examples of the data cleaning strategy.

Table 1. Affiliation Cleaning Strategy Examples.

	Original Record	*Cleaned Record*
Reporting Style	1. RAS, AM Prokhorov Gen Phys Inst; 2. Russian Acad Sci IOF RAN, Prokhorov Gen Phys Inst;	RAS Inst Gen Phys Prokhorov
Abbreviation	1. MISIS 2. State Univ Moscow Inst Steel & Alloys	Natl Univ Sci & Technol MISIS
Spelling	1. Alfa Akonis Res & Devices Enterprise 2. Alpha Akonis R&D Enterprise	Alpha Akonis R&D Enterprise
Change of Name	1. Leningrad State Tech Univ 2. St Petersburg State Tech Univ	St Petersburg State Tech Univ
Disambiguation	Tech Univ	1. St Petersburg Tech Univ 2. Tech Univ Moscow Inst Elect Technol 3. Tech Univ Berlin

Excessive aggregation of the data may lead to the loss of informative value. The Russian Academy of Sciences (RAS) presents the greatest challenge here. RAS is a large research

organisation that possesses more than 500 research institutes. However, the reported RAS affiliations are disordered, because research institutes often have long names and some of them do not issue guidelines for official English versions. Aggregating all these institutes under the domain of the "Russian Academy of Sciences" would yield analytical benefits in some circumstances, such as broad benchmarking. However, such a large agglomeration is not useful for detailed analysis. In our analyses of nanotechnology publications associated with RAS, we undertook disambiguation and identified 263 distinct affiliations, including research institutes of RAS, scientific centres and observatories.

We further grouped the data according to country, region, and type of affiliation. *Academy of Science* organisations are specific research entities that have wide government affiliations and heavily rely on government funding, that have a wide regional structure and hierarchical administrative division. We separately distinguished *Universities*. *Public Research Organisations* are private and state-owned research institutes that are neither academy of science institutions, nor universities. These also include research foundations and ministries. *Corporate actors* are privately and state-owned company affiliations. Organisations were usually labelled as 'corporate' actors if they had a distinctive property type word in their names (LLC, Ltd, GmbH, ZAO etc). *Other* included all other organisations that could not be attributed to any other category

In order to examine the internationalisation of Russian science we also separated publications into nationally collaborated publications (NCP) and internationally collaborated publications (ICP). The two groups are mutually exclusive and highlight the degree to which research produced in Russia only involves domestic actors (NCP), or there are also international partners (ICP).

Table 2. Grouping Results, number of publications.

Internalisation			*Domestic Affiliation Groups*			
				Orgs	Pubs	Share
			Acad of Sciences	3+1(263)	22927	68.5%
NCP	19098	56.9%	University	396	13868	41.4%
ICP	14440	42.8%	PROs	432	3781	11.3%
			Corporate	420	982	2.9%
			Other	3	3	0%

Results

The annual output of Russian nanotechnology publications steadily increased between 1990 and 2012. In 1998, there was a considerable jump in the number of publications; this probably reflects the fresh inclusion of a series of Russian journals within the WoS. Growth rates for domestic and international publications are almost identical starting from 1999 until 2012 and are about 1.1% per year. On average, domestic publications grow 2% faster than internationally collaborated publications.

The Academy of Sciences, 15 universities and four State Research Institutes are the leading organisations in terms of publication output. Some 68% of domestic publications are produced by the Russian Academy of Sciences and another 12% by Moscow State University. The top five organisations produced together 80% of all publications in 1990-2012 (Table 3). The top three organisations (RAS, MSU and St Petersburg State University) produced 78% of all publications. RAS is the dominant actor in producing nanoscience publications. However, in terms of annual publication outputs, university researchers have been catching up with RAS in the past decade.

Table 3. Biggest Publishers in Russian Nanoscience, 1990-2012.

	Organisation name	*Publications*	*Share*
1	Russian Academy of Sciences	22794	68.12%
2	Moscow MV Lomonosov State University	4007	11.98%
3	St Petersburg State University	1208	3.61%
4	Russian Research Centre Kurchatov Instute	613	1.83%
5	Nizhnii Novgorod State University	496	1.48%

Disambiguated, the bibliometric map of Russian science demonstrates a more nuanced picture of interactions in the nanotechnology research (Figure 1). One major research organisation, RAS Institute of Physics and Technology n.a. Ioffe, is a focal point for connecting various regional groupings of research centres, such as a cluster of four RAS institutes on Siberia that closely collaborate with one another, but do not have strong external links.

In terms of research performance, nanotechnology publications that only have Russian authors are cited on average 2.5 times per publication. Out of all domestic actors Russian Academy of Sciences publications collect the highest number of citations: 4.55 p/p. PRO publications, albeit being much smaller in number, collect 3.86 citations p/p. Universities collect on average 3.24 citations p/p, and publications produced by corporate actors collect 2.44 citations p/p.

Table 4. Shares of ICP and Average Citation Rate of Russia's Main Collaboration Partner Countries, 1990-2012.

Country	*Germany*	*USA*	*France*	*UK*	*Japan*	*Sweden*	*Italy*
ICP %	12.3%	8.2%	5.04%	3.4%	2.9%	2.08%	1.9%
Avg Cit	7.7	9.2	5.8	12.2	6.9	6.04	5.3
	Ukraine	*Poland*	*Spain*	*Netherlands*	*Belarus*	*Finland*	*South Korea*
ICP %	1.8%	1.5%	1.5%	1.4%	1.3%	1.1%	0.9%
Avg Cit	2.4	3.9	5.1	18.9	3.8	4.05	3.9

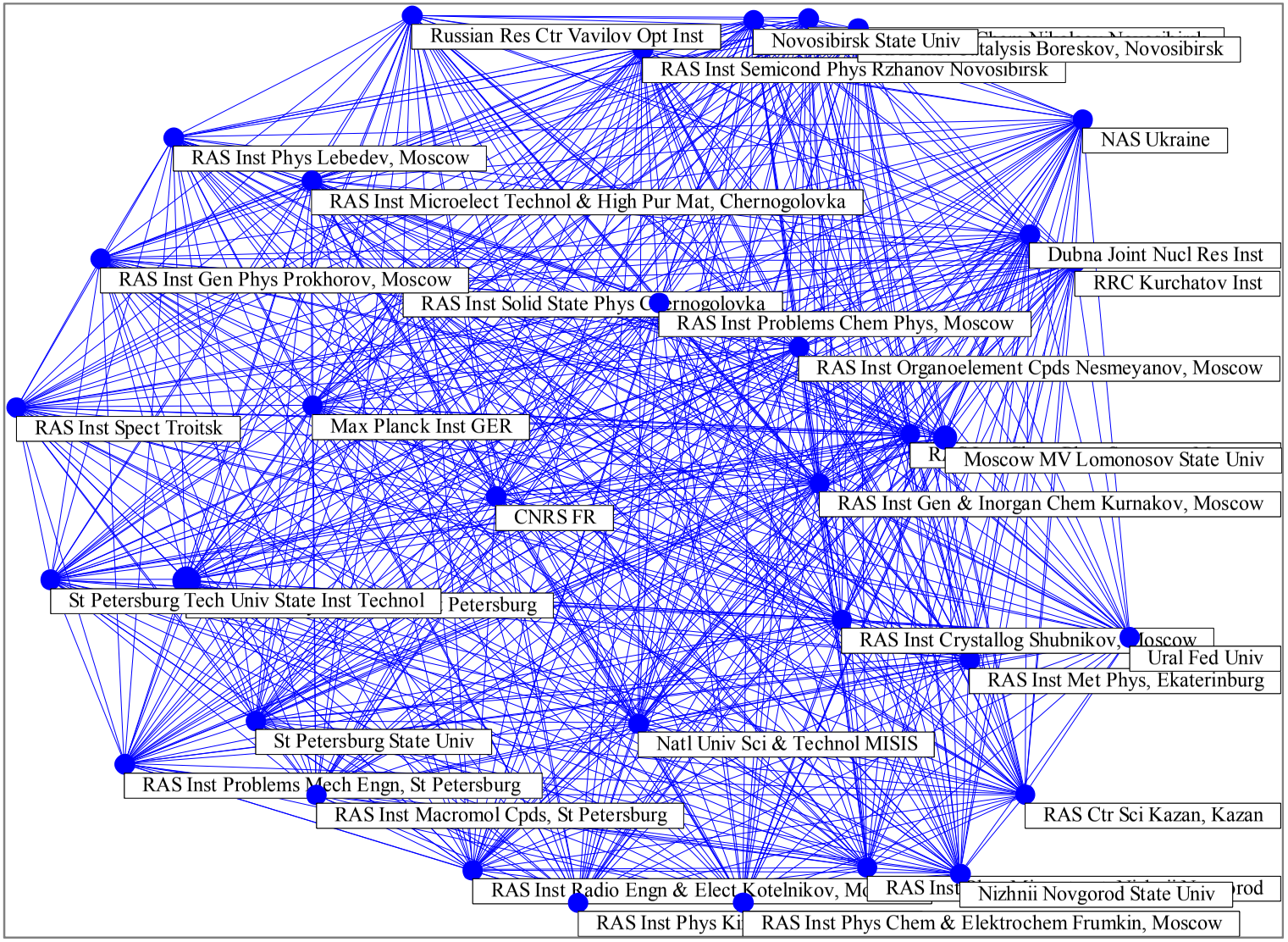


Figure 1. Bibliometric Map of Top 35 Publishers of Russian Nanoscience, 1990-2012.

Patterns of international collaboration seem to be connected to these structural differences. The average number of internationally cited publications is 4.33 times: international collaboration increases average citation by 1.7. There are, however, some regional variations in international collaboration performance outputs (Table 4). Russian international collaborations have strong European orientation, and there is evidence of recurrent path-dependent practices. It is noticeable that former Soviet states and influenced territories, such as Ukraine, Poland and Belarus factor highly in collaborative research. It implies research links are built on the older networks than the current political system and research takes place through these interactions. An impeding factor may be than average citation rates for these countries are significantly lower than for other countries with the same collaboration intensity (refer to Table 4). These 8.3% of CIS-collaborated ICPs represent collaboration patterns that may be detrimental to Russian science.

In the next section we pay particular attention to three elements of nanotechnology research that can highlight path-dependent dynamics of scientific knowledge production in Russia. We define them as journal gatekeepers, centralisation, and institutional diffusion. These all relate to structural features of the Russian science system that have persisted even after the Soviet Union broke apart.

Journal Gatekeepers

The data for journals in which Russian co-authored publications can be found, is available for 32844 publications, which constitutes 97% of the data. The majority of Russian publications in English were published in translated journals. Out of the top-10 journals with the biggest number of Russian publications, 7 are translated versions of Russian journals (refer to Table 5).

Translated versions of Russian journals are identified not by the publishing body (the rights to publish in most cases are owned by Springer), but by the contents of the journal and the editorial board. For example, Springer publishes *The Physics of the Solid State*. The description on the website says "The journal Physics of the Solid State presents the latest results from Russia's leading researchers in condensed matter physics at the Russian Academy of Sciences and other prestigious institutions" (Springer, n.d.). An analogous journal, called *Phyzika Tvyordogo Tela* (*The Physics of the Solid State*) is published in Russian by the Ioffe Institute in St.Petersburg (Ioffe Physical Technical Institute, n.d.). The Chief Editor of both journals is A.A. Kaplyanskii, and the editorial board matches both journal records. Tables of contents of issues match as well. Based on these we drew a conclusion that *The Physics of the Solid State* is a translated version of *Phyzika Tvyordogo Tela*, and the 'publishing body' is therefore an Institute within the Russian Academy of Sciences (the publishing body of the original), not Springer (the publishing body of the translated version). By doing manual analysis of the top journals in which Russian scientists publish we have identified that at least 25% of the entire publication volume was published in this manner (input of the Russian translated journals in the top-20 journal contributions). The overall contribution of the top-20 journals was 25%.

A paper is first published in a Russian peer-reviewed journal, and subsequently translated and published in the English version without an additional peer review. But it would also depend on the domestic peer reviewer whether a submitted article would be considered for publication and further translation for a WoS-indexed version of a journal. The publisher and the editorial board become important. As Table 5 demonstrates, vast majority of the translated Russian journals are published by the Russian Academy of Sciences and editorial boards mainly consist of members of RAS. This *status quo* is grounded in history: many of them were founded during the Soviet Union to inform the world about achievements of Soviet science.

Table 5. Top 20 Journals of Russian Nanotechnology.

	Journal	*Publishing Body*	*Records*	*Share*
1	Physical Review B	APS	1595	4.86%
2	Physics of the Solid State	RAS	1412	4.30%
3	Semiconductors	RAS	1255	3.82%
4	Technical Physics Letters	RAS	848	2.58%
5	JETP Letters	RAS	828	2.52%
6	Inorganic Materials	RAS	511	1.56%
7	Applied Physics Letters	American Institute of Physics	510	1.55%
8	Journal of Applied Physics	AIP Publishing	505	1.54%
9	Journal of Experimental & Theoretical Physics	RAS	490	1.49%
10	Russian Chemical Bulletin	RAS	411	1.25%

After the breakup of the Soviet Union, these established publication pathways and journals have been maintained and there has not been much impetus for change. Although an opportunity opened for Russian researchers to submit research publications to leading international journals, existing publication practices have persisted. Moreover, temporal dynamics highlight an increasing gap between publications submitted to translated Russian journals and international journals: the difference rose from twice as many translated journal publications as international journal publications in 2000 to 2.67 times in 2005 and to 3.8 times in 2011. In the earlier period this could have been explained by the lack of experience of researchers to publish abroad, or by poor knowledge of English. In the later period the English language problem continues, but it also has become prominent that internal domestic recognition for a Russian researcher can be even more important than international recognition in order to develop and continue a research career in Russia. Therefore, publishing in top domestic journals becomes a priority, and the English translation of these papers in journals that collect few citations is a by-product rather than the goal, because this research is anchored in Russian scientific discourse and debates.

RAS maintains the monopoly over acceptance of research outputs to the leading domestic journals, thus acting as a quality control body. It is also a gatekeeper in the Russian research system as to which domestic researchers are highlighted for international recognition. The domination of the Academy of Sciences constrains other research performers, such as universities and PROs, to develop and take advantage of publicly-provided research resources, for example through the Russian NNI. As a comparison, in their study of Chinese publication patterns Zhou and Leydesdorff (2006) recognised this 'gatekeeping' role as one of the main barriers to internationalisation of Chinese science in the early 2000s. However, this pattern has now changed with the emphasis in China in publishing directly in WoS journals.

Centralisation and the Academy

In our analysis, we observe two centralisation trends in publications within the Russian Academy of Sciences. These first of these is *geographical centralisation*. RAS has institutes in all 83 regions of Russia, but four regions (Moscow, St Petersburg, Novosibirsk, and the Moscow Region) produced the largest shares of publications in 1990-2012, contributing over 80% of the total amount. Moscow is the leader with almost 35% of all publications, together with the Moscow Region the agglomeration produced 45.2% of all Academy of Sciences publications. Previously, the high concentration of research in a limited geographical area and

with a large network of ineffective and low-performing institutes has been suggested to be one of the main reasons for the persistent problems of RAS (Graham, 1998).

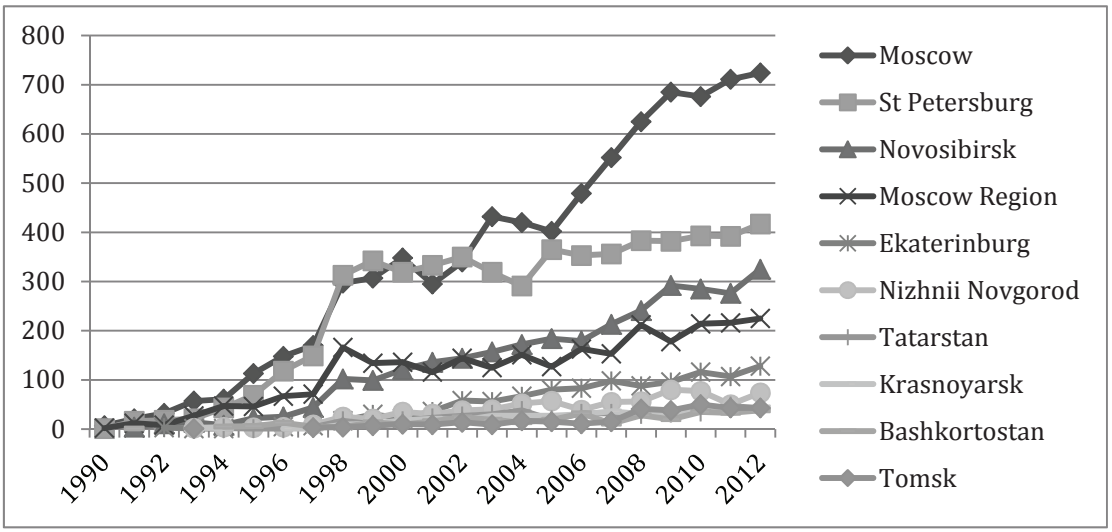


Figure 2. Temporal Dynamics of Geography of Nanoscience in Russia, 1990-2012.

Yet, while problems of RAS centralisation have long been observed, it seems that these trends have intensified in recent years: Academy research is becoming even more centralised (Figure 2). In nanotechnology, RAS institutes in Moscow surged upwards in the mid-2000s, producing almost twice as many publications in 2012 as the research cluster in St Petersburg. Many of these institutes have benefited from recent government science and innovation funding programmes, including specific nanoscience and nanotechnology funding programmes.

The *centralisation of high quality research* is a second persistent trend in Russian nanoscience. RAS has consistently contributed about 70% of the Russian annual publication output. In order to investigate whether quantity translates into quality, we assessed the performance of Russian domestic research system according to the criteria of (1) what affiliations of 10 top-cited ("star") scientists are, and (2) what affiliations of 100 top-cited publications are.

The top 10 most productive researchers coincide with the most cited researchers, with slight reversal in rank.[1] The majority of these "star" scientists are affiliated with RAS Ioffe Physical Technical Institute in St. Petersburg (Table 6). The Institute itself contributed about 14% of all publications and has an average citation of 6.13. The peak publication activity of all of the most productive scientists was between 1998-2000 after which the decline started. The most productive periods of the most productive Russian nanoscientists coincide with the most productive periods of Russian nanoscience: the contribution of "star" scientists was above 9% in 1996-2001, reaching a peak of 11.5% in 1998. A second, smaller, peak is reached in 2006, after which further decline occurs.

[1] The most highly cited Russian scientists are the ones who collaborated with colleagues at the University of Manchester in a paper in *Science* (Novoselov et al., 2005) that contributed to the award of the 2010 Nobel Prize in Physics to two Manchester researchers. This publication has 3541 citations. To include this exceptionally highly cited publication into the data would overshadow the underlying pattern of Russian nanotechnology performance, so this publication is not included in this part of the citation analysis.

Table 6. "Star" Scientists of Russian Nanoscience.

Rank	*Author Name*	*Affiliations*	*Times Cited*
1	Ledentsov, N	RAS Ioffe Physical Technical Institute	6033
2	Ustinov, Vr	RAS Ioffe Physical Technical Institute	5559
3	Alferov, Zh	RAS Ioffe Physical Technical Institute	5108
4	Kop'ev, P	RAS Ioffe Physical Technical Institute	5052
5	Zhukov, A	RAS Ioffe Physical Technical Institute	3504
6	Valiev, R	RAS Institute of Metals Superplasticity Problems; State Tech Univ of Aviation	3428
7	Egorov, A	RAS Ioffe Physical Technical Institute	2788
8	Morozov, S	RAS Institute of Microelectronics Technology & High Purity Materials	2323
9	Maximov, M	RAS Ioffe Physical Technical Institute	1909
10	Ruvimov, S	RAS Ioffe Physical Technical Institute	1812

The Post-Soviet period saw the rise and the peak of careers of scientists trained in the latter years of the Soviet Union. A drop in productivity coincides with the completion of the active research phase of their careers. There are few new 'rising stars' in the system, which explains the overall decline in performance. This data reinforces concerns about the 'generation gap' in nanotechnology where the average age of researchers is now in the mid-50s (Terekhov, 2011). RAS co-authored 81 out of the 100 most highly cited publications in Russian nanoscience.

Overall, it is notable that RAS dominates in quality as well as the quantity of research in Russian nanoscience. The productivity of RAS reached its peak in the late 1990s and has since then been in decline. The Russian government's support of the development of research universities and RAS reform in 2013 are expected to further contribute to decentralisation of the national research system and to the emergence of new centres of excellence. The trend towards concentration of research in the two capitals – Moscow and St Petersburg – is also a concern as government support to develop scientific research in other regions is limited.

Institutional Diffusion

The third and the final collaboration trend reflects the institutional diffusion of the Russian research system. Institutional theory proponents argue that institutions last and prosper when other elements of the system are dependent on them, e.g. when institutions are diffused well with other institutions (Clemens & Cook, 1999). In a research system this mainly takes form of inter-institutional collaborations. In order to examine the institutional relationships of the Russian research system we investigated (1) whether each organisation preferred to publish on its own; (2) if research was done through the collaboration of authors in one organisation; (3) whether the organisation engaged in collaborative activities with other organisations of the same type; (4) if organisations collaborated nationally; and (5) whether organisations collaborated internationally.

The results of this analysis demonstrate various patterns of domestic collaboration (Figure 3). For instance, corporate publishers have to rely heavily on collaborations, so they have higher rate of collaborations with all types of actors than the average. An asymmetric relationship among the system actors reflects institutional domination of the Academy of Sciences of Russia. The analysis of institutional collaboration patterns demonstrates that there are very weak collaboration links between the Academy of Sciences and other system actors.

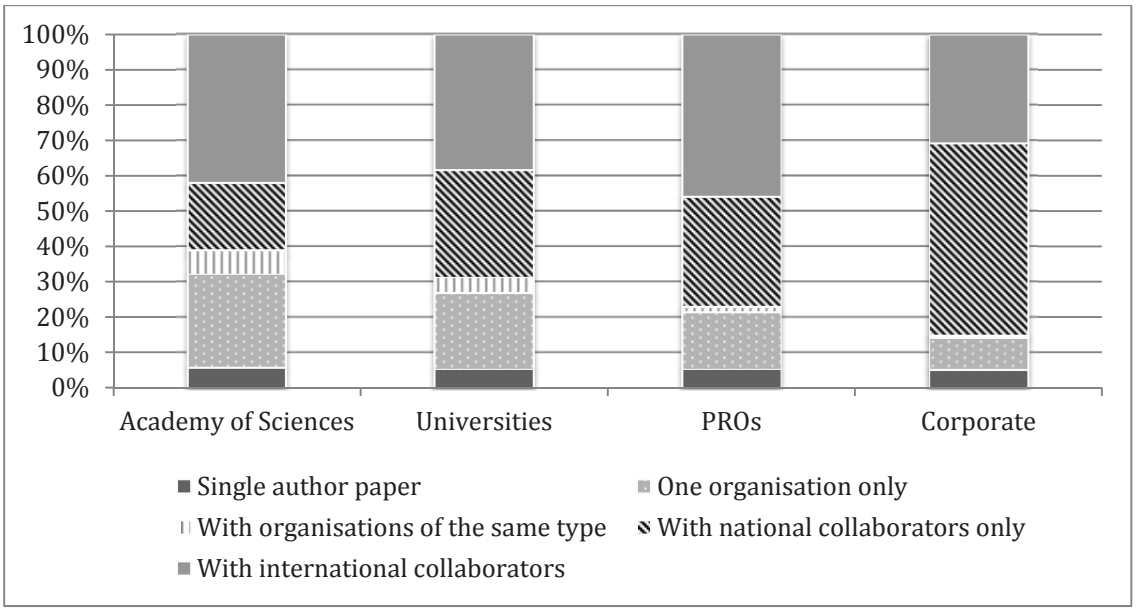


Figure 3. Institutional Diffusion of Russian Research System.

About two-fifths of academic publications are written either by a single author, or by a group of authors within RAS, and only 19% are collaborated with other Russian organisations. An international orientation is evident for PROs: over 46% of publications are internationally collaborated, but only 1.5% of publications are collaborated with other PROs. University organisations stand in the middle and have larger share of nationally collaborated publications than the Academy or PROs.

Weaknesses in international orientation and a reluctance to engage in national collaborative research projects is a particular concern for the Russian Academy of Sciences given that it dominates much of the Russian research system. In some RAS institutes, domestic collaboration rates with others outside of the home institute are noticeably low, for example just 11.6% in the Institute of Theoretical Physics RAS n.a. the Landau Institute of Theoretical Physics.

Conclusion

This exploratory study highlights three major path-dependent structural features of the Russian research system that are evident in Russia's nanotechnology research and publication activities. These structural features tend to be under-emphasized in other quantitative and qualitative studies, including those undertaken from within Russia itself. The available studies tend to focus on underfunding, deteriorating equipment, brain drain and other factors that, without a doubt, are very important in understanding the position of Russian science. In this research note, using bibliometric analysis in the case of nanotechnology, we draw attention to other less explicit but nonetheless important underpinning factors that frustrate the successful implementation of science and innovation policies and which may weaken returns on research investment. Reflecting upon and revising institutional practices of research that have remain largely unchanged since the breakup of the Soviet Union is an important challenge for Russian science policy. Some reform efforts have begun, but much more is likely to be needed to support the next generation of researchers.

Acknowledgements

This work was supported by the Economic and Social Research Council [grant number ES/J012785/1] as part of the project *Emerging Technologies, Trajectories and Implications of Next Generation Innovation Systems Development in China and Russia*.

References

Ananyan, M. (2005). Nanotechnology in Russia: from laboratory towards industry. *Nanotechnology Law & Business*, *2*, 194.

Appelbaum, R.P., Parker, R., Cao, C., (2011). Developmental state and innovation: nanotechnology in China. *Global Networks*, 11, 298-314. doi:10.1111/j.1471-0374.2011.00327.x

Bhattacharya, S. & Bhati, M. (2011). China's emergence as a global nanotech player: Lessons for countries In *Transition*. *China Rep*. 47, 243-262. doi:10.1177/000944551104700401

Clemens, E.S. & Cook, J.M. (1999). Politics and institutionalism: Explaining durability and change. *Annual Review of Sociology*, *25*, 441-466. doi:10.1146/annurev.soc.25.1.441

Gokhberg, L., Fursov, K., & Karasev, O. (2012). Nanotechnology development and regulatory framework: The case of Russia. *Technovation*, *32*, 161–162. doi:10.1016/j.technovation.2012.01.002

Graham, L.R. (1998). *What Have We Learned about Science and Technology from the Russian Experience?* Stanford University Press.

Ioffe Physical Technical Institute. (n.d.) Fizika Tvyordogo Tela / Physics of the Solid State Retrieved June 3, 2015 from: http://journals.ioffe.ru/ftt/

Karaulova, M., Shackleton, O., Gok, A., Kotsemir, M.N., & Shapira, P. (2014). Nanotechnology research and innovation in Russia: A bibliometric analysis (SSRN Scholarly Paper No. ID 2521012). Social Science Research Network, Rochester, NY.

Klochikhin, E.A. (2012). Russia's innovation policy: Stubborn path-dependencies and new approaches. *Research Policy*, *41*, 1620–1630. doi:10.1016/j.respol.2012.03.023

Liu, X., Kaza, S., Zhang, P. & Chen, H. (2011). Determining inventor status and its effect on knowledge diffusion: A study on nanotechnology literature from China, Russia, and India. *Journal of the American Society for Information Science and Technology*, *62*, 1166-1176. doi:10.1002/asi.21528

Liu, X., Zhang, P., Li, X., Chen, H., Dang, Y., Larson, C., Roco, M.C., & Wang, X. (2009). Trends for nanotechnology development in China, Russia, and India. *Journal of Nanoparticle Research*, 11, 1845-1866. doi:10.1007/s11051-009-9698-7

Lux Research. (2013). *Nanotechnology Update: Corporations Up Their Spending as Revenues for Nano-enabled Products Increase*.

Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Katsnelson, M.I., Grigorieva, I.V., Dubonos, S.V., & Firsov, A.A. (2005). Two-dimensional gas of massless Dirac fermions in graphene. *Nature*, 438, 197-200. doi:10.1038/nature04233

Schiermeier, Q. (2007). Russia pins its hopes on "nano". *Nature*, *448*, 233–233. doi:10.1038/448233a

Springer. (n.d.) Physics of the Solid State - Springer Retrieved June 3, 2015 from: http://www.springer.com/materials/journal/11451

Terekhov, A.I. (2011). Providing personnel for priority research fields (the example of nanotechnologies). *Herald of the Russian Academy of Sciences*, 81, 19-24. doi:10.1134/S1019331611010047

Terekhov, A.I. (2012). Evaluating the performance of Russia in the research in nanotechnology. *Journal of Nanoparticle Research*, 14, 1-17. doi:10.1007/s11051-012-1250-5

Terekhov, A.I. (2013). Russia's policy and standing in nanotechnology. *Bulletin of Science, Technology & Society*, 33, 96–114. doi:10.1177/0270467614524127

Zhou, P. & Leydesdorff, L. (2006). The emergence of China as a leading nation in science. *Research Policy*, 35, 83–104. doi:10.1016/j.respol.2005.08.006

Support Programs to Increase the Number of Scientific Publications Using Bibliometric Measures: The Turkish Case

Yaşar Tonta[1]

[1] *yasartonta@gmail.com*
Hacettepe University Department of Information Management, 06800 Beytepe, Ankara (Turkey)

"*Not everything that counts can be counted, and not everything that can be counted counts*." – William Bruce Cameron

Abstract
Bibliometric measures for scientific journals such as journal impact factor, cited half-life, and article influence score are readily available through commercial companies such as Thomson Reuters, among others. These metrics were originally developed to help librarians in collection building and are based on the citation rates of published papers. Yet, they are increasingly being used, albeit undeservedly, as proxies for peer review to assess the quality of individual papers; and research funding, hiring, academic promotion and publication support policies are developed accordingly. This paper reviews the use of such metrics by the Turkish Scientific and Technological Research Council (TUBITAK) in its Support Program of International Scholarly Publications and concentrates on the most recent policy changes. A sample of 228 journals was selected on the basis of stratified sampling method to study the impact of changing algorithms on the level of support that journals received in 2013 and 2014. Findings are discussed and some recommendations are offered to improve the existing algorithm.

Conference Topic
Country level studies

Introduction
Bibliometric measures such as journal impact factor (JIF) and cited-half life are based on citation rates of published papers in the literature and their aging. They were originally developed to help librarians in collection building and in making decisions as to how long the back issues of journals should be kept in stacks (San Francisco, 2012). Yet, such bibliometric measures are often used to assess the quality of individual papers, authors, and institutions. They are increasingly being used, albeit undeservedly, as proxies for peer review to assess the quality of individual papers; and research funding, hiring, academic promotion and publication support policies are developed accordingly. Algorithms used to rank authors, institutions or even countries are primarily based on such bibliometric measures as JIF and h index (Simons, 2008). This paper reviews the use of such metrics by the Turkish Scientific and Technological Research Council (TUBITAK) in its Support Program of International Scholarly Publications and concentrates on the most recent policy changes.

Literature Review
The drawbacks of citation-based metrics, especially JIF, for research assessment is well documented in the literature (e.g., Seglen, 1997; Guerrero, 2001; Simons, 2008; Browman & Stergiou, 2008; Lawrence, 2008; Todd & Ladle, 2008; Balarama, 2013; Kotur, 2013; Marks, Marsh, Schroer & Stevens, 2013; Marx & Bornmann, 2013; Casadevall & Fang, 2014; Jawaid, 2014). Convincing arguments supported by empirical data were brought forward as to why such measures should not be used to evaluate research (e.g., skewed citation distributions, different publication and citation practices in Science vs. Social sciences, and the manipulation of JIFs by editorial policies). Some researchers stressed the hidden dangers

of a "citation culture" (Todd & Ladle, 2008) while others drew attention to how measurement and "bean counting" harms science (Lawrence, 2008), as such metrics can easily be "gamed" (Marks et al., 2013). The title of the editorial of the special issue on "the use and misuse of bibliometric indices in evaluating scholarly performance" of the journal *Ethics in Science and Environmental Politics* says it all: "Factors and indices are one thing, deciding who is scholarly, why they are scholarly, and the relative value of their scholarship is something else entirely" (Browman & Stergiou, 2008).

The San Francisco Declaration on Research Assessment (DORA), signed by researchers, journal editors and publishers alike, strongly recommends not to use "journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions or in hiring, promotion, or funding decisions" (San Francisco, 2012). "[M]ost experts agree that the JIF is a far from perfect measure of scientific impact" (Bollen, Van de Sompel, Hagberg & Chute, 2009). Even Thomson Reuters, the publisher of such metrics through its Journal Citation Reports (JCR), is against using JIF to measure the quality of scientific papers (Marx & Bornmann, 2013, pp. 62-63). Yet, its use as "a tool of research assessment has reached epidemic proportions worldwide, with countries like India, China and the countries of Southern Europe being among the hardest hit" (Balaram, 2013, p. 1268). Some declared war on the impact factor (Balaram, 2013) and advised that its use should be abolished (Hecht, Hecht & Sandberg, 1998). Nonetheless, it is believed that, despite its misuse and abuse, JIF "will retain its impact and won't fade away" (Jawaid, 2014).

Consequently, policies developed for hiring, academic promotion, research funding, and monetary support to scientific publications in different countries tend to rely increasingly on metrics based on citation rates of published papers. Turkey is no exception (Tonta, 2014). The Higher Education Council of Turkey (YÖK) and the Turkish Scientific and Technological Research Council (TUBITAK) have been using journal impact factors for almost two decades in their academic promotion policies and incentive programs to support scientific papers, respectively.

The use of bibliometric measures for research assessment in Turkey along with their suitability as criteria to evaluate research quality has recently been reviewed (Tonta, 2014). This paper examines the most recent algorithmic changes introduced in 2013 and 2014 to rank the journals in the Support Program of International Scholarly Publications (UBYT) of TUBITAK and compares them with the earlier one (2012). The effects of year-to-year changes on the consistency of the ranks of journals are also studied. Note that, as the timeframe is short (2012-2014), we do not intend to study the impact of such changes on the authors' behaviour in terms of which journals they prefer to submit their papers to, journals' acceptance rates or the length of time it takes to publish therein. Rather, we try to understand the motives behind changes along with their effects on journal scores, which in turn determine the rank of each journal and thus the amount of monetary support that TUBITAK provides to the authors of papers that appeared in a specific journal.

TUBITAK's Support Program of International Scholarly Publications

Since 1993, TUBITAK provides monetary support to the authors of scholarly papers that appear in journals indexed by Thomson Reuters as an incentive to increase the number of such publications. The journal impact factor (JIF) was the sole criterion for support until 2013. As is well known, the impact factor (IF) of a journal is measured by the number of citations it gets in a given year to the papers published in it in the previous two years. Thomson Reuters publishes JCRs annually in which journals in each subject discipline covered by Science Citation Index (SCI) and Social Sciences Citation Index (SSCI) are ranked according to their JIFs. TUBITAK used JCRs to determine the eligible journals and

categorized the top 25% of journals in each subject discipline as Group "A", the next 25% of journals as Group "B" and the remaining 50% of journals as Group "C" (and "Group D" for social science journals—the bottom 10% of the remaining 50% of journals) (UBYT Programı, 2012).[1]

In 2013, TUBITAK has almost quadrupled the amount of support per paper. In parallel with this decision, TUBITAK also changed the rules to further classify journals with high IFs by developing its own "journal impact factor". Rather than simply classifying journals as A, B, C, and D on the basis of JCR's two-year JIF data, TUBITAK decided to use JCR's five-year JIFs and cited half-lives of journals in each discipline and multiplied the two figures to come up with its own JIF and ranked journals accordingly. (Cited half-life of a journal is the median—in years—of citations to papers published in it in a given year and depends on how fast the literature obsolesces in subject disciplines.) TUBITAK then took the average TUBITAK JIF of ranked journals and identified the journals with 2 standard deviations (SD) above and below the average to award them the maximum (5,000.00 Turkish Lira[2]) and minimum (500.00 TL) amount of support, respectively. Journals in between were awarded on the basis of a linear transformation formula taking the number of journals in each JCR discipline into account. This formula was criticized by some (Batmaz, 2013) as it happened to downgrade the ranks of some "A class" Archaeology journals considerably, thereby making them least supported ones. Similarly, the 2013 algorithm ranked 56% of Geology journals lower, including *Tectonics*, one of the most prestigious journals in this discipline (Yaltırak, 2014, p. 18).

Apparently, the new algorithm did not fulfill its objectives and TUBITAK, after using it for only one year, quickly replaced it in 2014 with the one that is based on JCR's article influence score. The 2013 transformation formula was used in 2014 to determine the exact amount to be paid to each journal (TUBITAK, 2013; 2014 Yılı, 2014). Comparable to IF, average influence score (AIS) is "a measure of the average influence, per article, of the papers in a journal" (Bergstrom, West & Wiseman, 2008) and is similar to Google's PageRank algorithm in that citations coming from papers in highly cited journals are weighted more heavily (Franceschet, 2010; Arendt, 2010). It is based on the number of citations, nonetheless. AIS is "the most stable indicator across different disciplines" (Franceschet, 2010) and can therefore be used for interdisciplinary comparisons (Arendt, 2010).

The drawbacks of metrics used by TUBITAK (JIF, TUBITAK's own JIF consisting of JCR's five-year IF and cited-half life and AIS) were discussed in detail elsewhere (Tonta, 2014). What follows is a survey based on a sample of 228 journals supported by TUBITAK to see the impact of changes introduced in 2013 and 2014.

Method

In order to find out the impact of most recent changes introduced in 2013 and 2014, we used TUBITAK's list of journals supported in 2012[3] to draw a sample. The list has a total of 11,562 journals. As explained earlier, TUBITAK categorized these journals in 2012 under Groups A, B, C and D according to JIFs reported in Thomson Reuters' JCR. The distribution of 11,562 journals under categories is as follows: Group A: 4,205 (or 36%) journals; Group B: 2,446 (or 21%) journals; Group C: 4,711 (or 41%) journals; and Group D: 200 (or 2%) journals. Social sciences journals constituted about one third of all journals. We selected a sample 232 journals (or 2% of the population) using stratified sampling method. Journals under Groups A, B, C and D formed the four strata. Two numbers between 1 and 100 were

[1] For more detail on TUBITAK's classification of journals, see Tonta (2014).

[2] *Circa* 2,000.00 USD.

[3] Available at http://ulakbim.tubitak.gov.tr/tr/hizmetlerimiz/ubyt-yayin-tesvik-programi.

identified (37 and 54) randomly and every 37th and 54th journal titles were selected. Table 1 provides population parameters and sample statistics.

The distribution of Science and Social science journals in the sample is quite similar to that of population. This can be interpreted as an indication of the generalizability of findings to the population with a calculated margin of error. The original sample size was 232 but 4 journals under Group D were later discarded to simplify the comparisons. Journals supported in 2013 and 2014 are not available as single lists but can be searched using a search engine available at the site.[4] All 228 journal titles in the sample were searched and their journal scores as well as the amount of support they would get were recorded. Six journals[5] in the 2012 list were no longer available in 2013 and 2014 among the supported journals and they were replaced with the next ones (e.g., 38th or 55th record) provided they were in the same category of Science and Social Science journals (e.g., Groups A, B, and C).

Table 1. Population parameters and sample statistics.

	Population parameters						*Sample statistics*					
	Science		*Social Science*		*Total*		*Science*		*Social Science*		*Total*	
Group	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*	*N*	*%*
A	2037	48	2168	52	4205	100	40	48	44	52	84	100
B	1824	75	622	25	2446	100	36	72	14	28	50	100
C	3763	80	948	20	4711	100	77	82	17	18	94	100
D	--	--	200	100	200	100	--	--	4	100	4	100
Total	7624	100	3938	100	11562		153		79		232	

It should be noted that the minimum and maximum amounts for 2012, 2013 and 2014 were fixed (433.00 TL and 1,300.00 TL for 2012 and 500.00 TL and 5,000.00 TL for 2013 and 2014). As journals in 2012 were awarded fixed amounts of support depending on which group they belonged to, the figure for each journal was obtained by checking its group (e.g., A, B, C) as well as its being a Science or Social science journal. Social science journals were paid twice the amount of what is determined for each group (e.g., the author of a paper published in a Social science journal under group A was awarded 2,600.00 TL instead of 1,300.00 TL).

Findings

Table 2 below provides descriptive statistics for 228 journal titles including the quartiles. Despite the fact that the amount of support was increased in 2013 to 5,000.00 TL, the mean and median values do not seem to be affected much from this increase. The percentage of increase for the journals in the 3rd quartile is noticeable (19%), the reasons for which will be discussed shortly.

Figure 1 provides the scatter graph of the amount of support given by TUBITAK in 2012, 2013 and 2014 to the authors of papers that appeared in 228 journals sampled. Note that the blue line represents the 2012 figures and ranked in descending order by the amount of support. The amount was fixed depending on which group the journal belonged to. The authors of articles that appeared in Groups A, B, and C journals were paid 1,300.00, 867.00, and 433.00 Turkish Lira (TL), respectively.[6] If the paper appeared in a Social science journal,

[4] http:// http://www.ulakbim.gov.tr/

[5] Or, they might have been discontinued or their names might have changed. Replaced journal titles are: *Journal of Dental Research, Tulsa Studies in Women's Literature, Journal of Electronic Imaging, Plasma Physics Reports, and Vie et Milieu – Life and Environment*.

[6] The authors of case studies, technical communications, letters to the editors, etc. received half this amount.

the amount of support is doubled so that the authors of Social science papers will be further encouraged. Therefore, the solid blue line at 2,600.00 TL and 1,733.00 TL represent both 43 Group A and 14 Group B Social science journals, respectively, whereas the blue line at 1,300.00 TL represents 41 Group A Science journals. The 867.00 TL band represents both 35 Group B Science journals and 17 Group C Social science journals. The 433.00 TL band represents 78 Group C Science journals.

Table 2. The amount of support (in Turkish Lira*).

	2012	*2013*	*2014*	*Increase 2013-2014 (%)*
Mean	1176	1317	1403	7
Minimum	433	500	500	0
1st quartile	433	533	558	5
Median	867	829	874	5
3rd quartile	1408	1518	1806	19
Maximum	2600	5000	5000	0

*Rounded to the nearest whole number.

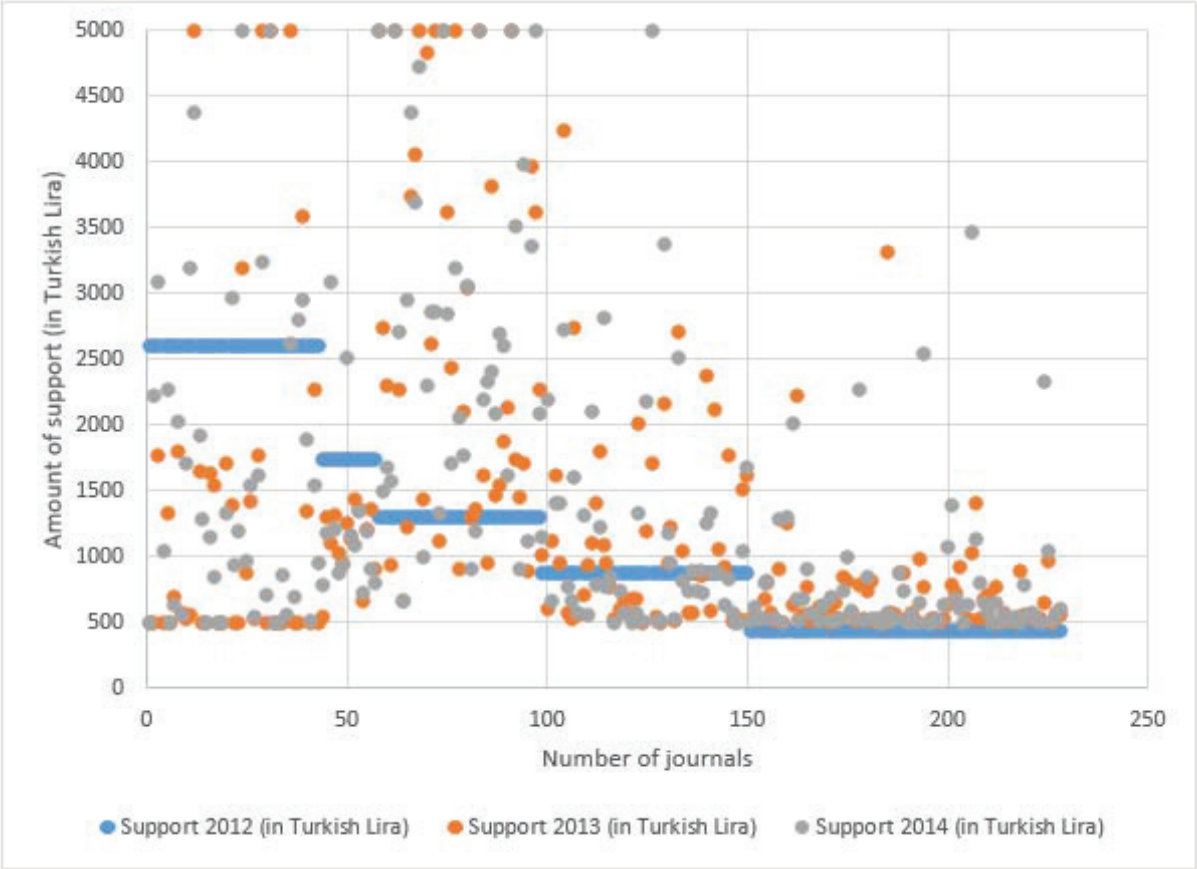


Figure 1. The scatter of journals by the amount of support in 2012, 2013 and 2014 (N = 228).

As indicated earlier, the maximum amount of support in 2013 was increased to 5,000.00 TL (the minimum being 500.00 TL). Note that the Group A journals of 2012 received relatively less support in 2013 and 2014. Out of 84 journals classified under Group A in 2012, only 15 (18%) maintained their top positions in the following years.[7] However, the positions of Social

[7] The amount between 500.00 TL and 5,000.00 TL was divided into three equal groups and the ones that were awarded between 3,500.00 TL and 5,000.00 TL are considered as top journals.

science journals classified under Group A fluctuated more than that of Science journals. Only 3 out of 43 Social science journals (7%) maintained their top positions as opposed to 12 out of 41 Science journals (29%).

Note that 2013 and 2014 figures are scattered without seemingly any discernible pattern (Fig. 1), as the 2012 figures are ranked in descending order by the amount of support and they do not necessarily correspond with the amounts in 2013 and 2014. Although statistically significant, the correlation between the amount of support to journals in 2012 and 2013 and that in 2012 and 2014 was rather low (Pearson's r = .289 and .231, p = .000, respectively). The correlation between the 2013 and 2014 journals was moderate (Pearson's r = .767, p = .000) (see Fig. 2).

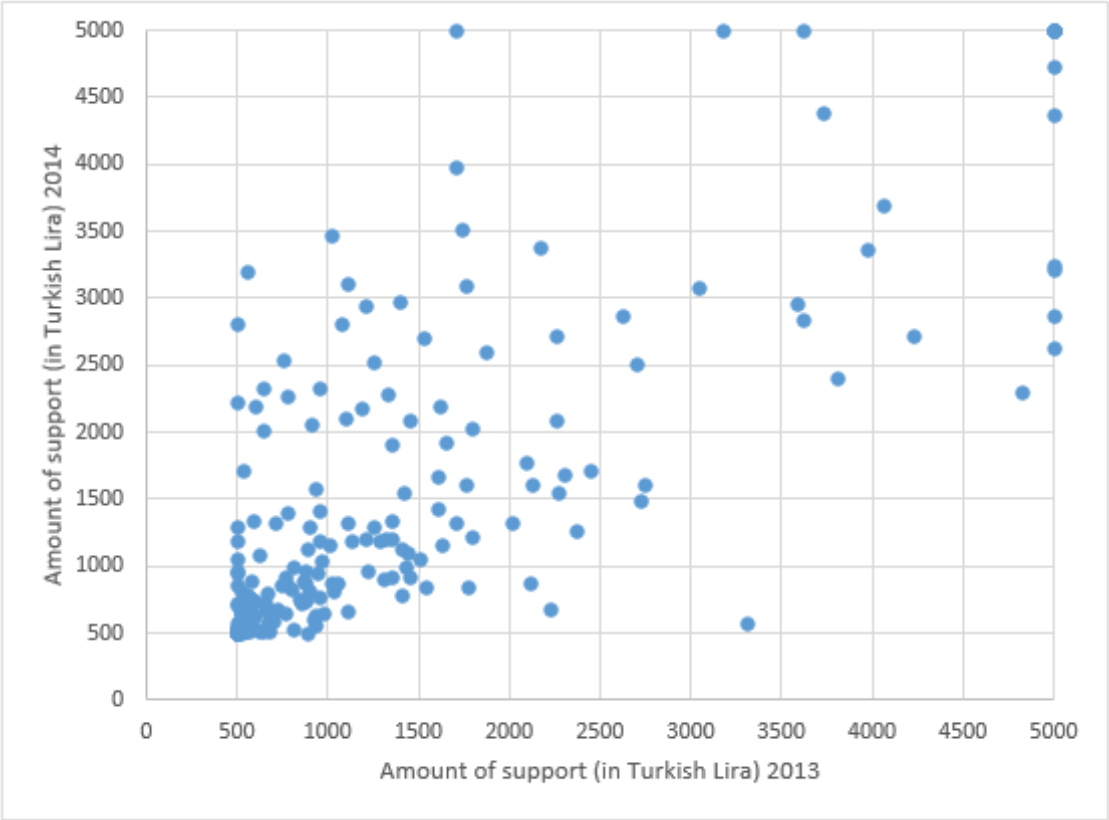


Figure 2. The scatter of journals by the amount of support in 2013 and 2014 (N = 228).

It is estimated that some 30,000 scholarly journals are published in the world. Thomson Reuters indexes about 12,000 of them and TUBITAK supports almost all of them (TUBITAK's 2012 journal list had 11,562 journal titles). It should be pointed out that TUBITAK's threshold for support is rather low. As Figures 3 and 4 below show, about one third of journals barely meet the minimum criteria and get the minimum amount of support (500.00 TL). It is reasonable to suggest that after careful consideration support to more than 3,000 journals can easily be discontinued.

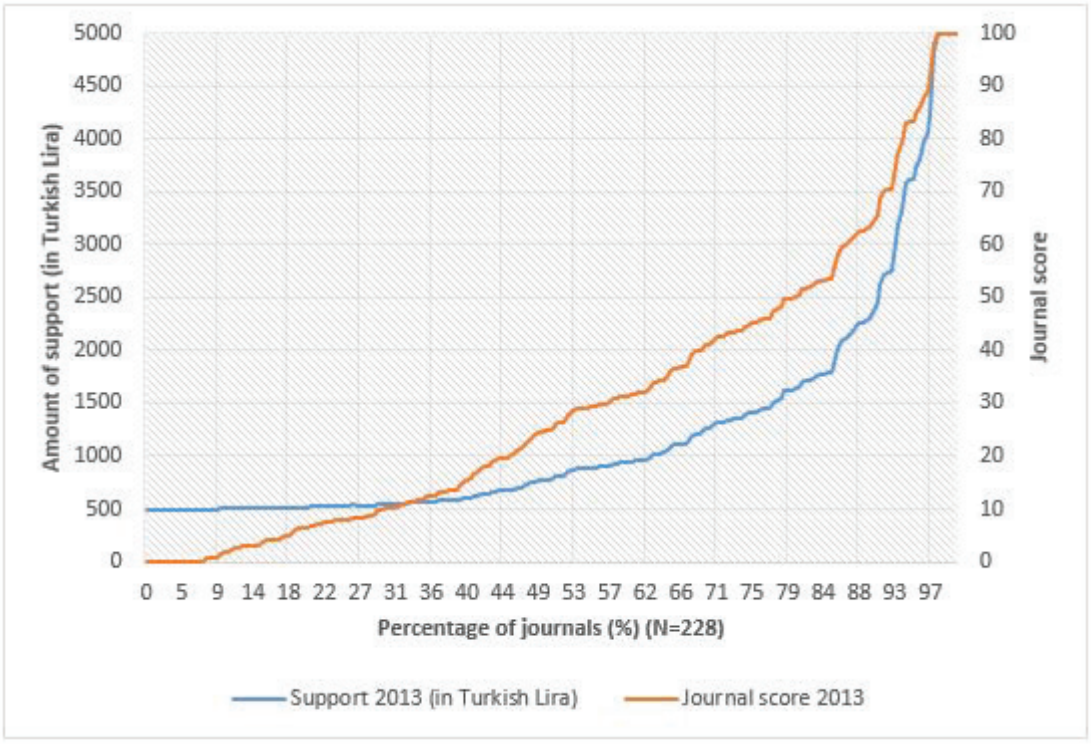


Figure 3. Relationship between journal score and the amount of support in 2013.

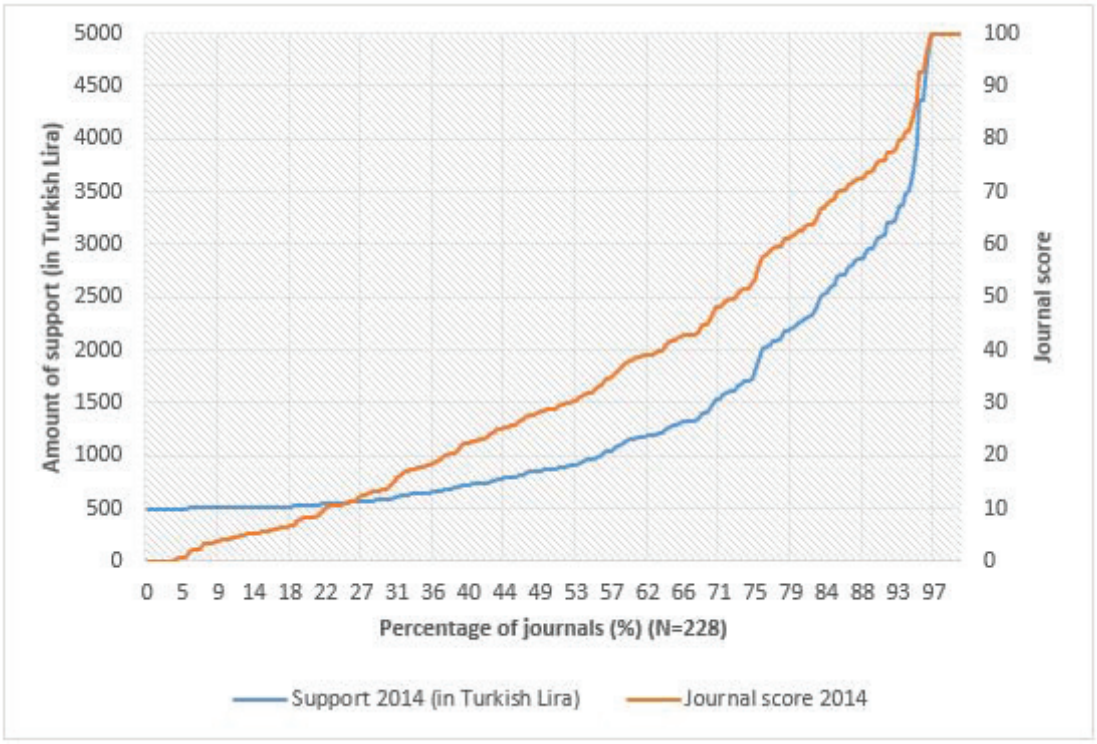


Figure 4. Relationship between journal score and the amount of support in 2014.

It should also be pointed out that the new policy discourages the authors of papers that appear in journals with low Article Influence Scores to seek support. As Figure 3 and 4 show, the gap between the journal scores and the amount of support starting from about 27%-35% gets widened. In other words, the amount of support is not that high for journals with relatively

lower AISs. More than 90% and 80% of journals received less than 2,500.00 TL (half the full amount of 5,000.00 TL) in 2013 and in 2014, respectively. Journals that received more than 4,000.00 TL support were about 5% of all journals in both 2013 and 2014. The situation was even worse for Social science journals (Fig. 5). This trend can also be followed from the last column of Table 2. The percentage of increase for the journals in the third quartile between 2013 and 2014 was 19% while it was only 5% for the journals in the first and second quartiles. This could be interpreted as a positive sign to encourage the authors to publish in more prestigious journals with higher AISs. Note that if the amount was less than 100.00 TL per co-author for papers with multiple authors, no support is provided. This is a further disincentive for authors not to claim the TUBITAK support for papers that appear in journals with low impact factors or article influence scores.

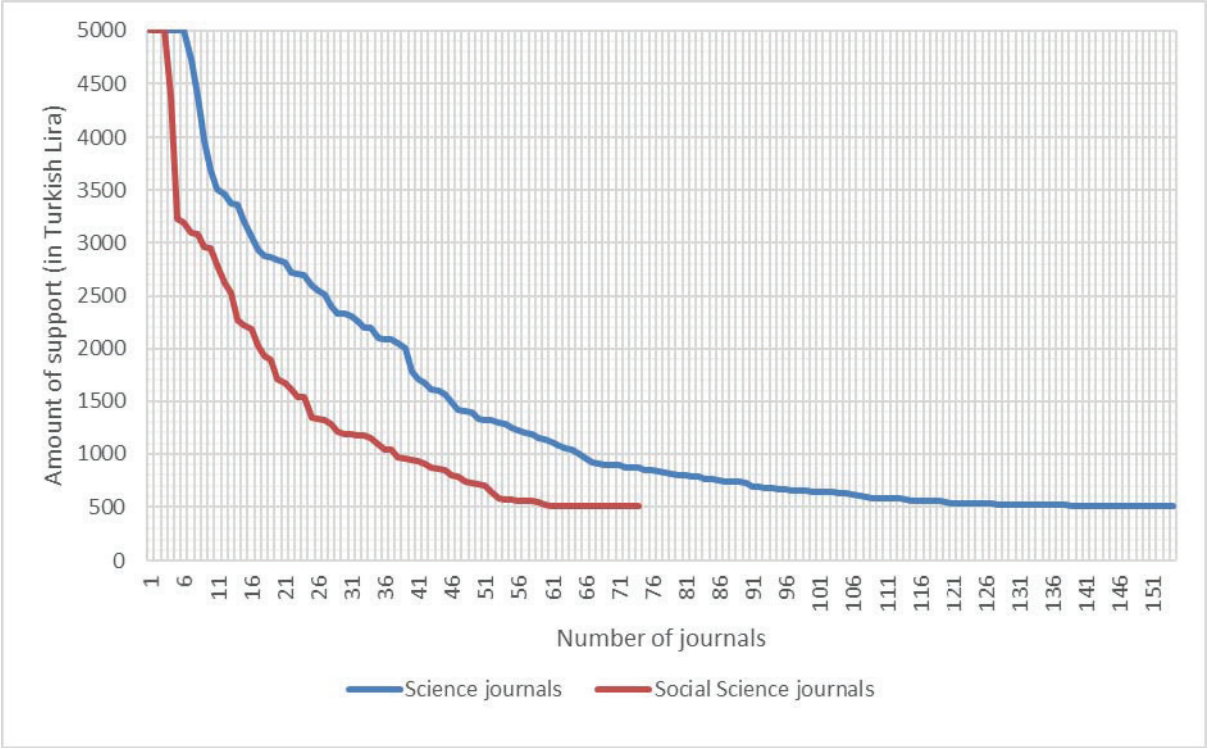


Figure 5. The amount of TUBITAK support for Science and Social science journals in 2014.

As we explained earlier, TUBITAK classified the second half of journals in Science disciplines listed in JCR under Group "C" and provided minimum support (433.00 TL per article) for these journals. (For Social Science disciplines, the second half of journals were divided into two: the top 40% of them being labeled as Group "C" and the remaining 10% as Group "D". Later, TUBITAK stopped supporting the authors of papers publishing in journals under Group "C" in Sciences (i.e., the last 50% of journals) and Group "D" in Social Sciences (i.e., the last 10% of journals) (UBYT Uygulama, 2012). As Group C Science journals constituted about one third of all journals supported in 2012, we wanted to see if they get supported after the policy changes in 2013 and 2014. Our sample included 77 Group C Science journals (one third of all sampled journals) (Table 1). It appears that all of them got supported both in 2013 and 2014. However, the overwhelming majority of them received very little support. As mentioned earlier, the 2013 algorithm was based on five-year JIFs and cited half-lives whereas the 2014 algorithm was based on article influence scores. Recall that the amount of support was increased almost four times starting from 2013. If TUBITAK were to continue supporting Group C Science journals, the amount would have been equal to 1,665.00 TL. Yet, the number of Group C Science journals receiving 1,665.00 TL (or higher) support

was only 2 in 2013 and 5 in 2014. The average amount of support in 2013 and 2014 were 701.00 TL (median=564.00 TL) and 770.00 TL (median=577.00 TL), respectively.

As JIFs and article influence scores are both based on the number of citations, it is not that surprising to see that journals that performed poorly in 2012 did so, too, in 2013 and 2014. What is surprising to see though is that TUBITAK seems to have nullified its earlier decision of not supporting Group C Science journals. A very few of those journals performed differently in 2013 and 2014 when new algorithms were used.

Discussion and Conclusion

It appears that the two algorithms used by TUBITAK in 2013 and 2014 are not that different from each other after all, even though the former was based on Thomson Reuters' JIFs and cited half-lives and the latter on article influence scores (AIS). However, as mentioned earlier, AIS is the most stable indicator and the average influence of journals can therefore be comparable across disciplines (Franceschet, 2010; Arendt, 2010). JIFs and AISs are highly correlated with each other and papers published in high impact journals usually have high AISs (Arendt, 2010; Rousseau & STIMULATE 8 Group, 2009). Arendt (2010) examined the relationship between the two metrics using 5,900 journals listed in JCR Science Edition (2007) and found that both JIFs and AISs vary by discipline. Moreover, the correlation between the two metrics was quite high (Pearson's r (172) = .896) and statistically significant ($p < .001$). Arendt (2010) cautioned that these two metrics should not be used formulaically for research assessment and for ranking scientific papers, authors or institutions.

This advice should be taken into account by TUBITAK as well. As the algorithm based on AIS is more stable and does not vary that much by scientific disciplines (Arendt, 2010; Franceschet, 2010), its use should be monitored closely by TUBITAK to see if it merits further refinement.

The support to journals in the lower end of the scale should be discontinued. Having decided in 2012 to discontinue support to Group C Science journals, it is not clear why TUBITAK reversed its decision the following year without monitoring how these journals performed with the new algorithms used in 2013 and 2014. In fact, the performance of all journals should be monitored to fine-tune the algorithms used.

TUBITAK is of the opinion that its support program caused to increase the number of scientific publications over the years. Turkey has indeed performed very well and became the 18th country in the world in terms of the number of scholarly papers published in ISI-indexed journals. However, the positive correlation between the amount of support provided by TUBITAK and the number of papers with Turkish affiliations is not a strong argument in and of itself[8] to justify the continuance of the support program because correlation does not necessarily mean causation. The existing support to papers published in low impact journals could very well be the main cause of this positive correlation. This merits further research because TUBITAK support does not seem to have encouraged the authors to publish in more prestigious journals.

In conclusion, bibliometric performance measures alone are not the sole criteria for research assessment and, as the Board of Directors of IEEE recently recommended, they "**should be applied only as a collective group (and not individually)**" (IEEE, 2013, original emphasis).

References

2014 yılı UBYT Programı teşvik miktarları hesaplama yöntemine dair bilgi notu (A note on the calculation of the amounts of support in TUBITAK's UBYT Program of 2014). (2014). Retrieved, January 25, 2015, from http://ulakbim.tubitak.gov.tr/sites/images/Ulakbim/ubyt_2014_hesap.pdf.

[8] The number of universities and researchers in Turkey have also increased tremendously during this period.

Arendt, J. (2010). Are article influence scores comparable across scientific fields? *Issues in Science and Technology Librarianship*, No. 60. Retrieved, January 25, 2015, from http://www.istl.org/10-winter/refereed2.html.

Balaram, P. (2013, May 25). Research assessment: Declaring war on the impact factor. *Current Science*, *14*(10), 1267-1268.

Batmaz, A. (2013, June 14). Türkiye'de bilim üretimi ve arkeoloji (Science production in Turkey and Archaeology). *Cumhuriyet Bilim ve Teknoloji*, (1369), 18. Retrieved, January 25, 2015, from http://www.arkeolojikhaber.com/?p=2569.

Bergstrom, C.T., West, J.D., & Wiseman, M.A: (2008). The Eigenfactor™ metrics. *The Journal of Neuroscience*, 28(45): 11433-11434. Retrieved, January 26, 2015, from http://www.jevinwest.org/Documents/Bergstrom_J_neurosci_2008.pdf

Bollen J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE, 4*(6), e6022. Retrieved, January 26, 2015, from doi:10.1371/journal.pone.0006022.

Browman, H.I. & Stergiou, K.I. (2008). Factors and indices are one thing, deciding who is scholarly, why they are scholarly, and the relative value of their scholarship is something else entirely (editorial). *Ethics in Science and Environmental Politics*, *8*, 1-3. Retrieved, January 26, 2015, from http://www.int-res.com/articles/esep2008/8/e008p001.pdf.

Casadevall, A. & Fang, F.C. (2014). Causes for the persistence of impact factor mania. *mBio*, 5(2). Retrieved, June 25, 2014, from http://mbio.asm.org/content/5/2/e00064-14.full.pdf.

Franceschet, M. (2010). Journal influence factors. *Journal of Informetrics*, 4(3), 239-248. Retrieved, January 26, 2015, from https://users.dimi.uniud.it/~massimo.franceschet/publications/joi10b.pdf

Guerrero, R. (2001, August). Misuse and abuse of journal impact factors. *European Science Editing*, 27(3): 58-59.

Hecht, F., Hecht, B.K. & Sandberg, A.A. (1998, July 15). The journal "impact factor": A misnamed, misleading, misused measure. *Cancer Genetics*, *104*(2), 77-81.

IEEE. (2013, September 9). Appropriate use of bibliometric indicators for the assessment of journals, research proposals, and individuals. Retrieved, April 7, 2015, from http://www.ieee.org/publications_standards/publications/rights/ieee_bibliometric_statement_sept_2013.pdf

Jawaid, S.A. (2014). Despite misuse and abuse, journal impact factor will retain its impact and won't fade away soon (editorial). *Journal of Postgraduate Medical Institute*, *28*(1), 1-4.

Kotur, P.F. (2013, August 10). Impact factor – the misnamed, misleading and misused measure of scientific literature. *Current Science*, *105*(3), 289-290. Retrieved, January 26, 2015, from http://www.currentscience.ac.in/Volumes/105/03/0289.pdf.

Lawrence, P.A. (2008). Lost in publication: how measurement harms science. *Ethics in Science and Environmental Politics*, *8*, 9-11. Retrieved, January 26, 2015, from http://www.int-res.com/articles/esep2008/8/e008p009.pdf.

Marks, M.S., Marsh, M., Schroer, T.A., & Stevens, T.H. (2013, June). Misuse of journal impact factors in scientific assessment (editorial). *Traffic*, *14*(6), 611-612. Retrieved, January 26, 2015, from http://onlinelibrary.wiley.com/doi/10.1111/tra.12075/full.

Marx, W. & Bornmann, L. (2013). Journal Impact Factor: "the poor man's citation analysis" and alternative approaches. *European Science Editing*, 39(2), 62-63. Retrieved, January 25, 2015, from http://www.ease.org.uk/sites/default/files/aug13pageslowres.pdf.

San Francisco Declaration on Research Assessment: Putting science into the assessment of research. (2012, December 16). Retrieved, January 25, 2015, from http://am.ascb.org/dora/files/SFDeclarationFINAL.pdf.

Simons, K. (2008, October 10). The misused impact factor. *Science*, 322, 165. Retrieved, January 26, 2015, from https://java-srv1.mpi-cbg.de/publications/getDocument.html?id=8a8182da238c39d10123955066a00100.

Rousseau, R. & STIMULATE 8 Group. (2009). On the relation between the WoS impact factor, the eigenfactor, the SCImago journal rank, the article influence score and the journal index. (Technical report). Retrieved, January 26, 2015, from http://eprints.rclis.org/16448.

Seglen, P.O. (1997). Why impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314, 497-502.

TÜBİTAK Türkiye Adresli Uluslararası Bilimsel Yayınları Teşvik Programı Uygulama Esasları (TUBITAK's Principles to Support International Scientific Publications with Turkish Affiliations). (2013). Retrieved, January 25, 2015, from http://www.tubitak.gov.tr/sites/default/files/esaslar_v_2_vers.2_2.pdf.

Todd, P.A., & Ladle, R.J. (2008). Hidden dangers of a 'citation culture'. *Ethics in Science and Environmental Politics*, *8*(1), 13-16. Retrieved, January 26, 2015, from http://www.int-res.com/articles/esep2008/8/e008p013.pdf.

Tonta, Y. (2014). Use and misuse of bibliometric measures for assessment of academic performance, tenure and publication support. *Metrics 2014: Workshop on Informetric and Scientometric Research (SIG/MET). 77th Annual Meeting of the Association for Information Science and Technology, October 31-November 5, 2014, Seattle, WA*. Retrieved, January 26, 2015, from http://yunus.hacettepe.edu.tr/~tonta/yayinlar/tonta-asist2014-seattle-sig-met-misuse-of-bibliometric-indicators.pdf

Yaltırak, C. (2014, June 21). TÜBİTAK yayın teşvik sistemini değiştirmeli! (TUBITAK has to change its publication support system!) *Cumhuriyet Bilim ve Teknoloji*, (1409), 18.

What's Special about Book Editors? A Bibliometric Comparison of Book Editors and other Flemish Researchers in the Social Sciences and Humanities

Truyken L.B. Ossenblok[1] and Mike Thelwall[2]

[1] *Truyken.Ossenblok@uantwerpen.be*
Centre for Research & Development Monitoring (ECOOM), Faculty of Political and Social sciences, University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium) (corresponding author)

[2] *M.Thelwall@wlv.ac.uk*
Statistical Cybermetrics Research Group, Faculty of Science and Engineering, University of Wolverhampton (United Kingdom)

Abstract
This paper examines the bibliometric characteristics of book editors and non-editors, focussing on gender, career stage, number of publications and collaboration practices. The data consist of 8970 Flemish affiliated researchers with at least one publication between 2000 and 2011 in the comprehensive Flemish academic bibliometric database (VABB-SHW). The analysis shows that most book editors are established male researchers while most non-editors are non-established male researchers. Moreover, males are more likely to be editors than are females. Half of the established editors edit more than 1 book, in contrast to only a small number of non-established editors. Overall, book editors publish more than non-editors, but, when controlling for career stage, book editors publish even more book chapters and monographs than do non-editors. Although editors are highly collaborative while editing a book, no significant differences were found in the number of collaborative articles, monographs, book chapters and proceedings written by editors and non-editors.

Conference Topic
Country-level studies

Introduction
Bibliometric studies have demonstrated the importance of books to many disciplines belonging to the Social Sciences and Humanities (SSH). There is a growing consensus among researchers and policy-makers that scholarly publication patterns and their underlying research cultures cannot be adequately analyzed without the inclusion of books (Hicks, 2004; Nederhof, 2006; Sivertsen, 2009). So far, this insight has resulted in a limited number of studies on books in the SSH, mostly focused on scholarly monographs. A book publication type that has received far less attention is the edited book. Editing a book often appears to be undervalued for academic careers (Edwards, 2012) but, in Flanders, from 2010 onwards, edited books are included in the funding system (Ossenblok & Engels, 2015) which gives incentives to individual researchers to take on book editorships (Gläser & Laudel, 2007).

We define an edited book here as a collection of chapters written by different authors, gathered and harmonized by one or more editors (Ossenblok & Engels, 2015) and identifiable by the presence of an ISBN. Edited books have been shown to comprise a sizeable share of the publication output of many SSH disciplines, especially in the humanities (Leydesdorff & Felt, 2012; Nederhof, 2006). In Flanders, the Northern Dutch-speaking part of Belgium, about 2% of all peer reviewed publications in the SSH are edited books, with up to 6% in Linguistics, Literature and Theology (Engels, Ossenblok, & Spruyt, 2012). Compared to monographs, edited books have significantly higher citation rates, especially in social science disciplines (Torres-Salinas, Robinson-Garcia, Cabezas-Clavijo, & Jiménez-Contreras, 2013).

This paper presents a bibliometric case study of the characteristics of book editors, for which, to the best of our knowledge, no previous studies exist. We analyse comprehensive

publication data and present four elements of a general profile of these scholars: career stage; gender; number of publications; and collaboration practices. We hypothesise that scholars tend to edit books only when they are established researchers that are at the forefront of scholarly collaboration.

Data and methods

The data set consists of 8970 authors affiliated with one of the five Flemish universities and who have published a minimum of one peer reviewed publication in the period 2000-2011: a journal article, monograph, edited book, book chapter and/or proceedings paper included in the VABB-SHW (for a full account see: Engels et al., 2012). Because of the use of this database for funding in Flanders, this database appears to be close to exhaustive in its coverage of Flemish research. In addition to the data found in the VABB-SHW, we also determined the gender of all authors. For this, two researchers independently divided all unambiguous first names into two groups: male names and female names. The remaining authors were looked up on the internet, resulting in an additional 1462 gender matches.

A comparison was made between two subsets: book editors (researchers who have published a minimum of 1 peer reviewed edited book in the period under study); and all other researchers, called here non-editors although they may be journal editors or may have edited books during other periods of time. Furthermore, we differentiated between established and non-established researchers. Established researchers are defined in this study as having a total of 12 publications or more and at least one publication in a minimum of 6 different years in the period 2000-2011. These heuristics were chosen after inspection of typical properties of authors in the database. Of course, non-established researchers may have many publications within up to five years, may have a prolific consistent set of outputs before or after the period analysed, or may have many outputs of a type not recorded in the database (e.g., book reviews, performances). Nevertheless, the criteria seem to be effective at differentiating between two sets of researchers, the first of which contains researchers that can reasonably be thought of as being established and the second of which probably contains a much lower proportion of established researchers. Cramer's V was used to measure the strength of the correlation between the different subsets, resulting in a number between 0 (no association) and 1 (maximum association). In addition the Mann-Whitney U test, a rank-based nonparametric test, was used to determine whether there were differences between the subsets on the different characteristics under study, using $p=0.05$ as the threshold for statistical significance.

Results

Career stage and gender

Figure 1 shows the proportion and number of established and non-established, male and female editors and non-editors in our study. In total, 676 (7.5%) researchers had published one or more edited books (i.e., editors), and 8970 (92.5%) researchers had not published an edited book (i.e., non-editors). Figure 1 demonstrates that 55.9% (n=378) of editors are established researchers whereas 13.3% (n=1102) of non-editors are established researchers. Furthermore, 74.3% (n=502) of editors are male whereas to 58.9% (n=4883) of non-editors are male. In addition, 9.3% of all male researchers are editors and 4.9% of all female researchers are editors. Furthermore, 25.5% of all established researchers are editors, whereas only 4% of all non-established researchers are editors. Altogether, 43.5% (n=294) are male established editors, 30.8% (n=208) are male non-established editors, 13.3% (n=90) are female non-established editors and 12.4% (n=84) are female established editors. Different proportions occur in the subgroup of the non-editors where 49.1% (n=4070) are male non-

established researchers, 37.6% (n=3122) are female non-established researchers, 9.8% (n=813) are male established researchers and 3.5% (n=289) are female established researchers.

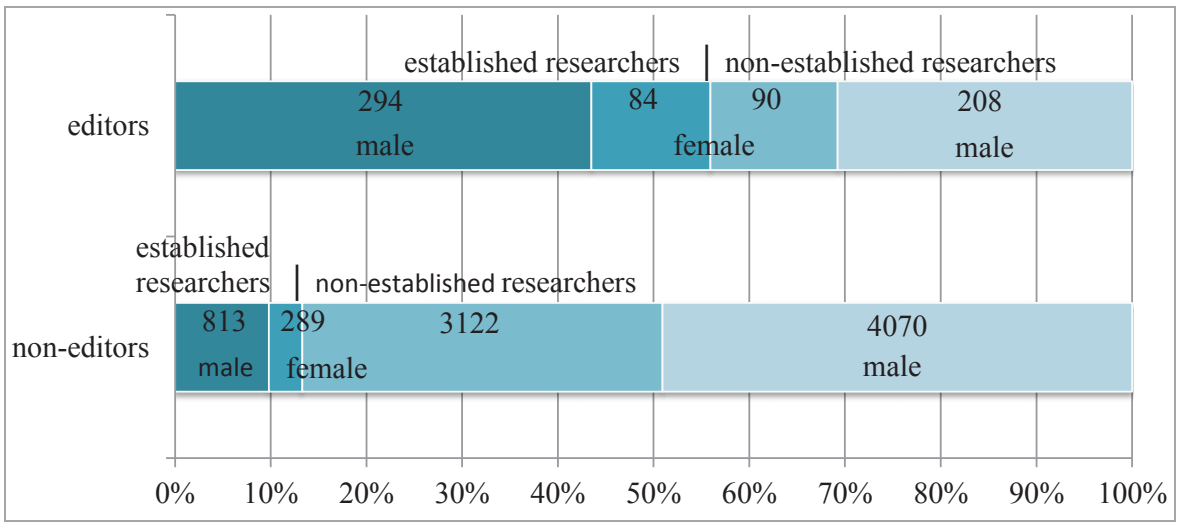


Figure 1: Share and number of established and non-established, male and female editors and non-editors (2000-2011).

There is a moderate association (Cramer's V=0.134; p=.000) between gender and career status overall (see also Figure 1). However, when looking at the different subsets, the correlation between gender and career status is stronger within the subset of non-editors (Cramer's V=0.119; p=.000) than within the subset of editors (Cramer's V=0.091; p=.000). Overall, though, career status has a stronger association with editorship than with gender (resp. Cramer's V=0.304; p=.000 and Cramer's V=0.083; p=.000). Therefore in the rest of this study we will focus on differences in career status rather than gender.

Number of publications

Table 1 shows the mean and median number of edited books, articles, book chapters, monographs and proceedings for all editors and non-editors. In addition, the table displays the difference between non-established and established researchers. Overall, editors publish on average a greater number of all publication types than do non-editors. However, established non-editors publish on average more articles than do established editors. Mann-Whitney U tests were run to test for differences in numbers of publications between editors and non-editors for all publication types except edited books. The distributions of all the publication types for editors and non-editors and for established and non-established researchers were visually similar. The differences between editors and non-editors are statistically significant for all publication types (all p=.000). When comparing established editors and established non-editors, all differences are significantly different (p=.000) except for the numbers of proceedings (p=.138). When comparing non-established editors with non-established non-editors, the differences for articles (p=.119) and proceedings (p=.911) were not significantly different, whereas the differences for book chapters and monographs were (both p=.000).

Furthermore, Table 1 shows that the median of numbers of edited books differ between established and non-established editors. Non-established editors are more likely to have (co−)edited one book whereas established editors are more likely to have more than 1 edited book. More specifically, 83.2% of all non-established editors have one edited book, whereas 48.4% of all established editors have one edited book, 24.3% have two edited books and 27.2% have three or more edited books.

Table 1: The mean and median (med) number of edited books, articles, book chapters, monographs and proceedings for all established and non-established editors and non-editors (2000-2011).

		edited books		*articles*		*book chapters*		*monographs*		*proceedings*	
		mean	*med*	*mean*	*med*	*mean*	*med*	*mean*	*med*	*mean*	*med*
Editor	established researcher	2.17	2	20.62	14	7.92	6	0.59	0	0.97	0
	non-established researcher	1.22	1	2.93	2	2.31	2	0.16	0	0.17	0
	total	1.76	1	12.82	7	5.44	4	0.40	0	0.62	0
non-editor	established researcher	-	-	26.00	18	1.57	1	0.22	0	0.82	0
	non-established researcher	-	-	3.00	2	0.29	0	0.03	0	0.16	0
	total	-	-	6.06	2	0.46	0	0.05	0	0.24	0

Collaboration practices

For both editors and non-editors, Figure 2 shows the proportion of their edited books, articles, book chapters, monographs and proceedings that have been published in collaboration (i.e., multiple authored versus single authored publications). Editors collaborate the most while editing a book (90.3%; n=1827), which is in agreement with previous research demonstrating that most edited books are co-edited (Ossenblok & Engels, 2015). Furthermore, established editors collaborate more than non-established editors for all publication types under study (p=.000). Altogether, though, non-editors seem to collaborate more for articles, book chapters, monographs and proceedings than do editors. Mann-Whitney U tests were run to determine if editors and non-editors differ significantly in their numbers of collaborative publications. The different distributions of all the publication types, except edited books, were visually similar. The numbers of collaborative publications of editors and non-editors were statistically significantly different for book chapters and monographs (both p=.000) but not for articles (p=.282) and proceedings (p=.116). Thus, non-editors collaborate significantly more in book chapters and in monographs than do editors. In addition, when comparing non-established editors with non-established non-editors, no significant difference in the number of collaborative publications was found for all publication types separately (but p=.000 for articles, monographs and book chapters; p=.005 for proceedings). However, when distinguishing between established editors and non-editors, the differences are significant for all publication types separately (p=.000) except for proceedings (p=.208). In sum, established non-editors collaborate more than do established editors for articles, monographs and book chapters.

Discussion and conclusions

Within a comprehensive collection of Flemish affiliated authors' publications for 2000-2011, this paper demonstrates that 7.5% of the authors have edited one or more books, that more than half of the book editors are established researchers, and that 3 in 4 editors are male.

Female researchers are less likely to be established than are male researchers and this difference is more pronounced for non-editor than for editors. As career status in this study is defined through numbers of publications and publication years, these findings confirm previous findings that male researchers are often more productive than are their female colleagues (Larivière et al., 2013; Puuska, 2010).

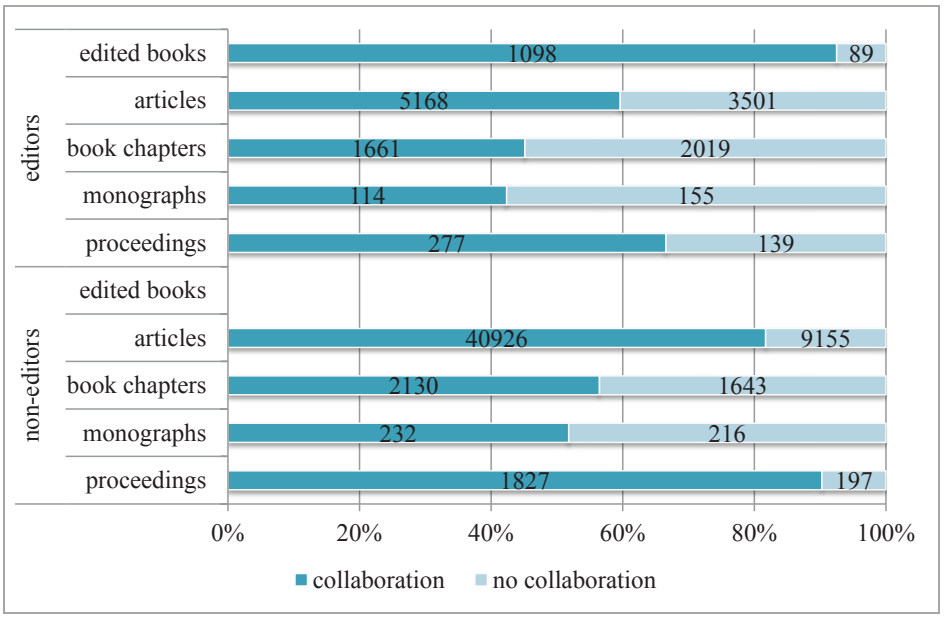


Figure 2: The proportion of collaborative and solo publications for all editors and non-editors by publication type.

Editors tend to publish significantly more articles, book chapters, monographs and proceedings than do non-editors. However, the differences are not statistically significant between the average number of proceedings of established editors and non-editors and between the average number of articles and proceedings of non-established editors and non-editors. Most non-established editors published only 1 edited book in the period under study, whereas more than half of the established editors published 2 or more edited books. This might be due to the need for a large network and good networking skills for gathering contributions from individual chapter authors for an edited book (Edwards, 2012; Thomas & Hrebenar, 1993). We therefore expected editors to be more collaborative than were non-editors for all publication types, but although 9 out of 10 editors collaborated while editing a book, non-editors collaborated significantly more for book chapters and monographs than did editors. Furthermore, no significant difference was found in the number of collaborative articles and proceedings between editors and non-editors. As edited books are more common in humanities disciplines (Engels et al., 2012) and the humanities have been known to collaborate less than the social sciences in articles and book chapters (Ossenblok, Verleysen, & Engels, 2014), the low level of collaboration of editors might be due to them tending to be humanities scholars.

Overall, the findings offer a first insight into some of the bibliometric characteristics of editorship. Future research will focus on disciplinary differences in collaboration practices between book editors and non-editors. A more detailed analysis of collaboration practices will involve not only the number of collaborative publications, but also the number of co-authors. As previous research (Ossenblok & Engels, 2015) has shown, edited books are often published in English, and so the study of the number of international co-authors and co-editors will broaden our knowledge about the international nature of the collaboration network of the editors. In addition, links between book editors and their chapter authors

would provide a more complete picture of the collaboration practices of book editors. This would contribute greatly to our understanding of collaborative practices in the SSH.

Acknowledgments

The authors thank their colleagues Nele Dexters, Tim Engels, Raf Guns and Frederik Verleysen for their useful comments.

References

Edwards, L. (2012). Editing academic books in the humanities and social sciences: Maximizing impact for effort. *Journal of Scholarly Publishing, 44,* 61-74.

Engels, T. C. E., Ossenblok, T. L. B., & Spruyt, E. H. J. (2012). Changing publication patterns in the social sciences and humanities, 2000-2009. *Scientometrics, 93,* 373-390.

Gläser, J. & Laudel, G. (2007). Evaluation without evaluators. In R.Whitley & J. Gläser (Eds.), *The changing governance of the sciences. The advent of research evaluation systems* (pp. 127-151). Dordrecht: Springer Science.

Hicks, D. (2004). The four literatures of social science. In H.F.Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative Science and Technology Research: The use of publication and patent statistics in studies of S&T systems* (pp. 473-496). Dordrecht: Kluwer Academic.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature, 504,* 211-213.

Leydesdorff, L. & Felt, U. (2012). Edited volumes, monographs and book chapters in the Book Citation Index (BKCI) and Science Citation Index (SCI, SoSCI, A&HCI). *Journal of Scientometric Research, 1,* 28-34.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66,* 81-100.

Ossenblok, T. L. B. & Engels, T. C. E. (2015). Edited books in the social sciences and humanities: Characteristics and collaboration analysis. *Scientometrics, Under review*.

Ossenblok, T. L. B., Verleysen, F. T., & Engels, T. C. E. (2014). Co-authorship of journal articles and book chapters in the social sciences and humanities (2000-2010). *Journal of the American Society for Information Science & Technology, 65,* 882-897.

Puuska, H.-M. (2010). Effects of scholar's gender and professional position on publishing productivity in different publication types. Analysis of a Finnish university. *Scientometrics, 82,* 419-437.

Sivertsen, G. (2009). Publication patterns in all fields. In F.Aström, R. Danell, B. Larsen, & J. W. Schneider (Eds.), *Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th birthday* (pp. 55-60). ISSI.

Thomas, C. S. & Hrebenar, R. J. (1993). Editing multiauthor books in political science: Plotting your way through an academic minefield. *Political Science and Politics, 26,* 778-783.

Torres-Salinas, D., Robinson-Garcia, N., Cabezas-Clavijo, Á., & Jiménez-Contreras, E. (2013). Analyzing the citation characteristics of books: Edited books, book series and publisher types in the Book Citation Index. *Scientometrics, 98,* 2113-2127.

Scientific Cooperation in the Republics of Former Yugoslavia Before, During and After the Yugoslav Wars

Dragan Ivanović[1], Miloš M. Jovanović[2] and Frank Fritsche[3]

[1] *dragan.ivanovic@uns.ac.rs*
University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, Novi Sad (Serbia)

[2] *milos.jovanovic@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis, Appelsgarten 2, 53879 Euskirchen (Germany)

[3] *frank.fritsche@int.fraunhofer.de*
Fraunhofer Institute for Technological Trend Analysis, Appelsgarten 2, 53879 Euskirchen (Germany)

Abstract
This paper presents an analysis of scientific research output of the republics of former Yugoslavia for the period 1970-2014. Thomson Reuters' Web of Science (WoS) database was used for data acquisition and 223 135 publications have been analyzed. The Yugoslav Wars were ethnic conflicts fought from 1991 to 1999 on the territory of former Yugoslavia, which accompanied the breakup of the country, and today, each republic of former Yugoslavia is an independent country, as well as the province of Kosovo. Results of the analysis are represented by four figures depicting cooperation networks between former Yugoslav republics and the province of Kosovo for the periods before the Yugoslav wars (from 1970 until 1990), during the wars (from 1991 until 1999), in the first decade after the wars (from 2000 until 2009), and in the last 5 years (from 2010 until 2014). The impact of the wars on scientific cooperation in the republics has been studied.

Conference Topic
Country-level studies

Introduction
The Socialist Federal Republic of Yugoslavia (SFRY) was established in 1946, after World War II. It was divided into six Republics (Serbia, Croatia, Slovenia, Bosnia & Herzegovina, Macedonia and Montenegro) and two autonomous provinces on the north and south of Serbia (Vojvodina and Kosovo). The Yugoslav Wars were ethnic conflicts fought from 1991 to 1999 on the territory of SFRY, which accompanied the breakup of the country. Today, each republic of former SFRY is an independent country. A Kosovo declaration of independence was adopted on 17 February 2008 by the Assembly of Kosovo, but the legality of this declaration have been disputed by the Serbian Government and other countries (e.g. the Russian Federation and China). This paper analyses the scientific cooperation in the republics of former SFRY and the province of Kosovo before, during and after the Yugoslav wars. The purpose of this analysis is to answer how the Yugoslav wars and social crises during and around those wars affected scientific productivity and scientific cooperation in these republics and whether this cooperation has recovered 15 years after the wars.

Related work
Bibliometric analysis is a useful method for characterising scientific research (Moravcsik, 1985; Fu & Ho, 2013) and this method can be used for analysing scientific cooperation in different countries and regions (Leta & Chaimovich, 2002; Wagner & Leydesdorff, 2005; Ho et al., 2010). Citations of a publication are not a direct measure of quality and significance, but they reflect the visibility and impact of the publication on the scientific community (Furlan & Fehlings, 2006; Baltussen & Kindler, 2004). The number of times an article was cited correlates significantly with the number of authors and the number of institutions

involved in collaboration (Figg et al., 2006) and highly cited articles are usually authored by a large number of scientists, often involving international collaboration (Aksnes, 2003). Thus, scientific cooperation is important for the further development of world science and for the further economic development of a region or country.

The impact of social aspects, economic and social crises, political crises and wars on scientific cooperation in some regions has already been studied. For example, de Bruin and colleagues (1991) stated that the cooperation between the Gulf States and former western and eastern bloc has been strongly affected by political crises, which culminated in the Operation Desert Storm in 1990. There are also studies that deal with the countries of the former SFRY like Lewison and Igic (1999), Igic (2002), Lukenda (2006), Đukić et al. (2011) and Kutlača et al. (2015). Furthermore, Jovanović et al. (2010). analysed the publications and cooperation between the republics of former SFRY and the province of Kosovo is analyzed for the years from 1970 until 2007. The authors found that the Yugoslav wars had a severe impact on the cooperation networks of former SFRY republics. Furthermore, they also found that the process of recovery started with the ending of the conflicts, but that scientific cooperation recovered faster in some of those republics. The current paper revisits the data and methods of this study by analysing publications of former SFRY republics and the province of Kosovo from 1970 until 2014, thus broadly extending the database and improving the methodology. Thus, the purpose of this analysis is to answer whether scientific cooperation in all former SFRY republics is fully repaired 15 years after the Yugoslav wars or whether the interpretation of the findings of the 2010 study has to be reformulated.

Methodology

Similar to the 2010 study, Thomson Reuters' Web of Science (WoS) database was used for data acquisition. This time, however, the Arts & Humanities Citation Index Expanded was not covered, because the authors' institutions did not have access. But in addition to the Science Citation Index Expanded (SCIE) and the Social Science Citation Index (SSCI) (which were also used in 2010), both conference proceedings citation indexes (Science and Social Sciences) were covered by the search queries. This was done in order to get a more complete coverage of the publication output of the former Yugoslav countries. Again similar to 2010, the search queries consisted of the names of cities from the former Yugoslav countries, since before 1990 all successor states belonged to SFRY. In 2010, a total of 133 city and town names were used in the search queries (including synonyms of city names). For the current study, we also used search queries that consisted of the country names (Yugoslavia and all successor states) in order to find city and town names (and synonyms), which were missing in our city search queries. In addition to that, the maximum number of 50 search arguments in WoS (still existing in 2010) is no longer limited which meant that we were able to use much longer search queries for the current study. Because of this, the new search query included 769 city and town names along with synonyms, misspellings etc. This has led to a much broader database and a better allocation of publications to their respective states, in comparison to the data used in 2010. In 2010, the data set consisted of 103 963 publications (for the years 1970 to 2007), the current study has 121 602 publications for this time period (20% more) plus 101 533 publications for the years 2008 to 2014, which brings the complete data set to a total of 223 135 publications. We rechecked whether these publications were all from the correct countries by using WoS exclude tool and removing all publications from the seven Yugoslav successor states. The remaining publications consisted of around 1% of the total data set and manual checks of these publications have shown that most of these were still relevant but wrongly indexed (for example publications from Kosovo which were attributed to Albania). This leads us to believe that our data set includes all publications from the former SFRY, which can be found in the WoS.

We analysed the data set using a proprietary bibliometry toolbox (programmed at Fraunhofer INT) and the following measures and method: (1) Absolute number of publications for each state (2) Absolute number of cooperation for each state and (3) Visualization of the Yugoslav cooperation network. In our future studies, we will add measures like Salton's measure and others.

Results

Results of the analysis are represented by four figures depicting cooperation networks between former Yugoslav republics and the province of Kosovo for the periods before the Yugoslav wars (from 1970 until 1990), during the wars (from 1991 until 1999), in the first decade after the wars (from 2000 until 2009), and in the last 5 years (from 2010 until 2014). Each republic's and the province of Kosovo's publications indexed by WoS have been represented in figures by a circle which size is proportional with the number of publications published by researchers from each respective republic. Lines between those circles represent cooperation of researchers in writing publications and line thickness is proportional with the number of collaborative publications of researchers from two republics whose circles are connected by the line. A cooperation was counted whenever more than one institution that published a paper was located on the territory of the former Yugoslavia and these institutions were not from the same republic. Cooperation between three or more republics are quite rare. These were enumerated as a set of multiple bilateral cooperation.

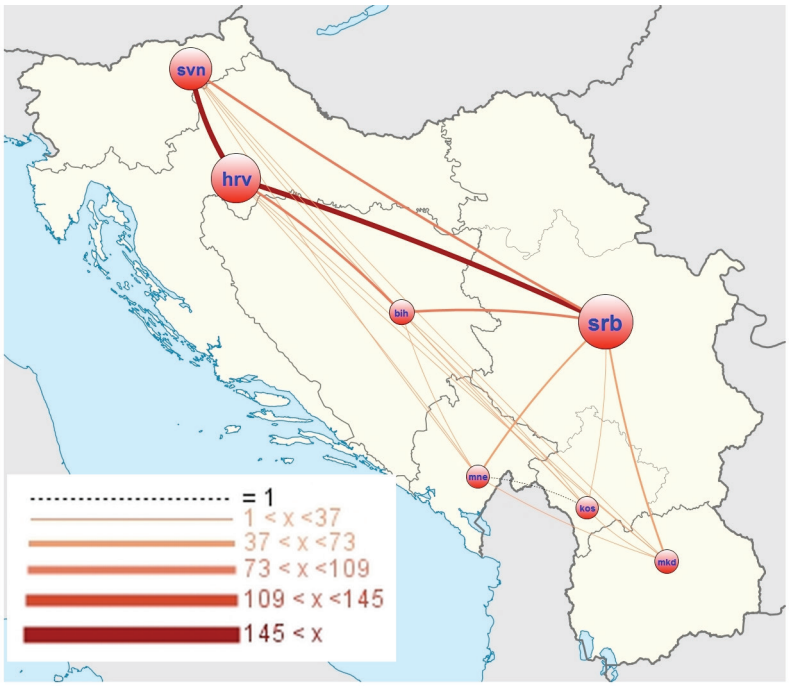


Figure 1. Visualisation of the cooperation network for 1970-1990 (before Yugoslav wars).

Figure 1 depicts the cooperation network for the period before the Yugoslav wars. Researchers from Serbia published the highest number of publications before the wars, followed by researchers from Croatia. Those two republics were the most productive republics and cooperated the most in former Yugoslavia. Slovenia, according to the productivity of its researchers and to the cooperation in this period, was in the middle between the groups of "big" republics by scientific productivity (Serbia and Croatia) and the group of "small" republics (Bosnia and Herzegovina, Macedonia, Montenegro and the province of Kosovo). Before the war, the most productive "small" republic was Bosnia and Herzegovina.

The Yugoslav wars started in 1991 and they led to a strong decrease of scientific cooperation in the republics in the 90's. Also, it affected the ratio of scientific productivity between republics during the wars. Figure 2 depicts the cooperation network for the period 1991-1999 which is the period of Yugoslav wars. Before the wars, Serbia was cooperating strongly with Croatia, Slovenia and Bosnia and Herzegovina. The cooperation triangle between Serbia, Croatia and Slovenia almost disappeared in the 90's, as well as the cooperation triangle between Serbia, Croatia and Bosnia and Herzegovina. However, scientific cooperation between Croatia and Slovenia was strengthened in this period. The reason for that is the fact that the conflict between Serbia, Croatia and Bosnia and Herzegovina during the wars was much stronger than the conflict between Croatia and Slovenia. Also, effects of the wars were much less on Slovenian economy than on the economies of other republics. War in Slovenia ended after ten days in 1991. Also, Macedonia remained at peace throughout the Yugoslav wars and declared its independence in September of 1991. Thus, the ratio of scientific productivity of Slovenian and Macedonian researchers in comparison to the other republics researchers had been changed in favour of Slovenia and Macedonia. In this period and in the followings periods Slovenia became a member of the group of "big" republics.

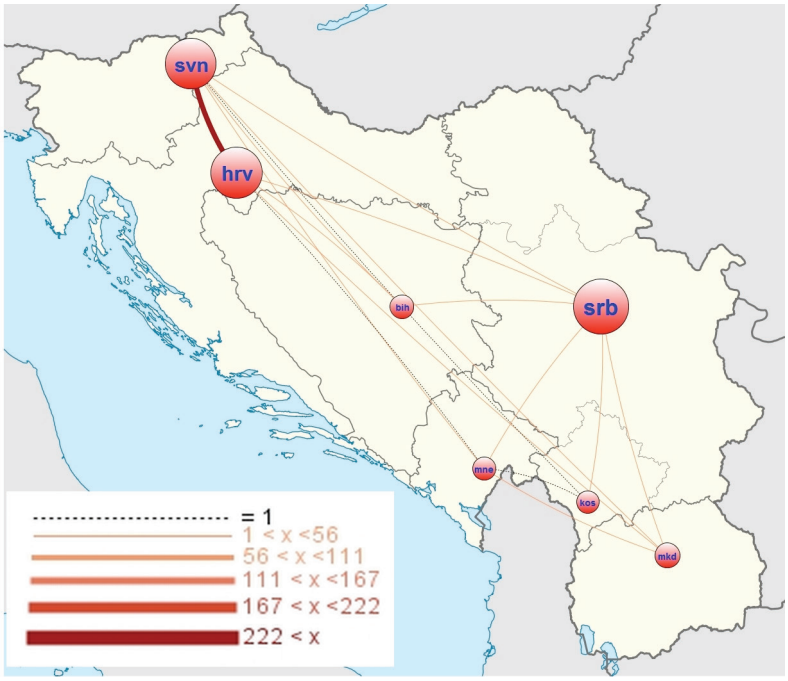


Figure 2. Visualisation of the cooperation network for 1991-1999 (during Yugoslav wars).

Figure 3 depicts the cooperation network for the period 2000-2009 which is the first decade after the Yugoslav wars. Scientific cooperation in this period between Serbia and Slovenia was strengthened again. The cooperation triangle between Serbia, Croatia and Slovenia was not as strong as before the wars (taking into account that the overall publication output increased), but it seems as if this cooperation triangle was resurfacing again.

Figure 4 depicts the cooperation network for the period 2009-2014. In this period Serbia has returned to having the most publications as before the Yugoslav wars. Reasons for this include introduction of a new rulebook for evaluation prescribed by the Ministry of Education, Science and Technological Development of the Republic of Serbia in 2008. That rulebook requires researchers must have articles published in journals in the Web of Science database for the promotion to scientific positions. In addition, the increase in the number of publications was influenced by the fact that several journals based in Serbia have, in recent years, started to be indexed by Web of Science: e. g. Vojnosanitetski Pregled, Archives of

Biological Sciences, Srpski Arhiv Za Celokupno Lekarstvo, Journal of the Serbian Chemical Society, etc. Those journals published a considerable number of articles written by Serbian researchers in the period 2010-2014 (Ivanović and Ho, 2014). The strengthening of the cooperation triangle between Serbia, Croatia and Slovenia started in the period 2000-2009 continues in the last five years. We conclude that this triangle is fully recovered.

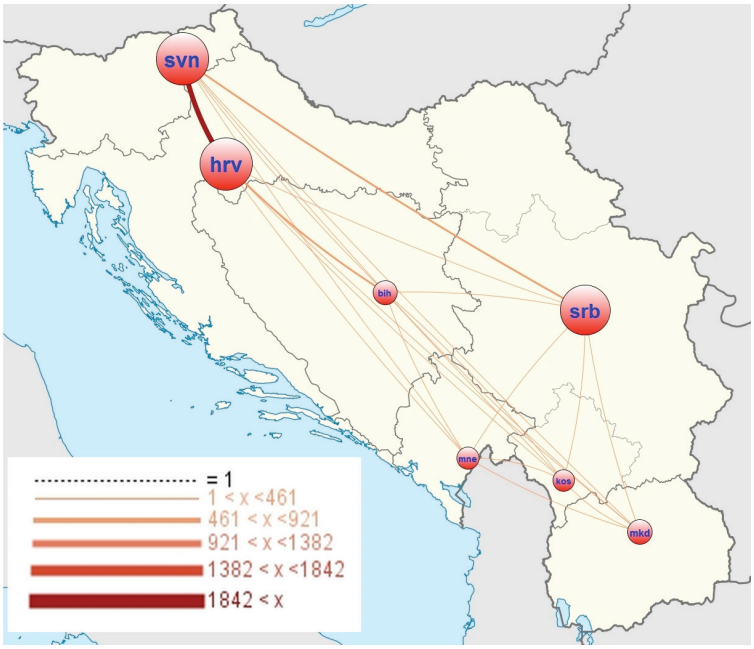


Figure 3. Visualisation of the cooperation network for 2000-2009 (1st decade after Yugoslav wars).

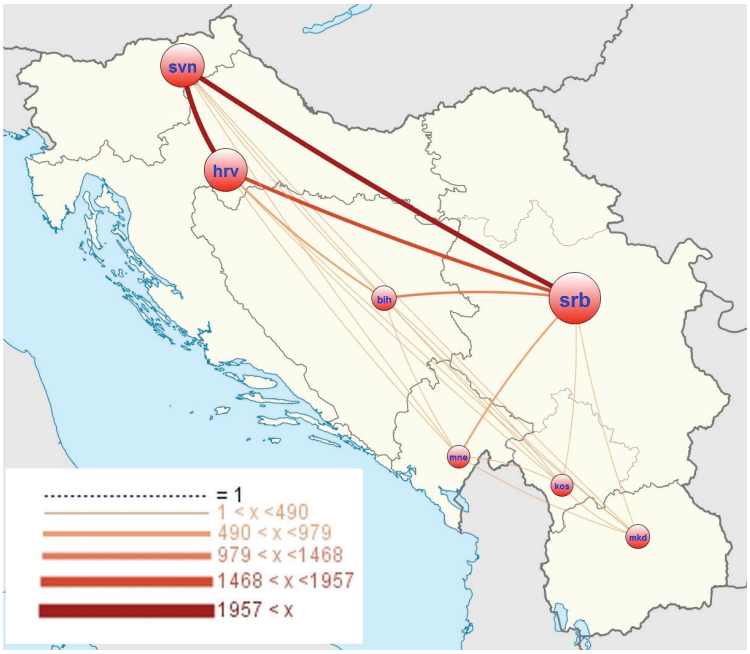


Figure 4. Visualisation of the cooperation network for 2010-2014.

Conclusion

The analysis of scientific-research outputs of the republics of former Yugoslavia for the period 1970-2014 has been presented in this paper. It reveals that civil Yugoslav wars affected the republics' productivities and scientific cooperation in different ways. The most

affected republics by wars and social crisis were Serbia and Bosnia and Herzegovina, while the least affected republics were Slovenia and Macedonia. However, it seems that in the last five years productivity and scientific cooperation look similar as before the Yugoslav wars. This result strengthens the results from the 2010 study. It would seem that old cooperation networks, which were disrupted during the Yugoslav wars, are in place again. However, our data cannot answer the question whether these are the same networks as before (i. e. the same researchers and/or institutions that are cooperating again) or whether new ones have taken the place of the old ones.

The presented results are the first part of our research. We are going to extend our research with following measures and methods: relative number of publications for each state and normalized cooperation score $R_i^{(cs)}$ (as described in Jovanović et al. (2010). Also, we are going to analyse the distribution of collaborative articles per the biggest Universities based in these states.

References

Aksnes, D. W. (2003). Characteristics of highly cited papers. *Research Evaluation*, *12*(3), 159-170.

Baltussen, A., & Kindler, C. H. (2004). Citation classics in anesthetic journals.*Anesthesia & Analgesia*, *98*(2), 443-451.

de Bruin, R. E., Braam, R. R., & Moed, H. F. (1991). Bibliometric lines in the sand. *Nature*, *349*, 559-562.

Đukić, V., Udiljak, N., Bartolić, N., Vargović, M., Kuduz, R., Boban, N., Pećina, M. & Polašek, O. (2011). Surgical Scientific Publication and the 1991-1995 War in Croatia. *Collegium Antropologicum,* 35(2), 409-412.

Figg, W. D., Dunn, L., Liewehr, D. J., Steinberg, S. M., Thurman, P. W., Barrett, J. C., & Birkinshaw, J. (2006). Scientific collaboration results in higher citation rates of published articles. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *26*, 759-767.

Fu, H. Z., & Ho, Y. S. (2013). Independent research of China in Science Citation Index Expanded during 1980–2011. *Journal of Informetrics*, *7*(1), 210-222.

Furlan, J. C., & Fehlings, M. G. (2006). A web-based systematic review on traumatic spinal cord injury comparing the" citation classics" with the consumers' perspectives. *Journal of neurotrauma*, *23*(2), 156-169.

Ho, Y. S., Satoh, H., & Lin, S. Y. (2010). Japanese lung cancer research trends and performance in Science Citation Index. *Internal Medicine*, *49*(20), 2219-2228.

Igić, R. (2002). The influence of the civil war in Yugoslavia on publishing in peer-reviewed journals. *Scientometrics*, *53*(3), 447-452.

Ivanović, D., & Ho, Y. S. (2014). Independent publications from Serbia in the Science Citation Index Expanded: a bibliometric analysis. *Scientometrics*,*101*(1), 603-622.

Jovanović, M. M., John, M., & Reschke, S. (2010). Effects of civil war: scientific cooperation in the republics of the former Yugoslavia and the province of Kosovo. *Scientometrics*, *82*(3), 627-645.

Kutlača, D., Babić, D., Živković, L. & Štrbac, D. (2015). Analysis of quantitative and qualitative indicators of SEE countries scientific output. *Scientometrics,* 102, 247-265

Leta, J., & Chaimovich, H. (2002). Recognition and international collaboration: the Brazilian case. *Scientometrics*, *53*(3), 325-335.

Lewison, G., & Igic, R. (1999). Yogoslav politics,“ethnic cleansing” and co-authorship in science. *Scientometrics*, *44*(2), 183-192.

Lukenda, J. (2006). Influence of the 1991-1995 war on Croatian publications in the MEDLINE database. *Scientometrics*, *69*(1), 21-36.

Moravcsik, M. J. (1985). Applied scientometrics: an assessment methodology for developing countries. *Scientometrics*, *7*(3-6), 165-176.

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research policy*,*34*(10), 1608-1618.

The Brazilian National Impact: Movement of Journals Between Bradford Zones of Production and Consumption

Rogério Mugnaini[1] and Luciano A. Digiampietri[2]

[1] *mugnaini@usp.br*
University of São Paulo, School of Communication and Arts (ECA), Av. Prof. Lúcio Martins Rodrigues 443, 05508-020 São Paulo (Brazil)

[2] *digiampietri@usp.br*
University of São Paulo, School of Arts, Sciences and Humanities (EACH), Av. Arlindo Bettio 1000, 03828-000 São Paulo (Brazil)

Abstract
A specific aspect of the scientific communication in non-English-speaking countries is the need for insertion in the global knowledge flows since a significant part of their publications occurs in national or regional journals. This had led many countries to create alternative ways to assess national journals, allowing a more trustworthy view of the national scientific production. This study aimed to characterize the journals used in the Brazilian scientific production in Web of Science and SciELO, in order to observe the dynamics along five triennia and across the Bradford Zones for both production and consumption in the different areas. Bradford zones showed to be an interesting relative indicator, when applied to evaluative purposes. Especially the joint analysis of production and consumption dimensions can bring a more complete view of the scientific communication system, and this study showed the flows of journals through zones in both dimensions.

Conference Topic
Country-level studies

Introduction
In the last years, several efforts were undertaken by the developing countries in order to improve their position in the global scientific scenario. However, as important as (or even more important than) improve their position is to formulate and implement initiatives for improving their research system, in which the scientific communication plays important role.

A specific aspect of the scientific communication in these countries, mainly in the non-English-speaking ones, is the need for insertion in the global knowledge flows (Ponomariov & Toivanen, 2014), because a significant part of their publications occurs in national or regional journals (Mugnaini et al., 2014). The researchers from these countries, many of them involved in scientific editing, face the dilemma between maximizing efforts to publish in *mainstream journals* and improve the national journals in order to internationalize them – and its negative consequences of such a process (Rego, 2014). Both aspects are typically treated as ways to internationalize the national science, but is this enough (Buela-Casal et al., 2006)? This duality comes from the national science policy, which in one hand valorizes the journals with high Impact Factor (IF) and, on the other hand, tries to attend the clamor for recognition of the national journals (Miranda & Mugnaini, 2013).

This had led many countries to create alternative ways to assess or classify the national journals, allowing a more trustworthy view of the national scientific production, identifying the role of the national journals. In order to do this, some countries built national citations indexes: SciELO Project (Packer et al., 1998), Chinese Science Citation Database (Jim & Wang, 1999), Korea Citation Index (Kim et al., 2013), Citation database for Japanese papers (Negishi et al., 2004) and Islamic World Science Citation Center (Mehrad & Arastoopoor, 2012). Other countries considered this kind of initiative as a solution only for the Humanities and Social Sciences, and are looking for different ways to include the national journals in their scientific evaluation process: Taiwan (Chen, 2004), Spain (Piñeiro & Ricks, 2015),

Poland (Winklawska, 1996), Serbia (Šipka, 2005), among other countries from Eastern Europe (Pajić, 2014) and a project originally european – European Reference Index for the Humanities and the Social Sciences-ERIH PLUS – which currently reaches worldwide.

By the way, despite being considered, national journals are minimally punctuated in comparison to journals indexed in WoS. One of the reasons of this non-recognition is the fact that many of these journals are not peer-reviewed, and, among the ones that are, some present and endogen editorial board (Packer, 2014). These facts explain the non-inclusion of these journals in the most recognized citation databases. Consequently, the commissions of researchers that tread the paths of the national research assessment exercise have to deal with these characteristics as extra factors. On the other hand, the creation of national data sources with defined selection process can be a solution.

The limited insertion of these countries' research in mainstream science finds no echo (Tijssen et al., 2006), since it lacks potential audience (MacRoberts & MacRoberts, 1996), indispensable to a consistent citation analysis. Thus, the evaluation is based strictly on productivity indicators, which impose even bigger challenge to establishing quality criteria. Therefore it became necessary the classification of the journals. A side effect of this is the need, for these researchers who work in a research area with local/regional focus (as typically occurs in Social Sciences and Humanities), to publish a significantly higher number of papers, inflating the entire scholarly communication system (Rego, 2014).

The journals evaluation performed by CAPES in Brazil fit these aspects and have considerably different criteria among the 48 areas (Miranda & Mugnaini, 2013). The most common criteria are (sorted in a decreasing way, according with the assigned importance): citation indicators (JCR Impact Factor, Scopus/SCImago or Google Scholar H-index, SCImago Journal Ranking, or a mix of more than one); indexing in databases with explicit selection criteria (such as Web of Science, Scopus, SciELO, thematic bases - e.g. MEDLINE, or regionals – such as, Redalyc, Latindex) or without explicit selection criteria (e.g. PASCAL); journals characteristics. All the journals where Brazilian researchers published their papers during the preceding triennium are classified. Some journals can receive different classifications from different areas (e.g. Cadernos de Saúde Publica).

Considering this scenario, stands out the need to complement the range of citation indicators for journals classification, providing a consistent view to the national context. In order to fulfill this need, in this paper a nationally recognized base - whose selection process considers explicit criteria – were created aggregating the national scientific production from SciELO and WoS (including the publications bibliographic references). The papers from this base were used to evaluate the national production and the references to evaluate the consumption. The former indicates the utility of each journal for its area; the latter indicates its impact. For both, the Bradford Zones (BZs) were calculated for each area and triennium.

This study aims to characterize the journals dynamics along five triennia and across the Bradford Zones for both production and consumption in the different areas. This study also searched for specific behaviors when comparing the journals from Brazil, from Latin America, and from the rest of the world. Other aspect analyzed was the temporal relationship in the climbs for the journals that presented climbs in both: production and consumption.

Methods

We retrieved the articles of Brazilian authors from Web of Science (WoS) and SciELO databases in a fifteen years period (1998 and 2012) - five triennia that match the national assessment exercise performed by CAPES. It was called production (PROD) data set, with 395,650 articles, published in 9,092 journals. WoS journals cover 56.4% of the articles, while 12.5% came from SciELO journals, and 28.8% from journals indexed in both databases. The remainder 23% came from journals indexed in SciELO in less than a half of a triennium

period, getting "not indexed" in such triennium - likewise, some SciELO journals turned SciELO/WoS in a triennial transition. We classified the journals using the Science Watch (2014) schema that relates WoS categories to 22 Essential Science Indicators categories, to which we added the Human Sciences. SciELO journals were classified at the same way.

Respectively, de consumption (CONS) data set was formed by 10,759,279 bibliographic references of the articles. In the case of SciELO, we just added references related to journals, but WoS data include references to proceedings, and sometimes, to thesis. These citations remained in such amount once it was discarded in the normalization process (described below) that resolved 71.3% of the references (7.67 million), as presented in Table 1.

For this first approach, we decided to restrict CONS information to citations directed to those titles that belong to PROD data set. The reason was the fact that we have almost 29% of total references not normalized automatically, and that PROD journals capture 90.3% of the normalized citation amount.

Table 1. Consumption data sets and its prevalence in the whole data set.

CONS data set (filters)	Citation window	Freq.	*% of* All citations	*% of* Normalized	*% of* Citations to PROD journals from any area	restricted to
All citations	all	10,759,279	*100.0%*			
	5 year	3,731,745	*34.7%*			
Normalized cited journal titles	all	7,666,238	*71.3%*	*100.0%*		
	5 year	2,777,013	*25.8%*	*36.2%*		
Citations to PROD journals, from any area	all	6,922,780	*64.3%*	*90.3%*	*100.0%*	
	5 year	2,655,547	*24.7%*	*34.6%*	*38.4%*	
Citations to PROD journals, restricted to its own area	all	3,748,044	*34.8%*	*48.9%*	*54.1%*	*100.0%*
	5 year	1,485,463	*13.8%*	*19.4%*	*21.5%*	*39.6%*

So we created four different CONS data sets (featured in bold in Tab. 1), resulting of crossing two dummy variables. The first one was the restriction or not of the citation window (all citations/5-year). The second concerns to the area from which the citation comes to one title. In one case we considered just the citation received from titles of the same ESI category (not too restrictive, since it aggregates lot of WoS categories). In the other case, we count the citations regardless the area. The former corresponds to 54.1% of the latter. To give an idea of our purpose on doing this, we calculated the share of citations each area receives on its own area. The first one in the list was Space Science (whose impact is the most endogenous, with 81.2%) and the last is Multidisciplinary (the least endogenous, as one can expect, with 2.3%).

The cited journal title normalization has been performed relating the ways a journal was cited by the papers' authors with a reference base which contains several variations of cited journal title for each journal obtained from different databases (ISSN, WoS, Scopus, SciELO and Lattes Platform). Thus, it was possible to identify the ISSN from the most of the cited journals. Whenever there were conflicts in this identification, i.e., the cited title could be referring to more than one journal, the year and volume of the publication was used. In order to do this, a database containing the valid years and volumes for each journal was created using information available from the citations were the normalization presented no conflict. If, even after the use of year and volume, the conflict persisted, the normalization was not performed for the respective citation.

Having the normalized data from PROD and CONS from the 9,092 journals, as well as their basic information (title, ISSN, classification area and citing and cited years) we identified BZs, with three partitions, for which of the 23 areas in each of the 5 triennia, totalizing 115 Bradford's distributions for PROD data set. In the case of CONS data sets we did the same,

but four times, resulting 460 distributions. Moreover, it was not assigned a BZ for the journals without production or consumption in a given triennium.

An initial analysis suggested some journals had to be discarded because there was not enough information to correctly identify the behavior of these journals along the triennia. It was the case of 2,376 journals that entered the PROD data set in the last two triennia (publishing less than ten papers per triennium). An opposite case consists of 39 journals that the community stopped publishing, having no publications in the last triennium. We also found 247 journals with no articles in four triennia, and no citation in four of five triennia. Without these exclusions, 6,492 journals remained in the analysis.

The dynamics of each journal across BZs in its area was assessed along the triennia. Journals without any change in the BZ along the five triennia were classified as Stable (S). The ones that climbed zones along the triennia without any fall were considered Up (U), and oppositely, journals that fell BZs across the triennia without any climb were considered Down (D). And a journal that had climbs and falls along the triennia was considered Oscillating (O).

Findings

The great amount of data demanded many cross-tabulations to define the way of treating the information of each variable. At this time, we decided not to differentiate if a journal climbed one (Z3 to Z2 or Z2 to Z1) or two (Z3 to Z1, in different triennium or in a unique double step). The same was proceeded in relation to journals that fell BZs.

As we needed to create a journal profile of change that combine both PROD and CONS, we aggregated it with the following ordered classification scheme: U, to any combination that occurred at least one Up, permitting one of them to be Stable (U-U, U-S or S-U, to both PROD and CONS, respectively); S-S, if the journal has being Stable in both dimensions; O, if it was found swinging in any of dimensions; and D, to any combination occurring a Down.

Table 2. Distribution of journals by profile of changes in Bradford zones of production and consumption, in the four CONS data sets – period 1998-2012.

Citation data sets	Journals (total)					
Publication country	U	S_S	O	D	%	Freq.
CONS, considering citations to PROD journals, from all areas						
all	***10.8%***	***76.1%***	***8.7%***	***4.4%***	***100.0%***	**6,492**
Other	*9.5%*	*77.3%*	*8.7%*	*4.4%*	*100.0%*	5,949
Latin Am. &Caribe	*2.6%*	*93.1%*	*3.4%*	*0.9%*	*100.0%*	233
Brazil	*41.0%*	*39.7%*	*11.0%*	*8.4%*	*100.0%*	310
5 year	***10.4%***	***73.4%***	***10.6%***	***5.6%***	***100.0%***	**6,410**
Other	*9.3%*	*74.5%*	*10.6%*	*5.7%*	*100.0%*	5,873
Latin Am. &Caribe	*3.1%*	*92.5%*	*3.9%*	*0.4%*	*100.0%*	228
Brazil	*38.2%*	*38.2%*	*14.9%*	*8.7%*	*100.0%*	309
CONS, considering citation to PROD journals, restricted to its own area						
all	***17.7%***	***65.0%***	***12.1%***	***5.3%***	***100.0%***	**6,430**
Other	*16.5%*	*65.8%*	*12.4%*	*5.2%*	*100.0%*	5,890
Latin Am. &Caribe	*3.9%*	*90.9%*	*3.9%*	*1.3%*	*100.0%*	232
Brazil	*50.3%*	*28.6%*	*12.3%*	*8.8%*	*100.0%*	308
5 year	***16.8%***	***60.9%***	***15.2%***	***7.0%***	***100.0%***	**6,310**
Other	*15.8%*	*61.6%*	*15.6%*	*7.0%*	*100.0%*	5,777
Latin Am. &Caribe	*4.8%*	*90.3%*	*3.5%*	*1.3%*	*100.0%*	227
Brazil	*45.8%*	*25.8%*	*17.6%*	*10.8%*	*100.0%*	306

So we first have looked to the general behavior of the journals, but focusing on the ones that improved across the triennia, at least in one of the dimensions. Tab. 2 shows that the great amount of journals (about 75%) are Stable in both dimensions, but we find 10% less journals

with this profile when we restrict the citations to the journals own area. It reveals that closing the context of citation to the specific area, we find more changes (and this tendency is even more evident in the 5-year citation window), especially for the journals that got climbed BZs.

Considering the publication country, we can realize that Brazilian journals present lesser stability, what is interesting to analyze changes, which is what we find abundantly: about 40% when considering citation from any area, and about 50% in the journals own area. Revealing the importance of studying the impact of these journals in their context.

Despite being less frequent, journals falling are more prevalent in the 5-year citation window.

All this tendencies have to be analyzed more carefully subsequently, since specific characteristics of the journals can help to understand such evidences.

Now focusing our analysis in U-U journals, it is important to mention that Clinical Medicine presents more journals (about 30), followed by Engineering (about 15), and in the opposite side is Physics (with 2). Another observation is that U-U Brazilian journals correspond to 14.5%, considering citations from all areas, and 18% in the journals own area. This is strongly different of journals out of Latin America & Caribe, whose correspondent percentage is about 3%. Among Brazilian journals, those indexed just in SciELO presents prevalence about 5% bigger than those indexed in both databases, when considering the citations in the journals own area. It reveals the growing importance of some journals in the national context, inside the area of specialty (data not shown).

Table 3. Distribution of journals U-U by triennium of first climb in Bradford zones of production and consumption, in the four CONS data sets – period 1998-2012.

Citation data sets		*% of journals: t* riennium of 1st climb in BZs (CONS)				Journals (total)	
U-U Journals		2	3	4	5	%	Freq.
CONS, considering citations to PROD journals, from all areas							
all		***11.4%***	***24.6%***	***29.8%***	***34.2%***	***100.0%***	**114**
Triennium of 1st climb in BZs (PROD)	**2**	*17.2%*	*44.8%*	*20.7%*	*17.2%*	*100.0%*	29
	3	*12.9%*	*29.0%*	*32.3%*	*25.8%*	*100.0%*	31
	4	*5.3%*	*10.5%*	*34.2%*	*50.0%*	*100.0%*	38
	5	*12.5%*	*12.5%*	*31.3%*	*43.8%*	*100.0%*	16
5 year		***14.1%***	***25.6%***	***30.1%***	***30.1%***	***100.0%***	**156**
Triennium of 1st climb in BZs (PROD)	**2**	*34.5%*	*37.9%*	*13.8%*	*13.8%*	*100.0%*	29
	3	*19.5%*	*41.5%*	*31.7%*	*7.3%*	*100.0%*	41
	4	*3.7%*	*11.1%*	*40.7%*	*44.4%*	*100.0%*	54
	5	*6.3%*	*18.8%*	*25.0%*	*50.0%*	*100.0%*	32
CONS, considering citation to PROD journals, restricted to its own area							
all		***18.5%***	***24.3%***	***27.2%***	***30.1%***	***100.0%***	**173**
Triennium of 1st climb in BZs (PROD)	**2**	*41.4%*	*31.0%*	*20.7%*	*6.9%*	*100.0%*	29
	3	*8.6%*	*51.4%*	*28.6%*	*11.4%*	*100.0%*	35
	4	*22.2%*	*20.4%*	*31.5%*	*25.9%*	*100.0%*	54
	5	*9.1%*	*7.3%*	*25.5%*	*58.2%*	*100.0%*	55
5 year		***22.9%***	***24.0%***	***29.6%***	***23.5%***	***100.0%***	**179**
Triennium of 1st climb in BZs (PROD)	**2**	*44.0%*	*32.0%*	*24.0%*	*0.0%*	*100.0%*	25
	3	*20.0%*	*48.6%*	*25.7%*	*5.7%*	*100.0%*	35
	4	*27.3%*	*20.0%*	*40.0%*	*12.7%*	*100.0%*	55
	5	*12.5%*	*10.9%*	*25.0%*	*51.6%*	*100.0%*	64

Attempting to the temporal relation between Ups in PROD and CONS BZs, we performed a bivariate analysis considering the triennium each journal had its first climb in BZs. Tab. 3 presents the distribution of journals of different triennia of CONS (columns), related to each triennium of PROD (lines). The row cells with bigger prevalence of journals are identified in grey scale The row cells with bigger prevalence of journals are identified in grey scale. In the first CONS data set, considering the first line, that respect to 29 journals that climbed BZs first time in the 2nd triennium, we see that most of the journals climbed in CONS in the 3rd,

followed by the 4th. It shows that most of them improved CONS BZs after (as to say, both of them above the principal diagonal). When we drop to the next lines the two more prevalent cells change to the diagonal and one before. The same can be observed in the second CONS data set (5-year citation window) and a little bit more concentrated in the principal diagonal when restricting the citation to the journals own area. Maybe in subsequent analysis we can verify properly if the increasing of consumption is pulling the increasing of production.

Final remarks

As we can observe in this first approach, a national system combining publications from both contexts (national and international) can be a useful tool to research evaluation. Bradford zones showed to be an interesting relative indicator, when applied to evaluative purposes. Especially the joint analysis of production and consumption dimensions can bring a more complete view of the scientific communication flow, considering the changes of journals through zones in both dimensions. National impact indicators can complement Impact Factor, in the sense it can add the local importance, as observed about SciELO journals.

Acknowledgments

We acknowledge FAPESP, that supports the *Young Investigators Awards* project, entitled *Scientific assessment in Brazil: study of scientific communication in scientific areas* (Grant number 2012/00255-6) and CNPq (Research Productivity Grant 477246/2013-3).

References

Buela-Casal, G. et al. (2006). Measuring internationality: Reflections and perspectives on academic journals. *Scientometrics*, *67*(1), 45-65.

Chen, K. (2004). The construction of the Taiwan Humanities Citation Index. *Online Information Review*, *28*(6), 410 - 419.

Jin, B. & Wang, B. (1999). Chinese Science Citation Database: its construction and application. *Scientometrics*, *45*(2), 325-332.

Kim, S. et al. (2013). Korea Citation Index and Its Macro Bibliometrics. *Asian Journal of Innovation and Policy*, *2,* 194-211.

MacRoberts, M. H. & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435-444.

Mehrad, J. & Arastoopoor, S. (2012). Islamic World Science Citation Center (ISC): Evaluating Scholarly Journals Based on Citation Analysis. *Acta Inform Med. 20*(1), 40-43.

Miranda, E. C. & Mugnaini, R. (2013). Scientific policy in Brazil: Exploratory analysis of assessment criteria. *PRO INT CONF SCI INF*, *14*, 1578-1586.

Mugnaini, R. et al. (2014). Comunicação científica no Brasil (1998-2012): indexação, crescimento, fluxo e dispersão. *Transinformação*, *26*(3), 239-252.

Negishi, M., Sun, Y., & Shigi, K. (2004). Citation database for Japanese Papers: A new bibliometric tool for Japanese academic society. *Scientometrics*, *60*(3), 333-351.

Packer, A. L. (2014). The emergence of journals of Brazil and scenarios for their future. *Educ. Pesqui*, *40*(2), 301-323.

Packer, A. L. et al. (1998). SciELO: a methodology for electronic publishing. *Ci. Inf.,* 27(2), 109-121.

Pajic, D. (2014). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics*, *102*(3), 2131-2150.

Piñeiro, C. L. & Hicks, D. (2015). Reception of Spanish sociology by domestic and foreign audiences differs and has consequences for evaluation. *Research Evaluation*, *24*(1), 78-89.

Ponomariov, B. & Toivanen, H. (2014). Knowledge flows and bases in emerging economy innovation systems: Brazilian research 2005–2009. *Research Policy*, *43*(3), 588-596.

Rego, T. C. (2014). Productivism, research and scholarly communication: between poison and medicine. *Educação e Pesquisa*, *40*(2), 325-346.

Šipka, P. (2005). The Serbian citation index: Context and content. *PRO INT CONF SCI INF*, *10*, 710-711).

Tijssen, R. J. et al. (2006). How relevant are local scholarly journals in global science? A case study of South Africa. *Research Evaluation*, *15*(3), 163-174.

Winclawska, B. M. (1996). Polish Sociology Citation Index (Principles for creation and the first results). *Scientometrics*, *35*(3), 387-391.

Sustained Collaboration Between Researchers in Mexico and France in the Field of Chemistry

Jane M. Russell[1], Shirley Ainsworth[2] and Jesús Omar Arriaga-Pérez[2]

[1] *jrussell@unam.mx*
National Autonomous University of Mexico (UNAM), Library Science and Information Research Institute (IIBI), Ciudad Universitaria, 04510 Mexico DF (Mexico)

[2] *shirley@ibt.unam.mx oarriaga@ibt.unam.mx*
National Autonomous University of Mexico (UNAM), Institute of Biotechnology (IBT), Av. Universidad #2001, 62210 Cuernavaca, Morelos (Mexico)

Abstract
We analyse the co-authorship and publication patterns of 863 mainstream WoS papers in the field of Chemistry co-authored between Mexican and French institutions from 1984 to 2013 with the purpose of identifying and characterizing the dynamics of sustained collaborative research partnerships in the field between the two countries. From a normalized set of the most productive authors with ≥ 5 co-authorships we selected three Mexican scientists for a detailed analysis of their co-authorship network visualized using Gephi software and its development over time. The first was the most productive Mexican author from the main national university whose collaboration with France spanned the period from 1987-2012, while the second and third researchers work in provincial universities and whose collaboration with France is more recent but lasting 10 and 15 years respectively, and also continues up to the present day. Preliminary results suggest that sustained partnerships are driven by a strong central bond between the Mexican researcher and their foreign partner. In the first two cases, the bond is with directly with a French scientist but in the third, is stronger with an Italian rather than with the French counterpart.

Conference Topic
Country level studies

Introduction
A recent paper examining the main research thrusts and future challenges facing research into scientific collaboration mentions the need to characterize the factors underpinning successful collaborations and to ascertain how collaboration can benefit scientific development in the less developed countries (González Alcaide & Gómez Ferri, 2014). International collaboration is known to be especially important for countries whose scientific infrastructure and capacity can benefit from forging alliances with researchers from institutions abroad. Colombian researchers for instance were found to increase team output by almost 40% by co-authoring with overseas partners (Ordóñez-Matamoros, Cozzens & García, 2010).

We know little about the duration of international research collaboration between individual researchers in terms of the number and timeline of co-authored papers. Two decades ago a study looked at the production and duration of collaboration between researchers from institutions in Mexico and France in all scientific areas (Narvaez-Berthelemot & Russell, 1996). Chemistry was the subject of the greatest number of bilateral publications as well as having the highest continuity index defined as the number of articles (>2) in a given period, in this case 1980-1989, that were co-authored by the same groups. More recently an analysis of co-publications between the two countries from 1984 to 2010, showed that Chemistry gradually lost ground with respect to other disciplines notably Physics, even though the number of papers increased with time (Ainsworth et al., 2014).

The present research in progress sets out to characterize the publication dynamics of sustained collaborative research partnerships between Mexico and France in Chemistry in the period 1984-2013. We take as our starting point, the most productive authors in papers with at least

one author from both Mexico and France. Considering that interpersonal links are the key drivers of collaboration (Gaillard et al., 2013) we are also interested in analysing the relationship between co-authors and tracing the development of their networks over time. Another aspect of the collaboration we consider is the level of importance of the relationship with Mexico in the case of the French scientists or France for the Mexicans, for the total body of work of the key players during the same period and who might be the senior partner in the bilateral relationship. We adopt two approaches when analysing our publication and co-authorships data based on the following assumptions: 1. Sustained collaboration is characterized by a central relationship established between one Mexican and one French scientist. 2. Sustained collaboration between the two countries is characterized by a series of relationships forged with different French scientists and institutions.

Data source and methods

Data source was the Web of Science searching France and Mexico in the country field, covering the period 1984-2013, in the discipline of Chemistry. WoS journal subject categories were adapted to the RFCD classification scheme for the assignment of the discipline (Butler, Henadeera y Biglia, 2006). Records were downloaded to a local MySQL database. Author names with ≥ 5 co-authorships were normalized and assigned (often several) Scopus author ids and affiliations, given that author identification in WoS proved less than adequate for our purpose. Case studies were selected from the group of the most prolific Mexican authors with bilateral France-Mexico collaboration. For this preliminary presentation of results we have selected three case studies based on our initial analysis of their collaboration dynamics. These include the most prolific Mexican researcher and two other productive researchers from established groups with substantial French collaboration from two provincial state universities, namely Cecilio Álvarez y Toledano from the Institute of Chemistry at the big national Mexican university, Universidad Nacional Autónoma de México (UNAM), Ricardo Navarro-Mendoza from the Universidad de Guanajuato (UG) and Claudio Marcelo Zicovich-Wilson from the Universidad Autónoma del Estado de Morelos (UAEM).

The interactive visualization open source software Gephi was used to select and represent these collaborations and to show sub-networks within clusters. Co-authors involved in each of the papers were examined to characterize the temporal collaboration, and separately the normalized author information from Scopus was used to represent the importance of the Mexico-France collaboration in the main authors' output. The corresponding author of each paper was also identified.

Overall panorama of Mexico-France co-authorship in Chemistry

The number of co-authored papers in Chemistry between Mexico and France showed a steady rise from a mere two in 1984 to 54 in 2013 (Figure 1). Social network graphs (not shown here) show an increasing dense and complex series of relationships when comparing the first 15 years (1984-1993) with the second period (1994-2013).

Publication dynamics of sustained partnerships

Figure 2, divided into three decades, shows the dense network of co-authorships of Cecilio Álvarez y Toledano with French institutions during our period of study. The strongest link is with Henri Rudler of the Université Pierre et Marie Curie, Institut Parisien de Chimie Moléculaire starting in 1987, and to a lesser extent with Andrée Parlier of the same laboratory except during the middle period 1994-2003. Rubén Alfredo Toscano works in the same institute as Cecilio Álvarez y Toledano as a highly specialized technician and is a regular co-author. Of the 29 papers of Álvarez y Toledano in co-authorship with a French institution, 23 were published in co-authorship with Rudler. There was a notable pause in their collaboration

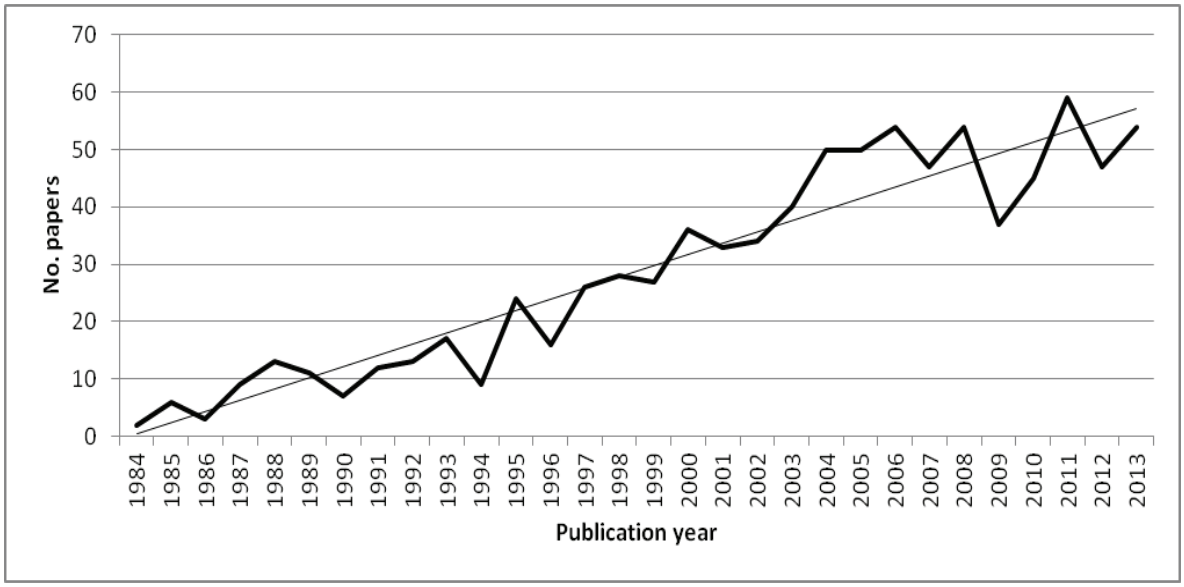


Figure 1. Papers in collaboration between Mexico and France in Chemistry 1984-2013.

from 1996 to 2004 when Álvarez y Toledano co-authored two papers with two other French authors, Henri Arzoumanian, Aix-Marseille Université, and Bruno Donnadieu now of the newly formed Université de Montpellier but at the time of the Universite Montpellier 2, respectively and involving a different set of co-authors. Nonetheless, Andrée Parlier and Henri Rudler continued their collaboration without Álvarez and Toledano during this period, together with Jacqueline Vaissermann, also from the same laboratory.

During the first two periods four clusters of co-authors are apparent, while in the most recent period 2004-2013, co-authorships are concentrated in two with Rudler and Parlier at the centre, respectively. A strong central bond with Henri Rudler is evident in the collaboration of Álvarez y Toledano over the whole period suggesting that this bilateral partnership is the motor driving this example of sustained co-authorship between Mexico and France.

Data taken from Scopus using the author id field for papers co-authored by Rudler and Álvarez-Toledano in Chemistry show Rudler to be senior (corresponding) author in 11 of these 29 papers as compared to 6 in the case of Álvarez-Toledano, which would seem to show that Rudler is the senior partner in this collaboration. The issues of authorship order are discipline-specific, but in many scientific areas it is accepted that the principal investigator is named as the corresponding author (Frandsen & Nicolaisen, 2010). These 29 papers represent 26% of all Rudler's papers as represented in Scopus, compared to 20% of those of Álvarez-Toledano suggesting that the bilateral partnership is of significance for the output in Chemistry for both researchers.

The network of collaboration with French institutions starting in 1998 around Ricardo Navarro Mendoza from the Universidad de Guanajuato appear in Figure 3 with strong links to Eric Guibal from the École des Mines d'Alès. Fourteen of the 15 papers published from 1998-2012 appear with both authors. Imelda Saucedo Medina, also from the Universidad de Guanajuato, is a co-author in 11 of these papers. In one article at the beginning of the period in 1998, there is a collaboration with other French authors, Denise Bauer and Gérard Cote, both from the École Nationale Supérieure de Chimie de Paris, and in two articles, 2000 and 2001 with Thierry Vincent from École des Mines d'Alès.

This suggests a consolidated partnership, though perhaps also an unequal one. Scopus data for papers co-authored by author Ricardo Navarro Mendoza and Eric Guibal in Chemistry show Navarro-Mendoza to be corresponding author in 8 of the 11 instances, compared to 3 for

Guibal. This would suggest that in this case the Mexican is the senior partner. These 11 papers represent 33% of all Navarro Mendoza's papers in Scopus, but only 7% of Guibal's.

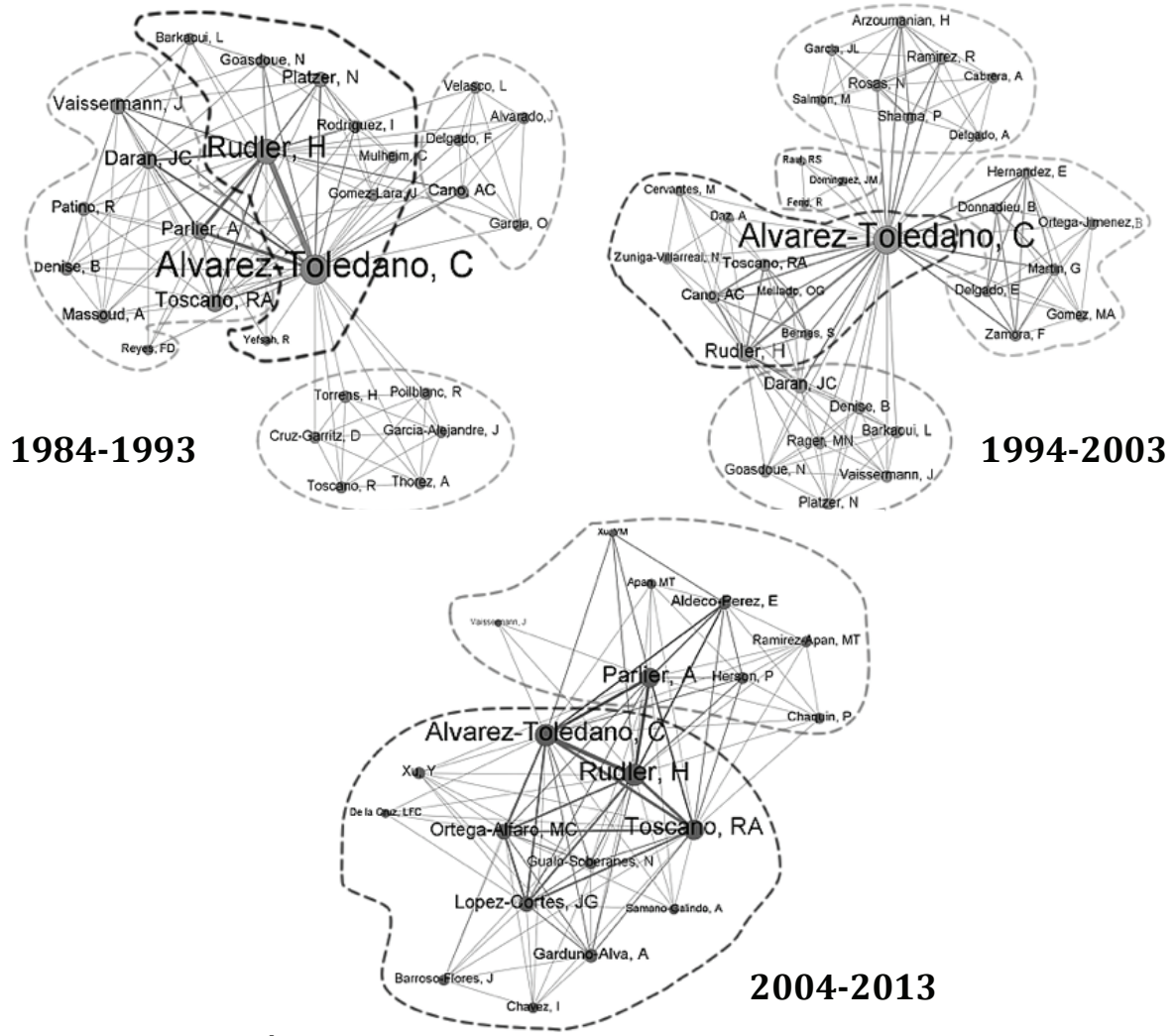


Figure 2. Álvarez y Toledano: Network of ≥ 3 co-authorships 1984-2013.

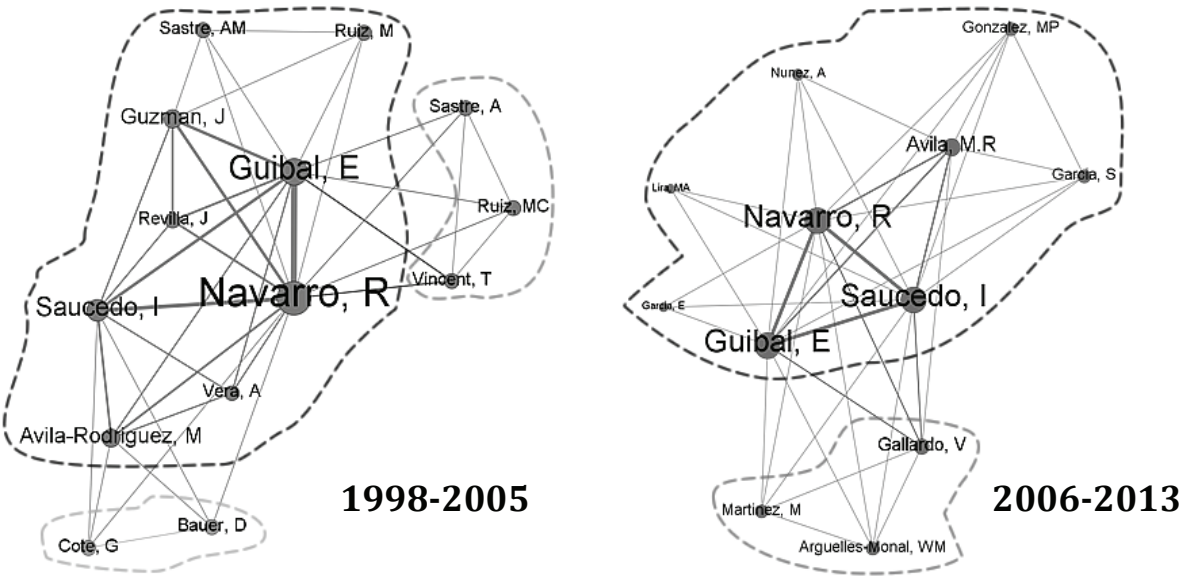


Figure 3. Navarro Mendoza: Network of ≥ 3 co-authorships 1998-2013.

The co-authorship of Claudio Marcelo Zicovich Wilson from the Universidad Autónoma del Estado de Morelos with researchers from France that began in 2004, is reflected in Figure 4, as is also the importance of a group of Italian authors for this collaboration. Roberto Dovesi from the Universita degli Studi di Torino appears as co-author in 13 of the 16 papers of Zicovich Wilson where there are also authors from French institutions in the period 2004-2013. Other researchers from the same Italian institution such as Roberto Orlando (6 papers), Piero Ugliengo (4 papers) Loredana Valenzano (3 papers 2006-2008) and Raffaella Demichelis (also 3) appear together with Dovesi, the latter co-author during 2010-2011. The predominant French author is Fabien de Pascale, at the time of Université Henri Poincaré - Nancy I, who is a co-author in 8 of the 16 papers during 2004-2010, Yves Noël, CNRS Institut des Sciences de la Terre de Paris with 5 papers 2007 then 2010-2012, together with Michel Rérat, Université de Pau et des Pays de L'Adour form a separate French collaboration, albeit together with Roberto Dovesi. The central role of Roberto Dovesi in the Mexico-France collaboration seems evident from the data taken from WoS. Data from Scopus for papers in Chemistry co-authored by Zicovich Wilson and Pascale reveal that the Mexican is corresponding author in only one of these, and Pascale not in any of them. (Pascale appears as first author in three of them.) The role of Roberto Dovesi in this collaboration seems to be confirmed in that he is corresponding author in 6 of these 10 papers. These papers correspond to 9% of all Zicovich Wilson's papers, 40% of Pascale's but only 4% of those of Dovesi. These data imply that Pascale is the junior partner here.

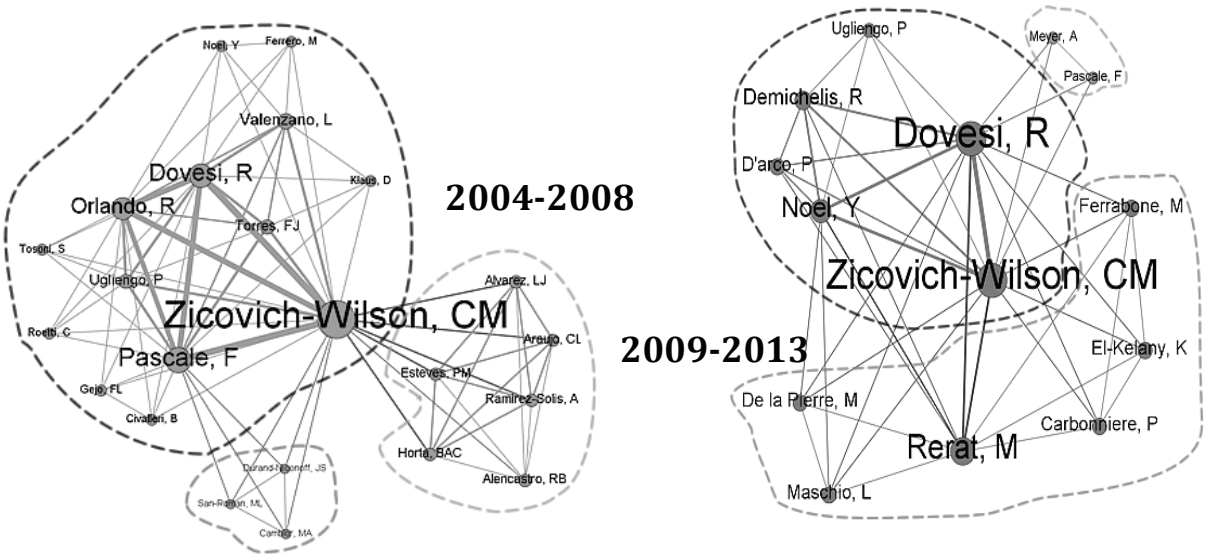


Figure 4. Zicovich Wilson: Network of ≥ 3 co-authorships 2004-2013

Preliminary Conclusions

Our detailed analyses of the co-authorship networks of three Mexican scientists, one from the large national university located in Mexico City where the national scientific research effort is centred and two from provincial universities, with ≥ 5 co-authorships with France in mainstream Chemistry journals during our period of study, lend support to our initial assumption that sustained collaboration is characterized by a central relationship established between two individual scientists but not necessarily directly between a Mexican and a French scientist. In the first two cases the bond is with a French scientist but in the third, is stronger with an Italian rather than with the French co-author. These central relationships are

strengthened and supported by frequent co-authorship from both Mexican and French groups in the first two cases and in the third, by the Italian group. A substantial number of one-time co-authors was evident in all three cases. We found differences with respect to the importance of the bilateral collaboration for the Mexican and French authors and with respect to which of the two could be considered the senior author. These preliminary conclusions will be tested by analysing further case studies of sustained partnerships between Mexican and French chemists.

References

Ainsworth, S., Russell, J.M. Narvaez-Berthelemot, N. & Arriaga-Pérez, J.O. (2014). Mapeo de la colaboración en ciencia y tecnología entre México y Francia a través de un análisis de co-publicaciones 1984-2010. In D. Villavicencio & M. Kleiche (Coords.), *Cooperación, Colaboración Científica y Movilidad Internacional en América Latina*. (pp.49-74). Buenos Aires: Consejo Latinoamericano de Ciencias Sociales-CLASCO. Retrieved January 3, 2015 from: http://www.clacso.org.ar/libreria-latinoamericana/ libro_detalle.php?orden=&id_libro=916& pagenum_rs_libros=0&totalrows_rs_libros=885

Butler, L., Henadeera, K. & Biglia, B. (2006). State and Territory based assessment of Australian research. Technical paper. Retrieved January 8, 2015 from http://www.pc.gov.au/inquiries/completed/science/technicalpaper1

Frandsen, T.F. & Nicolaisen, J, (2010). What is in a name? Credit assignment practices in different disciplines. *Journal of Informetrics*, 4, 608-617.

Gaillard, A.M., Gaillard, J., Russell, J.M., Galina, C.S., Canesse, A.A., Pellegrini, P., Ugartemendra, V. & Cárdenas, P. (2013). Drivers and outcomes of S&T international collaboration activities. A case study of biologists from Argentina, Chile, Costa Rica, Mexico and Uruguay. In J. Gaillard & R. Arvanitis (Eds.), *Research Collaborations between Europe and Latin America. Mapping and Understanding Partnership.* (pp. 157-191). Paris: Editions des Archives Contemporaines,

González Alcaide, G. & Gómez Ferri, J. (2014). La colaboración científica: principales líneas de investigación y retos de futuro. *Revista Española de Documentación Científica*, 37, 1-15.

Narvaez-Berthelemot, N. & Russell, J.M. (1996). La continuite dans la colaboration scientifique internationale: Le cas de la France et du Mexique. In R. Arvanitis & J. Gaillard (Eds.), *Memoires du Colloque ORSTOM UNESCO sur "les Sciences hors d'Occident au 20o siècle"*, Vol. 7 Coopérations Scientifiques Internationales (pp.39-52). Paris: ORSTOM.

Ordóñez-Matamoros, H.G., Cozzens, S.E. & García, M. (2010). International co-authorship and research team performance in Colombia. *Review of Research Policy*, 27,415-431.

Innovation and Economic Growth: Delineating the Impact of Large and Small Innovators in European Manufacturing

Jan-Bart Vervenne and Bart Van Looy

{jan-bart.vervenne, bart.vanlooy}@kuleuven.be
KU Leuven, Faculty of Business and Economics, Department of Managerial Economics, Strategy and Innovation, Naamsestraat 69, B-3000 Leuven (Belgium)

Abstract
In the course of the past decades, the link between innovation and economic growth has become a well-established one in the economic literature. In the current study an attempt has been provided to complement this line of research with an assessment of the wealth implications of the 'entrepreneurialisation' of innovation systems. Relying on a 9 year panel of post-millennial observations for 22 European countries and using stock based patent indicators, it was found that on top of the positive productivity impact of innovative activity growth, a premium effect can be observed when the stake of small firms in it increased at the same time. These findings can be interpreted as confirming Baumol's (2004) assignment of different roles to large and small firms in innovation systems: the former as provider of the technological breakthrough that the latter improves in a range of incremental steps. The entrepreneurialisation of manufacturing as a whole, measured by the stakes of small businesses in employment, yields a productivity discount: outside of innovative activities, economies of scale outweigh co-occurring diseconomies of scale. Distinct country groups in different stages of economic development form the main drivers of both entrepreneurialisation effects: a core of North-Western European countries that has attained the innovation-driven stage against a periphery of Southern and Eastern European countries around them that have not transcended the more preliminary efficiency-driven stage. Further rationales explaining the additional explanatory power of entrepreneurial innovation were found in the weakening of the link between innovation measured by patents and added value in large firms.

Conference Topic
Country-level studies; Patent analysis

Introduction
Substantial agreement exists among economists and policymakers that technological innovation is a key driver of sustainable economic growth. Technological innovation implies the implementation of inventions in the production of final goods or services and as such yields productivity gains for the innovating economy. Using knowledge capital to transform existing knowledge into such inventions, the amount of research and development (R&D) efforts is an important determinant of the pace of technological innovation.

Endogenous growth scholars have shown that technological innovation is an endogenous component of the process of long-run economic growth, both theoretically (Romer, 1986) as well as empirically (Nadiri, 1993). As opposed to their neoclassical counterparts (Solow, 1956), they postulate that technological innovation is an inherent component of the growth process: profit-maximising firms purposely allocate resources towards R&D in the presence of sufficient perspectives suggesting that they will be capable to appropriate the gains from it.

The analysis in this paper contributes to the mentioned line of research by complementing the measurement of overall technological innovation effects using patent statistics with an additional, patent-based indicator capturing the footprint of small, more entrepreneurial firms in the countries' stock of knowledge capital.[1] Further explanation for the rationale triggering

[1] Note that throughout this excerpt alternately we describe the firms of our interest as entrepreneurial or small. As Wennekers and Thurik (1999) argue, smallness and entrepreneurship can only be synonymous when management and ownership are not distinct. Subsidiaries of large business groups can qualify as small as well when shareholder information is not taken into account. This remark is of concern to us given the definition of small firms we will use in the empirical part (cf. below). However, given that small firms pertaining to larger

our interest to differentiate between innovation induced by small and large firms follows next. Subsequently methodology and results are reported, followed by some concluding notes. The focus on Europe in this study is justified among others by referring to the entrepreneurial innovation deficit Europe faces in comparison with the US (Veugelers, 2009).

Delineating the entrepreneurial contributions to innovation

The rationale to differentiate between incumbent and entrepreneurial innovation draws extensively from research on entrepreneurial innovation by Audretsch (2001), Baumol (2004) and Veugelers (2009). Whereas Schumpeter in 1942 predicted the gradual replacement of the entrepreneurial inventor - naturally associated with the small start-up - by routinized innovation organized by large industrials, Baumol (2004) emphasized the complementary relationship of both types of players within innovation systems. Their organizational design has induced them to specialize in different components of society's innovation process. Over the past decades revolutionary breakthrough inventions in the US have continued to come predominantly from small entrepreneurial enterprises whereas large industry have provided ever-increasing streams of incremental improvements to them multiplying capacity and speed and increasing reliability and user-friendliness. This is the result of the oligopolistic competition this relatively limited amount of very large firms, particularly in high-tech industries, engage in. It forces them to keep innovating in order to survive, but in a very risk-free and thus path-dependent way, avoiding the risks of the unknown that the revolutionary breakthrough entails. As such, inert incumbents leave plenty of room to explore for the enterprising entrepreneur. Unaffected by concerns relating to existing products and markets, the latter can pick up the ideas the former would deem too risky (Audretsch, 2001; Baumol, 2004). The other way around, incumbents are more suited to follow-up and improve those breakthrough innovations in more mature stages of the technology life-cycle (Baumol 2004).

Plugging the level of 'entrepreneurialisation' of innovation into a growth model

Methodology

The neo-classical growth model (Wong et al., 2005) we use to test a number of research questions distilled from the context described above is based on an augmented Cobb-Douglas production function:

$$Y = A^{O} K^{\propto} L^{\beta}$$

Where Y = output, A^{O}= total factor productivity, K = stock of physical capital and L = labor employed. Assuming constant returns to scale, $\alpha + \beta = 1$, both sides of the equation are then divided by labour. Taking natural logs the resulting model to estimate economic productivity per employee goes as follows:

$$\ln\left(\frac{Y}{L}\right) = lnA^{o} + \propto ln\left(\frac{K}{L}\right)$$

Following the approach by Wong et al. (2005), we assume that the stock of knowledge capital is the main determinant of total factor productivity, A^{O}. The stock of knowledge capital is captured using technological innovation statistics, among which patent based-indicators comprise one of the best proxies. More specifically, the level of innovation (*INNO)* is measured using stocks of patent applications depreciating at a rate of 20% per year as the

conglomerates in the countries of our sample never comprise a majority, on average our population of small firms can be described as 'more entrepreneurial'.

effects of investment in innovation transcend the short run.[2] The technological innovation variable was normalized by employment to capture its intensity and limit the effects of country size as much as possible. As suggested in the previous section, as factor of total productivity the general intensity of technological innovation is complemented by a patent-based indicator, measuring the degree of small firm engagement in innovative activity, and an equivalent employment-based indicator to control for overall small firm activity. The latter to make sure increased innovative activity of small firms is not simply capturing the potential productivity effects of an increase in entrepreneurial activity in general.

Determining the degree to which national innovation systems have ran on entrepreneurial initiative was based on the assignment of patents to small and large firms using the methodology presented in Eurostat (2014).[3] Due to shortcomings in the matching methodology and data gaps in the financial database - among others the result of country-specific disclosure exemptions rewarded to certain company types - only for approximately 62% of the corporate applicants in Europe firm size could be determined. We assume however that these country-level constraints equally hold for all years of the sample and as such are coped with by estimating coefficients using country fixed effects (cf. infra).

The effects of entrepreneurial and incumbent engagement in innovation could not just be measured by plugging raw stocks of their respective patent applications into the equation: R&D clustering dynamics within countries result in a high correlation – more than 0.97 even when removing country effects – with the annual innovative activity deployed by the national innovation system as a whole, that is already captured in the core variable measuring technological innovation. Given our main interest towards the benefits of entrepreneurial innovation and to avoid multicollinearity, the degree of 'entrepreneurialisation' of corporate technological innovation (*ENTR_INNO*) was measured by computing the share of small firms in the stock of patents assigned to firms with identified size.

The within variance of this share value captures to what extent small firms have shown relative over- or underactivity in R&D in comparison with their large counterparts. Given the large level of correlation among the small firm, large firm and overall patent stocks it is safe to assume that entrepreneurial and incumbent innovation do not have an opposite effect on economic productivity which would hamper a straightforward interpretation of *ENTR_INNO*. At most one of them can have a relatively larger impact on productivity. In line with the rationale elaborated above we expect that to be the small innovators. The result of that should

[2] All patent statistics were extracted from EPO's Worldwide Patent Statistical Database 'PATSTAT' (Autumn version 2014). In general we relied on EPO patent applications, including granted and non-granted patents, with the idea that counting both yields a relatively more input-oriented measure capturing the level of R&D spending than if one would stick to grants only (Ernst, 2003). Depreciation of the patent stock at a rate of 20% per year is based on the perpetual inventory method described in Ulku (2004). The patent stock variable incorporates annual EPO patent counts from 1970 onwards. The restriction of our attention to EPO patents can be easily justified given the geographical reach of our dataset and their costliness, which is a direct result of their supra-national character. Being that expensive, especially for more financially constrained SMEs, counts of them at the macro-level bear the potential to be good signals of R&D input & output levels per country over time.

[3] The lack of dynamic shareholder data in BvD's Amadeus (a database gathering annual account information) withheld us from determining firm size at the business group level. In contrast with the matching exercise presented in Eurostat (2014), firm size was determined dynamically by linking patents to financial information from the financial years that corresponded with the patent application filing year. In addition financial account data from Amadeus 2012 was enriched with equivalent information from earlier versions (2004 and 2007) to dispose of financial information in the earliest years of the matched sample (1999-2011) and to account for the BvD rule to discard companies not filing accounts for 5 years in a row. Firm size – or rather entity size – classification for patenting companies from 1999 onwards was based on the European Commission SME definition (2005): enterprises that employ fewer than 250 employees and which have an annual turnover not exceeding 50 million euro, and/or an annual balance sheet total not exceeding 43 million euro.

be *ENTR_INNO* exerting a positive effect on productivity, which would imply the existence of a productivity premium to an increased entrepreneurial stake in corporate innovation.

Given that the large majority of patents in Europe can be assigned to the manufacturing industry (Fraunhofer, 2003), downloads of observations for the non-patent based variables of country *c* in year *t* were restricted to that sector. Indicators for value added at factor cost (*VAFC*), the number of persons employees (*NPE*), gross investment in tangible goods (*GITG*) and the share of small firms in corporate employment (*ENTR_EMP*) were extracted from the Eurostat website.[4] [5] Furthermore, a quadratic year trend is included to capture time effects.[6] Conform previous research all R&D related indicators are lagged since it is assumed that the effects of R&D on economic performance take a couple of years to surface. In line with Ulku (2004) and given the limited time-series at our disposal we opted for a 2-year time lag. Following an equivalent rationale, the physical investment and share of entrepreneurial employment variables were also lagged by 1 year.

The resulting equation to be estimated using panel data techniques is:

$$lnVAFC/NPE_{ct} = \propto + \beta_1 lnGITG/NPE_{c,t-1} + \beta_3 INNO/NPE_{c,t-2} + \beta_4 ENTR\_INNO_{c,t-2} + \beta_5 ENTR\_EMP_{c,t-1} + year + year^2 + u_c + \varepsilon_{ct}$$

Results

Coefficients are estimated using fixed effects OLS.[7] Table 1 reports the estimation results, including robust standard errors, for the overall set of European countries (panel 1: ALL) and split sets of countries that lead (panel 2: LEADERS) or lag behind (panel 3: LAGGARDS) in terms of innovation according to the European Commission's (EC) Innovation Union Scoreboard (2015). The left hand of each panel contains estimates for the basic model as expressed in the equation above. The right hand side in addition reports an additional interaction effect between the technological innovation intensity and its degree of 'entrepreneurialisation'.

Conclusion and directions for future research

Apart from confirming previous findings regarding the positive impact of technological innovation on economic output, overall results (ALL) reveal that there is an additional productivity premium to a larger share of entrepreneurial engagement in the development of new, patented technology. The entrepreneurialisation of employment on the other hand, a broader measure of corporate activity, appears to be negatively associated with productivity.

---

[4] The resulting set of 22 countries consists of: Austria, Belgium, Germany, Denmark, Finland, France, United Kingdom, the Netherlands, Norway, Slovenia, Sweden (LEADERS), Czech Republic, Cyprus, Estonia, Greece, Hungary, Italy, Latvia, Poland, Portugal, Slovakia and Spain (LAGGARDS). Other European countries were discarded for multiple reasons: a lack of employment, investment or gross added value statistics available to the public or a too low rate of patenting companies matched to companies in the financial database, as such, hampering a representative image of the distribution of patents between incumbents and small businesses. Unusual annual productivity growth induced by preferential tax regimes for foreign firms, inciting those to shift profits to local subsidiaries, resulted in elimination of Ireland and Luxemburg from the sample as well.

[5] All currency-based series – expressed in Euro – were deflated using per country GDP price deflators (World Bank WDI website). Due to the lack of availability of stock variables capturing the total amount of outstanding fixed capital, in line with Ulku (2004) we used the flow variant.

[6] Preferably time dummies are included but using a functional form, in this case a quadratic trend allowing for one up and one down trend, can be an alternative in order to preserve degrees of freedom. Results turned out to be largely consistent for trend- and dummy-based models.

[7] Correlations among demeaned variables suggest that multicollinearity is not an issue for within-transformed variables.

Table 1. OLS fixed effects regression results.

	ALL		LEADERS		LAGGARDS	
ln_GITG/NPE (1y lagged)	0.676	0.529	0.686	-0.014	0.578**	0.825***
	(1.23)	(1.11)	(1.0)	(0.03)	(2.51)	(4.81)
INNO/NPE (2y lagged)	0.872**	-1.936	-0.736	-3.615*	-1.378	7.178
	(2.11)	(1.60)	(0.63)	(2.17)	(1.03)	(1.12)
ENTR_INNO (2y lagged)	0.003	-0.011	0.018	-0.114**	0	0.007
	(0.56)	(1.29)	(0.53)	(2.30)	(0.06)	(1.31)
INNO/NPE * ENTR_INNO (both 2y lagged)		7.873**		13.545***		-16.253
		(2.66)		(3.40)		(1.29)
ENTR_EMP (1y lagged)	-0.044**	-0.040**	-0.046	-0.053	-0.039**	-0.037**
	(2.67)	(2.38)	(1.06)	(0.90)	(2.82)	(2.56)
year	0.800**	0.699*	0.925	0.692	0.572**	0.590**
	(2.17)	(2.01)	(1.32)	(1.07)	(2.58)	(2.75)
year²	0.000**	0.000*	0.000	0.000	0.000**	0.000**
	(2.17)	(2.01)	(1.32)	(1.06)	(2.58)	(2.75)
_cons	-803.722**	-701.719*	-930.145	-695.108	-574.378**	-593.032**
	(2.18)	(2.01)	(1.33)	(1.07)	(2.59)	(2.76)
# observations	177	177	92	92	85	85
# groups	22	22	11	11	11	11
F statistic	38.62	51.13	39.55	44.54	29.68	149.8
R-squared Within	0.49	0.53	0.48	0.54	0.77	0.79
R-Squared Between	0.54	0.58	0.19	0.24	0.3	0.23

* $p<0.1$; ** $p<0.05$; *** $p<0.01$

The dynamics behind these observed effects could be explained among others by referring to a mix of economies and diseconomies of scale (Brock & Evans, 1989). The observation of an entrepreneurial innovation premium could be attributed to the higher likelihood that patents introduced by small businesses will be high impact ones, making the average small firm patent more technically and thus more economically important. This finding complies with Baumol's (2004) assignment of different roles to small and large firms in innovation systems with the former being relatively better at the introduction of radical new technologies and the latter in perfecting those by incremental improvements. The observed discount observed on the entrepreneurialisation of employment suggests that in the non-innovation-related aspects of business operations the economies of scale outweigh the diseconomies of scale. This observation counters earlier findings underlining the increasing importance of non-technologically oriented scale diseconomies that result from growing markets valuing specialized products, increasing advantages to flexibility in a globalized world, the rising availability of educated labour to recruit from and decreasing standard fixed costs of running a business (Brock & Evans, 1989).

Separate results for countries tagged by the EC as innovation leaders and laggards further reveal some of the potential deeper dynamics behind this. Not surprisingly, the innovation leaders turn out to be the driving force behind the productivity premiums to technological innovation in general and entrepreneurial innovation. The former and latter can be seen as highly intertwined: established knowledge-based economies possess the critical mass that is necessary to produce knowledge that matters. Knowledge stock growth in turn increases the potential for spill-overs of various ideas to entrepreneurs. On top of that, local rivalry between high-tech entrepreneurial ventures capturing the same localized knowledge flows increases their respective efficiency (Furman et al., 2002). The laggard countries appear to be the

driving force behind the productivity discounts associated with small firm employment share growth. The distinct geographic origins of the premium effect on entrepreneurial innovation and discount effect on entrepreneurial employment confirm the heterogeneous nature of the European economic landscape. Relying on Porter et al.'s (2002) framework of economic development to explain differences between split dataset results one could claim that it consists of less developed countries in a 'preliminary' efficiency-driven stage and more advanced countries in the 'final' innovation-driven stage (Porter et al., 2002; Acs et al., 2008).

In a complementary attempt to explain the additional explanatory power of entrepreneurial innovation in general we refer to the increasing disjunction between patents as measure of innovation and productivity in large firms: the availability of in-house IP departments increase their propensity to patent low-value inventions and tax optimization strategies applied by multinationals blur the value of license fees as proxy for added value.

Future research is necessary to further disentangle the mechanics behind the observed effects. Measurement of knowledge spill-overs could help to provide insights about their nature, origins and the direction in which they are heading. Adding proxies capturing the distinct drivers of scale diseconomies is another potential direction for future research. Further inquiry is also needed to list the policy implications of our findings.

References

Acs, Z. J., Desai, S., & Hessels, J. (2008). Entrepreneurship, economic development and institutions. *Small Business Economics,* 31, 219-234.

Audretsch, D. B. (2001). The dynamic role of small firms: evidence from the US, *Small Business Economics,* 18 (1/3), 13-40.

Baumol, W. J. (2004). Entrepreneurial enterprises, large established firms and other components of the free-market growth machine, *Small Business Economics,* 23, 9-21.

Brock, W. A. & Evans, D. S. (1989). Small business economics, *Small Business Economics* 1 (1), 7-20.

Ernst, H. (2003). Patent information for strategic technology management, *World Patent Information* 25, 233-242.

European Commission (2005). The new SME definition: user guide and model declaration, *Enterprise and industry publications,* European Commission.

European Commission (2015). Innovation Union Scoreboard 2015.

Eurostat (2014). Patent statistics at Eurostat: mapping the contribution of SMEs in EU patenting.

Fraunhofer Institute, Systems and Innovation Research (2003). Patents in the service industries. Final report European Commission Contract No. ERBHPV2-CT-1999-06.

Furman, J. L., Porter, M.E., & Stern, S. (2002). The determinants of national innovative capacity, *Research Policy,* 31, 899-933.

Nadiri, I. (1993). Innovations and technological spillovers, NBERWorking Paper 423.

Porter, M., Sachs, J., & McArthur, J. (2002). Executive summary: Competitiveness and stages of economic development. In M. Porter, J. Sachs, P. K. Cornelius, J. W. McArthur, & K. Schwab (Eds.), *The global competitiveness report 2001-2002* (pp. 16-25). New York: Oxford Univ.Press.

Romer, P. M. (1986). Increasing returns and long run growth*, Journal of Political Economy,* 94, 1002-1037.

Schumpeter, J. A. (1942). *Capitalism, socialism and democracy,* Harper and Row: New York.

Solow, R. M. (1956). A contribution to the theory of economic growth, *Quarterly Journal of Economics,* 70, 65-94.

Ulku, H. (2004). R&D, innovation, and economic growth: an empirical analysis, IMF Working Paper 04/185.

Veugelers, R. (2009). A lifeline for Europe's young radical innovators, Bruegel Policybrief 1.

Wong, P.K., Ho, Y.P., and Autio, E. (2005). Entrepreneurship, innovation and economic growth: evidence from GEM data, *Small Business Economics,* 24, 335-350.

Chemistry research in India: A bright future ahead

Swapan Deoghuria[1*] Gayatri Paul[2] and Satyabrata Roy[3]

[1]ccsd@iacs.res.in * *Corresponding Author*
Scientist-III, Indian Association for the Cultivation of Science, Jadavpur, Kolkata – 700032 (India)

[2]*libgp@iacs.res.in*
Sr. Doc. Assistant, Indian Association for the Cultivation of Science, Jadavpur, Kolkata – 700032 (India)

[3] *saturoy@gmail.com*
Guest Faculty, LIS Department, Jadavpur University, Jadavpur, Kolkata – 700032 (India)

Introduction

Chemistry is the most preferred research area among Indian scientists for quite some time in terms of total number of publication, global share, visibility and citation impact are concerned. Growth rate of India in chemistry research area is more than that of global growth rate as evidenced from the data covered in Web of Science database (WoS). The trend of research output in chemistry clearly indicates that India is steadily putting stiff challenge to traditionally established countries like Japan and Germany and even surpassed them in 2014 to acquire 3rd position in global ranking. From this study we predict that India will grow further in chemistry research area and even can put challenge to USA and China in long run.

The output and trend of science & technology (S&T) research in India are of considerable interest to scientometricians from all over the world for quite some time. Gupta and Dhawan (2009), Glänzel and Gupta (2008) and Gunasekaran, Batcha and Sivaraman (2006) have studied different aspects of S&T research in India.

Methodology

Data sources and processing

All bibliometric data have been extracted from WoS Core Collection of Thomson Reuters till April 30, 2015. The period for publication activity has been taken for six years (2009-2014) as findings till 2008 are available in literature.

Results and Discussions

In chemistry research area a total 1,045,343 number of papers has been published during the period 2009-2014. USA and China are leaders in this field in terms of number of publications with global share of 22.502% and 20.792% respectively. India is at 5th position with global share of 5.767%. Chemistry research output of ten most productive countries excluding USA and China in terms of global share has been shown in Figure 1. India's growth is very steady during this period and acquired 3rd position in 2014 followed by USA and China, with global share of 6.456%. India has published maximum number of research papers in Chemistry compared to other research areas and its global share in chemistry research has been increased steadily during 2009 to 2014.

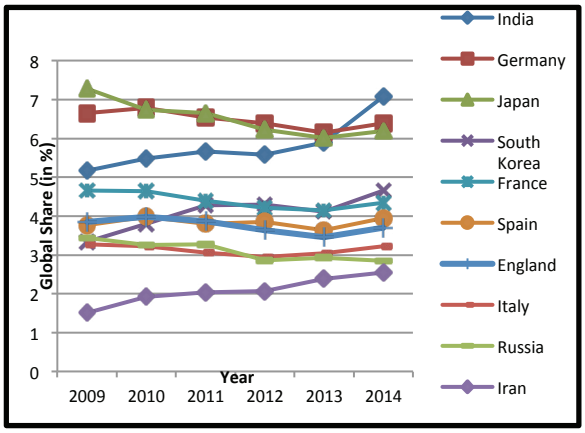


Figure 1. Global share of countries in chemistry.

It is evident from Figure 1 that global share of Japan has been decreased during 2009-2014 and its positions in global ranking have been fallen from 3rd position in 2009 to 5th position in 2014. Global share of Germany in Chemistry research has been decreased slightly during this period but Germany has managed to keep its position at 4th during the entire period. South Korea and Iran have increased their research output in chemistry steadily in terms of global share during this period. Research output of other countries (France, England, Spain, Italy and Russia) shown in this Figure are comparable to each other in chemistry and they are placed in between 7th to 11th positions during this period.

Table 1 shows India's ranking in major research areas covered in WoS during 2009-2014. In terms of number of publications and global share, India's performance is the best in Chemistry.

In Table 2 we have shown the h-index and average citation per article in chemistry during 2009-2012. We see that h-index and average citation per article are comparable with that of Japan and Germany.

Table 1. India's Position in major research areas in terms of global share.

Research Areas	2009	2010	2011	2012	2013	2014
Physics	10	9	8	8	7	**7**
Chemistry	5	5	5	5	5	**3**
Materials Science	7	6	6	6	**5**	6
Engineering	11	12	11	6	**4**	6
Computer Science	12	12	9	**3**	4	11
Biochemistry Molecular Biology	12	11	11	11	10	**9**
Neuroscience Neurology	18	17	17	16	**16**	17

Table 2. Comparison of citation and h-index of chemistry publications during 2009-2012.

Countries	2009		2010		2011		2012	
	h- index	Avg Citation	h- index	Avg Citation	h- index	Avg Citation	h- index	Avg Citation
India	83	11.54	78	10.39	66	8.41	57	6.49
Japan	106	15.83	97	13.81	89	11.63	66	7.94
Germany	118	19.86	125	19.29	95	14.04	74	9.94

Conclusions

This study clearly indicates the trends in chemistry research during 2009-2014 for most productive countries in terms of number of publications and global share. It is evident from the results that India has done remarkable progress in chemistry research area during this period. One of the reasons for this progress is that quite a few key persons in science policy makers in India are having chemistry background. Indian scientists working in the field of chemistry are more focused and recognized worldwide as many of them have been awarded TWAS prize and fellowship, FRS, and other distinguished international fellowships and medals. Strong collaboration between India and other countries in chemistry research is worth mentioning as 10,941 numbers of papers out of total 60,285 are published in collaboration. As a traditional subject, most of the Indian universities teach chemistry and around 40% of total publications is contributed by the universities. Research laboratories also get a steady flow of trained students with chemistry background from universities. Looking at the distribution of the publications to the institutes we see that CSIR laboratories publish most (11,037) followed by IITs (7,382) in chemistry. Some of the most productive laboratories in chemistry research in India are BARC (2,394), IICT (2,210), IISc (2,065), IACS (1580) and NCL (1,508). Prominent universities in chemistry research are JU (1,262), DU (1,182) and BHU (1,136). We see that there is almost no role of industries as per the funding of research is concerned in the field of chemistry in India. CSIR, DST and UGC are the major sponsors in chemistry research in India. As per the topic or subject category is concerned where Indian scientists publish more, we see Physical chemistry is the most focused (29%) followed by Organic (20%), Inorganic (11%), Analytical (10%), Applied (7%), Nanoscience (6%) and Atomic-Molecular (5%) respectively. The bright side of chemistry research in India is also reflected in the number of patents granted in this subject area. From Derwent Innovations Index of WoS, we see that out of total 462 numbers of patents granted to Indian innovators during 2009-2014, 330 numbers i.e. 71% are in the field of chemistry. Interestingly, DRDO, India holds most (79%) of the patents. The picture is not much different in Indian patent database (http://ipindiaservices.gov.in/publicsearch/), where we see 4,801 numbers of patents (i.e. 37%) have been granted in chemistry research area out of total 12,982 patents granted in all fields during 2009-14. India has a large consumer base. As a result chemical industries in different sectors like fertilizer, pesticide, plastic, paint, petro-chemical, medicine, cosmetics and health care products are thriving in India. So career as research scientist in chemistry is attractive for better placement in the R&D labs of those industries. India's contribution in chemistry research has been recognized by ACS and designated IACS, Kolkata on 15/12/1998 as International Historic Chemical Landmark for C V Raman and the Raman Effect.

References

Glänzel, W & Gupta, B. M. (2008). Science in India. A bibliometric study of national and institutional research performance in 1991-2006. *Proc. WIS.*

Gunasekaran, S., Batcha, M. S. & Sivaraman, P. (2006). Mapping chemical science research in India: A bibliometric study. *Annals of Library and Information Studies, 53*, 83-95.

Gupta, B. M. & Dhawan, S. M. (2009). Status of India in science and technology as reflected in its publication output in Scopus Int. databases, 1996 – 2006. *Scientometrics, 80*(2), 473-490.

Main Institutional Sectors in the Publication Landscape of Spain: The Role of Non-profit Entities

Borja González–Albo[1], Javier Aparicio[1], Luz Moreno-Solano[2], María Bordons[2]
[1] *borja.gonzalezalbo@cchs.csic.es; javier.aparicio@cchs.csic.es*
Transversal Support Research Unit (UTAI), Centre for Humanities and Social Sciences (CCHS), Spanish National Research Council (CSIC). Albasanz 26–28, 28037 Madrid (Spain)

[2] *luz.moreno@cchs.csic.es; maria.bordons@cchs.csic.es*
ACUTE Group, IFS, Centre for Humanities and Social Sciences (CCHS), Spanish National Research Council (CSIC). Albasanz 26–28, 28037 Madrid (Spain)

Introduction

The study of national efforts in R&D by institutional sector is a matter of great concern because sectors differ in their main activities, accounting systems, orientation towards research and type of R&D (OECD, 2003). However, bibliometric analyses at the level of institutional sectors are not very common because the assignation of centres to sectors is not free of difficulties and the resulting sectors may entail a certain degree of heterogeneity. The role of institutional sectors in the scientific activity of countries, either for the total country (Godin & Gingras, 2000; Moya et al., 2013) or in a given field (Lander, 2013), has been analysed in the literature, although studies dealing with specific sectors such as universities or companies are much more frequent.

In most countries, main institutional sectors in publications include universities, hospitals and public research centres, while papers from non-profit entities (NPE) are usually scarce. Although this applies in Spain, an impressive increase in papers from NPE has been observed in the last fifteen years. This paper aims to analyse the research performance of non-profit entities in Spain with regard to activity, impact and collaboration; to locate them in the national context; and to identify main types of active organisations.

Methods

Spanish publications (original articles and reviews), hereafter papers, covered by Web of Science (WoS, 2000-2011), search strategy CU=Spain and PY=2000-2011, are analysed. Six institutional sectors are identified in all addresses through a semi-automatic process (Morillo et al., 2013) followed by a manual revision to assess validity: companies, health sector, non-profit entities, public administration, public research centres and university. A full counting method is used.

The impact of publications is analysed through the percentage of papers in first quartile journals within each field (%Q1), normalised position (NP) (Bordons & Barrigón, 1992), relative impact factor (RIF), % non-cited papers and citations relative to country average (RC) (three-year citation window). The orientation of sectors towards collaborative research is explored through the number of authors per paper, number of institutions per paper and collaborative pattern (percentage of papers with a single institution, percentage of papers with national collaboration, percentage of papers with international collaboration). An in-depth analysis of NPE is carried out. The NPE's activity index (AI) in ten broad thematic areas is obtained to gain insight into the specialisation profile of these entities as compared to Spain.

Results

Main institutional sectors in Spanish papers in WoS (2000–2011) include university (66%), public research organisations (22%) and the health sector (18%). Non–profit entities amount to 10% of the papers, and show the highest increase during the period (3% of the country output in 2000 vs. 18% in 2011). This sector shows high specialization in Biomedicine (AI=1.59) and Clinical Medicine (AI=1.67). Collaboration in NPE is above the country average in terms of team size (11 vs. 8), number of institutions per paper (5 vs. 3) and share of collaborative papers (91% vs. 68%). NPE show also the highest shares of both nationally and internationally co-authored papers (75% vs. 41% and 45% vs. 40%, respectively). NPE display the highest percentage of papers in high-quality journals and the highest impact through relative citations (Table 1).

From the inspection and categorization of the NPE, the following organisational types emerge: foundations (50.3%), research networks (24.6%), consortia (16.0%), research management entities (12.2%), associations (6.5%), and scientific parks (1.0%). The highest increase during the period corresponds to research management entities and research

networks. Research management entities stand out because of their high figures in both the percentage papers in high impact factor journals and relative citations (Table 2).

Research management entities show the lowest proportion of papers with a single institution (2%), a high share of papers with national (89%) and international collaboration (68%), and the highest average team size. The highest share of papers in Q1 journals is observed for co-authored activity between national and foreign partners for all sectors except associations and research networks.

The specialization of NPE varies according to the organisational type: Biomedicine and Clinical Medicine for networks, consortia and foundations; Physics for research management entities; Biomedicine and Chemistry for scientific parks; and Engineering for associations.

Conclusions

The in-depth analysis of the NPE in Spain shows the rising trend of different organisational types which differ according to the field and respond to specific strategic procedures to manage research (creation of foundations in the context of medicine, networks for clinical research, scientific parks to link basic and applied research in the university context, etc.). Interestingly, some of these organisational types (research networks, consortia, parks) include cross-sector and cross-discipline collaboration which is supposed to lead to major discoveries in science and even to radical innovation. Collaboration in the context of the structured and stable framework provided by these organisational forms is more effectively enhanced than through occasional collaborative projects. Our data indicate the success of these emerging organisations in supporting/conducting high impact research.

Acknowledgements

The financial support of the Spanish Ministry of Science and Innovation (CSO2011-25102) is acknowledged.

References

Bordons, M. & Barrigón, S. (1992). Bibliometric analysis of publications of Spanish pharmacologists in the SCI (1984-89). Part II. Contribution to subfields other than "Pharmacology and Pharmacy", *Scientometrics*, 25 (3): 425–446.

Godin, B., & Gingras, Y. (2000). The place of universities in the system of knowledge production. *Research Policy*, 29 (2), 273–278.

Lander, B. (2013). Sectoral collaboration in biomedical research and development. *Scientometrics*, 94 (1), 343–357.

Morillo, F., Aparicio, J. González–Albo, & B., Moreno, L. (2013). Towards the automation of address identification. *Scientometrics*, 94 (1), 207–224.

Moya–Anegón, F., Chinchilla–Rodríguez, Z., Corera–Álvarez, E., González–Molina, A., López–Illescas, C., & Vargas–Quesada, B. (2013). *Indicadores bibliométricos de la actividad científica española 2010*. Madrid: FECYT.

OECD (2002). *Frascati Manual*. Paris: OECD.

Table 1. Number of papers and impact indicators by institutional sector in Spain (WoS 2000-2011)

	No. Papers	NP	%Q1	%Non cited papers	RC	RIF
Universities	271399	0.66	47.93	23.45	0.85	0.89
Public Research Centres	91095	0.74	62.41	12.94	1.31	1.24
Health sector	74337	0.59	39.66	21.32	1.20	1.16
NPE	41605	0.74	62.59	10.56	1.75	1.57
Public Administration	17238	0.66	49.04	20.65	1.01	0.96
Companies	15682	0.63	43.72	22.15	0.81	0.84

Table 2. Number of papers and impact indicators of the NPE by organisational type (WoS 2000-2011)

	No. Papers	NP	%Q1	%Non cited papers	RC	RIF
Foundations	20934	0.76	65.50	9.71	1.82	1.67
Research Networks	10249	0.75	63.16	7.18	1.83	1.74
Consortia	6651	0.73	60.83	9.88	1.69	1.55
Research Management Entities	5074	0.81	76.47	6.42	2.71	1.96
Associations	2692	0.66	47.73	20.84	0.90	0.94
Scientific Parks	310	0.76	66.11	8.71	1.21	1.55
Other NPE	1204	0.60	35.35	27.99	0.75	0.76

Reform of Russian Science as a Reason for Scientometrics Research Growth

Andrey Guskov

guskov@spsl.nsc.ru
The State Public Scientific Technological Library of Siberian Branch of the Russian Academy of Science, Voshod Str. 15, Novosibirsk (Russia)
Novosibirsk State University, Pirogova Str. 2, Novosibirsk (Russia)

Introduction

After the USSR had fallen down in 1990, there was a steady stagnation of Russian science for fifteen years. Iron curtain that separated soviet researchers from the international science disappeared, but research funds sharply decreased due to the economic problems. As a result, the number of publications registered in Web of Science, stayed between 30 000 and 34 000 per year. Thus, Russian science moved from the group of leading countries to the second dozen.

Restoration of Russian Science started in 2006 after government had introduced a new model of the research process. Essential part of the model was wide application of the formal scientific results assessment. This approach triggered a rapid growth of scientometrics publications written by mathematicians, physicists, philosophers and others. The main goal of this paper is to make a review of new Russian scientometrics landscape, which could help to determine its strengths and weaknesses and launch new collaborations.

Method

In this paper basic set of scientometric articles produced by Russian scientists is analysed. It consists of two periods: 1988-1999 and 2000-2014. The data for the first part (99 publications) was extracted from Russian Institute for Scientific and Technical Information database, abstract journal "Informatics" (Penkova, O. & Tyutyunnik V., 2011) Publications from 2000 until 2014 were requested from Russian Science Citation Index (national bibliometric database) by using context search with terms "bibliometric", "scientometric", and "webometric" (in Russian) in titles and annotations.

For every article in this set we identified topic category according to its title, annotation and, in some cases, full text. Afterwards, we analysed the distribution and dynamics of the categories and of the whole set.

Dynamics of Russian scientometric researches

Noteworthy, scientometrics in Russia has very meaningful historical background. It was Russian philosopher and mathematician V. Nalimov, who in 1969 introduced the term "Scientometrics" in his famous book. In 1973 Marshakova and Small simultaneously introduced co-citation analysis, which is used for research front findings now. Dutt, Garg & Bali in 2003 analysed fifty volumes of journal Scientometrics during 1978 to 2001 and examined the distribution of the output of different countries. According to their paper, former USSR contributed 59 of 1317 articles that are emphasized on history of science, theoretical studies and scientometrics distribution. Despite these go-ahead results, scientometric researches became a trend in Russia only after 2006 (Fig.1).

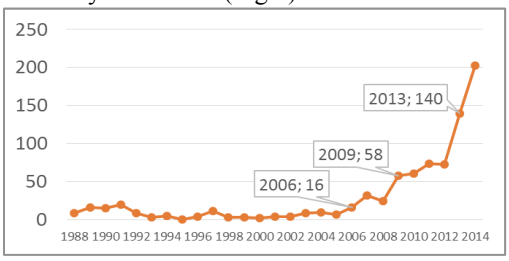


Figure 1. Dynamics of scientometric publications in Russian journals.

There are three sharp increases at Fig 1: in 2006, 2009, and 2013. The first growth in 2006 relates to the reformation of salary system, which implied significant dependency of the payment bonuses upon publication scores for every single scientist. Facing this new challenge, a number of researchers considered its fairness; some of them noticed the helpfulness of the bibliometric methods and started to apply it for their subject area. The second wave started in 2009th after the end of the salary system reformation. From that moment, every researcher became financially interested in improving his scientometric indicators. Research society had to analyze these changes, thus we can observe sharp increase in 2009th at Fig 1.

Despite the rapid growth before, in 2013th the number of scientometric publications had doubled. The reason is clear: in May 2012, President of Russia V.V. Putin proclaimed that the fraction of publications of Russian researches indexed by Web of Science in 2015th has to be greater than 2.44%. This was quite a big challenge for national science, because it literally meant that the annual number of articles has to be increased from 32-33 thousands in 2010-2011 to 46-50 in the next 3 years. The

reasons, the ways and the possibilities of that breakthrough were the main topics for discussion over the year. After that, in June 2013 another dramatic event occurred: restructuring of the Russian Academy of Science (RAS), headquarters of fundamental sciences. This tough stage was accompanied by criticism of the Academy for low scientometric indicators. Unfortunately, scientometrics has been used as an instrument for a radical transformation of management of Russian science.

Directions of researches

We defined 16 categories and analyzed the articles distribution (Fig.2). 33% of researches were devoted to a specific subject area investigation. It is followed by: development and applying of indicators (13%), general discussions about scientometrics and its place in research management (11%), impact-factors and journal improvement issues (7%), positions of Russian science in a global scope (6%). According to our estimates, from 50% to 75% of publications were made using bibliometric methods, principally in categories: "Subject areas", "Journals", "National science", "Dissertations", "Regional research", "Leading scientists research", "Science in HEI", "Conferences", "Organizations", "Collaborations", "Patents".

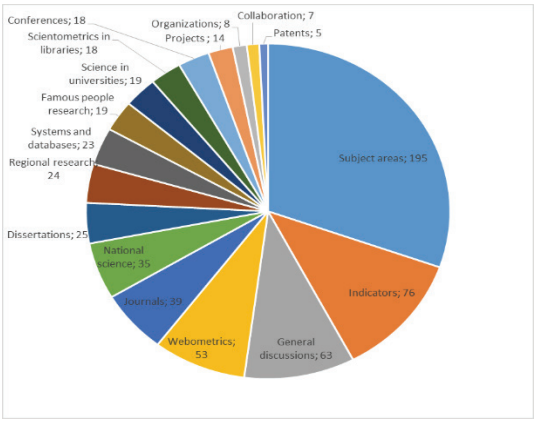


Figure 2. Distribution of scientometric researches by categories (number of publs.)

We determined the most developing categories and analyzed the dynamics. The main contribution to publication rise, shown at Fig.1, was made by "Subject area" category from 2007 to 2012. The second contributing category "Indicators" contains a number of articles about publications and citations amount, impact factor and Hirsh index. The third category supports general scientific discussion about scientometrics, started in 2009. Three more categories significantly increased in 2013: "Journals", "Science in universities", "Systems and databases".

Conclusion

Figure 1 can be thought of as an indirect measure of the influence of the State on Russian Science. Indeed, there was a lack of scientometricians and poor scientometric publication activity in Russia before 2006th, the very beginning of reformation. The following alterations made many researches slow down or suspend what they had been doing before and start making their own scientometric investigations. The more severe were the changes, the more scientists were influenced. Furthermore, it seems there were no other reasons for the mentioned breakthrough. At first glance, scientometrics is supposed to benefit from it. That would be so, excepting two facts. First, concerning scientometrics as an instrument of reformation, many scientists consider it primarily as a stick for punishment and do not trust it. This creates quite a negative environment for further development, but this story has already happened. "When a system of assessing and funding researchers was introduced in South Africa, there were cases when scientists attacked scientometrics…" (Pouris, 1994). Second, the most of the scientometric researches, which were published in Russia the last years, relate to one of the groups: 1) Position of the scientometrics and its indicators in the processes of the management of Russian science. 2) Bibliometric researches of science disciplines and Russian science as a whole. 3) Bibliometric and webometric researches of various sources of publications: journals, organizations (incl. universities), famous scientists, conferences, projects, dissertations sets and so on. Since those three groups include up to 90% of publications, there is not much space left for more complicated and go-ahead researches, such as collaboration studies, research fronts detecting, R&D cycle analysis, altmetrics, society impacts, etc. At the moment, scientometrics in Russia remains the "product for internal use" mostly. Still, we expect the internalization of this research field and the increase of the visibility of Russian publications worldwide.

Acknowledgements

The author acknowledges Denis Kosyakov, Irina Selivanova and Mikhail Tsentalovich.

References

Dutt, B., Garg K. & Bali, A. (2003) Scientometrics of international journal Scientometrics. *Scientometrics*, Volume 56, No. 1, pp 81-93. DOI: 10.1023/A:1021950607895

Penkova, O. & Tyutyunnik V. (2001) Informetrics, scientometrics and bibliometrics: the sciento-metrics analysis of current state. *Vestnik Tomskogo gosudarstvennogo universiteta.* 6(1) 86-87.

Pouris, A. (1994) Is scientometrics in a crisis? *Scientometrics*, Volume 30, Nos 2-3, pp 397-399. DOI: 10.1007/BF02018111

Leadership among the Leaders of the Brazilian Research Groups in Marine Biotechnology

Sibele Fausto[1] and Jesús P. Mena-Chalco[2]

[1] *sifausto@usp.br*
University of São Paulo, Rua da Biblioteca, s/n, Complexo Brasiliana,
São Paulo, SP, CEP 05508-050 (Brazil)

[2] *jesus.mena@ufabc.edu.br*
Federal University of ABC, Av. dos Estados, 5001, Santo André, SP, CEP 09210-580 (Brazil)

Introduction

The Marine Biotechnology (MB) research area is gaining increasing relevance in Brazil. Its analysis is a challenge owing to the inherently multidisciplinary nature, and the study of research groups (RGs) may support this work. The task of analysing RGs is facilitated in Brazil, which has a national source gathering the country's RGs, maintained by the National Council for Scientific and Technological Development (*Conselho Nacional de Desenvolvimento Científico e Tecnológico* - CNPq): the Directory of Research Groups of the Lattes Platform (*Diretório dos Grupos de Pesquisa da Plataforma Lattes,* http://lattes.cnpq.br/web/dgp), with information from RGs related to: i. institutional headquarters; ii. Research Group name; iii. First leader name, iv. Second leader name (if any), and v. Predominant area. This source allows automatic data extraction already made available by research groups, allowing for full and systematic exploitation. This work aims to present first findings from exploitation on research groups in MB existing in Brazil registered in the Directory of RGs of the Lattes Platform, checking the collaboration networks formed by the leaders of these groups, mainly highlighting the natural influence that leaders have on other peers, meaning a leadership, focusing on research groups through the topological properties of networks with the use of Social Network Analysis (Abbasia, Wigand & Hossain, 2014), in order to behold their evolution and the role of the RGs' leaders in MB in Brazil and testing if it is possible to establish a relationship between the degree of leadership of the leaders considering topological information from networks.

Methods

This initial approach is focused on three points: 1. networks characterization in number of RGs involved, the active institutions and their location, and the dominant areas in multidisciplinary research; 2. description of the dynamic aspect of the network formed by these RGs through its evolution over the last 15 years, distributed in three five-year periods; and 3. determination of the "degree of leadership" of these networks' leaders, as measured by AuthorRank indicator, which is a numerical value that indicates the impact of a member in collaboration graph. This measurement is similar to PageRank for directed graphs (with weights) (Liu et al., 2005). Thus, the aim was to consider this indicator as an attribute of the leadership for the leaders of these RGs in the analyzed period.

Data collection and analysis

First, the MB research groups were identified by search using 37 MB terms raised in the related literature. Following, it was obtained data related to RGs such as institutions involved, 1st Leader name, and Main Area, allowing identify the Lattes ID (researcher identification number registered in the Lattes Platform) of the groups' leader. Second, we used scriptLattes tool (Mena-Chalco & Cesar Junior, 2009) in order to extract information associated with all the investigated leaders during the period of 15 years (1999-2013). We obtained data from the scientific production of each leader related to total articles, books, book chapters, and conference papers. For data analysis, we consider the professional addresses recorded for each leader to obtain the geographic location of each group through Google Maps tool. We obtained lists of full papers (solely) of the groups' leaders published in journals, and with scriptLattes tool we identify all publications in co-authorship. In addition, there were obtained the endogenous networks (internal collaboration) of the leaders. The AuthorRank was calculated for each actor. This indicator is commonly used for measuring the impact of members of an academic collaboration network (Liu et al., 2005). Our analysis was outlined considering four time periods: A global period (1999-2013) and three five-year periods: 1999-2003, 2004-2008, and 2009-2013. This division into different periods allows to study distinct topological characteristics of the network and its evolution.

Results

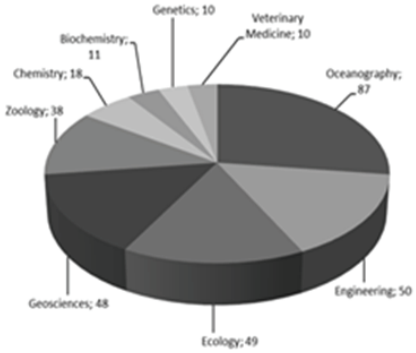


Figure 1. Main subject areas of the Brazilian research groups in Marine Biotechnology

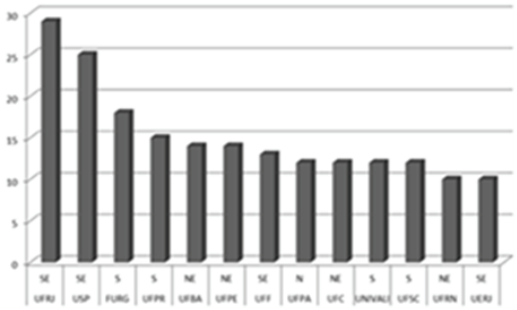


Figure 2. Brazilian institutions with over ten research groups in Marine Biotechnology

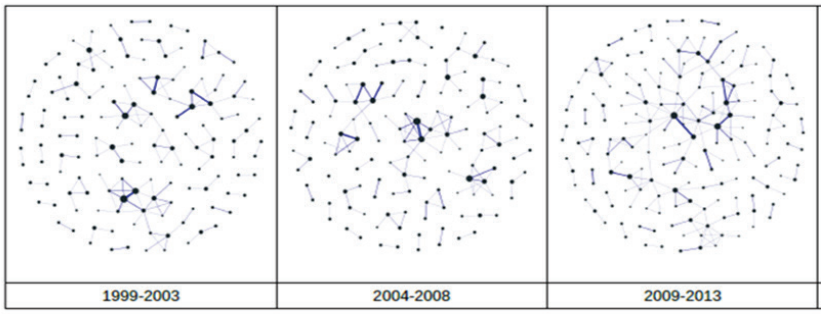


Figure 3. Co-authorship networks among leaders associated with the Brazilian research groups in Marine Biotechnology

Table 1. AuthorRank of the Leaders of the Brazilian research groups in Marine Biotechnology

Author Rank	Leader	Institution/Region
4.10	Teixeira, VL	UFF/SE
3.59	Colepicolo Neto, P	USP/SE
3.46	Rörig, LR	UFSCar/SE
3.33	Pereira, RC	UFF/SE
2.94	Pinto Jr, E	USP/SE
2.75	Mantelatto, FLM	USP/SE
2.67	Amado Filho, GM	JBRJ/SE
2.55	Sampaio, LAN	UFRN/NE
2.34	Bianchini, A	UFRN/NE
2.33	Berlinck, RGS	USP/SE

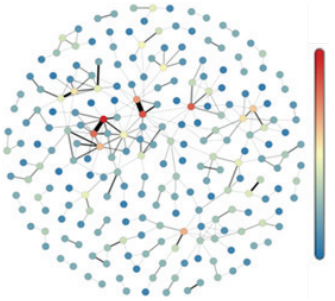


Figure 4. AuthorRank of the Leaders of the Brazilian research groups in Marine Biotechnology: co-authorship network

Discussion and conclusion

There are 402 RGs working in one or more topics related to the MB field from 34 different subject areas, main ones showed in Figure 1. RGs are from 110 institutions geographically concentrated along the Brazilian coast (South and southeast prevailing in number of institutions and research groups – Figure 2). We identified the leadership of the ten most active researchers in the co-authorship networks, with AuthorRank varying between 2.33 and 4.1 (Table 1). It was observed that there is a systematic increase in academic interactions during the considered period (Figure 3) and that academic leadership is not uniform among the leaders (Figure 4). The task of characterizing the emerging area of research in MB has grown in importance in Brazil, and this work relates to this issue.

References

Abbasia, A., Wigand, R. T. & Hossain, L. (2014). Measuring social capital through network analysis and its influence on individual performance. *Library & Information Science Research*, 36, 66–7.

Liu, X., Bollen, J., Nelson, M. L., & van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information processing & management*, 41(6), 1462-1480.

Mena-Chalco, J. P. & Cesar Junior, R. M. (2009). ScriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, 15(4), 31-39.

An Empirical Study on Utilizing Pre-grant Publications in Patent Classification Analysis

Chung-Huei Kuan[1] and Chan-Yi Lin[2]

[1] *maxkuan@mail.ntust.edu.tw*
National Taiwan University of Science and Technology, Graduate Institute of Patent Research, No. 43 Sec. 4 Keelung Rd., Taipei City, 106 (Taiwan, R.O.C.)

[2] *u9703220@gmail.com*
National Taiwan University of Science and Technology, Graduate Institute of Patent Research, No. 43 Sec. 4 Keelung Rd., Taipei City, 106 (Taiwan, R.O.C.)

Abstract
Patent classification analyses are usually conducted using issued patents. Issued patents however suffer lengthy examination and the derived analytic results reflect R&D activities occurring considerable time in the past. The only option for an analyst to reduce such observational time delay is to use the so-called pre-grant publications (PGPubs) that are open to public 18 months after patent applications are filed. The PGPubs and their corresponding issued patents are both assigned classification symbols. If the two sets of symbols are very different, using patent classification analysis on PGPubs to observe R&D activities is dubious. This study therefore compares the United States Patent Classification (USPC) symbols assigned to about 235,000 pairs of U.S. utility patents issued in 2012 and their PGPubs in three ways, each corresponding to an approach of a conventional patent classification analysis: (1) considering only the class codes of the main classification symbols; (2) considering only the main classification symbols; and (3) considering both main and auxiliary classification symbols. The study finds that only the class codes of the PGPub main classification symbols are reliable enough for patent classification analysis as there are about 78% of the PGPubs have identical class codes as their corresponding issued patents.

Conference Topic
Patent analysis

Introduction
A patent application is classified during its prosecution process based on its inventive content by an examiner and one or more classification symbols are assigned in accordance with a standard scheme such as International Patent Classification (IPC), Cooperative Patent Classification (CPC), U.S. Patent Classification (USPC), etc. Patent classification analysis (PCA) is a popular practice by patent analysts using the patent classification symbols, and it is so popular that, to the authors' knowledge, all commercial patent analytic systems/services, such as Thomson Innovation® and WIPS Global®, have various types of PCA built-in.

A common type of PCA is to investigate the R&D focuses of an entity (i.e., a company, an institute, a country, a technical field, etc.). An analyst gathers the patents affiliated with the entity, collects the classification symbols assigned to these patents, counts the number of times each classification symbol is assigned to these patents, and usually produces a diagram such as a histogram, a heat map, etc., to visually manifest the assignment frequencies of the classification symbols. By observing the diagram, the analyst then claims that the entity has its R&D focused in a few technical areas denoted by the most frequently assigned classification symbols.

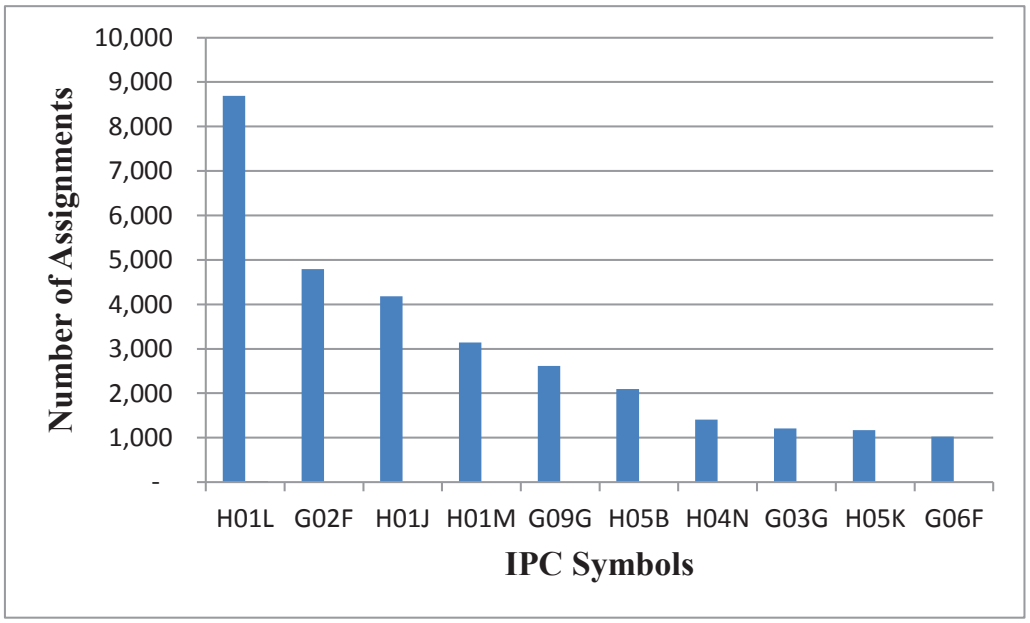


Figure 1. A sample histogram from a fictitious PCA.

A sample histogram from a fictitious PCA using IPC symbols for a company is shown in Figure 1. As illustrated, the company is considered to have its R&D effort mainly focused in the field Semiconductor Devices denoted by the most frequently assigned IPC symbol H01L.

Other than the real-life application described above, patent classification symbols are considered as a viable source of technological information by researchers, and various types of PCA have been proposed in the literature. To mention just a few, the number of different classification symbols assigned to an entity's patents is used as a proxy to the entity's technological diversity (cf. Lerner, 1994), the co-classification of patents (i.e., patents assigned one or more identical classification symbols) is used to investigate the linkage among technologies (cf. OECD, 1994), or the relationships among organizations (cf. Leydesdorff, 2008). There are also studies investigating the technological relatedness of two entities using the classification symbols assigned to their patents (cf. Jaffe, 1986; 1989). In addition, the classification symbols of a patent's forward and backward citations are used to evaluate the patent's "generality" and "originality" (cf. Henderson, Jaffe, & Trajtenberg, 1997). However it should be noted that there are opinions considering the existing patent classification schemes are "never intended to provide conceptual delineations of technology areas, but instead identify inventions by function at very low levels of abstraction in order to serve as aids to prior art searching" (Allison et al., 2004).

As described above, PCA can be used to observe the focus of an entity's R&D activities up to the time of analysis or, if the entity's latest patents are gathered, of the entity's recent R&D activities. However, what is revealed by the latter is actually not the R&D activities happened around the time of analysis but a considerable amount of time in the past. To see this, the curve with diamond marks in Figure 2 depicts the distribution of U.S. utility patents issued in the year 2012 according to their application years. About three quarters of the 2012-issued utility patents are actually filed between 2007 and 2010. In other words, if a histogram similar to Figure 1 is derived from these 2012-issued patents, the revealed R&D focuses actually occur and disperse in a period of time quite in the past.

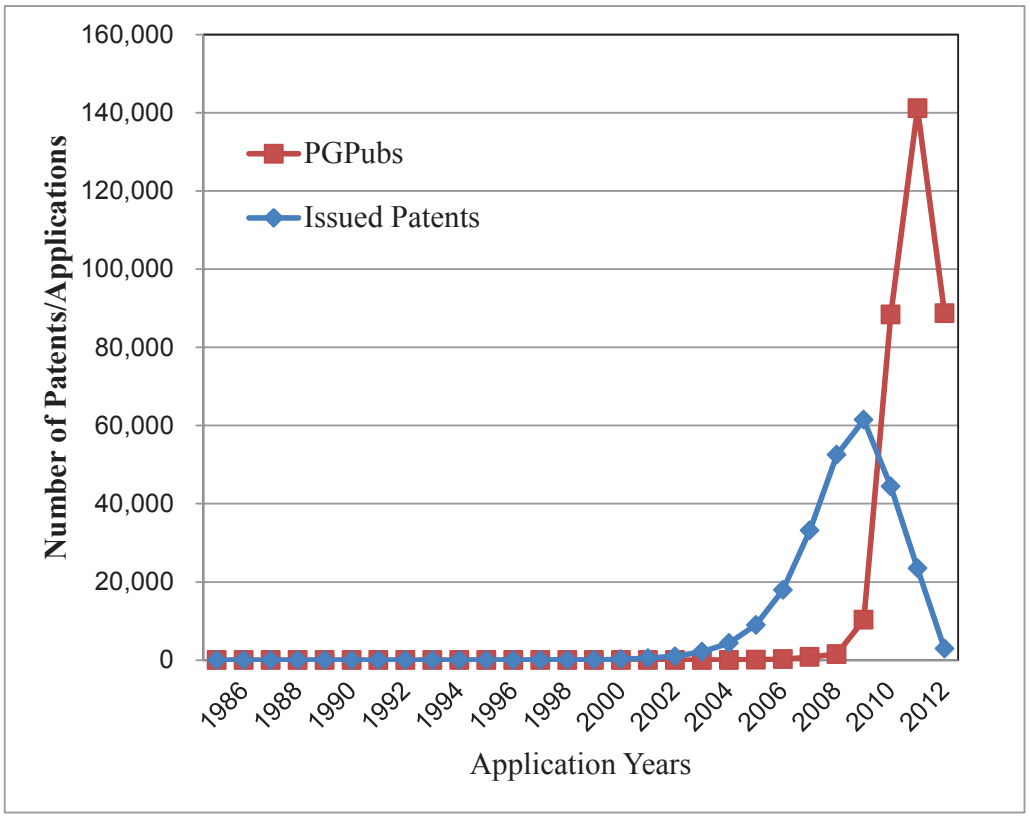


Figure 2. Distributions of 2012 issued patents and PGPubs based on application years.

The only possible way to reduce this time delay is to use the so-called *pre-grant publications* (PGPubs), instead of the issued patents. A patent application usually undergoes an early publication process before the patent is issued by the authority or before the patent application is given up by the applicant. Again taking Figure 2 as example, the curve with square marks depicts the distribution of U.S. PGPubs published in the year 2012 according to their application years. As illustrated most PGPubs are filed between 2010 and 2012, which are concentrated in a more limited period of time and in a more recent past.

The early publication process is a common practice for authorities across various nations and regions. For example, U.S. Patent Act (35 U.S.C. § 122(b)) specifies that, "each application for a patent shall be published … promptly after the expiration of a period of 18 months from the earliest filing date for which a benefit is sought under this title." There are indeed exceptions that an application is not early published if the application is (i) no longer pending; (ii) subject to a secrecy order; (iii) a provisional application; (iv) an application for a design patent; or (v) requested by the applicant. These exceptions are not common and, for utility patent applications, which are the most common type of patent applications, it is very possible that an issued utility patent is early published. According to our statistics, there are 253,580 utility patents issued in the year 2012 and 17,993 of them (7.1%) do not have corresponding PGPubs.

When a patent application is filed, the patent application is initially classified and classification symbols are assigned so as to route the patent application to an appropriate examiner team (USPTO, 2004). Then, after the patent application has undergone substantive examination, its examiner may alter the initial classification and assign different classification symbols (USPTO, 2005). As such a PGPub and its subsequently issued patent have their respective classification symbols and their classification symbols may not be identical.

PCAs usually utilize the issued patents, instead of PGPubs, most likely due to that the PGPubs have not undergone substantive examination, and their classification symbols may

not fully reflect their inventive contents. Yet PGPubs are better subjects for investigating the latest R&D focuses as they do not suffer lengthy pendency and strict screening by the examination process as reflected in Figure 2.

This study therefore tries to investigate the adequacy of using PGPub classification symbols for PCA. If the answer is yes, analysts can effectively reduce the time delay of their analytic observations to about 18 months, which is a significant improvement. On the other hand, even if the answer is no, analysts would know that PGPub classification symbols are not reliable, and they should avoid using them or at least be cautious about PCAs based on PGPub classification symbols.

Methodology

To investigate the adequacy of PGPub classification symbols for PCA, we collected U.S. utility patents issued in 2012 and their corresponding PGPubs for comparison. Utility patent is chosen because, for the three types of U.S. patents, utility patent is the most common and numerous one, design patents do not undergo the early publication process, and there are only a small number of plant patents. According to our statistics, there are only 868 plant patent applications filed each year between 1992 and 2011 on the average.

Each U.S. utility patent/PGPub is classified with three classification schemes: IPC, CPC, and USPC, and we choose the USPC symbols for comparison. This is because USPC is the default scheme for United States Patent and Trademark Office (USPTO) (USPTO, 2012), the IPC symbols are most likely machine-converted from the USPC symbols, and the CPC are not popular yet. Most importantly, USPC scheme does not have versions as it is updated every two months and the USPC symbols of all documents contained in USPTO databases are thoroughly and automatically re-classified accordingly (Wolter, 2012). In other words, when the USPC symbols of an issued patent are compared against those of its PGPub, whether the USPC symbols are of the same version is not an issue. One may question that USPC, as a domestic scheme, may not be representative. However, we believe that what this study observes from using U.S. patents and USPC could provide us at least some hint when dealing with patents of different countries and using different classification schemes.

Like all other classification schemes, USPC provides a hierarchical taxonomy of technical areas. Each USPC symbol contains a class code and a subclass code separated by "/." For example, a USPC symbol 623/2.1 has class code 623 and subclass code 2.1. The class code (e.g., 623) represents a highest level of non-overlapping technical area whereas the subclass code (e.g., 2.1) represents a lower level of technical area belonging to the one denoted by the class code. For subclass codes under the same class code, they may have hierarchical relationship among themselves. For example, 623/2.11 and 623/2.12 represent parallel technical areas but the two technical areas both belong to the technical area denoted by the symbol 623/2.1 (USPTO, 2012).

A U.S. utility patent/PGPub is assigned one or more USPC symbols. Among them, one and only one is expressed in boldface in the patent/PGPub documents. For issued patents, the official name for the bold-faced symbols is *original classification* symbols and, for the normal-faced symbols, *cross-reference classification* symbols by USPTO. As to PGPubs, the official name for the bold-faced ones is *primary classification* symbols and, for the normal-faced ones, *secondary classification* symbols. For simplicity's sake, we refer to the bold-faced symbols as the *main classification* symbols whereas the rest of the normal-faced symbols as the *auxiliary classification* symbols, whether or not they are from issued patents or PGPubs. The main classification covers the novel and non-obvious information contained in a patent/PGPub whereas the auxiliary classification covers other information considered to be valuable for searching (USPTO, 2012).

To determine whether PGPub classification symbols is adequate for PCA, we use the classification symbols assigned to the corresponding issued patents as reference as they are assigned by examiners after substantive examinations and therefore assumed to have better reflected the inventive contents of the patents.

Table 1 provides a number of examples where the sets of classification symbols assigned to three U.S. utility patents issued on 2015/02/10 and their PGPubs are listed side by side for comparison. As illustrated in Table 1, the two sets of classification symbols may not be identical, and the set assigned to the issued patent indeed seems to be more detailed than that assigned to the corresponding PGPub.

Table 1. The classification symbols assigned to three sample pairs of PGPubs/patents.

PGPub no./Patent no.	*PGPub symbols*	*Patent symbols*
20140289912/8,955,161	**850/18**	**850/1**; 250/339.11; 250/339.14; 73/105; 850/5; 850/50; 850/6
20120124680/8,955,160	**726/34**	**726/34**
20110252484/8,955,159	**726/32**	**726/32**; 380/201; 705/57; 726/27; 726/31; 726/33

There are quite some researches involving the measurement of similarity between nodes in a hierarchical taxonomy of concepts, which can be applied to classification symbols as well. For example, in one so-called edge-based approach, the similarity between two nodes is calculated based on the numbers of edges from the root of the hierarchical structure to the two nodes and to their nearest common ancestor node (Slimani, Yagahlane, & Mellouli, 2008). Similar edge-based approaches can be found in McNamee (2013). There are also so called node-based approaches, which capture a node's feature in the hierarchical structure as a vector and calculate a similarity measure based on the concept vectors of two nodes (cf. Liu, Bao, & Xu, 2012).

These studies do have their academic merit but cannot directly tell us whether PGPub classification symbols is reliable or not for PCA. We therefore adopt a different and practical treatment to the comparison of the classification symbols. First, we notice that existing commercial analytic systems/services conduct PCA using one of three simple approaches:

- PCA using Approach 1 counts only the class codes of the patent or PGPub main classification symbols so as to obtain a broad picture of the distribution of R&D activities;
- PCA using Approach 2 counts only the main classification symbols and ignores all auxiliary classification symbols of patents or PGPubs, considering that the main classification symbols are the most representative ones; and
- PCA using Approach 3 counts all patent or PGPub classification symbols with no distinction between main and auxiliary classification symbols, believing all classification symbols are equally important.

To demonstrate the three approaches, using the Patent Symbols column listed in Table 1 as example:

- Approach 1 counts the class codes 850 as being assigned once, 726 being assigned twice;
- Approach 2 counts each of the main classification symbols 850/1, 726/34, and 726/32 as being assigned once; and
- Approach 3 counts each of the 14 classification symbols as being assigned once.

Please note that, to the authors' knowledge, commercial analytic systems/services ignore the hierarchical relationship between classification symbols. For the above example, 850/6 is actually a technology area belonging to that of 850/5 but commercial analytic systems conducting PCA using Approach 3 treat 850/5 and 850/6 as denoting distinct technology areas probably for simplicity's sake.

Then, to see whether PCA using one of the above approaches on PGPubs classification symbols would deliver trustworthy result, we conduct three analyses as follows, each corresponding to one of the approaches above:

- Analysis 1 compares the main classification class codes of PGPubs to those of the corresponding issued patents.
- Analysis 2 compares the main classification symbols of PGPubs to those of the corresponding issued patents and calculates the consistency rate.
- Analysis 3 compares the sets of classification symbols of PGPubs to those of the corresponding issued patents.

Then all three analyses calculate the percentage of PGPubs having *identical* main classification class codes, main classification symbols, and sets of classification symbols to their corresponding issued patents. Since commercial analytic systems/services ignore the hierarchical relationship between classification symbols, our three analyses follow the same practice.

A 100% percentage indicates that PCA on PGPubs using one of the approaches would yield a result identical to that using their issued patents, meaning that using PGPubs can achieve reduced time delay with total accuracy. But a 0% percentage implies that PCA on PGPubs using one of the approaches delivers totally incorrect result. We therefore specifically refer to the percentage as *consistency rate* so as to avoid confusion with the general term *percentage*.

If statistically there is a very high consistency rate or similarity from the PGPubs, a histogram such as Figure 1 obtained from PGPubs using Approach 1, 2, or 3 would be very close to one from the corresponding subsequently issued patents. An analyst then can confidently utilize the PGPubs for PCA by Approach 1, 2, or 3 and achieve a reduced time delay.

To demonstrate the three analyses, again using the three sample pairs of PGPubs/patents listed in Table 1 as example:

- Analysis 1 shows that PCA using Approach 1 on PGPubs has a 100% consistency rate (i.e., all three pairs' PGPubs have identical main classification class codes to those of their issued patents);
- Analysis 2 shows that PCA using Approach 2 on PGPubs has a 66% consistency rate (i.e., except the first pair, the other two pairs' PGPubs have identical main classification symbols to those of their issued patents); and
- Analysis 3 shows that PCA suing Approach 3 on PGPubs has a 33% consistency rate (i.e., only the second pair's PGPub has an identical set of classification symbols to that of its issued patent).

For PCA using Approach 3, the simple consistency rate described above is too narrow to give us a complete picture. For example, even though the two sets of classification symbols from the third pair of patent/PGPub listed in Table 1 are different, the PGPub classification symbol {726/32} is actually a proper subset of the issued patent's classification symbols {726/32, 380/201, 705/57, 726/27, 726/31, 726/33} and therefore still captures a portion of the inventive content. The calculation of the consistency rate however ignores this condition.

Therefore in conducting Analysis 3, we divide the PGPub-patent pairs into 5 categories based on the relationships between their sets of classification symbols so as to gain more insight.

- Category 1: their sets of classification symbols are identical (i.e., $\{\text{PGPub}\} = \{\text{Patent}\}$).
- Category 2: their sets of classification symbols are entirely different (i.e., $\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{PGPub}\} \cap \{\text{Patent}\} = \emptyset$).
- Category 3: the PGPub's set of classification symbols is a proper subset of that of the corresponding patent (i.e.,$\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{PGPub}\} \subset \{\text{Patent}\}$).
- Category 4: the patent's set of classification symbols is a proper subset of that of the corresponding PGPub (i.e.,$\{\text{PGPub}\} \neq \{\text{Patent}\}$ and $\{\text{Patent}\} \subset \{\text{PGPub}\}$).

- Category 5: their sets of classification symbols are not entirely different, do not belong to each other, and have a non-empty intersection (i.e.,$\{\text{PGPub}\} \neq \{\text{Patent}\}$,$\{\text{PGPub}\} \not\subset \{\text{Patent}\}$, $\{\text{Patent}\} \not\subset \{\text{PGPub}\}$, and $\{\text{Patent}\} \cap \{\text{PGPub}\} \neq \emptyset$).

Then, for the patent/PGPub pairs belonging to each category, we calculate an average Jaccard Coefficient (Jaccard, 1901) as expressed in (1) where {PGPub} and {Patent} are the two sets of classification symbols assigned to the PGPub and the corresponding issued patent, respectively. Jaccard Coefficient, or Jaccard Index, or Jaccard Similarity Coefficient, was originally designed for comparing similarity between sample sets, and has already been applied in patent bibliometrics such as co-citation analysis (Small, 1973). Here we use it to capture the degree of discrepancy between {PGPub} and {Patent}.

$$J = \frac{|\{\text{PGPub}\} \cap \{\text{Patent}\}|}{|\{\text{PGPub}\} \cup \{\text{Patent}\}|} \quad (1)$$

Findings

We collected 253,580 utility patents issued in the year 2012 from USPTO database. After removing those having no corresponding PGPub, those having no classification symbol (e.g., these patents are withdrawn and withdrawn patents do not have patent classification symbols recorded in the USPTO database), and for unknown reason those having no main classification symbols, there are total 234,966 patents eligible for analysis. As mentioned in the previous section, USPC is updated every two months and all patents are re-classified accordingly. We collected the USPC symbols assigned to the 234,966 patents and their corresponding PGPubs under the USPC scheme up to 2013/10/31.

An initial statistics shows that the 234,966 patents have average 3.9 USPC symbols and their corresponding PGPubs have average 2.2 USPC symbols, and that 64.16% of the 234,966 patents have a greater number of USPC symbols than that of the corresponding PGPubs, indicating that issued patents seem do have more careful assignment of classification symbols than their PGPub counterparts. In some extreme cases, PGPub No. 2010/0316607 has the greatest number of USPC symbols (48) among all PGPubs whereas patent No. 8,179,540 has the greatest number of USPC symbols (65) among all patents. The latter is also the case having the greatest difference (63) between the issued patent and the corresponding PGPub.

Analysis 1

For each pair of the 234,966 PGPubs and corresponding issued patents, we compared the class code of the PGPub's main classification symbols against that of the corresponding issued patent, and we found that the consistency rate is 77.89%. That is, 183,024 out of the 234,966 pairs of PGPubs and patents have identical main classification class codes, and the remaining 51,942 pairs (22.11%) have difference main classification class codes. In other words, there is a 22.11% probability that a PGPub's main classification class code does not accurately reflect the inventive content of the corresponding patent.

Analysis 2

For each pair of the 234,966 PGPubs and corresponding patents, we compared the main classification symbol of the PGPub against that of the corresponding issued patent, and we found that the consistency rate drops to only 36.42%. That is, 85,584 out of the 234,966 pairs of PGPubs and patents have identical main classification symbols, and the rest 149,382 pairs (63.58%) have different main classification symbols. In other words, there is a very

significant 63.58% probability that a PGPub's main classification symbol does not accurately reflect the inventive content of the corresponding patent.

Analysis 3

For the 234,966 pairs of PGPubs and corresponding patents, we categorized them into 5 categories based on the relationships between their sets of classification symbols, and calculated the average Jaccard Coefficient for each category. The result is summarized in Table 2.

Table 2. Comparison result from Analysis 3.

Category	*Pairs*	*Percentage*	*Avg. Jaccard Coefficient*	*Std. Deviation*
1	14,958	6.37%	1	0
2	89,981	38.30%	0	0
3	63,057	26.84%	0.34	0.16
4	10,693	4.55%	0.45	0.15
5	56,277	23.95%	0.22	0.11

As illustrated, PGPubs in Category 1 are those having identical sets of classification symbols to their issued patents and their share (6.37%) among the 234,966 PGPubs is exactly the consistency rate of Analysis 3.

PGPubs in Category 2 are those having totally different sets of classification symbols from their issued patents and, for a PCA on these Category-2 PGPubs using Approach 3, the analytic result would be totally incorrect, but PGPubs of this category has the greatest share (about 38%) among all PGPubs.

PGPubs in Category 3 are those having sets of classification symbols being proper subsets to those of their issued patents, and cover about 27% of all PGPubs. For these Category-3 PGPubs, their classification symbols capture only 34% of the inventive content as reflected by their average Jaccard Coefficient. We can imagine that, for a PCA on Category-3 PGPubs using Approach 3, a histogram such as Fig. 1 would miss a significant amount of information.

Category 4 is a special case where PGPubs have sets of classification symbols that are proper supersets to those of the corresponding issued patents, and therefore covers the smallest share (less than 5%). For these Category-4 PGPubs, their classification symbols capture all inventive content but unfortunately provide on the average 55% (1-0.45) surplus and erroneous information. Again we can imagine that a histogram from PCA on Category-4 PGPubs using Approach 3 would contain too much noise.

Category 5 is a combination of Categories 3 and 4, meaning these 24% of the PGPubs have sets of classification symbols that not only miss significant amount of information but also provide significant amount of erroneous information, as reflected by the very limited average Jaccard Coefficient (0.22).

Conclusion

This study arises out of an attempt to use PGPub classification symbols for PCA so as to investigate an entity's latest R&D focuses with limited time delay. It is however speculated that the PGPub classification symbols are not carefully assigned and their adequacy for PCA has to be determined first.

We therefore gathered 234,966 pairs of issued patents and corresponding PGPubs, and compared their classification symbols in accordance with the three approaches that a commercial patent analytic system/service usually employ.

Assuming that the classification symbols of the corresponding issued patents better reflect the inventive contents of the patents and as such using them as reference, we find that, if the commercial patent analytic systems/services count the main classification symbols, or the entire sets of classification symbols of the PGPubs for PCA, only 36.42% of the PGPubs have identical main classification symbols, and only 6.37% of the PGPubs have identical sets of classification symbols to those of the corresponding issued patents. PCA using PGPubs as described can hardly be considered as reliable.

The best candidate for using PGPubs in PCA is the PGPubs' main classification class codes. We find that as high as 77.89% of the PGPubs have identical main classification class codes to those of the corresponding issued patents. The main classification class codes, however, represent the broadest technical areas and using them to investigate R&D focuses would provide only limited insight.

This study can be further carried out as follows. In order to make the main classification class codes even more useful for PCA, the consistency rate for each individual class can be determined. For some classes that have statistically very high consistency rate, PGPubs assigned with these class codes can be used for PCA with high confidence whereas, for classes of low consistency rate, an analyst should avoiding using them for PCA.

Additionally, one may be curious about why some class codes reveal higher consistency rates than the others. We speculate that, for some well-developed technical fields, the consistency rates of their class codes would be high as the classification of the related technology should be familiar to the examiners whereas for emerging technical fields, the consistency rates of their class codes would be low as the examiners may have different opinions on what the related technology should be classified. The investigation of this speculation is currently under way.

If both reduced time delay and better analytic insight are required, an analyst would require a better tool that can take the hierarchical relationship among classification symbols into consideration. If this kind of tool is available, we speculate that some specific technical areas may reveal a high consistency rate or similarity measure even for PCA using Approaches 2 and 3. The identification of these specific technical areas and how reliable the PGPub classification symbols are in these specific technical areas can be further investigated.

Acknowledgments

This study is funded by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant No. MOST 103-2221-E-011-115.

References

Allison, J.R., Lemley, M.A., Moore, K.A., & Trunkey, R.D., (2004), Valuable Patents. Georgetown Law Journal, 92, 435–479.

Henderson, R.M., Jaffe, A., & Trajtenberg, M., (1997), University versus Corporate Patents: A Window on the Basicness of Invention. *Economics of Innovation and New Technology, 5*(1), 19–50.

Jaccard, P., (1901), Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, 547–579.

Jaffe, A.B., (1986), Technological opportunity and spillovers of R&D: Evidence from firms' patents, profits and market value. *The American Economic Review, 76*(5), 984–1001.

Jaffe, A.B., (1989), Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy, 18*(2), 87–97.

Leydesdorff, L., (2008), Patent Classifications as Indicators of Intellectual Organization. *Journal of the American Society for Information Science and Technology, 59*(10), 1582–1597.

Lerner, J., (1994), The Importance of Patent Scope: An Empirical Analysis. *RAND Journal of Economics, 25*(2), 319–333.

Liu, H.Z., Bao, H., & Xu, D., (2012), Concept Vector for Similarity Measurement Based on Hierarchical Domain Structure. *Computing and Informatics, 30*(5), 881–900.

McNamee, R.C., (2013), Can't see the forest for the leaves: Similarity and distance measures for hierarchical taxonomies with a patent classification example. *Research Policy, 42*(4), 855–873.

OECD (1994), The Measurement of Scientific and Technological Activities Using Patent Data as Science and Technology Indicators: Patent Manual. Paris: OECD Publishing. DOI: 10.1787/9789264065574-en.

Slimani, T., Yagahlane, B.B., and Mellouli, K., (2008), A new similarity measure based on edge counting. *Proceedings of the World Academy of Science, engineering and Technology, 23*, 773–777.

Small, H., (1973), Co‐citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science, 24*(4), 265–269.

USPTO (2004), Pre-Grant Publicaiton (PGPub) Global Concept of Operations. USPTO. Available on-line at http://www.uspto.gov/web/offices/dcom/olia/aipa/PGPubConOps.pdf.

USPTO (2005), Handbook of Classification. USPTO. Available on-line at http://www.uspto.gov/web/offices/opc/documents/handbook.pdf.

USPTO (2012), Overview of the U.S. Patent Classification System (USPC). USPTO. Available on-line at http://www.uspto.gov/patents/resources/classification/overview.pdf.

Wolter, B., (2012), It takes all kinds to make a world–Some thoughts on the use of classification in patent searching. *World Patent Information, 34*(1), 8–18.

The New Development Trend of Chinese-funded Banks and Internet Financial Enterprises from Patent Perspective

Zhao Qu, Shanshan Zhang and Kun Ding

qz_31@sina.cn, shann1027@163.com, dingk@dlut.edu.cn
School of Administration and Law, WISE Lab, Dalian University of Technology, Dalian, 116085
(People's Republic of China)

Abstract
Relying on the perfect integration of Internet technology, new business format and financial services, the Internet finance is developing at an unexpected speed, bringing impacts to Chinese-funded banks in the traditional business and emerging areas such as customization. Based on the preliminary study of the close contact between Chinese-funded banks and Internet financial enterprises as well as the necessity of patent protection, the paper proposes a comprehensive analytical framework and makes statistical comparison between 5 well-known Chinese-funded banks and Alibaba Group's patents from the perspective of annual trend, collaboration, application organizations, citation and other characteristics with data up to 2014 collected from Derwent Innovations Index(DII). It builds a Derwent Manual Code co-occurrence network with time coordinate by combining with visual tools and quantized the respective patent focuses of banks and Internet financial enterprises from the perspective of frequency and burst. After analysing the patents' contents, the paper discusses the mode of patent assignment. Finally, according to the status of patents, the paper concludes the strategic layout of domestic banks and Internet financial enterprise's intellectual property protection to predict the trend of further competition and alliance.

Conference Topic
Patent Analysis

Introduction
The data of British magazine "Banker" showed that in 2014, 13 Chinese banks ranked among the world's top 100 banks. Among them, Industrial and Commercial Bank of China ranked No.1 with the fund scale of 2,076.14 billion U.S. dollars, followed by China Construction Bank, Bank of China and other Chinese-funded banks, highlighting the fast growth and significant expansion of Chinese-funded banks. Nevertheless, the rates of return on assets of these banks were less than 3%, indicating that although the overall profit scale of China's banking ranked No.1 in the world, its profitability was not the case. With the slowdown of economic growth, substantial promotion of interest rate liberalization and further standardization of banking regulation, it is difficult for banks to maintain rising profit by relying on traditional channels. Like a huge dam, commercial banks store the saving deposits and collaborative deposits, but now there is a gap in the dam and the initiator is Internet finance. In the extensive penetration of Internet technology, traditional financial industry is undergoing dramatic changes: financial services have become the area competed by major institutions. Investors' "financial outlook" is corrected and the process of interest marketization has been promoted virtually (SOHO, 2014). The release of small and micro enterprises and individual consumer market's demand for loan is accelerated and the financing market presents a thriving prospect. With huge dividends of reform as well as the progress of big data and cloud computing technology, the Internet financial innovation is increasingly deepening. The rapid rise of Internet financial enterprises obliges Chinese – funded banks to face the continuous overlapping business, increasing demand for product service, competition and challenges brought by the application of innovative technologies.
In the new era, the competition between Chinese-funded banks and Internet financial giants does not only stay in the extent of business coverage, and more importantly, it is a rigid form

of innovation, which has been highly concerned by famous financial institutions, especially international banks, and produced historical and substantial effect on financial markets, services, products and management (Chen, 2006). Meanwhile, as an important link of financial products and intellectual properties, patents reflect the high degree of innovation of bank and Internet financial enterprises in service and product development. Meanwhile, in the period of patient protection, the banks exclusively enjoy the market of the innovative product, increase extra profits and safeguard fundamental interests. Events including the determination of the United States on the patentability criteria of bank business methods in 1998 or the patent bulk purchase of Alibaba Group before the listing in the United States in 2014 indicated that the field of financial patent protection has always been a focus of people. With the constant innovation of e-commerce and in-depth integration of Internet and mobile communication network, transaction platforms and payment means represented by e-banking, online banking and mobile banking will be bound to become the main form of future financial services. This control of the patents closely related to high-tech may become constitutor of financial market rules.

Theoretical basis and analytical framework

The slight decline of net interest margin posed no threat to large banks like ICBC, and the real blow came from the endogenous market force, the counterattack of Internet financial enterprises. For example, Ali Group's financial system has fundamentally broken the ice of the domestic credit loan by the "one-stop" service of customer absorption, credit assessment, loan review and issuance via e-business platform, providing more possibilities to the SME's problem of "difficult financing and expensive financing". In addition, Ali Group does not only involve in traditional fields of commercial banks including deposits and loans, financing, payment and settlement, but resulting in profound impact on commercial banking services and business philosophy. The formal establishment of Zhejiang E-business Bank ("Ali Bank") in 2014 intensified the potential threat to traditional banks. The strengthening of intellectual property protection strategy fired the first shoot of the competition between domestic banking industry and Internet financing; meanwhile, to defend the intellectual property disputes with foreign companies, especially under the circumstances of Ali's listing in the United States, Chinese companies will be exposed to a wider range of patent competition, so the enhancement of information sharing, innovative alliance building (Feng, 2013), and especially the optimization of patent protection become particularly important.

Overseas research on the relationship between Internet finance and banks was significantly earlier than China. Chou, et al, believed the in-depth integration of Internet and bank caused a revolutionary upheaval to the banking sector (Chou & Chou, 2000); Tsai, et al held the customers of Internet financial enterprises and traditional commercial banks varied in age, which was related to the degree of acceptance of innovative technologies and uncertain risk factors (Tsai, Huang & Lin, 2005). Meyer pointed out compared with commercial banks, P2P platform has lower operating costs and higher utilization of funds (Meyer, 2007); Ocean believed Internet financial enterprises provided more convenient credit business than bank process (Tess, 2013).

Chen believed the pressure of commercial banks caused by Internet finance should not be overlooked, forcing commercial banks to accelerate the pace of reform and strengthen customer customization (Chen, 2014); according to the status quo of competition between Internet financial enterprises and traditional commercial banks, Wang proposed four competitive strategies such as growth-orient strategy and aggressive strategy (Wang & Wang, 2014) by using the SWOT analysis; Gong thought the Internet financial model would not shake the traditional business model and earning way of commercial banks in a short term, and commercial banks should seek new development opportunities by using the Internet

(Gong, 2013). The above literature study involved the impact of Internet finance on traditional commercial banks as well as the business model based discussion on how commercial banks deal with Internet finance. However, its analysis of the relationship between commercial banks and Internet finance from the perspective of patent and technological innovation is still a blank area. This paper makes econometric analysis of the patents of Chinese banking industry and Internet financial giants, providing important reference basis for the development and improvement of the related patent protection system and patent strategy, the comprehensive analytical framework is proposed as shown in Figure 1.

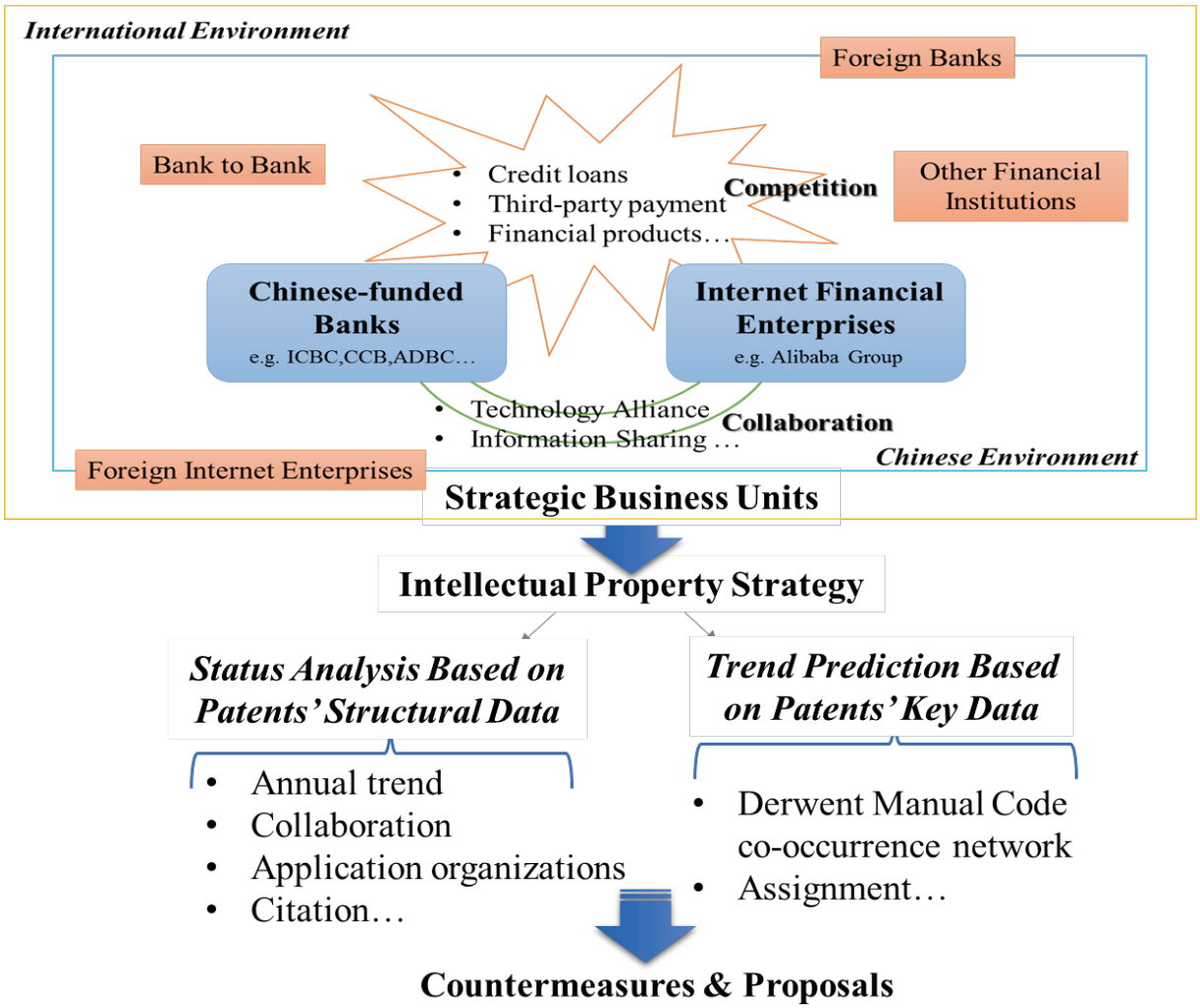


Figure 1. Patent analytical framework of Chinese-funded banks & Internet financial enterprises.

Data collection and analysis approach

The paper acquires the patents of the five representative Chinese-funded banks (ICBC, CCB, ADBC, BOC and BOCOM) and Alibaba Group Holding Limited on Jan.7, 2015 in DII by the way of Assignee Name and Assignee Code complex retrieval mode (Assignee Name and Assignee Code is connected by "OR" internally and by "AND" between two), the time span is from 1963 to 2014. After manual screening and exclusion, 917 Chinese bank patents and 1088 Ali patents are finally obtained.

The paper generalizes the patent development status and trend prediction of Chinese-funded banks and Internet financial enterprises by approaches of patent quantity statistical analysis

and patent content measurement in combination of visual tools, and proposes strategies and measures for the two sectors to improve patent protection, enhance technological innovation capacity, share information and build technology-business alliance if necessary, providing reference for the new development layout.

Results

Results of status analysis based on patents' structural data

Although the five Chinese-funded banks were built significantly earlier than Alibaba Group, they didn't occupy a striking advantage in the patent protection starting year, and lagged behind Ali in the total number of patents. In 2002, ICBC's patent of bank-card with dual account's processing device and method (PN: CN1397916-A) started the bank patent applications. Three years later, Alibaba carried out comprehensive patent protection and gradually exceeded the banks at an amazing growth. The annual patent application amount is shown in Figure 2.

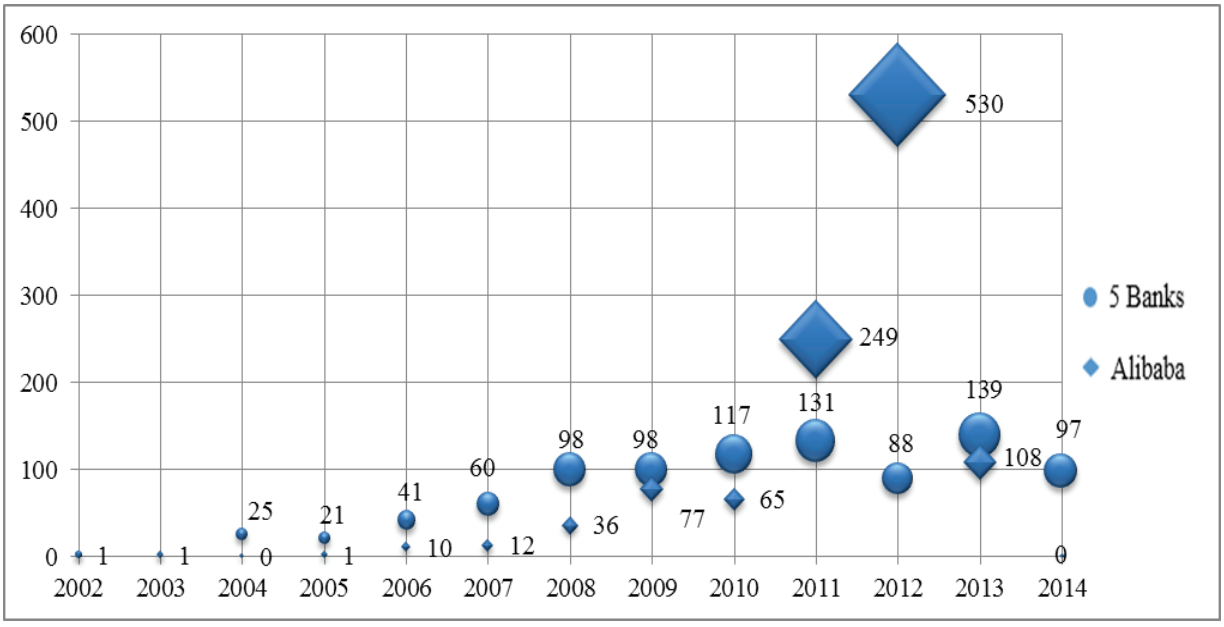


Figure 2. Annual trend of five Chinese-funded banks and Alibaba Group's patent quantity.

Figure 2 shows that the patent application amount of the selected banks has entered into fast growth since 2004. Though with slight fluctuation, but the overall situation is stable and the annual application number is relatively balanced. ICBC (549 patents) and CCB (253 patents) occupied a dominant position and led domestic banks to quickly engage in the patent development gradually integrating high-tech into the enterprise strategic level. In contrast, Ali Group's patent application was almost in exponential growth trend. The number of patent in 2012 was as high as 530, and the growth declined since 2013. The rapid deployment of domestic banks and financial enterprises was inseparable from the guidance of a series of policy documents including "National Intellectual Property Strategy" and also inseparable from the continuous expansion of Chinese enterprises and high-tech application.

By making statistics according to the patentee, we found all the 2005 patents were independently applied by banks and Ali Group. Few patents were produced via internal cooperation, and the branches concentrated in Zhejiang and Jiangsu. This phenomenon indicated that Chinese-funded banks and Internet financial enterprises didn't have close external relation in the patent activities, with a low degree of cooperation. To some extent, it

indicated that in the scope of finance, domestic enterprises have the relatively independent R&D team and were not positive enough in the flow and share of knowledge and information. If the external cooperation characterizes the degree of openness of proprietary technology, the geographical distribution of patent pending organizations is the indicator of measuring the corporate strategic deployment breadth. By the patent geological layout, we can learn and predict the key development areas of banks and Internet financial enterprises as well as the market distribution status of financial products and services (Luan, 2012). This paper makes analysis based on the connotation of the patent pending areas and organizations represented by the first two bits of code, we find only three patents of the Chinese-funded banks are applied in the non-Chinese mainland pending organizations, which are held by ICBC and distribute in WIPO, Taiwan and Russia. Although ICBC ranked No.1 in the world by a higher core capital and positively promoted international business strategy by means of organization application, mergers and acquisitions (till 2014, ICBC set up more than 330 overseas establishments in 41 countries and regions), its patent strategy failed to achieve the corresponding expansion (People, 2014). In contrast, Aliaba's patent has a wider geographical distribution; up to 71.7% (780pcs) of the patents were applied in organizations out of China. The average number of non-Chinese mainland patent application is 2.4 times (non-Chinese mainland application number/ non-Chinese mainland patent application number 1879/780), and the application of a number of patents has covered the range of over 6 organizations, and the pending mechanisms mainly distribute in Hong Kong, the United States and Europe (Table 1). Since the expansion of overseas business (since the establishment in 1998, Ali Group has set international headquarters in Hong Kong, offices in the United States, European and Japan), maintaining a highly consistent direction.

Table 1. Distribution of Ali's patent applications (outside of mainland China).

Region	*QTY*	*PCT(%)*	*Region*	*QTY*	*PCT(%)*
HK	631	33.58%	JP	186	9.90%
US	337	17.94%	KR	2	0.11%
WO	321	17.08%	SG	1	0.05%
EP	201	10.70%	AU	1	0.05%
TW	196	10.43%	DE	1	0.05%

Furthermore, the paper analyses status of two sections with patent citation data. These citations open up the possibility of tracing multiple linkages between inventions, inventors, scientists, firms, locations, etc. (Hall, Jaffe & Trajtenberg, 2001). 171 and 101 patents of Chinese banks and Ali Group were cited by other patents, respectively; patents with high citing frequency (top 5) were selected for analysis by combining with the cited patent information, and Table 2 is derived. Data showed that all the highly cited patents of Chinese banks were from ICBC, highlighting its outstanding R&D level among the peers.

Table 2. Highly cited patents of Ali and ICBC (Top 5).

ICBC		*Ali Group*	
PN/Freq. (cited patents)	AE/Freq. (citing patents)	PN/Freq. (cited patents)	AE/Freq. (citing patents)
CN1556449-A/19	BEIJ-Non-standard/10	CN101562543-A/7	GOOG-C/5
CN101183456-A/7	INCO-Non-standard/3	CN101662460-A	SALE-Non-standard/4
CN1588846-A/7	TNCT-C/3	CN101662460-A/6	IPCU-Non-standard/3
CN101119202-A/6	JIED-Non-standard/2	CN1835438-A/6	HUAW-C/2
CN101393671-A/5	SONG-Individual/2	CN101685516-A/5	TNCT-C/1

The patents of ICBC and Ali Group were mainly cited by enterprises, and a small number distributed in the patents held in the name of individuals and universities. Enterprises cited the patents of ICBC including categories of marketing, communications, telecommunications, network equipment, data security, authentication and other related categories, of which the citing frequency of BEIJING FEITIAN CHENGXIN SCI & TECHN CO (a world leading professional software protection and authentication of high-tech intelligence company), indicating the important of the authentication–related technology included in ICBC patents and also reflecting the close relation between the company products and ICBC business. Enterprises' citations of Ali Group involved customer consulting, Internet, software, communications (communications equipment), electronics, telecommunications, investing and financing, and the patent citers distributed in the United States and Japan. It is noteworthy that enterprises with similar business as Alibaba like Google, Tencent, are also among the citing group, showing Ali's patent technology is playing a guiding role in the Internet industry. In addition, Beijing Institute of Technology and Taiyuan University of Technology cited the patent of Ali and ICBC once, respectively.

Results of trend prediction based on patents' key data

Compared to other classification system, Derwent manual code (MC) outlines more detailed indexing information in retrieval of patent's theme and core content based on the uses and applications of an invention, rather than just a straight forward description of what the invention is (Stembridge, 1999).

Table 3. High frequency Derwent Manual Codes (Top 10).

Alibaba freq			*Five Chinese-funded banks*		
Freq	MC	Content	Freq	MC	Content
310	T01-J05B4P	Database applications	175	T01-J05A1	Financial
230	T01-N01D3	From remote site or server	140	T01-N01A1	Eft/banking
184	T01-S03	Claimed software products	139	T01-N01D3	From remote site or server
172	T01-N02A3C	Servers	134	T01-J05B4P	Database applications
154	T01-N03A2	Search engines and searching	81	T05-L03C1	General control system
126	T01-J05B3	Search and retrieval	75	T01-N02A3C	Servers
123	T01-N01D2	Document transfer	69	T01-D01	Data encryption and decryption
77	W01-A07G1	Transmission control procedure	67	T01-N01A	Financial/business
74	T01-N01A	Financial/business	59	T01-N01D2	Document transfer
65	T01-N02A2C	Client/server system	57	T01-N02B2B	System and fault monitoring

Further, we transforms the bibliographic data of all the 2005 patents into WoS logging data and introduced into the CiteSpace, and set the analysis interval as 1 year, then drawing the maps (Figure 3 and Figure 4). By depicting the association and combination between the MCs, it can analyse the correlation between patents and even technologies, and can also facsimile the internal technology composition and structure (Shen, Gao & Teng, 2012). Timeline visualization provides a directly temporal overview of technologies, columns are time periods of co-occurrence of technologies and rows are clusters (Gong, Jiang, Yang& Wei, 2011). The dynamically changing course of banks and Internet finance patent technologies can be revealed by combining with the attribute changes in timeline axis. Moreover, the development trend can be predicted through their restive business characteristics. The top 10 high-frequency manual codes of Chinese-funded banks and Ali Group (Table 3) were intercepted respectively to explore the hot fields.

It can be seen from the analysis that the technical research of both subjects was carried out by centring the category of "T01", showing the Chinese banks and Internet financial enterprises are very concerned about the application of digital computer in financial services. A series of patent activities were conducted by combining with the research of "database applications" and "application originating from remote sites or remote servers". It is noteworthy that in the distribution of the top 10 high-frequency bank patents, Internet financial patents showed a high degree of overlap in some technical contents. In addition to "database applications" and "remote service", "document transfer" and "Financial/business" were also included in the key content of their patent developments. In contrast, the patents of banks are more inclined to the study of financial, banking, system monitoring and related technology; Ali Group makes innovation and protection based on the contents of search engine and software.

As the largest cluster in the bank MC network, "bank background" demonstrated the general picture of banking business featuring electronic funds transfer point of sale equipment, currency handling systems, smart media and the Internet and information transfer, which occupied the central position in the entire time chain.

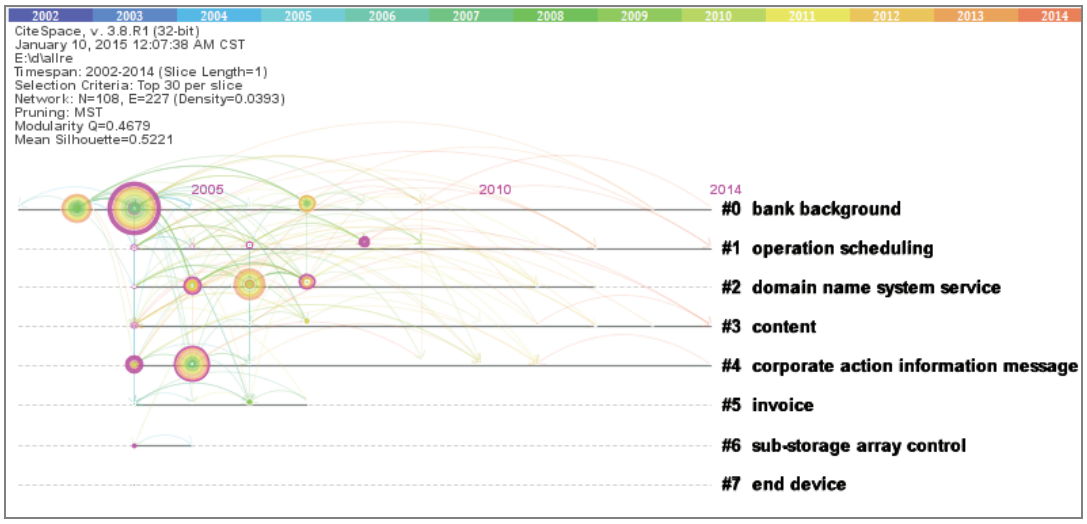


Figure 3. Five banks' Derwent Manual Code co-occurrence network (Timeline view).

Figure 4. Alibaba Group's Derwent Manual Code co-occurrence network (Timeline view).

In Ali's network, the cluster "outgoing message" constituted by the close connection of digital information transmission, Internet and messaging, data processing systems and process control comprehensively summarized the business flow carried out by Ali Group based on Internet data. Second, the cluster "central account registration system" composed by audio / video record and Internet-based information processing and transfer, and nine clusters including data and communications. The overall technology relevance and research contents are similar to these shown in the MC of Chinese bank patents, but more emphasis was made on the application of Internet in business.

On this basis, codes with high frequency change rate with the time sequence (burst term, Table 4&5) further determined the technology frontier and development trend of Chinese banks and Ali (Huang, Wang &Wang, 2014).

Table 4. Bursts of Banks' Derwent Manual Codes

Burst	*MC*	*year*	*Content*
5.77	T05-L03	2002	Cash dispensing and depositing machines
6.18	T05-L02	2003	Electronic funds transfer
5.21	T01-N01A1	2003	Eft/banking
3.06	T01-N01A2A	2004	E-shop, e-auction, e-mall, and e-services
2.94	T01-J05A1	2004	Financial
2.93	T05-L01D	2004	Data transfer and network aspects
2.76	T01-J12C	2004	Security
2.76	T01-J05B4P	2005	Database applications
5.7	T01-F05	2006	Arrangements for executing specific programs and system management software
4.93	T01-N01D	2006	Data transfer
3.53	T01-J05A2	2006	Administration and management tools
4.08	W01-A07G1	2011	Transmission control procedure
2.99	W01-A06C4	2011	Radio link
2.68	T01-N03A2	2011	Search engines and searching
3.04	T04-K03B	2012	Rfid/transponder

Table 5. Bursts of Ali' Derwent Manual Codes

Burst	*MC*	*year*	*Content*
5.12	T01-N01A1	2005	Eft/banking
3.14	T01-N02A3C	2006	Servers
5.02	T01-E01A	2007	Sorting
4.48	T01-S03	2007	Claimed software products
2.86	T01-J16C3	2007	Natural and pictorial language processing
4.58	T01-M02	2008	Multiprocessor systems
6.29	T01-E01	2009	Sorting, selecting, merging or comparing data
4.63	T01-J20C	2011	Software test, verification, debug, optimization
2.73	W01-A06E	2013	Network control and software

The patented technology burst of Chinese banks are more evenly dispersed in 2002~2012, following the development course of bank reserves appliances → electronic funds / bank → online business and data processing → database applications → specific project management and data transfer → search engine, control → wireless communications, showing the trend of gradual evolution from traditional banking to Internet financial sector. Since 2005, Ali's patent started from e-funds/e-bank technologies, and then underwent a series of technology evolution of data processing from server, data sorting, and software to graphic language processing, which is currently in the data processing optimization and study of Internet control technology. Although the related technologies of e-transaction technology appeared earlier in the patent of Chinese banks, but Ali Group is more sustainable in the ongoing online transactions, which continues to carry out the research based on big data and gradually establish technology chain in the field of Internet finance.

Technological evolution is the exploration on the development route and trend of bank and Internet financial enterprises based on patent, and the conclusion of patent assignment information can provide references to the patent development mode of the two. In 2014, Alibaba Group made IPO financing amounted to 25 billion U.S. dollars, which was the largest IPO. The United States is a country with frequent patent disputes, to avoid the patent infringement issues encountered by Facebook or Twitter in IPO, Ali Group has made significant patent deployment in the U.S. since 2013, where a lot of patents have been reserved. Till the retrieval date of this paper, 399 U.S. patent family cases were found and more than 50 have been authorized (Chinaip, 2014). In addition to independent application, Alibaba purchased 21 patents from IBM in 2013, and one of which was for Amazon, the largest U.S. e-commerce platform, and also prepared for coping with the patent competition and litigation. We made inquiry of the operating data of Ali Group and five Chinese banks in Chinese patent database and found that Ali Group started to purchase the patents of other organizations since 2012 onwards, but only limited to the category of invention patents. Patent seller expanded from domestic organizations to international institutions, such as Shanghai Yiren Information Technology Co., Ltd. and IBM; in addition to enterprises, Ali also purchased patents from Chinese Academy of Science Institute of Computing Technology; the change of some patent was caused by the changes of the corporate nature, such as Alibaba to Alibaba Group Holding Limited. The aforementioned technical fields of patent change included electric digital data processing, transmission of digital information, arrangements of circuit components or wiring on supporting structure and coin-freed or like apparatus. However, the patent purchased by Chinese banks included patent, utility models and appearance design, and the patents with internal change were almost 1/2 of the total patent transfer amount. These patents mainly came from the bank branches and individuals, and only CCB had one patent purchase from enterprise (Shandong Confucian Culture Communication Co., Ltd.), and the technical fields of patent change mainly involved the bank cards, security cards, teller settings and other contents, no transactions concerning goods and services of bank financial commodities and services were made.

Discussion and conclusions

General comments

In a long term in the past, Chinese banks made huge profits by relying on monopoly advantages and policy bonus, and occupied the position on the top of financial ecology. However, the single channel and curing product business model can no longer work. In China, the rapid development trend of Internet finance represented by Alibaba does not only occupy a significant share in domestic financial sector, but also causes widespread concern in the overseas business expansion. Traditional profit making channels of banks have been

hindered in a variety of aspects, including the competition of domestic and overseas banking industries and the pressure caused by the enhancement of overlap ratio with Internet finance business. With the development of commodities and services based on big data, Internet financial enterprises are inseparable from the application of technology. In the new situation, it faces the transfer from purely financial products to technical competition; whether banks or Internet financial enterprises, technology innovation and application have been upgraded to a new strategic plan.

By the comparison of patents of 5 Chinese banks and Alibaba Group Holdings Limited, we found that the patent activities of Chinese banks started late, with limited number, especially in key business areas like e-commerce. Most of the bank patents were independently applied in China, and their overseas IPR protection does not match their development of business, which may become a potential hazard for patent disputes arising from overseas promotion of financial products and services. Although the banks have higher patent citing frequency, the citing parties are mostly in China and the all the highly cited patents are held by ICBC. In contrast, Ali Group has achieved rapid progress of patent activities, with advantages in the total number, patent geographical distribution and the composition of citing groups. However, like banks, Ali Group also has low degree of external cooperation, indicating their closure and limitations in patent research and development. We can learn from MC co-occurrence network that banks and Internet financial enterprises have relatively concentrated technology, which were the patent R&D centred by computer and showed a high degree of overlapping in database use, financial/commercial and remote control, etc. The patent contents of Chinese patents tend to the research of digital communication, hardware equipment and banking business operation, whereas Alibaba pays more attention to search engine and software-related innovation and protection. From 2002 to 2014, bank patent technology showed the shift from bank reserves appliance to e-funds/banking, online services and data processing. Currently, it is in the stage of network and wireless communications, whereas the research of Alibaba has undergone a series of technology evolutions from e-funds/e-banking, data processing from server, data processing, software to graphic language processing. Patent assignment data showed that independently developed ones are still the main source of banks and Internet financial enterprises' patents, while the patent purchase of Internet financial enterprises are quietly rising, and may form a new patent development mode of "independent R&D and purchase".

Countermeasures & Proposals

Based on the abovementioned patent status and future development direction of banks and Internet financial enterprises, China's banking industry shall attach important to the development, protection, management and utilization of bank patents at all levels. Moreover, it is essential to set up product and service technology early warning, make technical prediction and selection in fields with priority. At the same time, cooperation with high-tech industries represented by information technology shall be emphasized to improve the patent technical quality. At the same time, on the basis of full study of international regulations and overseas local laws and regulations, Chinese banks shall learn from Alibaba's international patent strategies to increase the overseas patent application quantity, expand market share and gain competitive advantages. After the listing in the United States, as the leader of Internet financial industry, Alibaba shall not only strengthen the risk control effort, promote the innovation of financial products and services and customer participation as well, but shall accelerate the deployment of intellectual property, take the mode of simultaneous patent purchase and independent R&D, to avoid patent disputes with overseas companies and win market opportunities by appropriate use of patents. In addition to strengthening their competitive advantages, banks and Internet financial enterprises shall strengthen cooperation

to make best use of the advantages and bypass the disadvantages, so as to form a new finance-technology alliance. Banks can use the network resources, information data and cloud computing of Internet financial enterprises to play their professional administration, thus introducing customers to the professional advantages via network channel. Likewise, by relying on the financial background of banks, Internet financial enterprises shall set up long-term, stable relationship with mutual trust to expand the scope of commercial exchanges, strengthen financial risk management and control, thereby providing a cooperation and win-win opportunity to both parties.

Further research

In the process of researching the status quo and future trend of Chinese-funded banks and Internet financial enterprises, this paper only took into account of their competition and cooperation. In fact, we can learn from the framework of this paper that factors affecting the development of them are multifaceted and complex. Hence, in the following study, the author will put overseas companies into the comparison to explain the development situation of banks and Internet financial enterprises in detail.

Acknowledgments

We would like to thank the anonymous reviewers for their important comments. Thanks to Dr. Liu and Dr. Tang for his assistance in data cleaning.

References

Chen, J, M. (2006). The influence of science and technology revolution on the development of financial sector - our country uses technological innovation to promote financial innovation. *Journal of Dialectics of Nature*, 27(6), 99-100.

Chen, W. Y. (2014). What does commercial bank learn from Internet financial business? *Management and Administration*, *11*, 014.

China Intellectual Property (2014). Alibaba purchased patents to deal with patent litigation risk. *China Intellectual Property Platform*. Retrieved January 10, 2014 from: http://www.chinaipmagazine.com/news-show.asp?id=12219.

Chou, D. C., & Chou, A. Y. (2000). A guide to the Internet revolution in banking. *Information Systems Management*, *17*(2), 51-57.

Feng J.J. (2013). Research on competitive strategy of commercial bank under the background of the Internet finance. *Contemporary Finance*, *4*, 14-16.

Gong, X., Jiang, L., Yang, H., & Wei, F. (2011). Mapping Intellectual Structure: A Co-citation Analysis of Food Safety in CiteSpace II. Gene, 412.

Gong, X. L. (2013). The influence of Internet financial mode on traditional banking. *South China Finance*, *5*, 86-88.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). National Bureau of Economic Research.

Huang L.C., Wang K., & Wang K.K. (2014). Technology Hot Spots and Fronts of Household Air Conditioner: Identification and Trend Analysis Based on CiteSpace. *Journal of Intelligence, 33*(2),

Luan C.J. (2012). Emperical Study on the Measuring Indicators of Generic Technology of Emerging Industries of Strategic Importance. *Forum on Science and Technology in China*, *6*, 73-77.

Meyer, T., Heng, S., Kaiser, S., & Walter, N. (2007). Online P2P lending nibbles at banks' loan business. *Deutsche Bank Research*.

People. (2014). Industrial and commercial bank branch opened in London, and has been set up more than 330 overseas agencies. *People.cn-Bank channel*. Retrieved December 30, 2014 from: http://finance.people.com.cn/money/n/2014/1202/c218900-26132904.html.

Shen J., Gao J., & Teng L. (2012). Derwent Manual Code Co-Occurrence: A Practical Method in Patent Map. *Science of Science and Management of S. & T.*, *33*(1), 12-16.

SOHU. (2014). Report on Top 16 Chinese Internet Financial Enterprises. *SOHO Media platform*. Retrieved December 10, 2014 from: http://stock.sohu.com/20140821/n403633305.shtml.

Stembridge, B. (1999). International patent classification in Derwent databases. *World Patent Information, 21*(3), 169-177.

Tess Ocean. (2013) Online Personal Loans: Access Easy Finance At Cheap Interest Rates By. Retrieved December 10, 2014 from: http://www.Internetmonetary.com.

Tsai, H. T., Huang, L., & Lin, C. G. (2005). Emerging e-commerce development model for Taiwanese travel agencies. *Tourism Management*, *26*(5), 787-796.

Wang, R. C. & Wang, J. Z. (2014). SWOT analysis of the commercial banks to deal with the Internet financial enterprise competition. *Small and medium-sized enterprise management and technology*, *8*, 62-63.

Who Files Provisional Applications in the United States?

Chi-Tung Chen[1] and Dar-Zen Chen[2]

[1] *d94522022@ntu.edu.tw*
Department of Mechanical Engineering, National Taiwan University, Taipei, Taiwan

[2] *Corresponding Author: dzchen@ntu.edu.tw*
Department of Mechanical Engineering and Institute of Industrial Engineering, National Taiwan University, Taipei, Taiwan

Abstract
This paper employed the US Patent Application Database to find out who files provisional applications in the United States. Preference rates, use rates, and provisional application to non-provisional application rates were used to evaluate the filing behaviour of provisional applications with respect to non-provisional applications. Factors weighing toward filing provisional applications include filing date sensitivity, patent term sensitivity, and necessity of promoting. Factors weighing against filing provisional applications include cost sensitivity and English abilities. These factors were discussed in order to explain the filing behaviour of provisional applications with respect to non-provisional applications. Applicants form English speaking countries are more likely to file provisional applications than applicants from other countries. We reasoned that the English ability of applicants might be the cause for such a result. Applicants from the fields of Computers and Communications and Drugs and Medical are more likely to file provisional applications than applicants from other fields. We reasoned that patent term sensitivity and filing date sensitivity might be the cause for such a result.

Conference Topic
Patent Analysis

Background and purpose
A provisional application for patent (hereafter referred to as 'provisional application') is a US national application filed in the United States Patent and Trademark Office (USPTO) that has been offered to applicants since June 8, 1995 and was designed to provide a lower-cost first patent filing in the United States. A provisional application is not required to have a formal patent claim or an oath or declaration. Provisional applications also should not include any information disclosure (prior art) statement since provisional applications are not examined. A provisional application provides the means to establish an early effective filing date in a later filed non-provisional patent application (hereafter referred to as 'non-provisional application'). It also allows the term "Patent Pending" to be applied in connection with the description of the invention. A provisional application has a pendency lasting 12 months from the date the provisional application is filed. The 12-month pendency period cannot be extended. Therefore, an applicant who files a provisional application must file a corresponding non-provisional application for patent during the 12-month pendency period of the provisional application in order to benefit from the earlier filing of the provisional application. By filing a provisional application first, and then filing a corresponding non-provisional application that references the provisional application within the 12-month provisional application pendency period, a patent term endpoint may be extended by as much as 12 months. (USPTO, 2014).

Although the provisional application filing approach has been offered to applicants for almost two decades, the USPTO does not make its database of provisional applications publicly available other than the individual files in Patent Application Information Retrieval (PAIR). Therefore, it is still difficult to answer the following two crucial questions: (1) Who files provisional applications in the United States? (2) Why do applicants file provisional applications in the United States?

Dennis Crouch (2008) studied approximately 15,000 utility patents issued in April and May 2008 and found out that only 21% of issued patents claiming priority from a provisional application, only 5% of the patents that associated with a provisional application were assigned to international applicants while 30% of the patents that associated with a provisional application were assigned to a U.S. applicant, Israel and Canada filed the highest proportion of provisional parent claims, only 2% of the Japanese & Korean patents included provisional parent claims, new drug inventions have the highest rate of association with a provisional application, and patents on electrical and electronic applications had the lowest rate of provisional filing. Dennis Crouch provided a rough first look of provisional application filings in the United States, but the dataset used by Dennis Crouch was rather small and time-limited (approximately 15,000 utility patents issued in April and May 2008). Therefore, it seems that the dataset used by Dennis Crouch was not sufficiently large to guarantee the results; and moreover, Dennis Crouch provides the results but lacked to explain the results.

The purpose of this paper is to address the two questions identified with sufficient dataset and detailed analyses to guarantee the results and to fully understand the filing behaviour of applicants. First, we employ the US Patent Application Database for 2005-2013 to find out who files provisional applications by checking the provisional application filings in different countries of origins, technological categories, assignee types, and assignees. Second, we explain why applicants file provisional applications in the US According to the USPTO, most obvious advantages of filing a provisional application are: (1) obtaining an effective filing date with a lower cost and an easily prepared application; (2) extending the statutory patent term up to one year; and (3) the ability to use the term "patent pending" (USPTO, 2014). Therefore, we assume that the following factors are weighing toward filing provisional applications: (1) filing date sensitivity; (2) patent term sensitivity; and (3) the necessity of promoting. Although the provisional application is designed to provide a lower-cost first patent filing in the US, an applicant still needs to spend extra money to file a corresponding non-provisional application in order to obtain a patent. In addition, although the provisional application was supposed to be an easily prepared application as it may be filed in a foreign language, an applicant still requires the English ability to prosecute the provisional application. Therefore, we assume that the following factors are weighing against filing provisional applications: (1) cost sensitivity; and (2) the English ability of applicants.

Trends in filing provisional applications

Since the database of provisional applications is not published, the filing numbers of the provisional applications can only be obtained from annual fiscal reports by the USPTO. Moreover, since the USPTO has never made publicly available the provisional applications that are not relied on for claiming priority by non-provisional applications, we employed the USPTO Patent Application Database to find out the number of provisional applications that have been claimed for priority by at least one non-provisional application.

Figure 1 shows the trends in filing provisional applications. The black bars represent the number of utility applications (non-provisional applications) filed each year from 2005 to 2013; the hatched bars represent the number of provisional applications filed each year from 2005 to 2013; and the grey bars represent the number of provisional applications filed each year from 2005 to 2013 that are relied on as priority documents in non-provisional applications. Please note that the USPTO only reported the number of provisional applications by fiscal year. So in Figure 1, the hatched bars were calculated by the fiscal year (October 1 to September 30), not by the calendar year (1 January to 31 December).

As shown in Figure 1, from 2005 to 2013, over 4.29 million non-provisional applications and over 1.27 million provisional applications have been filed. Among the 1.27 million

provisional applications, over 0.71 million provisional applications have been converted to non-provisional applications. It can be inferred that both non-provisional application filings and provisional application filings continued to rise, with over 570,000 and 170,000 filed in 2013. There was a drop in each of the non-provisional application filings and the provisional application filings in 2009. A possible explanation for such a drop could be attributed to the financial crisis of 2008.

Figure 1 also shows the provisional applications that have been relied on for claiming priority by non-provisional applications. It is observed that the number of provisional applications that have been relied on for claiming priority by non-provisional applications is growing. Although the provisional applications continued to be more popular, applicants have abandoned more of the provisional applications without relying upon them for claiming priority. The difference between each pair of the hatched bar and the grey bar is the number of provisional applications abandoned without being used as priority documents each year.

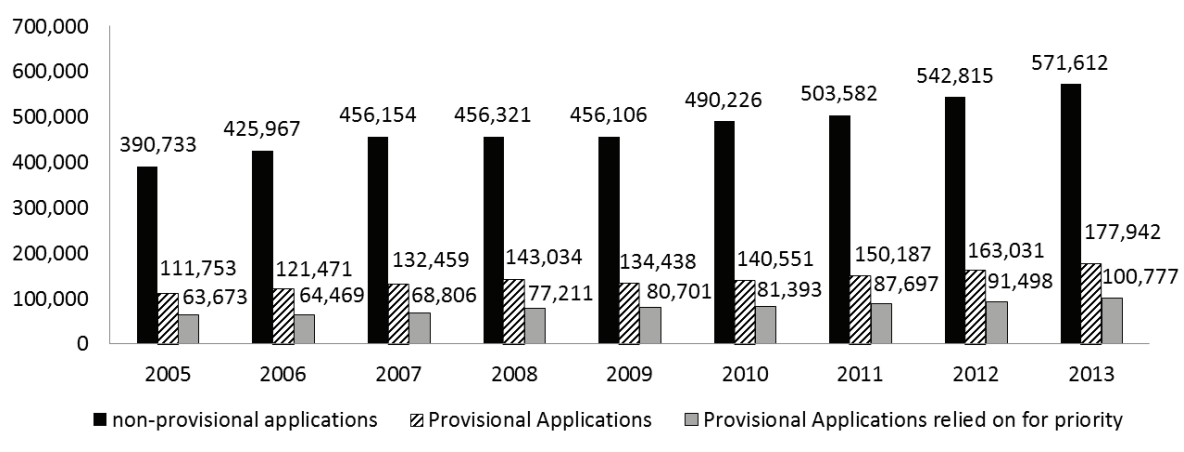


Figure 1. Non-provisional applications, provisional applications, and provisional applications relied on for priority filed each year for 2005-2013.

Rates of provisional applications/non-provisional applications

Rates of provisional applications/non-provisional applications (hereafter referred to as preference rates) show the preference of applicants in filing provisional applications with respect to non-provisional applications. The preference rate represents the percentage of a provisional application being filed in proportion with a non-provisional application in deciding filing patent applications in the United States. In Figure 2, the dotted line shows the preference rate of all provisional applications filed each year from 2005 to2013. It is clear that the preference rate remained steady during the period, except for 2009-2010, and the preference rate continued to slightly rise to 31.13 % in 2013.

Rates of provisional applications relied on for priority /provisional applications

As mentioned above, a provisional application has a pendency lasting 12 months from the date the provisional application is filed. An applicant who files a provisional application must file a corresponding non-provisional application for patent during the 12-month pendency period of the provisional application in order to benefit from the earlier filing of the provisional application (USPTO, 2014); otherwise, the provisional application will be automatically abandoned. Therefore, it is interesting to find out the use rate of the provisional applications that have been used for claiming priority by non-provisional applications (hereafter referred to as use rate). The use rate represents the usage of provisional applications. The result is shown in Figure 2, where the first solid line represents the use rate of all provisional applications filed each year from 2005 to 2013. As shown in Figure 2, the use rate

of provisional applications was located between about 52% and about 60% in 2005-2013, that is, about 40% to about 48% of the provisional applications were abandoned without being converted to non-provisional applications each year during 2005 and 2013.

Rates of provisional applications relied on for priority/non-provisional applications

Rates of provisional applications relied on for priority/non-provisional applications (hereafter referred to PA to NPA rate) show both the filing preference and the usage of provisional applications. The PA to NPA rate can be calculated by the preference rate times the use rate. Since the USPTO has never mad publicly available the provisional applications that are not relied on for claiming priority by non-provisional applications, the PA to NPA rate became the only practical rate for evaluating the provisional application filings with respect to non-provisional application filings in different countries of origins, technological categories, and assignees. As shown in Figure 2, the second solid line represents the PA to NPA rate of all the provisional applications filed each year between 2005 and 2013. It can be seen that the PA to NPA rate remained steady during the period, except for 2009-2010, and it continued to slightly rise to 17.63% in 2013. In other words, approximately one in six non-provisional applications was expected to claim priority upon a provisional application.

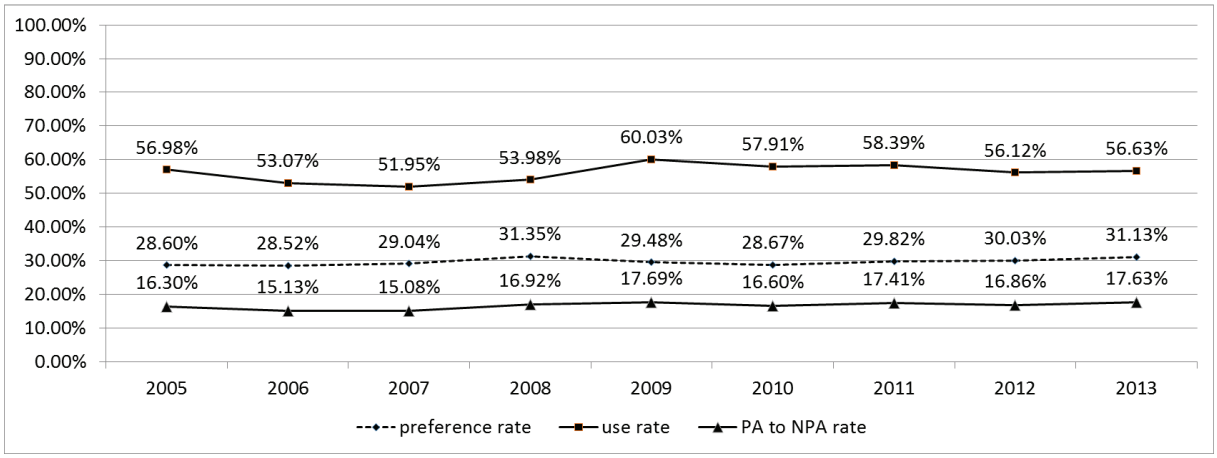


Figure 2. Preference rate, use rate and PA to NPA rate each year from 2005-2013.

Provisional applications by different countries of origins

The date of the filing of the provisional patent application can also be used as the foreign priority date for applications filed in countries other than the United States. Therefore, the need is identified for a foreign applicant to file a patent application as a provisional application in the United States first, and then to claim the priority of the provisional application to file a regular patent application in the United States as well as in the countries other than the United States.

Table 1 shows the ranking of the top 10 countries of origins where applicants filed provisional applications and non-provisional applications in the US in 2005-2013. During this period, the top 10 countries were: United States of America (US), Canada (CA), Germany (DE), Japan (JP), Israel (IL), Netherlands (NL), Korea (KR), Taiwan (TW), France (FR), and Switzerland (CH). It can be seen in Table 1 that the ranking of provisional applications and that of non-provisional applications varied for some countries. For example, JP was ranked second in non-provisional applications but fourth in provisional applications; KR was ranked fourth in non-provisional applications but seventh in provisional applications; TW was ranked fifth in non-provisional applications but eighth in provisional applications; FR was ranked sixth in non-provisional applications but ninth in provisional applications; and CN (China) was

ranked seventh in non-provisional applications but was not ranked in the top ten in provisional applications. It can be concluded that applicants in JP, KR, TW, FR and CN prefer filing their first applications in the United States as regular non-provisional applications rather than provisional applications. On the contrary, applicants in the US, CA and IL very much prefer filing their first applications in the US as provisional applications.

Table 1. Ranking of the top 10 countries of origins where applicants filed provisional applications and non-provisional applications in the US in 2005-2013.

ranking	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
provisional applications	US	CA	DE	JP	IL	NL	KR	TW	FR	CH
non-provisional applications	US	JP	DE	KR	TW	FR	CN	NL	CA	GB

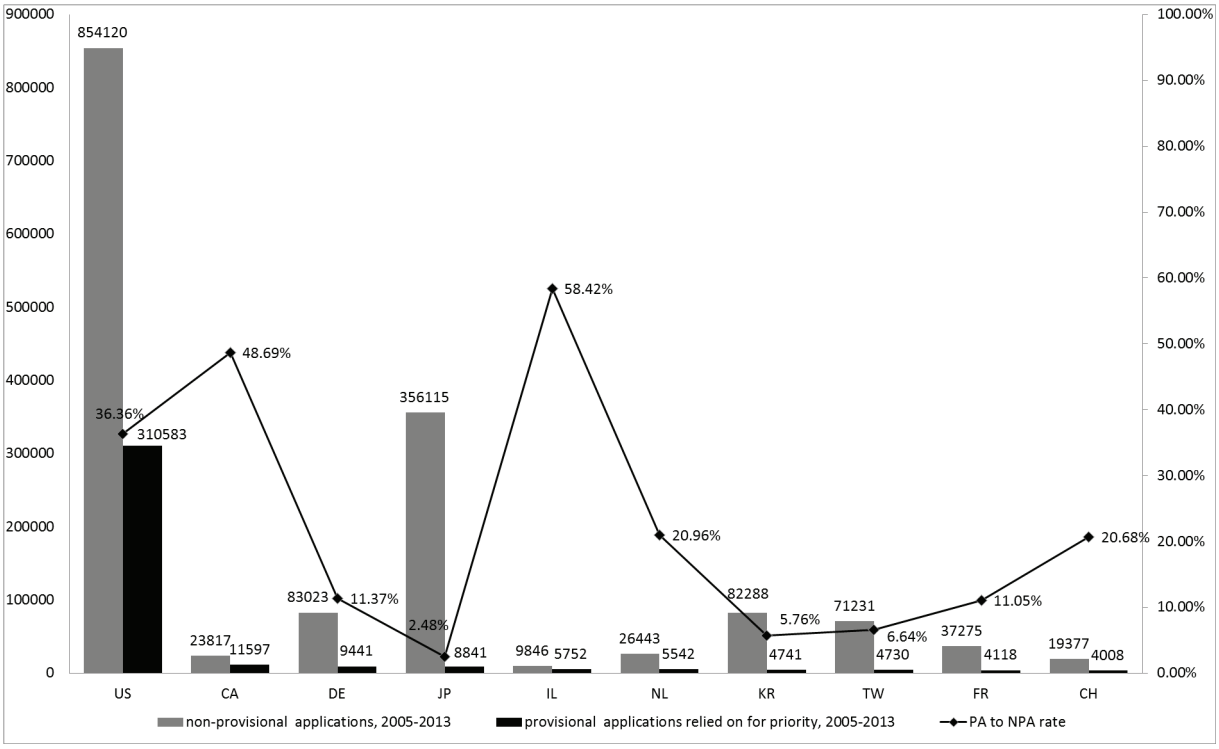


Figure 3. Top 10 countries of origins where applicants filed provisional applications with respect to corresponding non-provisional applications and the PA to NPA rate in the US in 2005-2013.

Furthermore, we checked the PA to NPA rate in order to find out the preference of filing provisional applications for applicants in different countries of origins. Figure 3 shows the top ten countries of origins, where applicants filed provisional applications with respect to corresponding non-provisional applications and the PA to NPA rate in the US in 2005-2013.

In Figure 3, the black bars represent the number of provisional applications filed by applicants from each country in the US in 2005-2013; the grey bars represent the number of non-provisional applications filed by applicants from each corresponding country in the US in 2005-2013; and the solid line represents the PA to NPA rate of each corresponding country in 2005-2013. Figure 3 shows that the PA to NPA rates of the US (36.36%), CA (48.69%) and IL (58.42%) were very much above the average percentage (about 17%). Contrarily, the PA to NPA rates of JP (2.48%), KR (5.76) and TW (6.64%) were far less than the average percentage. We reasoned that the English ability of applicants might be the cause for such a result. Comparing to applicants from JP, KR and TW, applicants from the US, CA and IL are either native English speakers or having good English abilities, so it is relatively easy for applicants in these countries to prepare a provisional application that is suitable for being

relied on for claiming priority by a non-provisional application. Moreover, some foreign laws limit the filing of patent applications abroad before a national patent application filing or authorization occurs. So the PA to NPA rate is expected to be low for applicants from those countries. For example, CN has this kind of law, and its PA to NPA rate was only 2.75%.

Provisional applications by different technological categories

In this paper, we used the six main technological categories (i.e. Chemical, Computers & Communications, Drugs & Medical, Electrical & Electronic, Mechanical, and Others) developed by The National Bureau of Economic Research (NBER) (Hall et al., 2001) to analyse provisional applications by technological categories.

Figure 4 shows the provisional applications relied on for priority filed each year from 2005 to 2013 divided by the NBER main technological categories. As shown in Figure 4, Computers and Communications and Drugs and Medical were the most popular main technological categories, in which applicants filed provisional applications and further converted them to non-provisional applications by claiming priority.

Sukhatme and Cramer (2014) suggested that an applicant who cares about the patent term will seize an opportunity to increase the term if it is offered to him/her. Applicants in industries in which the patent term is especially important would be more likely to file provisional applications than applicants in industries in which the term is less important. In the Drugs & Medical industry, the patent term is critical, i.e. applicants consider the patent term sensitivity, so the applicants tend to extend the statutory patent term up to one year by filing provisional applications first instead of non-provisional applications. In the Computers & Communications category, technologies change rapidly, i.e. applicants consider filing date sensitivity, so obtaining an early effective filing date is important to inventions in this category.

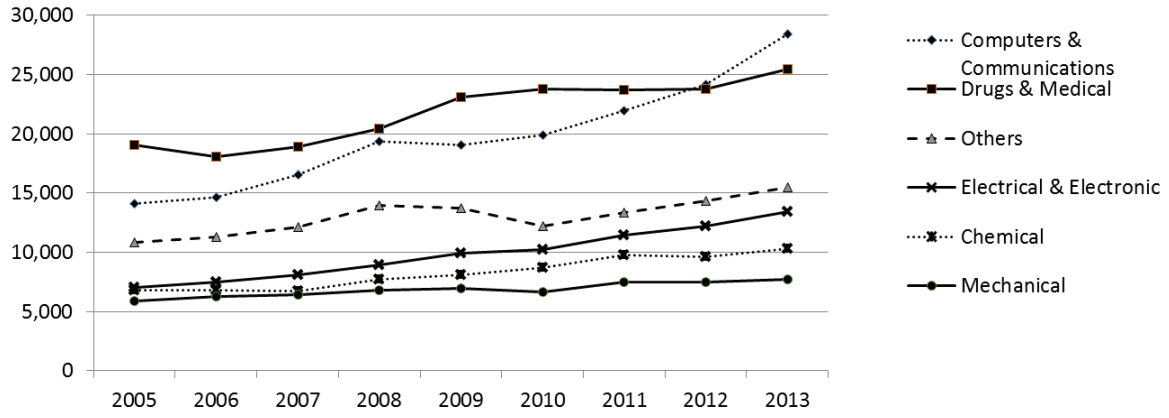


Figure 4. Provisional applications relied on for priority filed each year from 2005-2013, by NBER main technological categories.

Provisional applications by different assignees

Table 2 displays the top ten assignees filing provisional applications that were relied on for priority in the US in 2005-2013. Table 2 also shows the corresponding non-provisional applications by the top ten assignees, and their PA to NPA rates. It is clear that except for Samsung (5.68%) and Microsoft (9.27%), the PA to NPA rate of each of the other assignees was very much above the average percentage (about 17%). Take California University as an example, its PA to NPA rate was up to 81.28%. That is, in about every ten non-provisional

applications, over eight non-provisional applications claimed priority based upon early filing provisional applications.

Table 2. Top ten assignees filing provisional applications that were relied on for priority in the US in 2005-2013, the corresponding non-provisional applications, and the PA to NPA rates.

Assignee	*provisional applications relied on for priority*	*non-provisional applications*	*PA to NPA rate*
Qualcomm	6291	10018	62.80%
California University	3426	4215	81.28%
Broadcom	2876	4963	57.95%
Samsung Electro-Mechanics	2771	48814	5.68%
Koninklijke Philips Electronics N.V.	2519	12386	20.34%
Microsoft	2483	26799	9.27%
DuPont	2429	3286	73.92%
Texas Instruments	2353	5943	39.59%
LG Electronics	2318	9211	25.17%
Apple	1772	5124	34.58%

Table 3 shows main patent areas of each of the top ten assignees. For example, Qualcomm focused on the Computers & Communications field. So among all the 6291 provisional applications that relied on for priority, 5612 applications (about 89%) filed in the category of Computers & Communications. Broadcom, Samsung Electro-Mechanics, Microsoft, Texas Instruments, LG Electronics, and Apple also focused on the field of Computers & Communications.

Table 3. Provisional applications filed by the top ten assignees in the US in 2005-2013 by technological categories.

Assignee	*Chemical*	*Computers & Communications*	*Drugs & Medical*	*Electrical & Electronic*	*Mechanical*	*Others*
Qualcomm	0	5612	0	483	74	106
California University	514	269	1702	716	102	123
Broadcom	0	2264	0	441	0	145
Samsung Electro-Mechanics	0	2187	0	318	0	206
Koninklijke Philips Electronics N.V.	0	852	761	649	53	175
Microsoft	0	1880	0	107	0	474
DuPont	1007	0	509	394	95	374
Texas Instruments	0	1439	0	792	60	0
LG Electronics	0	2041	0	122	0	140
Apple	0	1169	0	429	38	107

It appears that applicants in the Computers and Communications field tend to file more provisional applications than those in other fields. We checked provisional applications that

were relied on for claiming priority filed by the top ten assignees in the Computers & Communications field each year between 2005 and 2013 and all provisional applications that were relied on for claiming priority filed by each of the top ten assignees each year between 2005 and 2013. The result was shown in Figure 5. For all the ten assignees, provisional applications filed in the Computers and Communications field were very close to all provisional applications. It indicates that, applicants in the Computers & Communications field only focused on one field.

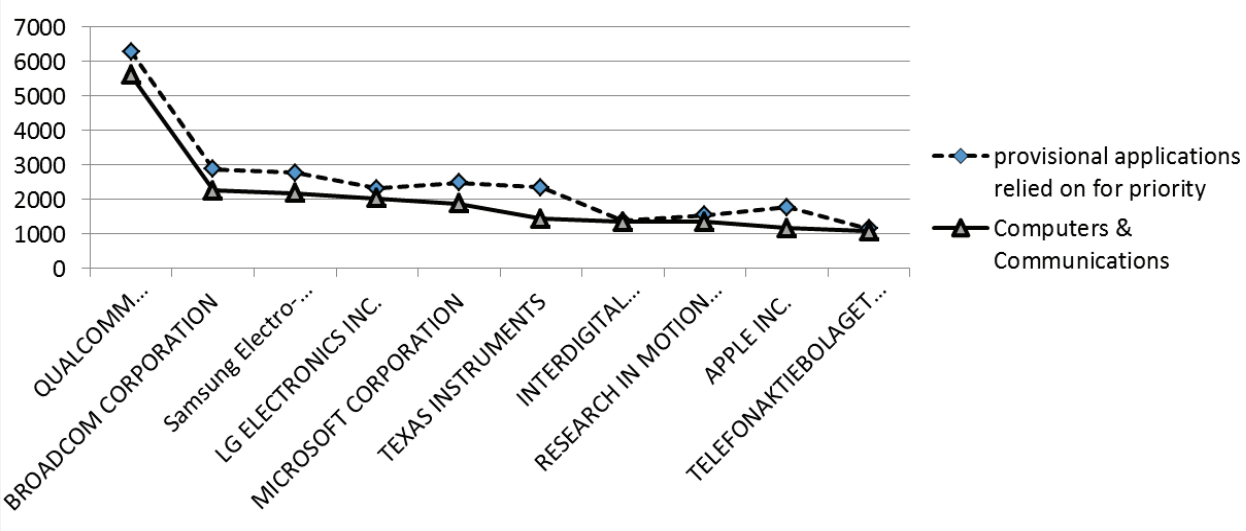


Figure 5. Provisional applications that were relied on for claiming priority filed by the top ten assignees in the Computers & Communications field each year between 2005 and 2013 and all provisional applications that were relied on for claiming priority filed by each of the top ten assignees each year between 2005 and 2013.

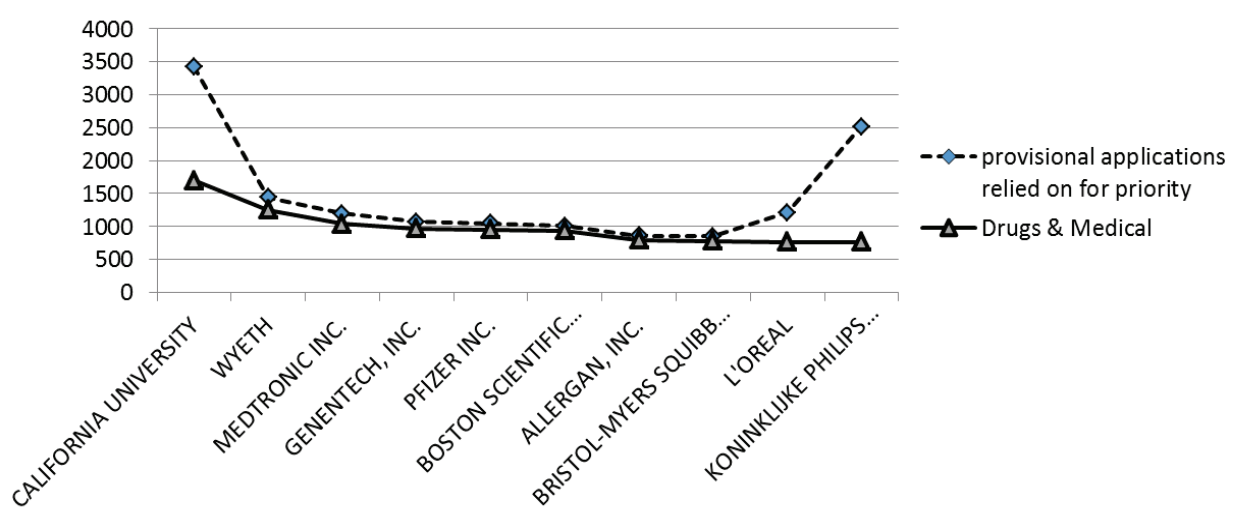


Figure 6. Provisional applications that were relied on for claiming priority filed by the top ten assignees in the Drugs & Medical field each year between 2005 and 2013 and all provisional applications that were relied on for claiming priority filed by each of the top ten assignees each year between 2005 and 2013.

Furthermore, we checked the provisional applications that were relied on for claiming priority filed by the top ten assignees in the Drugs and Medical field each year between 2005 and 2013 and all provisional applications that were relied on for claiming priority filed by each of the top ten assignees each year between 2005 and 2013. The result was shown in Figure 6.

Except for California University and Koninklijke Philips Electronics N.V., assignees filing provisional applications in Drugs & Medical also performed similarly to those in Computers & Communications, i.e. they had less diversity and only focused on one field.

Conclusion

It was found that provisional application filings continued to rise with an increase of non-provisional application filings between 2005 and 2013. The preference rate remained steady with a slight increase. The use rate of provisional applications was about 52% to 60% each year between 2005 and 2013. The PA to NPA rate can be used to evaluate the provisional application filings with respect to non-provisional application filings in different countries of origins, technological categories, and assignees. Filing date sensitivity, patent term sensitivity, and the necessity of promoting were regarded as factors weighing toward filing provisional applications. Cost sensitivity and English abilities were regarded as factors weighing against filing provisional applications.

For provisional applications by different countries of origins, applicants from Eastern Asian countries, including Japan, Korea, Taiwan and China, were less likely to file provisional applications in the US Contrarily, applicants form English speaking countries, including the US, Canada and Israel, were more likely to file provisional applications in the US. Therefore, applicants' English ability might be a major factor that influenced whether or not they would like to file provisional applications in the US.

For provisional applications by different technological categories, applicants in the fields of Computers and Communications and Drugs and Medical were more interested in filing provisional applications in the US.

For provisional applications by different assignees, most of the top ten assignees came from the Computers and Communications field.

References

Anderson, M.H., Cislo, D., Saavedra, J., & Cameron, K. (2014). Why International Inventors Might Want to Consider Filing Their First Patent Application at the United States Patent Office & the Convergence of Patent Harmonization and E-Commerce, *30 Santa Clara High Tech. L.J. 555*.

Crouch, D. (2008). *A First Look at Who Files Provisional Patent Applications.* Retrieved April 10, 2015 from: http://patentlyo.com/patent/2008/06/a-first-look-at.html

Crouch, D. (2012). *Provisional Patent Applications as a Flash in the Pan: Many are Filed and Many are Abandoned.* Retrieved December 31, 2014 from: http://patentlyo.com/patent/2012/11/provisional-patent-applications-as-a-flash-in-the-pan-many-are-filed-and-many-are-abandoned.html

Crouch, D. (2013). *Abandoning Provisional Applications*. Retrieved December 31, 2014 from: http://patentlyo.com/patent/2013/01/abandoning-provisional-applications.html

Crouch, D. (2014). *Claiming Priority to Provisional Applications.* Retrieved December 31, 2014 from: http://patentlyo.com/patent/2014/04/priority-provisional-applications.html

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). The NBER patent citation data file: Lessons, insights and methodological tools (No. w8498). National Bureau of Economic Research.

Sukhatme, N.U & Cramer, J.N.L. (2014). *Who Cares About Patent Term? Cross-Industry Differences in Term Sensitivity.* Manuscript submitted for publication.

USPTO (2014). *Provisional Application for Patent.* Retrieved December 31, 2014 from: http://www.uspto.gov/ patents/ resources/types/provapp.jsp.

A Preliminary Study of Technological Evolution: From the Perspective of the USPC Reclassification

Hui-Yun Sung[1], Chun-Chieh Wang[2] and Mu-Hsuan Huang[3]

[1] *hsung@dragon.nchu.edu.tw*
Graduate Institute of Library and Information Science, National Chung Hsing University, Taichung (Taiwan)

[2] *chunchiehwang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taipei (Taiwan)

[3] *Corresponding Author: mhhuang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University, Taipei (Taiwan)

Abstract
This study aimed to investigate technological evolution from the perspective of the USPC reclassification. The results showed that there existed significant differences among five types of patents based on the USPC reclassification: Patents reclassified to Class 001, Patents with Inter-field Mobilised Codes, Patents with Intra-field Mobilised Codes, Patents with Abolished Codes, and Patents with Original Codes. Patents reclassified to Class 001, mostly related to the topic of "Data processing", performed better than other patents in novelty, linkage to science, technological complexity and innovative scope. Patents with Inter-field Mobilised Codes, related to the topics of "Data processing: measuring, calibrating, or testing" and "Optical communications", involved broader technology topics but had a low speed of innovation. Patents with Intra-field Mobilised Codes, mostly in the Computers & Communications and Drugs & Medical fields, tended to have little novelty and a small innovative scope. Patents with Abolished Codes and patents with Original Codes performed similarly – their values of patent indicators were low. It is suggested that future research extend the patent sample to subclasses or reclassified secondary USPCs in order to understand the technological evolution within a field in greater detail.

Conference Topic
Patent Analysis

Introduction
For patented inventions, their technological novelty is indicated through their U.S. Patent Classification (USPC) assigned by the U.S. Patent Office. However, patent technology codes are an underutilized data resource for research on technological capabilities, technological novelty, technological complexity and technological change (Strumsky, Lobo & van der Leeuw, 2012). In order to fill the research gap, this study takes a first step towards using the USPC reclassification to trace technological evolution in the past two decades. This section introduces basic information regarding the USPC reclassification and sets out the research aim for investigation.

Reclassification of the U.S. Patent Classification (USPC)
The USPC is a system for organizing all U.S. patent documents and many other technical documents into relatively small collections based on common subject matter (USPTO, 2012b, I-1). A combination of a *class* (i.e. a major component) and a *subclass* (i.e. a minor component) is used to indicate every subject matter division in the USPC system. Based on the technology used, each patent is assigned specific USPC technology code(s) to reflect their technological topics. In order to distinguish from other patent classification schemes, this study only focuses on the USPC classification.

According to the USPTO (2012b, I-15), "[r]eclassification is the process of changing classifications assigned to documents classified in the USPC." There are different types of

modification of the USPC codes originally assigned to patents, including: creating, abolishing or modifying USPC class schedules. The USPC reclassification is seen necessary to reflect the evolving technological changes. For instance, Strumsky, Lobo and van der Leeuw (2012) used patent technology codes to study technological change.

Five types of patents based on the USPC reclassification

In order to keep pace with knowledge, modification/updates of classes and subclasses have been made to the Dewey Decimal Classification (DDC) system regularly. For instance, one of the new features in the DDC (Edition 23) was an update of "004–006 Computer science (and parallel provisions in 025.04 Information storage and retrieval systems and 621.39 Computer engineering) to reflect current technical trends" (Online Computer Library Center, 2013, p.3). Therefore, this study aims to investigate technological evolution from the perspective of the USPC reclassification.

As a result of the USPC reclassification, technology codes assigned to patents were created, modified and abolished. To this end, this study divided the utility patents into the following five types, according to the types of the modification of their original USPC:

- **Class 001:** If the record for a patent is incomplete and contains no *Primary Classification*[1], or if the USPTO is unable to assign specific technology codes to the patent, then the patent is reclassified to class 001, titled "CLASSIFICATION UNDETERMINED" (USPTO, 2012b).
- **Intra-field Mobilised Code:** A patent's newly assigned codes are derived from the same technological field as its original codes. Six technological fields are discussed in this paper, which are defined by Jaffe, Trajtenberg and Romer (2005).
- **Inter-field Mobilised Code:** A patent's newly assigned codes are derived from a different technological field from the original codes.
- **Abolished Code:** A patent's original technology codes are abolished and reclassified to new codes based on the Current USPC.
- **Original Code:** A patent's original technology codes remain the same as the newly assigned codes based on the Current USPC.

Based on the aforementioned five types of the utility patents, this study conducts a 20-year trend analysis and compares their variances using six patent indicators.

Methodology

Patent bibliometrics

In this study, patent data were collected solely from the United States Patent and Trademark Office (USPTO) database, which is generally accepted and is accessible to the researchers. While there exist different categories of patents (e.g. plant patents, design patents, reissues, and continuations), this study, based on the recommendations offered by Narin (2000), collected the number of regular U.S. utility patents to keep the focus of the database on the key category of patents, which contributes to corporate technological strengths. In order to observe the recent development of patents with the USPC reclassification, this study covered the past two decades. This study used the following six patent indicators to analyse the differences between different types of USPC reclassified patents.

- **Technology Cycle Time (TCT)** indicates the speed of innovation of a patent. Companies with a shorter cycle time than their competitors in a given technology area

[1] According to the USPTO (2012b), U.S. PGPub documents classified in the USPC are assigned one, and only one, principal mandatory classification, known as the *Primary Classification* (PR).

may be advancing more quickly from prior technology to current technology (Narin, 2000).

- **Non-Patent Reference (NPR)** indicates a patent's linkage to science. Narin (2000) proposed that the average rate of citations to scientific papers can be used to indicate the patent's science linkage. Other scholars (Gupta, 2006; Lo, 2010) also regarded the average rate of citations to NPRs as the patents' linkage to science. Therefore, this study used the number of NPRs to indicate the strength of linkage between the patent and science.
- **Patent Reference** indicates the novelty of a patent. A higher number of patent references generally indicate a reduction of invention novelty.
- **USPC Count** indicates the breadth of the technology topics of a patent. If a patent has broader technology topics, it tends to belong to a more highly applicable technological field.
- **Patent Term Extension** indicates the technological complexity of a patent. If the term of a patent is extended, it usually means that the patent involves a higher level of technological complexity and therefore requires more time for examination (Pantros IP, 2013).
- **Patent Claim** indicates the innovative scope of a patent. Patents containing a higher number of claims have been shown to have a wider innovative scope (Pantros IP, 2013).

Data collection

The empirical data analysed in this study were collected from the USPTO Granted Patent Database. The sample was restricted to the utility patents granted from 1994 to 2013. According to the classification system of Jaffe, Trajtenberg and Romer (2005), the U.S. patents were classified into six technological fields: Chemical, Computers and Communications (C&C), Drugs and Medical (D&M), Electrical and Electronics (E&E), Mechanical, and Others. The six fields were used to form the basis for an analysis of the patents with USPC reclassified inter-field or intra-field. USPC patents (with/without reclassification) were identified through the use of XML to compare Original USPC (i.e. USPC codes before reclassification) and Current USPC (i.e. USPC codes after reclassification). USPC reclassified patents in the recent 20 years were collected. In order to conduct a comparison analysis, the sample was randomly selected from the patents with Original USPC Codes that had the same patent count with Current USPC Codes each year.

Descriptive statistics

Descriptive statistics provide brief summaries about the sample and the observations made. Such summaries may be either quantitative (i.e. summary statistics) or visual (i.e. clear graphs). These summaries may either form the basis of the initial description of the data as part of a further statistical analysis, or they may be sufficient in and of themselves for a particular investigation. This study used the Line Chart to analyse the trends of patent counts for all types of the USPC reclassified patents granted each year. For the characteristic differences of each type of the USPC reclassified patents, this study used One-Way ANOVA to conduct significant difference tests on the patents' TCT, NPR, Patent Reference, USPC Count, Patent Term Extended, and Patent Claim.

In statistics, one-way analysis of variance (abbreviated one-way ANOVA) is a technique used to compare means of three or more samples (using the F distribution). The ANOVA tests the null hypothesis that samples in two or more groups are drawn from populations with the same mean values. To do this, two estimates are made of the population variance. If the group means are drawn from populations with the same mean values, the variance between the group means should be lower than the variance of the samples, following the central limit

theorem. A higher ratio therefore implies that the samples were drawn from populations with different mean values (Wikipedia, 2014).

Results

Trends of the USPC reclassified patents

There were 3,342,076 U.S. utility patents granted between 1994 and 2013. Among them, 102,204 patents belonged to the main class in Primary USPC reclassification, which accounted for 3.1% of the total utility patents. Calculations of those patents by their types showed that patents with Abolished Codes accounted for the majority (42.53%), which was followed by patents with USPC Intra-field Mobilised Codes. Patents with Class 001 or Inter-field Mobilised Codes accounted for appropriately 15% respectively. See Table 1.

Table 1. Counts of patents with/without USPC reclassification.

Patent with/without USPC Reclassification	Count
Main class in Primary USPC Reclassification	**102,204(100%)**
A. Class 001	15,862(15.52%)
B. Abolished Code	43,465(42.53%)
C. Inter-field Mobilised Code	15,740(15.40%)
D. Intra-field Mobilised Code	27,137(26.55%)
E. Random selection of patents with Original Code	**102,204**

Observed from the yearly distribution of the patent counts of various types of USPC reclassification, it was found that the number of USPC reclassified patents tended to be higher in the early stage, which indicated that the USPC was revised in accordance with the evolution of technologies. From the perspective of the Current USPC, some Original USPC appeared inappropriate in today's context and therefore the count of the USPC reclassified patents has increased. Furthermore, when the advance of newer technologies adopted the Original USPC that was similar to the version of October 2014, the number of USPC reclassified patents decreased in tandem.

The number of patents with Abolished Codes dramatically increased prior to 2000 but dramatically dropped after 2001, meaning that the elimination of main class did not occur after 2001. The number of patents with USPC Intra-field Mobilised Codes was above 1,000 before 2009 and started to decrease after 2010, which was considered relevant to "Technological development for stability". The numbers of patents with USPC Inter-field Mobilised Codes and with Class 001 tended to decrease in 2010, which was also considered relevant to "Technological development for stability".

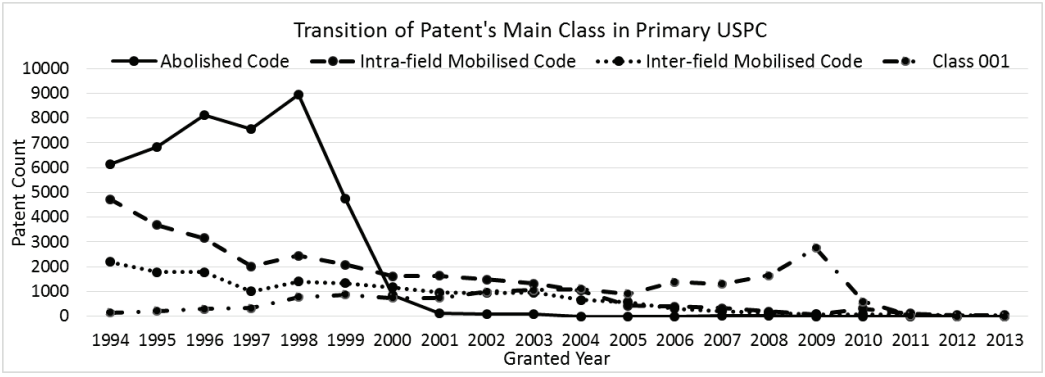


Figure 1. Transition of patents' main class in primary USPC.

Average citation rates were used to represent the quality of patents. This study calculated patents' average citation rates from 1994 to 2013, as shown in Figure 2. Due to the fact that the citation window of patents has become shorter each year, patents' average citation rates also decreased gradually. Figure 2 shows that the average citation rates of patents with Class 001 were the highest, which was followed by patents with USPC Inter-field/Intra-field Mobilised Codes. (They performed similarly in terms of their average cited rates recently.) The average citation rates of patents with Abolished Codes were higher than patents with Original Codes before 2002, but their average citation rates became the lowest among all types of patents.

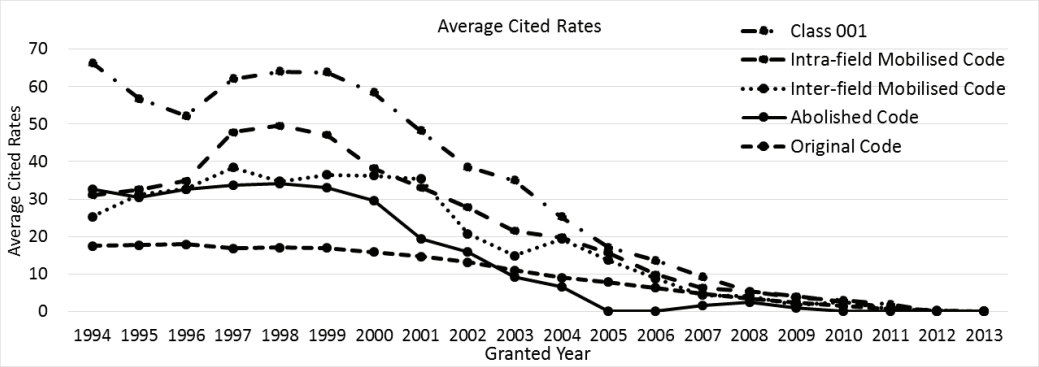


Figure 2. Average cited rates of USPC reclassified patents.

USPC reclassified patents among fields

Table 2. Patent counts in technological fields with USPC Reclassification.

	Patent Reclassified to Current Tech Field (%)						
Original Tech Field	**1.**	**2.**	**3.**	**4.**	**5.**	**6.**	**Sum**
1. Chemical	3,303	62	276	816	684	252	**5,393**
	(61.25)	(1.15)	(5.12)	(15.13)	(12.68)	(4.67)	**(100)**
2. Computer & Communication	135	11,649	16	1,201	81	1,913	**14,995**
	(0.90)	(77.69)	(0.11)	(8.01)	(0.54)	(12.76)	**(100)**
3. Drugs & Medical	958	13	6,260	44	23	96	**7,394**
	(12.96)	(0.18)	(84.66)	(0.60)	(0.31)	(1.30)	**(100)**
4. Electrical & Electronic	155	1,627	49	1,187	124	1,273	**4,415**
	(3.51)	(36.85)	(1.11)	(26.89)	(2.81)	(28.83)	**(100)**
5. Mechanical	979	3,037	74	172	2,773	237	**7,272**
	(13.46)	(41.76)	(1.02)	(2.37)	(38.13)	(3.26)	**(100)**
6. Others	756	94	111	159	323	1,965	**3,408**
	(22.18)	(2.76)	(3.26)	(4.67)	(9.48)	(57.66)	**(100)**
Sum	**6,286**	**16,482**	**6,786**	**3,579**	**4,008**	**5,736**	**42,877**
	(14.66)	**(38.44)**	**(15.83)**	**(8.35)**	**(9.35)**	**(13.38)**	**(100)**

Table 2 displays the U.S. utility patents granted from 1994 to 2013 with USPC reclassified inter/intra-field. It was found, through calculating the variances in the patent count in the original and current technological fields that patents in C&C were reclassified most among all the USPC reclassified patents. Among the patents in original technological fields in C&C, 77.69% belonged to the main class in the Primary USPC Intra-field Mobilised Code, with 12.76% reclassified to Others. Another variance occurred to D&M. 84.66% of the patents belonged to the main class in Primary USPC Intra-field Mobilised Code, with 12.76%

reclassified to Chemical. The last variance occurred to Mechanical. 38.13% of the patents belonged to the main class in Primary USPC Intra-field Mobilised Code, with 41.76% reclassified to C&C. 36.85% of patents in E&E were reclassified to C&C, 28.83% reclassified to Others, and only 26.89% reclassified intra-field.

Statistical differences among five patent groups

Six one-way between subjects ANOVAs were conducted to compare the effect of patents with different USPC reclassification types on patent performance in TCT, NPR, Patent Reference, USPC Count, Patent Term Extended, and Patent Claim. There were all significant differences of indicators on patent performance at the $p<.001$ level for the five types of patents with/without USPC reclassification. Post hoc comparisons using the Dunnett T3 test (Dunnett, 1980) showed significant differences in the mean scores of the six indicators for the patents in different types of the USPC reclassification.

- **TCT Performance:** When the value of TCT is lower, it means a patent involves more fast-moving technologies and a patent tends to cite recently issued patents. Results derived from statistical tests showed: B. Abolished Code (5.7 year) < C. Inter-field Mobilised Code (6.3 year) < E. Original Code. (7.8 year). Short TCT of the patents with Abolished Codes indicated that patents of this kind involved the most fast-moving technologies and the speed of their technological innovation was clearly faster than patents with Inter-field Mobilised Codes. On the contrary, patents with Original Codes tended to be slower in term of their speed of the technological innovation.
- **NPR**: When the number of NPR is higher, it means the linkage of technology to science is stronger. Results derived from statistical tests showed: A. Class 001 (10.4), C. Inter-field Mobilised Code (11.7) & D. Intra-field Mobilised Code (10.7) > E. Original Code (7.9) & B. Abolished Code (5.5). When calculating Science Linkage, the more NPRs were, the stronger the linkage of technology to science was. Therefore, patents reclassified to Class 001, patents with Inter-field Mobilised Codes and Intra-field Mobilised Codes had stronger linkages to science, compared to patents with Original Codes and Abolished Codes.
- **Patent Reference:** When the number of Patent References is low, it indicates the novelty of technology is high. Results derived from statistical tests showed: B. Abolished Code (11.6) < E. patent with Original Code (14.2) < A. Class 001 (19.3) < C. Inter-field Mobilised Code (15.0). It can be inferred that the technological novelty of patents with Abolished Codes was much higher than that of patents with Original Codes. Clearly, the technological novelty of patents with Class 001 or with Inter-field Mobilised Codes tended to be low.
- **USPC Count:** Patents with more USPC counts indicate they involve broader technologies. Results derived from statistical tests showed: C. Inter-field Mobilised Code (5.2) > E. Original Code (4.4) > B. Abolished Code (3.9). The technology breadth of patents with Inter-field Mobilised Codes was the largest. The technology breadth of patents with Abolished Codes was smaller than that of patents with Original Codes.
- **Patent Term Extended:** When the term extension lasts longer, it indicates that a patent involves more complicated technologies. Results derived from statistical tests showed: A. Class 001 (416) > C. Inter-field Mobilised Code (341), D. Intra-field Mobilised Code (307) > E. patent with Original Code (300) > B. Abolished Code (168). It can be inferred that patents with Class 001 involved a higher level of technological complexity than patents with Inter/Intra-field Mobilised Codes. However, the term extension of patents with Abolished Codes was the shortest, indicating that they involved the lowest level of technological complexity.

- **Patent Claim:** When the value of patent claims is higher, it indicates that a patent's innovation scope is wider. Results derived from statistical tests showed: A. Class 001 (22.2) > C. Inter-field Mobilised Code (17.6) > B. Abolished Code (16.5), E. patent with Original Code (15.1). It can be inferred that the innovation scope of the patents with Class 001 or patents with Inter-field Mobilised Codes was obviously wider than that of patents with Abolished Codes and patents with Original Codes.

Technological evolution from the USPC reclassification perspective

This study divided patents granted in the last two decades into two groups, i.e. 1994-2003 and 2004-2013. Observations were made from the evolution of USPC codes as a result of the USPC reclassification. Table 3 shows the USPC with top three most patent counts in the two periods respectively. If a patent was reclassified to Class 001, it meant that there was no specific technology code suitable for the patent. To some extent, it indicated that the patent belonged to emerging technologies or original USPC codes assigned were not appropriate for the patent, which required a new code. Table 3 shows in both periods, the majority of patents reclassified to Class 001 came from Class 707 in the C&C field. This phenomenon reflected the technological uncertainty of patents originally assigned to Class 707, the majority of which were therefore reclassified to Class 001. In the first period, there were 19.7% of patents originally assigned to Class 395 and then reclassified to Class 001. However, due to the abolition of Class 395, their technological description remained unknown.

Table 3. Patents with USPC reclassified in the Class 001 and the Abolished Code groups.

USPC	1994-2003	2004-2013	USPC Description
Original class reclassified to 001 (Class 001)			
707	4,884 (79.6%)	9,684 (99.5%)	Data processing: database and file management or data structures
395	1,206 (19.7%)	0 (0.0%)	(Abolished)
364	19 (0.3%)	0 (0.0%)	(Abolished)
705	0 (0.0%)	18 (0.2%)	Data processing: financial, business practice, management, or cost/price determination
714	0 (0.0%)	7 (0.1%)	Error detection/correction and fault detection/recovery
Current class of original abolished (Abolished Code)			
438	4,895 (11.3%)	1 (4.3%)	Semiconductor device manufacturing: process
714	4,179 (9.6%)	0 (0.0%)	Error detection/correction and fault detection/recovery
710	3,448 (7.9%)	0 (0.0%)	Electrical computers and digital data processing systems: input/output
703	1,314 (3.0%)	2 (8.7%)	Data processing: structural design, modeling, simulation, and emulation
477	2 (0.0%)	2 (8.7%)	Interrelated power delivery controls, including engine control

For patents with Abolished Codes, it meant that their original codes did not align with the technological evolution any more, and thus the codes were abolished and the patents were reclassified to new codes. As shown in Table 3, the majority of patents with Abolished Codes

occurred in the first period, with only 23 patents of this kind in the second period. In the first period, the majority of patents whose original USPC codes were abolished were reclassified to Classes 438 (11.3%), 714 (9.6%), 710 (7.9%), and 703 (3.0%). Patents reclassified to Class 438 were about semiconductor device manufacturing in the E&E field, and those reclassified to Classes 714, 710 and 703 focused on technologies in the C&C field. Based on the patents reclassified to Class 001 and with Abolished Codes, it was found that the USPC reclassification tended to occur in the C&C and E&E fields in the first period and in the C&C field in the second period.

According to Table 2, patents with Intra-field Mobilised Codes mainly occurred in the C&C (77.69%) and D&M (84.66%) fields. Therefore, Table 4 focuses on the top three Intra-field Mobilised Codes, and Figures 3 and 4 present the flow of the patents between USPCs in the two fields, where the flow occurred more than ten patents. In the C&C field, the USPC reclassification in both periods mainly occurred from Class 345 to Class 715 (28.8% and 26.6%), which was about "Operator interface processing" and from Class 369 to Class 720 (11.8% and 5.1%), which was about "Information storage or retrieval". Additionally, in the first period, there remained 10.2% of patents reclassified from Class 707 to Class 715, which was also about "Operator interface processing". In the second period, there remained 6.2% of patents reclassified from Class 707 to Class 709, which was about "Multicomputer data transferring". In the D&M field, the USPC reclassification occurred from Class 128 to Class 600 (68.0%) which was about "Surgery" in the first period, and from Class 514 to Class 424 (76.4%) which was about "Drug, bio-affecting and body treating compositions" in the second period. The code mobilisation within the same field occurred due to the extension of the original USPC.

Table 4. USPC reclassification: the Intra-field Mobilised Code group.

Main Class of USPC		Count	
Original	**Current**	**1994-2003**	**2004-2013**
Intra-field Mobilised Code in C&C			
345	**715**	2,793(28.8%)	516(26.6%)
369	**720**	1,144(11.8%)	98(5.1%)
707	**715**	991(10.2%)	96(5.0%)
707	**709**	68(0.7%)	120(6.2%)
Intra-field Mobilised Code in D&M			
128	**600**	3,909(68.0%)	4(0.8%)
514	**424**	626(10.9%)	389(76.4)
606	**623**	227(3.9%)	18(3.5%)
514	**435**	17(0.3%)	26(5.1%)
435	**424**	49(0.9%)	21(4.1%)

345: Computer graphics processing and selective visual display systems; **369:** Dynamic information storage or retrieval; **707:** Data processing: database and file management or data structures; **709:** Electrical computers and digital processing systems: multicomputer data transferring; **715:** Data processing: presentation processing of document, operator interface processing, and screen saver display processing; **720:** Dynamic optical information storage or retrieval

128: Surgery; **424:** Drug, bio-affecting and body treating compositions; **435:** Chemistry: molecular biology and microbiology; **514:** Drug, bio-affecting and body treating compositions (an integral part of Class 424); **600:** Surgery (an integral part of Class 128); **606:** Surgery (an integral part of Class 128); **623:** Prosthesis (i.e., artificial body members), parts thereof, or aids and accessories therefor

Observed from the patents with Intra-field Mobilised Codes, it showed that in the C&C field those patents were related to "Operator interface processing" in both periods. In the D&M field those patents were related to "Surgery" in the first period and "Drug, bio-affecting and body treating compositions" in the second period. Observed from the patents with Inter-field Mobilised Codes, it showed that the USPC codes were mainly mobilised from the E&E and Mechanical fields to the C&C field, as seen in Table 2. Statistics on the top three USPC mobilisation were detailed in Table 5, and Figures 5 and 6 present the flow of the patents between USPCs among the three fields, where the flow occurred more than ten patents. In the first period, the USPC reclassification mainly occurred from the E&E field to the C&C field, for example from Class 348 to Class 375 (64.6%) about "Pulse or digital communications", and from Class 346 to Class 374 (20.6%) about "Thermal measuring and testing". However, in the second period, inter-field code mobilisation was not obvious. It can be seen that the topics of technological evolution were different in the two periods.

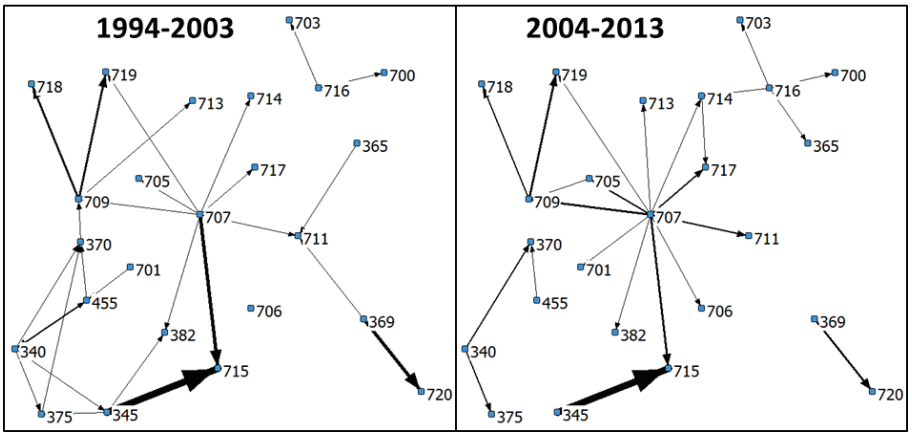


Figure 3. The flow of patents between USPCs in the C&C field.

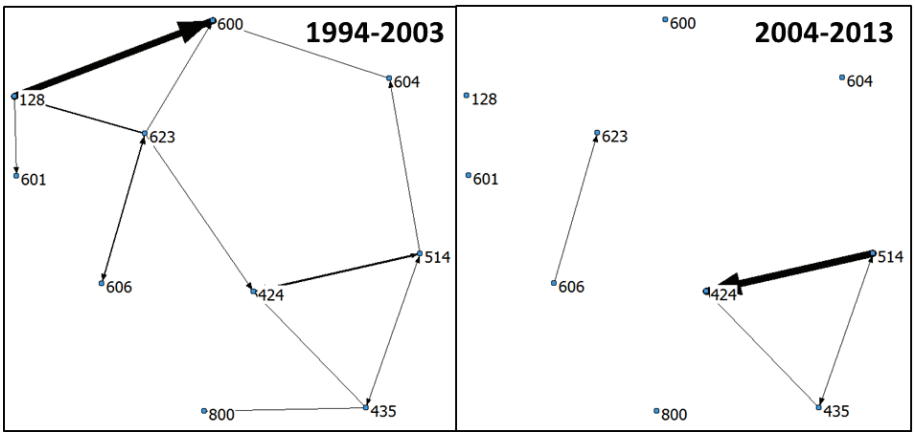


Figure 4. The flow of patents between USPCs in the D&M field.

Looking at patents with Inter-field Mobilised Codes from the Mechanical field to the C&C field, the flow of the mobilisation tended to occur from Class 359 to Class 398 (94.8% and 37.8%) about "Optical communications" in both periods.

Observed from the patents with Inter-field Mobilised Codes, it showed that patents with the USPC reclassification from the E&E field to the C&C field focused on the technology topics of "Pulse or digital communications" and "Thermal measuring and testing" in the first period, but focused on "Data processing: measuring, calibrating, or testing" in the second period. As

for patents with USPC reclassification from the Mechanical field to the C&C field, they tended to be related to "Optical communications" in both periods.

Table 5. USPC reclassification: the Inter-field Mobilised Code group.

Main Class of USPC		Count	
Original	**Current**	**1994-2003**	**2004-2013**
Inter-field Mobilised Code from E&E to C&C			
348	**375**	989(64.6%)	2(2.1%)
346	**374**	316(20.6%)	0(0.0%)
257	**365**	21(1.4%)	3(3.2%)
257: Active solid-state devices (e.g., transistors, solid-state diodes); **346:** Recorders; **348:** Television; **365:** Static information storage and retrieval; **374:** Thermal measuring and testing; **375:** Pulse or digital communications			
Inter-field Mobilised Code from Mechanical to C&C			
359	**398**	2,837(94.8%)	17(37.8%)
235	**705**	22(0.7%)	0(0.0%)
359	**369**	15(0.5%)	0(0.0%)
235: Registers; **359:** Optical: systems and elements; **369:** Dynamic information storage or retrieval; **398:** Optical communications; **705:** Data processing: financial, business practice, management, or cost/price determination			

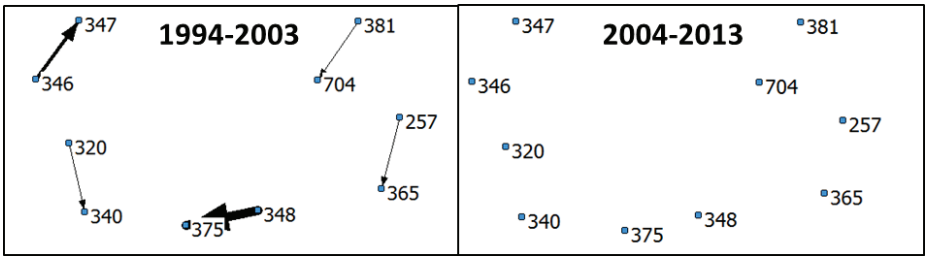


Figure 5. The flow of patents between USPCs from the E&E to the C&C field.

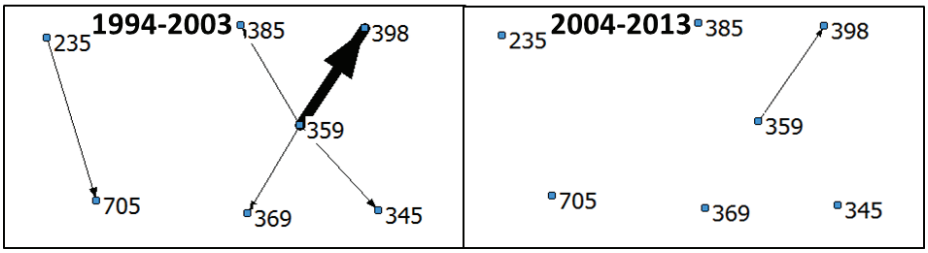


Figure 6. The flow of patents between USPCs from the Mechanical field to the C&C field.

Conclusion and Discussion

The majority of USPC reclassified patents occurring prior to 2000 and in the Computer & Communications field

With the advance of new technologies, the USPC system is updated quarterly in March, June, September and December (USPTO, 2012a). Newly granted patents were assigned with technology codes derived from the latest version of the USPC. Accordingly, their original USPC technology codes were less likely to be reclassified. This study found that the number of patents with main class in primary USPC reclassification hit the highest prior to 2000 and began to decrease every year after 2001. Patents with Abolished Codes accounted for 42.53%

and the majority of the patents were granted prior to 2000. Next were patents with Intra-field Mobilised Codes, which accounted for 26.55%. For the average citation rates every year, patents reclassified to Class 001 were ranked as top, and patents with Original Codes were ranked as bottom. Due to the USPC reclassification, patents with Intra-field Mobilised Codes occurred most frequently in the C&C field, and patents with Inter-field Mobilised Codes occurred most frequently from the Mechanical field to the C&C field.

USPC reclassified patents showing significant differences in patent indicators

Six one-way between subjects ANOVAs were conducted to compare the effects of patents in different groups by the USPC reclassification, according to their patent performance in TCT, NPR, Patent Reference, USPC Count, Patent Term Extended, and Patent Claim. Different results were obtained for the different types of patents, as below.

- **Patents reclassified to Class 001:** They got higher values of NPR, Patent Reference, Patent Term Extended and Claims Count, indicating that they performed better than other patents (whether they were reclassified or not) in novelty, linkage to science, technological complexity and innovative scope. Therefore, USPTO needs to re-examine appropriate USPC technology codes for them or assign appropriate codes to them when the new codes are created.
- **Patents with Inter-field Mobilised Code**: Compared to patents reclassified to Class 001, they got more USPC counts and longer TCT, indicating that they involved broader technology topics and therefore their codes assigned were mobilised inter-field. Their longer TCT meant that their technology had a low speed of innovation.
- **Patents with Intra-field Mobilised Code**: They tended to have low novelty and a small innovative scope; therefore, their codes assigned were mobilised intra-field.
- **Patents with Abolished Code**: They were mainly granted prior to 2000. Patens of this type and patents with Original Code performed similarly – their values of patent indicators were low.

Technological evolution from the perspective of the USPC reclassification

This study investigated different groups of patents based on the USPC reclassification. Statistical analysis was conducted on the technology codes and comparisons were made between two ten-year periods. Based on the results derived, different types of technological evolution were found.

- **Emerging technologies in Class 001:** In both periods, a large portion of the emerging technologies were about "Data processing: database and file management or data structures" in the C&C field. This reflects the uncertainty of the development of the emerging technology, and thus patents originally assigned to Class 707 needed to be continually redefined and reassigned with specific technology codes.
- **Technological transition in Inter-field Mobilised Code:** Technologies from the E&E and Mechanical fields tended to be transferred and applied to the C&C field. Technologies about "Television" in E&E was transferred and applied to "Pulse or digital communications" in the C&C field. Technologies about "Recorders" in E&E were also transferred and applied to "Thermal measuring and testing" in the C&C field. In the Mechanical field, technologies related to "Optical: systems and elements" were transferred and applied to "Optical communications" in the C&C field in both periods.
- **Technological cohesion or spread in Intra-field Mobilised Code:** Technologies in this group tended to focus on the C&C and D&M fields. In the C&C field, technologies related to "Computer graphics processing and selective visual display systems" and "Data processing: database and file management or data structures" were combined together and applied to "Data processing: presentation processing of document, operator

interface processing, and screen saver display processing". Figure 3 shows not only technological cohesion but also technological spread. For example, technologies about "Data processing: database and file management or data structures (Class 707)" were spread to other technologies in different fields. Patents with original USPC 707 were reclassified to eight different codes in the first period, and then spread to other ten codes in the second period.

- **Technological substitution in Abolished Code:** Technologies in this group tended to occur in the first period. This indicates that the USPC scheme in the second period has been adapted to the recent technological development. In the first period, technologies of this kind mainly occurred to those related to "Semiconductor device manufacturing", which were reclassified to Class 438 with their original USPC 437 being abolished. Technologies related to "Error detection/correction and fault detection/recovery" which were reclassified to Class 714 with their original USPC 371 and 395 being abolished. This indicates that the mature technologies have caused the biggest impact on the USPC scheme.

It is suggested that future research extend the sample to patents with reclassified USPC subclasses or patents with reclassified secondary USPCs in order to observe recent intra-field technological changes in great detail. The Radical (Leaps) Innovation of technologies is only applied to the minority, but the majority of patents are embedded with Incremental Innovation. Incremental Innovation tends to occur inside fields. Through extending the patent sample to subclasses or secondary of USPC, it helps understand more technological evolution within a field. Besides, understanding the establishment, abolishment and movement of technology codes recorded in the Classification Orders Archival Report (USPTO, 2013) helps understand the trajectories of technological evolution more detail. Although this study focused on the reclassification of USPC schemes, it is argued that the same research model could be applied to trace the changes in the class schemes in International Patent Classification (IPC) or Cooperative Patent Classification (CPC) and changes in classification codes in their counterpart patents.

References

Dunnett, C. W. (1980). Pairwise Multiple Comparisons in the Homogeneous Variance, Unequal Sample Size Case. *Journal of the American Statistical Association*, 75, 789-795.

Gupta, V. K. (2006). References to literature in patent documents: A case study of CSIR in India. *Scientometrics*, *68*(1), 29-40.

Lo, S. S. (2010). Scientific linkage of science research and technology development: A case of genetic engineering research. *Scientometrics*, *82*(1), 109-120.

Narin, F. (2000). Tech-line® background paper. In Tidd J. (Ed.), *From knowledge management to strategic competence* (pp. 155–195). London: Imperial College Press.

Online Computer Library Center. (2013). New Features in DDC Edition 23. Retrieved April 10, 2015 from: https://www.oclc.org/content/dam/oclc/dewey/versions/print/new_features.pdf

Pantros IP. (2013). Patent Factor Reports. Retrieved January 7, 2015 from: http://admin.patentcafe.com/reports/pantrosip_reports/patentfactor_terms.pdf

Strumsky, D., Lobo, J., & van der Leeuw, S. (2012). Using patent technology codes to study technological change. *Economics of Innovation and New Technology*, *21*(3), 267-286.

USPTO. (2012a). Classification Orders Index (COI). Retrieved January 10, 2015 from: http://www.uspto.gov/patents/resources/classification/orders/coi.jsp.

USPTO. (2012b). Overview of the U.S. Patent Classification System (USPC). Retrieved January 7, 2015 from: http://beta.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf.

USPTO. (2013). Classification Order Archival Report. Retrieved January 7, 2015 from: http://www.uspto.gov/patents/resources/classification/archiverpt.pdf.

Wikipedia. (2014). One-way analysis of variance. Retrieved January 7, 2015 from: http://en.wikipedia.org/wiki/One-way_analysis_of_variance.

Cognitive Distances in Prior Art Search by the Triadic Patent Offices: Empirical Evidence from International Search Reports

Tetsuo Wada

tetsuo.wada@gakushuin.ac.jp
Gakushuin University, Faculty of Economics, Mejiro, Toshima-ku, Tokyo 171-8588 (Japan)

Abstract
Despite large numbers of empirical studies are conducted on examiner patent citations, few have scrutinized the cognitive limitations of officials at patent offices in searching for prior art to add citations during patent prosecution. This research takes advantage of the longitudinal gap between International Search Reports (ISRs) required by the Patent Cooperation Treaty (PCT) and subsequent examination procedure in national phase. It inspects whether several kinds of distances actually affect the probability that a piece of prior art is caught at the time of ISRs, which is much earlier than national phase examinations. Based on triadic PCT applications for all of the triadic patent offices (EPO, USPTO, and JPO) between 2002 and 2005 and their citations made by the triadic offices, evidence shows that geographical and organizational distances negatively affect the probability of prior patents being caught in ISRs, while lag of prior art positively affects the probability. Also, technological complexity of an application negatively affects the probability, whereas the size of forward citations of prior art affects positively.

Conference Topic
Patent Analysis (foundation of examiner patent citations, in particular)

Introduction
Patent citations have been widely utilized for empirical studies of patent systems, particularly for such issues as economic value and knowledge flows. Several empirical studies have examined whether examiner citations are different from inventor citations. One of the studies on the subject was conducted by Alacer and Gittleman (2006), who showed the similarity between examiner citations and inventor citations with respect to geographical distance in particular. While previous studies have compared examiner citations and inventor citations in other aspects such as the relationship with renewal rates, there have not been enough analyses concerning how patent offices are influenced by several kinds of "distances" that can limit cognitive boundary during prior art search. This study focuses on ISRs as a basis for measuring the search obstacles of the triadic patent offices, and tests how officials are bounded by "distances," including similar kinds of cognitive obstacles against prior art search, without relying on comparison with inventor citations. In conducting the analyses, we consider applicants' self-selection, since applicants from the U.S. and Japan can choose the European Patent Office as their search agency, where the EPO has reputation for its complete search (applicants who seek stringent search may choose the EPO ex ante).

The methodology: PCT and ISR as the basis of empirical measurement
This project proposes and implements a method of measuring the search obstacles, namely binding conditions on search capability, of the triadic patent offices by focusing on ISRs issued by different ISAs, specifically the patent offices in Europe, the U.S. and Japan, according to the PCT. In particular, binary choice models are employed for each of cited patents (which are added in the national phase in all of three jurisdictions) about whether or not they were already caught at the earlier time of ISR issued by the triadic offices. We limit our samples to those PCT applications made to and examined at all of the three offices. There are advantages to employ this methodology.

First, ISRs are issued under the common search criterion imposed by the WIPO under the PCT system. Under the PCT, "an applicant must file an application with a receiving office and choose an international searching authority to provide an international search report and a written opinion on the potential patentability of the invention." "The applicant generally has at least 30 months from the filing (priority) date to decide whether to enter the national phase in the countries or regions in which protection is sought" (WIPO, 2014). The guideline at the WIPO applies to every ISA when issuing ISRs, whereas applicants in some countries are allowed to choose ISAs. The same criterion for prior art search is applied over different patent offices, while national phase examinations do not have such standardized rules.

Second, the lag mentioned above between ISRs and national phase examinations allows a "level" testing ground for search completeness. While ISRs are issued at an early stage, more searches are conducted in national offices later. Since knowledge is geographically localized (Jaffe et al. 1993; 1999), and knowledge diffusion takes time, additional time between ISRs and national phase search facilitates more complete search in the later stage. We limit our samples to those PCT applications that are examined at all of the three triadic offices, meaning that localized knowledge in any of these areas at the time of ISRs is more likely to be caught by the offices at the national phase in a less localized way. See Figure 1 below for the lag and collective searches made at later stages in national phase.

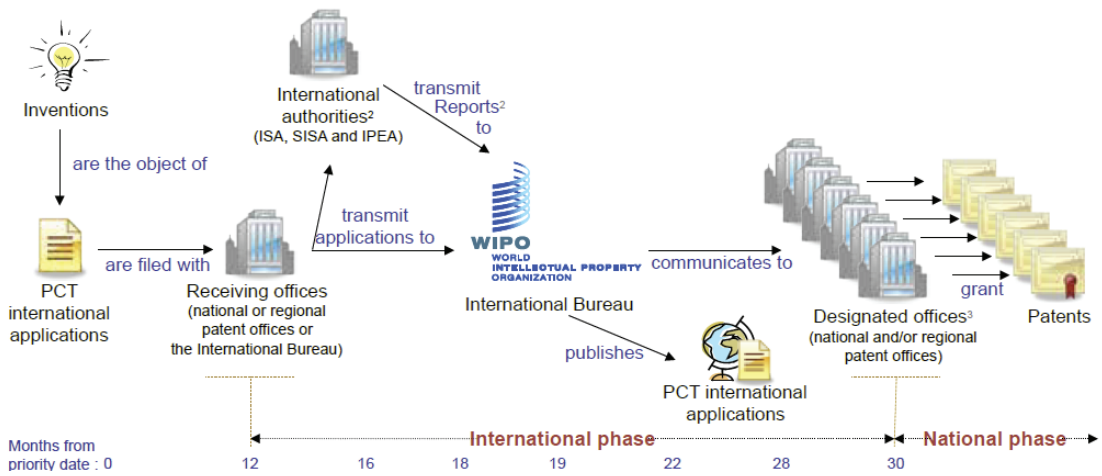


Figure 1. PCT procedure (replicated from WIPO, 2014, p.13).

Following the logic above, we retrospectively define the probability of every cited patent depicted in national phase, identified at the INPADOC family level, to have been already caught in the ISR of the originating PCT application. Taking this probability (a binary variable *found_in_ISR*, empirically) as the dependent variable, we implement PROBIT analyses at INPADOC family level with explanatory variables representing the various "distances" between citing and cited patents, including technological complexity of originating applications, and other related indicators.

Applicants' (inventors') citations are excluded from the analysis, since the objective is to evaluate the determinant of search completeness by the ISAs. However, self-selection of the U.S. and Japanese applicants to choose the EPO as their ISA is considered in the analyses, since the EPO has high reputation of examination standard and therefore applications with higher quality from the U.S. and Japan may choose the EPO as the ISA.

Although actual ISR search is sometimes outsourced to non-PTO agencies, we consider ISRs as a basis of evaluating PTOs, since they are issued under the name of the patent offices, not private search agencies. Only citations made by the triadic offices are considered in the current analyses. Since PATSAT, our primary data source, records non-patent literature in

non-standardized formats, we could not consolidate the same non-patent literature across different records. For this reason, we employ patent citations only at this time.

Hypotheses

Since ISR searchers (examiners/searchers for patent offices) are affected by cognitive obstacles from various "distances," we hypothesize that a prior patent (that was found in ISR or national phase) is more likely to be found in ISR when "distances" are less problematic, i.e.,
H1) a relevant prior patent is closer in geography (physical distance),
H2) prior patent is older (knowledge diffusion time),
H3) prior patent is from the same applicants (organizational distance),
H4) prior patent has more number of forward citations (knowledge diffusion probability), and
H5) application for which an ISR is issued has less scope, less number of claims, less number of inventors, and less number of international family (complexity against diffusion).
In addition, we consider if applicants' self-selection of ISAs affects the outcome variable.

Data source

The empirical domain of analysis is the triadic patent applications through PCT, with their earliest priority date within its international family between 2002 and 2005. Triadic PCT patent applications are defined here as INPADOC families that contain all of EPO, USPTO and JPO applications recorded on EPO's PATSTAT database, with only one "WO (PCT)" application in a family, meaning that a single PCT application initiates international phase for all applications in a family. The number of international families for the analysis is 97,828. Although international applications to and from China and Korea has increased dramatically in the last ten years, the triadic patent offices of the EPO, the USPTO and the JPO represented the vast majority before 2005, which is our observation period.
EPO PATSTAT (2013 OCT version) is used, and INPADOC family is the unit of analysis. Citation data also comes from PATSTAT (2013 OCT), although JPO citation data is augmented by Seiri-Hyojunka data (JPO's standardized patent prosecution data). US citations are not complete as well on PATSTAT, since citations for rejected applications are not registered on PATSTAT. The lack of the US citations for rejected applications may affect the result of the analysis, but this has not been verified yet. Applicant identifiers are consolidated by the EEE-PPAT database developed by ECOOM (Du Plessis et al., 2009; Magerman et al., 2009; Peeter et al., 2009).

Variables

We employ several categories of explanatory variables, representing each of hypotheses above, in PROBIT analyses taking the probability of a cited patent being caught in the previous ISR as the binary dependent variable ("*found_in_ISR")*. The unit of analysis is a pair of citing and cited international families, both consolidated at INPADOC family level.
For H1, three variables of *euro_cited* (cited family has its 1st priority, i.e., the earliest date, in EPC countries within a family, derived from tls201 and tls219 tables of PATSTAT), *us_cited* (cited family has its 1st priority in the U.S.), and *jp_cited* (cited family has its 1st priority in Japan) are defined. When a cited family has its origin in the same region where ISR is issued, the ISA of the region is expected to have geographical advantage over the relevant technology. Expected sign is positive for each region, e.g., positive *jp_cited* coefficients for applications originating from Japan.
For H2, citation lag between the 1st priority of a citing family and that of a cited family is defined as *fam_cite_lag* (derived from tls201 and tls219 tables of PATSTAT). The longer the lag is, the easier the prior art will be to be found at the time of ISR.

For H3, *self* is defined as a binary variable, taking the value of one if one of patents in a cited family and one of patents in a citing family belongs to the same applicant, based on PATSTAT (tls207) combined with EEE-PPAT, using "L2" id. Patent office will find it easier to locate prior relevant art within the same applicant.
For H4, *fwd_cite_of_the_cited* is defined and obtained from PATSTAT (tls217) as the number of forward examiner citations, counted at publication level (but consolidated at family level), and made out to the cited patent family.
For H5, we first use scope indicators. *IPC4_count* is the total net count of IPC subclasses (4-digit IPC, derived from tls209) assigned in a citing INPADOC family. Since patent classification of an application may change during prosecution process both in international phase and in national phase, we include all IPC subclasses to capture the breadth of a family. The number of claims of a patent is correlated with the complexity of the technological content. As an indicator of the number of claims, we obtain *publn_claims_max_tls211*, which is the maximum number of claims registered on PATSTAT (tls211 table) in a citing INPADOC family. We do not simply rely on claims data from a single office such as from the EPO, since an application can be modified during its prosecution internationally. We also employ *invt_nr*, the maximum number of inventors in an application included in a citing INPADOC family, from PATSTAT (tls207). The size of international family, *family_size*, is a count variable of applications in different countries in a citing INPADOC family (tls211/219).
In addition to the variables above, which are used to test hypotheses directly, we define three variables to address self-selection of ISAs by applicants. The first two represent the potential of the applicant. The first of the two is *total_count*, which is the number of total applications that an applicant has made, taken from EEE-PPAT. The second one is *applicant_avg_cited*, which is the number of average forward citations that an applicant has received, calculated by PATSTAT (tls212) and EEE-PPAT. Both are supposed to represent the experience level of the applicant, and are used as instrument variables for instrumented PROBIT on the variable *ISA_CHANGED*. This binary variable *ISA_CHANGED* indicates that the U.S. and Japanese applicants choose the EPO as their ISA (the EPO can be chosen from the U.S. and Japanese applicants, but not vice versa). This information can be obtained for PCT applications on PATSTAT, since the citation table tls212 has a field on "citation origin" where "ISR" is shown for PCT applications. Since first application country (RO) in a family is available from tls201, switching from RO to a different ISA can be coded. The correlation coefficient between *ISA_CHANGED* and the dependent variable *found_in_ISR* is low at 0.0348.
Control variables for originating areas, which are *JP_app* and *US_app* (applications from Japan and the U.S., respectively), are used. Technology class is controlled by thirty-five WIPO technology classification dummies (results not shown for space reason).

Estimation results

The result shown in the Model 1 of Table 1 employs all samples from the triadic regions. As is evident from the negative sign for *JP_app* and *US_app*, the baseline ISA (EPO) is found to be advantaged in finding prior art at the time of ISR. The positive sign of *ISA_CHANGED* also indicates that prior art is easier to be identified at the time of ISR if applicants from the U.S. or Japan choose the EPO as their ISA (for which robustness is checked in Model 4 and 5). These are consistent with the EPO's good reputation from international applicants. H1 is supported from the positive sign of *euro_cited*. Likewise, H2, H3, H4 and H5 are all supported o this model, except that the number of inventors has an insignificant coefficient.
Model 2 uses applications from Japan only in order to examine the locality of knowledge in Japan. As is expected in H1, *jp_cited* has a positive and significant sign, whereas *us_cited* has negative and significant sign. Other variables show similar results with the Model 1 and are consistent with hypotheses, except *self* indicates the negative sign. Model 3 uses U.S.

applications only, and the results are just consistent with the hypotheses. Model 4 and 5 limit the citation data to non-self citations only for robustness checks, while employing two instrument variables on the variable *ISA_CHANGED*. For Japanese applications, the coefficient for *ISA_changed* lost the significance in the Model 4, suggesting that the advantage provided by the ISA change from JPO to EPO is due to the applicants' self-selection. However, this effect is not observed for the U.S. applications in the Model 5.

Table 1. PROBIT analyses on the probability of ISR coverage; dep. var.=*found_in_ISR.*

Model 4 and 5 use "*total_count*" and "*applicant_avg_cited*" as instruments for "ISA_CHANGED."

****<0.001 ***<0.01 **<0.05 Robust standard errors are in the parentheses (clustering on citing family).

Model & sample	Model 1 (all of triadic samples/ baseline=EP_app)	Model 2 (JP app only)	Model 3 (US app only)	Model 4 (JP app & non-self only)	Model 5(US app & non-self only)
method	Probit	Probit	Probit	IV Probit	IV Probit
euro_cited	0.1419984**** (0.0080393)	-0.031025 (0.0160179)	0.1776262**** (0.0120059)	0.0203394 (0.0174625)	0.148418**** (0.0253879)
us_cited	-0.0620007**** (0.0078305)	-0.3377195**** (0.0155267)	0.050351**** (0.0114757)	-0.2974986**** (0.0169034)	0.0777813**** (0.0159886)
jp_cited	0.0393056**** (0.0082601)	0.8054234**** (0.0151802)	-0.4295359**** (0.0121628)	0.8367819**** (0.0175193)	-0.3751166**** (0.0427623)
fam_cite_lag	0.0030127**** (0.000212)	0.0023379**** (0.0004175)	0.0046464**** (0.000329)	0.0005303 (0.0004425)	0.0026492**** (0.0005495)
self	0.2091817**** (0.0047187)	-0.1759722**** (0.0082345)	0.1123806**** (0.0076398)		
fwd_cite_of_the_cited	0.0000359**** (0.00000321)	-0.00000566 (0.00000781)	0.0000573**** (0.00000437)	-0.00000566 (0.00000799)	0.0000551**** (0.00000526)
IPC4_count	-0.0165033**** (0.0013614)	-0.0176023**** (0.002381)	-0.0215867**** (0.0022476)	-0.0170435**** (0.0026306)	0.0099131 (0.011092)
publn_claims_max_tls211	-0.0080901**** (0.0001942)	-0.0029271**** (0.0003468)	-0.0094453**** (0.0002733)	-0.0033284**** (0.0004149)	-0.0081833**** (0.0010323)
invt_nr	0.0000932 (0.0011831)	-0.0007108 (0.002112)	-0.0058672*** (0.0018111)	0.0008906 (0.0023144)	-0.0089979*** (0.0026535)
family_size	-0.006626**** (0.0007439)	-0.0142835**** (0.0021553)	-0.0053694**** (0.0011327)	-0.0091501*** (0.0032126)	-0.0138593**** (0.002496)
JP_app	-0.0667862**** (0.0069462)				
US_app	-0.2808785**** (0.0072769)				
ISA_CHANGED	0.3096426**** (0.0066579)	0.2758815**** (0.0169662)	0.380766**** (0.0074961)	0.0109491 (0.1314658)	1.35421**** (0.3121653)
Technology class dummies	included	included	included	included	included
n	1031127	325990	455830	264805	363328

Discussion and further development

Overall results are consistent with the hypotheses, suggesting that examiners (and searchers working for the PTOs) are bound by various kinds of "distances," including technological complexity of applications. These are intuitive, and are supported by the novel methodology for the first time. An interesting interpretation is that examiners (unlike inventors) are

required to find prior art by law, but that they are naturally bound by informational horizons they have. This has policy implications, since Patent Prosecution Highways (PPH) rely on outcomes from previous patent offices. Most prior studies using examiner citations do not incorporate these informational obstacles born by examiners, but they cannot be ignored. For example, prior studies on the difference of examination outcomes between patent offices (Jensen et al., 2005; Webster et al., 2007, 2014) do not explicitly consider them, but the cost of prior art search may affect the results. The results with instrument variables suggest the self-selection is working, but is evident for the Japanese samples only. Further scrutiny is needed.

Acknowledgments

This interim output is drawn from the collaborative project with Professor Setsuko Asami (Tokyo University of Science) and Professor Yoshimi Okada (Hitotsubashi University). The entire project is supported by RISTEX/JST. The comparison of search quality between PTOs at aggregated level was previously presented at the International Workshop on Patent System Design for Innovation at Hitotsubashi University (Wada and Asami, 2014) and at the 2014 Annual Conference of the Asia-Pacific Innovation Network. The idea of the probability of ISR coverage of this paper evolved out of the idea of aggregate ISR coverage ratio, which Professor Asami first thought of. The author also acknowledges the support from the MEXT/JSPS (Grant #22330122) for the analyses of citations and firm boundaries.

References

Alcacer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, *88*(4), 774-779.

Du Plessis, M., Van Looy, B., Song, X & Magerman, T. (2009). Data Production Methods for Harmonized Patent Indicators: Assignee sector allocation. EUROSTAT Working Paper and Studies, Luxembourg.

Jaffe, A., & Trajtenberg, M. (1999). International knowledge flows: evidence from patent citations. *Economics of Innovation and New Technology 8*, 105-136.

Jaffe, A., & Trajtenberg, M. & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, *108*, 577-598.

Jensen, P.H., Palangkaraya, A. & Webster, E. (2005). Disharmony in international patent office decisions. *Federal Circuit Bar Journal*, *15*, 679.

Magerman T, Grouwels J., Song X. & Van Looy B. (2009). Data Production Methods for Harmonized Patent Indicators: Patentee Name Harmonization." EUROSTAT Working Paper and Studies.

Peeters B., Song X., Callaert J., Grouwels J., & Van Looy B. (2009). Harmonizing harmonized patentee names: an exploratory assessment of top patentees. EUROSTAT working paper and Studies.

Wada, T., & Asami, S. (2014). Quality comparison of International Search Reports (ISRs) by selectable International Search Authorities (ISAs) under the Patent Cooperation Treaty (PCT) system, *a paper presentation at the 2014 International Workshop on Patent System Design for Innovation,* Hitotsubashi University, Tokyo, Japan.

Webster, E., Jensen, P. H. & Palangkaraya, A. (2014). Patent examination outcomes and the national treatment principle. *The RAND Journal of Economics*, *45*, 449–469.

Webster, E, Palangkaraya, A & Jensen, P.H. (2007). Characteristics of international patent application outcomes. *Economics Letters* 95, 362-368.

World Intellectual Property Organization. (2014). Patent Cooperation Treaty Yearly Review: The International Patent System. *WIPO Economics and Statistics Series*.

A Collective Reasoning on the Automotive Industry: A Patent Co-citation Analysis

Manuel Castriotta and Maria Chiara Di Guardo

{manuel.castriotta, diguardo}@unica.it,
University of Cagliari (Italy)

Abstract
While collective cognition has received increasing attention in the broader field of organization, academic research has largely overlooked its potential role on shaping innovation trajectories and technological change adaptation at a firm and industrial levels. Through a strategic lens and based on the patent bibliometrics and patent co-citation methods, we integrate and extend the cognition and technology strategy literatures by proposing an invention behavior map of leading companies and groups in the automotive industry. How collective cognition influence patent strategies? How economic trends impact on patent paths? Empirical evidence for these reasons is drawn from a longitudinal patent analysis quantitative approach of the period 1991-2013 considered overall and consequently subdivided into three sub periods of seven years each 1991-1997, 1998-2004, 2005-2013. About 443.000 patents, 1.108.356 citations and 1.234.623 co-citations of 49 automotive assignees were collected from Derwent Innovation Index (DII), the largest world patent and innovation database. Multi dimensional scaling and cluster analysis techniques are employed to detect embryonic cognition homogeneity measures and provide an overview of groups technology composition and companies innovation strategies trends. Finally, explorative findings are discussed below with suggestions about how they might be translated into managerial implications.

Conference Topic
Patent Analysis

Introduction
The empirical literature on technological regimes argues that firms within an industry behave in correlated ways because they share sources of information and technology (suppliers, universities, other industries), and perceive similar opportunities for innovation. The existence of a collective cognition shared by firms within a sector can also influence how inventions arise and how quickly and completely they diffuse, and can give us another key to better understand the collective failure of some industries as a result of surprisingly unexpected technological changes, or the innovation trajectories that have characterized some sectors. Yet, while collective cognition has received increasing attention in the broader field of organizational theory (Johnson & Hoopes, 2003; Nadkarni & Narayanan, 2007), research on innovation and patent strategies has been largely silent about the cognition's role (Kaplan, 2011, 2012; Kaplan & Tripsas, 2003, 2008) and empirical studies thus far have not questioned how industry boundaries truly define patent strategies and how economic trends impact on technological trajectories.

To take the first steps at going beyond these limitations and embryonically understand how industry structure and interaction among players can shape technological trajectories, we examine the case of the automotive sector from 1991 to 2013 and identify the dynamic evolution of patent paths among the principal actors in this sector. We chose the automotive sector for several reasons: first, the ability of firms to innovate is crucial to commanding a competitive advantage in this industry (Norhia & Garcia-Pont, 1991); second, all relevant players in this industry must routinely patent their innovations; and third, the automotive market is characterized by high entry barriers able to isolate new entrants and incumbents' dynamic noise.

In order to understand the phenomenon at stake, we analyze the evolution of the technological trajectory in the automotive sector by utilizing bibliometric information such as patent co-citations (Lai & Wu, 2005; Wang, Zhang & Xu, 2011). This approach displays a larger picture of the overall innovation structure and the patent linkages among players and groups' technology positioning, thereby shedding light on the patterns of patent strategies within an industry.

In total, a 21-year period, subdivided as three sets of years in seven-year time spans from 1991 to 1997, 1998 to 2004, and 2005 to 2013, are visualized. About 443.000 patents, 1.108.356 citations and 1.234.623 co-citations of 49 automotive assignees were collected from Derwent Innovation Index (DII), the largest world patent and innovation database. Multidimensional scaling and cluster analysis techniques are employed to detect the embryonic cognition homogeneity measures and to provide an overview of the groups' technology composition and companies' innovation strategy trends.

This study adds to the literature in multiple ways. First, it contributes to the patent literature showing the evolutionary patterns of patent strategies inside a specific industry using patent co-citation analysis. Second, it contributes to innovation literature by enhancing our understanding of how technological firms and group positioning evolve and are influenced by collective cognition. Third, it also contributes to the still-inadequate understanding of the drivers of patent strategies and innovation trajectories.

The paper is organized as follows. In section two, we describe the patent co-citation methodologies employed; in section 3, we present the bibliometric results and provide a graphical representation of firms' and groups' proximities performed by multidimensional scaling (MDS) and cluster analysis; in section 4, we discuss embryonic results and offer some conclusions;

Theoretical background

Bibliometrics and patent citation analysis

Patent citation analysis is an academic set of bibliometric methods directly derived from methodology that seeks to link patents in the same way that science references link papers. Papers and patents are both research instruments that adopt citation-count measurement systems (Narin, 1994). Moreover, in bibliometrics, the use of a citation approach for the assessment of similarity for the classification of documents is a mature methodology, and for this reason, it is feasible to apply the citation analysis of bibliometrics to patent analysis (Zhao & Guan, 2013).

Patent co-citation analysis

Co-citation analysis is a measure of the frequency of how many times A and B units are co-cited by third earlier units such as papers, authors, institutions, and in our study patents, inventors, or assignees (Lai & Wu, 2005; Wang et al., 2011). The assumption of co-citation analysis is that documents that are frequently cited together cover closely related subject matter (Small, 1973; Narin, 1994). In this vein, the co-cited frequency of patents can be used to assess the similarities or relatedness and to post evaluation and less-subjective unobtrusive patent maps and classification systems (Lai & Wu, 2005). In bibliometrics, it is used to assess document similarities in order to analyze the intellectual structure of science studies and identify cluster specialties and sub-fields (McCain, 1990; Di Guardo & Harrigan, 2012; Di Stefano, Gambardella & Verona, 2012).

Methodology

Sample and unit of analysis selection

Our analysis, following the bibliometric co-citation and patent co-citation methods prescriptions (McCain, 1990; Wang et al., 2011; Di Guardo & Harrigan, 2012) and in order to correctly select the unit of analysis started by tracing the history of most relevant M&As and alliances automotive industry milestones. This allow us to consequently identify in Derwent database the standard and non standard assignees codes for the overall and intermediate periods and correctly formulate compound Derwent Innovation Index and Derwent World Patent Index search queries (Wang et al., 2011). We retrieved assignees patent bibliometrics and assignees patent citation counts and finally co-citation frequencies. Operationally, the compilation of the raw co-citation matrix and its conversion to correlation matrix allow us to run multivariate analysis and consequently interpreting the findings. In the case of academic bibliometric studies, the unit of analysis may consist of scientific articles, authors and institutions (Small, 1973). Symmetrically, in the study of citation behavior in the patent analysis, the unit of analysis can be identified by single patents, inventors, institutions or assignees (Lai & Wu, 2005). Our research aims to show the strategic positioning and similarities between the leading automotive companies by displaying and then comparing the entire period of time with three different timespans. For these reasons we adopted assignees as unit of research.

Starting from the OICA 2013 report ranking, we selected the top 80 global companies in the automotive industry of manufacturers based on the number of commercial, passenger, and industrial vehicles produced. We examined the companies' websites and identified the number of brands for each company and its automotive groups. In the Derwent database, we checked individually for brands, single companies and groups, and the number of patents of the application date for the period 1991 to 2013. In this way, we divided the commercial brands by independent enterprises capable of producing technology. Then we looked back across the brands' histories, alliances, and M&As that occurred in the years between 1991 and 2013. In addition, in order to avoid the traditional limitations due to strategic and formal changes in companies and group structures, Derwent provides a comprehensive data set of joint ventures drawn up within industries in the period considered. From the operational point of view and following the correct search strategy proposed by Wang et al. (2011), we did a screening of all potential Derwent codes, including those with a different denomination than the main automotive group, related to joint ventures and M&As. In the research, we took into consideration 14 joint ventures formalized during the period among 18 companies.

Then, we launched an investigation of patent bibliometrics and identified the number of citations of the top 60 car manufacturers. Furthermore, in the hope of exploring the potential effects of the crisis in the strategic positioning of technology groups, we considered these in conjunction with the Asian crisis of 1997 - 98 and just before the start of the crisis of 2007–2008. Moreover, we took into account the M&A histories that showed that in these three periods, the most influential automotive group changes were concentrated. By analyzing the three periods, it was possible to visualize the structural change trends of automotive world industry. Finally, through the multidimensional scaling, a methodology that reduces the complexity and allows the matrices of proximity of certain objects to be studied (Mc Cain, 1990), we displayed the shape and measure the density of automotive sector conformation.

Discussion of results

Patent co-citation

The analysis of co-citations highlights the strategic positioning of the 49 major technological automotive companies in the global market in the period 1991 to 2013, 28 of the main groups in the periods 1991 to 1997 and 1998 to 2004, and finally the 34 major groups between 2005 and 2013. During the full period, the unit of analysis is the single automaker, while in the three time spans it is the automotive group through the extraction of aggregate data. The analysis of the complete map and the trends and changes in technology portfolios in the three time spans, considering the M&A histories and joint ventures, are discussed below through the results of multidimensional scaling and cluster analysis.

MDS and Cluster Analyses

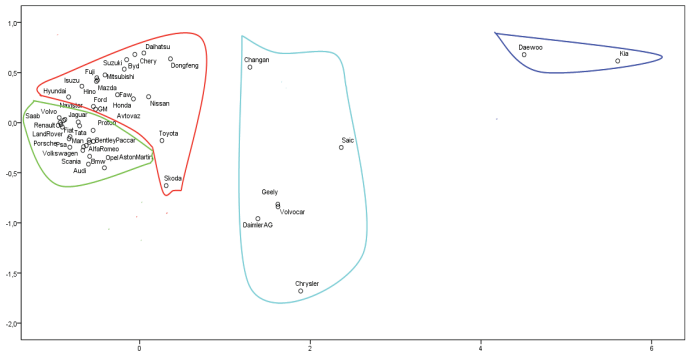


Figure 1. 1991-2013.

On the left of Figure 1 shows an area of high concentration and high technological similarities, while on the right, the distances among firms increase. In this scenario, cluster analysis clearly highlights four groups. The Japanese firms Toyota, Honda, and Nissan are the most central companies and belong to a larger international group comprised of Japanese, Chinese, Korean, and US companies. On the bottom left of the map, European manufacturers emerge, such as Volkswagen, Fiat, Porsche, Renault, BMW, PSA, and MAN, among which are India's Tata and the Soviet Avtovaz and the Malaysian Proton and its Lotus brand. Ford, GM, and Hyundai represent a technological bridge between the two areas. An important peculiarity of some company outliers such as Chrysler, Daimler AG, Geely, Volvo, and Chinese Saic and Dongfeng that belong to cluster 3 is seen, while peripheral positioning is occupied by Daewoo and Kia at the top right.

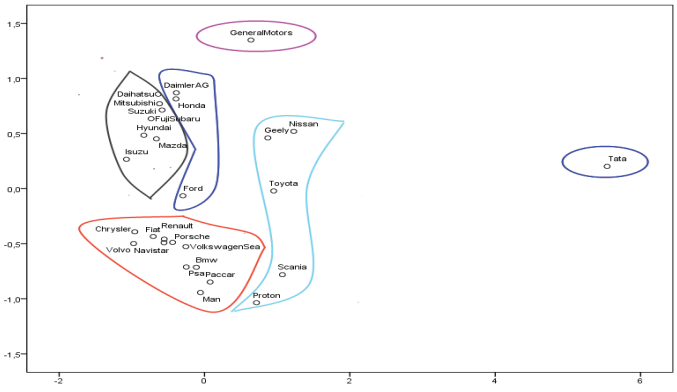


Figure 2. 1991-1997.

Figure 2 shows a major cognition concentration among firms, with the exception of the Indian company Tata on the right side. Ford, Toyota, and Renault are the major groups of centrality. Geely is the only Chinese enterprise present. Cluster analysis clearly shows six groups. General Motors is highly decentralized, a symptom of the uniqueness of its patent portfolio. Daimler and Hyundai are central, positioned in the two groups at the top along with the major Japanese companies, while at the bottom are MAN, Navistar, Volvo, and Paccar, which are all specialized in truck production, just below the European Union automakers. Interesting is the proximity of technology for Fiat and Chrysler, now belonging to the same group, and vice versa, the distance between Toyota and Daihatsu as separate companies at that time and since 1999 part of the same group. Of note is the proximity between Porsche and Volkswagen. Finally, the Volvo Group, at this stage not yet divided between truck and car production, is positioned at the left side near Navistar.

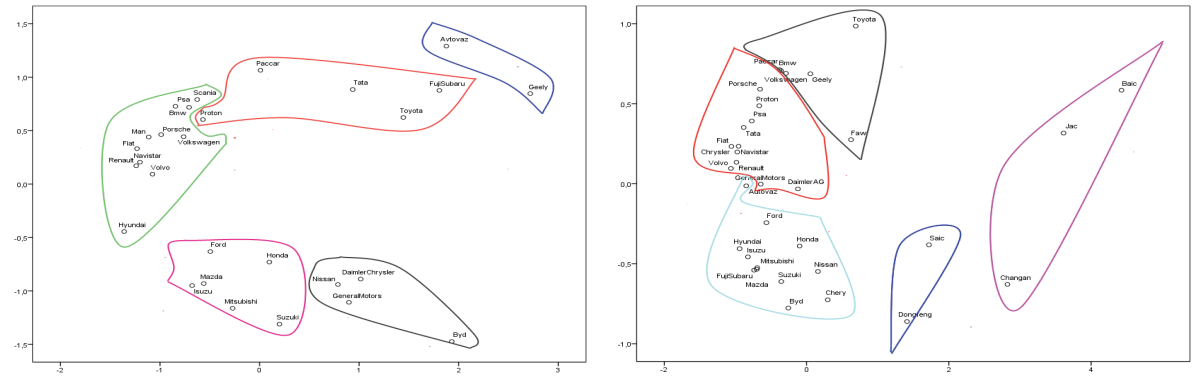


Figure 3. (a) 1998–2004. (b) 2005-2013.

Figure 3(a) transposes the effects of the Asian crisis of 1997-1998 and has a strong dispersion compared to the previous period's technology structures. The distances between companies are larger. To highlight the lack of a technological leader and a high level of technological heterogeneity, the central part of the map is empty.

Figure 3(b) includes the effects of the strong economic performance and global sales of the previous five years to have a stronger concentration symptomatic of technological proximity than in the previous period. During this period, Daimler AG, Ford, and GM occupy the most central locations on the map. General Motors, in particular, takes a decidedly opposite path in the three periods compared to Toyota. The American company tends to centralize its positioning technology, while Toyota tends to move within the confines of the map.

Conclusion and Limitations

This exploratory study increases the awareness of scholars by detecting and visualizing the cognitive structure, operationalized as companies' technological distances, of the automotive sector between 1991 and 2013. It reveals innovation similarities, technology positioning, and trends of assignees and groups, and makes it possible to hypothesize patent strategies and latent relationships among them. A contribution to the patent strategy and cognition literature has emerged on the basis of differences in positioning among companies and groups during the entire period and divided into time spans. In the overall map, this has emerged as some groups are composed of firms with heterogeneous positioning and consequently heterogeneous patent portfolios, while other groups have steadily increased over the years by acquiring high map closeness with companies with similar technological characteristics.

Second, the analysis of the three subdivided periods has highlighted how the level of similarity or distance among the groups, namely the collective cognition, changes continuously. The high concentration level that characterizes the first period is changed in the

second, which is more dispersed and where there are not central or technological leader groups. Yet the third one returns to a concentration level similar to the first period. Such behavior of the map, if considered in relation to the economic performance of the production and sales of the industry, reveals how, in times of crisis, companies tend to look for a heterogeneous technology portfolio to obtain competitive advantages, while in positive economic periods, conformity tends to prevail. It is as if the collective cognition profoundly affects the technology positioning and behavior of firms at the expense of objective assessments of patent strategy decisions. Third, research has highlighted significant strategic differences in positioning in the various periods in which such central enterprises move to the suburbs and vice versa, and some change their technology cluster membership by moving into another and finally emerge or disappear because of a failure or because of an M&A.

Fourth, an explorative contribution originates from the evaluative study of the groups' conformation in terms of brands and partnership formal contracts. In fact, it opens new horizons to researchers who want to analyze the impact of M&As or JVs on technological map positioning and, for example, in Foreign Direct Investments (FDI) and technology strategy literature. Finally, explorative findings of this study might be translated into managerial implications from the point of view of the companies strategic positioning planning. In fact, by detecting the heterogeneous technologies adoption (displayed by the more distant nodes in MDS), manager can potentially create innovative patent recombination strategies and consciously determine innovative future technological positioning scenarios.

References

Di Guardo, M. C., & Harrigan, K. R. (2012). Mapping research on strategic alliances and innovation: a co-citation analysis. *The Journal of Technology Transfer*, *37*(6), 789-811.

Di Stefano, G., Gambardella, A., & Verona, G. (2012). Technology push and demand pull perspectives in innovation studies: Current findings and future research directions. *Research Policy*, 41(8), 1283-1295.

Johnson, D. R., & Hoopes, D. G. (2003). Managerial cognition, sunk costs, and the evolution of industry structure. *Strategic Management Journal*, *24*(10), 1057-1068.

Kaplan, D. M. (2012). How to demarcate the boundaries of cognition. *Biology & Philosophy*, *27*(4), 545-570.

Kaplan, S. (2011). Research in cognition and strategy: reflections on two decades of progress and a look to the future. *Journal of Management Studies*, *48*(3), 665-695.

Kaplan, S., & Tripsas, M. (2003). Thinking about technology: understanding the role of cognition and technical change. Division of Research, Harvard Business School.

Kaplan, S., & Tripsas, M. (2008). Thinking about technology: Applying a cognitive lens to technical change. *Research Policy*, *37*(5), 790-805.

Lai, K. K., & Wu, S. J. (2005). Using the patent co-citation approach to establish a new patent classification system. *Information Processing & Management, 2*, 313-330.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American society for information science*, *41*(6), 433-443.

Nadkarni, S., & Narayanan, V. K. (2007). Strategic schemas, strategic flexibility, and firm performance: the moderating role of industry clockspeed. *Strategic management journal*, *28*(3), 243-270.

Narin, F. (1994). Patent bibliometrics. *Scientometrics*, 30(1), 147-155.

Nohria, N., & Garcia-Pont, C. (1991). Global strategic linkages and industry structure. *Strategic management journal*, *12*(S1), 105-124.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, *24*(4), 265-269.

Wang, X., Zhang, X., & Xu, S. (2011). Patent co-citation networks of Fortune 500 companies. *Scientometrics*, *88*(3), 761-770.

Zhao, Q., & Guan, J. (2013). Love dynamics between science and technology: some evidences in nanoscience and nanotechnology. *Scientometrics*, *94*(1), 113-132.

Statistical Study of Patents Filed in Global Nano Photonic Technology

Zhang Huijing[1], Zhong Yongheng and Jiang Hong

[1] *zhanghj@mail.whlib.ac.cn*

Wuhan Documentation and Information Centre, Chinese Academy of Sciences, West 25, Xiaohongshan, Wuhan, Hubei, PR, 430071 (China)

Key words

nano photovoltaic technology (NPT); patent analysis; review; photoelectron device; semiconductor material; industry layout

Introduction

As one of leading core technology in the 21th century, nano photonic technology (NPT) is highly interdisciplinary, involving physics, chemistry, biology, materials science, and the full range of the engineering disciplines (Picraux, 2014). NPT is a study of the interaction of electrons and photons and its components in nano structure based on the great development and popularization of nanometre semiconductor materials (Liu, 2005). In 2011, NPT was identified as one of Key Enabling Technologies (KETs) for its vital role in strengthening Europe's industrial and innovation capacity (European Commission, 2011). It is widely used in telecommunications, optical interconnects, display, lighting, photovoltaic, sensors, data storage, imaging, and testing, etc (AIRI/Nanotec IT, 2008).

Patent analysis, which involves statistical, analytical, and comparative methods for examining information in patent documents, has been widely applied in studies examining R&D capacity, technological fields, industrial departments, and company levels (Pavit, 1988). Careful analysis of NPT-related patents can assist in elucidating technological details and relationships, identifying business trends, inspiring novel industrial solutions, and developing investment policies. Therefore, this study performed a statistical analysis of patent data to explore the technological developments of NPT. The technology life cycle and regional distribution of the patents were studied, and the top ten patent assignees were also explored.

Methodology

The searching for NPT patents from the Derwent World Patent Index (DII) database, keywords search were performed for the term appearing in titles, abstracts, or claims. The search strategy of DII database based on NPT was as follows: TS=(((solar or photovoltaic or "optoelectronic integrated device" or OEIC or "optic switch" or "holographic memory" or "light amplifier" or "optical amplifier" or ROADM or "optical add-drop multiplexer" or "optoelectronic display") and nano) or (optoelect* and (semiconductor or GaAs or "gallium arsenide") and nano) or (("quantum well" or "quantum wire" or "quantum dot") and (laser or "photoelectric effect")) or "micronano laser" or "nano laser" or Nanophot* or "Nanowire laser" or "Uv nm laser" or "microcavity laser" or (nano same LED) or (nano same "light emitting diode")). After querying, filtering, and organizing the search results, 8168 NTP-related patents were obtained on December 12, 2014, and the data were analyzed using Thomson data analyzer (TDA).

Results and discussion

Figure 1 showed the evolution of the number of patents relative to the assignees, which is a typical value for exploring the technology life cycle base on patent data. It was showed that the number of patents and assignees increased gradually before 2000, indicating that the technology life cycle was in the introductory stage. This trend implied that few manufactures and institutions were investing in the R&D of NPT before 2000. By contrast, the number of patents and assignees increased rapidly after 2000, particularly during the 2007-2013 periods, indicating that the technology had entered the growth stage. Specifically, the number of patents (assignees) increased from 378 (558) in 2007 to 1006 (843) in 2013.

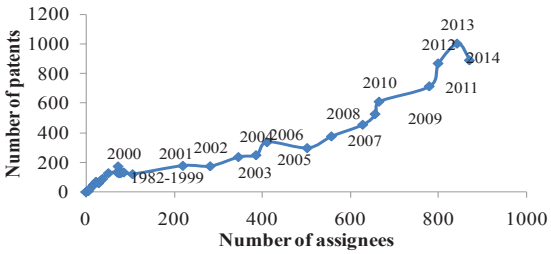


Figure 1. Technology life cycle

Figure 2 showed the number of patents filed in various countries/offices, as well as the trend of the number of patent applications. China (CN), Japan (JP), United State (US), WIPO (WO), and Korea (KR) were the top five countries/offices, with the number of patent applications of 2133, 1964, 1946, 970 and 656. The number of patent applications filed in CN was the highest, indicating that the NPT market in CN might offer the most potential for future development. Compared with other countries, the filing of NPT-related patents commenced only recently in CN, although the number of patent applications increased markedly in 2004-2014.

Moreover, the NPT-related patents were filed earliest in US and WO, and the number of patent applications of these two countries grew rapidly since the beginning of 2004.

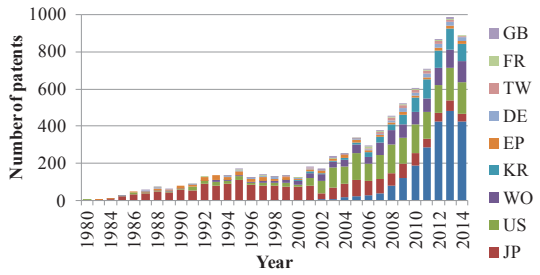


Figure 2. Number of patents and its evolution by country/office. The initialisms "WO" and "EP" indicate that the patent was filed in the WIPO and EPO, respectively.

Table 1 showed a summary of the top ten patent assignees. It was found that all of top ten patent assignees were from JP except *Semiconductors Institute of Chinese Academy of Sciences* and *Samsung Electronics Company, Limited*. In addition, the JP assignees were all companies, and these JP companies had already manufactured commercial NPT products. Furthermore, the JP and KR assignees were filed their patents in many countries/offices for the global layout of NPT. By contrast, *Semiconductors Institute of Chinese Academy of Sciences* filed patents only in CN.

Table 1. Top 10 patent assignees.

Assignee (nationality)	No. of patents	No. of application countries	Times cited (average)
NEC Corporation (JP)	280	4	425 (1.5)
Mitsubishi Denki K.K. (JP)	188	7	402 (2.1)
Fujitsu Limited (JP)	179	5	210 (1.2)
Sharp KK (JP)	170	6	430 (2.5)
Hitachi Limited (JP)	156	4	187 (1.2)
Samsung Electronics Company, Limited (KR)	153	6	137 (0.9)
Semiconductors Institute of Chinese Academy of Sciences (CN)	143	1	71 (0.5)
Furukawa Electric Company, Limited (JP)	138	6	423 (3.1)
Nippon Telegraph & Telephone Corporation (JP)	132	3	30 (0.2)
Matsushita Denki Sangyo KK (JP)	115	5	286 (2.5)

Conclusion

This study analyzed patent data to explore the technological developments of NTP. After querying, filtering, and organizing the search results, this study analyzed 8168 NTP-related patents. The primary findings of this study were detailed as follows.

(1) Based on the analysis results, the technology life-cycle status of the NPT is currently in the growth stage, indicating that many products were sufficiently developed for commercialization.

(2) US assignees were the most prominent assignees, although the most patent applications were filed in CN, indicating that the market for NPT in CN might offer the most potential for future development.

(3) All of the top ten assignees were from JP, KR, or CN. The JP and KR assignees were all companies, and the assignees were filed their patents in many countries/offices for the global layout of NPT and products. By contrast, *Semiconductors Institute of Chinese Academy of Sciences* is academic institution and filed patents only in CN.

Future studies should consider evaluating the current state of NPT developments in a specific field to identify application areas for new patents.

Acknowledgments

We deeply appreciate the financial supports to this research from Youth Innovation Promotion Association of Chinese academy of sciences.

References

AIRI/Nanotec IT (2008). Roadmmaps at 2015 on nanotechnology application in the sectors of: materials, health & medical systems, energy. Retrieved December 12, 2014 from: http://www.iva.se/upload/Verksamhet/Projekt/Nano/internetionellt/EU%20Nano%20Roadmaps%20SYNTHESIS.pdf.

European Commission (2011). High-Level Expert Group on Key Enabling Technologies. Retrieved December 12, 2014 from: http://ec.europa.eu/enterprise/sectors/ict/key_technologies/kets_high_level_group_en.htm.

Liu W. L. (2005). New advancement and developmental trend of nanometer optoelectronic devices. Sensor World, 11, 6-9.

Pavitt K. (1988). Uses and abuses of patent statistics. In Van Raan AFJ (Ed.), Handbook of quantitative studies of science and technology (pp: 509-536). Amsterdam: Elsevier Press.

Picraux S.T. (2014). Nanotechnology. Retrieved December 12, 2014 from: http://global.britannica.com/EBchecked/topic/962484/nanotechnology.

A SAO-based Approach for Technology Evolution Analysis Using Patent Information: A Case Study on Graphene Sensor

Zhengyin Hu[1,2] and Shu Fang[1]

huzy@clas.ac.cn
[1]Chengdu Documentation and Information Center, Chinese Academy of Sciences,
No.16, Nan'erduan, Yihuan Road, Chengdu (China)
[2]University of Chinese Academy of Sciences,
No.19A Yuquan Road, Beijing (China)

[1] *fangsh@clas.ac.cn*
[1]Chengdu Documentation and Information Center, Chinese Academy of Sciences,
No.16, Nan'erduan, Yihuan Road, Chengdu (China)

Introduction

The Subject-Action-Object (SAO) structures are composed of Subject (noun phrase), Action (verb phrase) and Object (noun phrase), which can represent technology information with more details in a simple manner and have been widely applied in patent text mining (Cascini, Lueehesi, & Rissone, 2001; Sungchul et al., 2012; Zhang et al., 2014a). This paper presents an approach for technology evolution analysis based on SAO. SAO structures are extracted and cleaned from patent text. The technology information of patents such as problems, solutions, functions and effects are stated by SAO. By calculating the distributions of problems over solution groups, a technology evolution map of problems can be drawn. Graphene sensor patents are selected as a case study.

Methodology

Extracting SAO Structures

After collecting patents, some national language processing (NLP) tools are used to extract raw SAO structures from patent text fields. Normally, the fields such as "Title" and "Abstract" are precise and meaningful for NLP (Sungchul et al., 2012).

Cleaning SAO Structures

The number of raw SAO structures is huge and they need to be cleaned. Text mining tools and domain thesauri are used to carry out Subject and Object cleaning by following a term clumping framework (Zhang, et al., 2014b). The verb phrases of Action are normalized and categorized by experts.

Tagging SAO Structures

According to a classification model learned from a training data, the cleaned SAO structures are tagged with 4 kinds of labels of *problem*, *solution*, *function* and *effect*.

Clustering SAO of Solution

After tagging the semantic type of each SAO, those with *solution* label are clustered into different *solution* groups. Each solution group with similar SAO can be considered as a *solution* topic.

Drawing technology evolution map of problems

Kim, Suh and Park (2008) approached a method that can be used to draw technology evolution map of keywords by calculating the distributions of keywords over the keyword cluster groups. We draw technology evolution map of problems based on Kim, Suh and Park's (2008) research. Firstly, we calculate the distributions of problems over the *solution* groups. If the co-occurrence frequency of two problems is above a threshold, we draw a directed line segment between them to show their relevance. Then the occurrence frequency of each problem in *solution* groups is counted. Finally, by adding the earliest filling date of each problem, a technology evolution map of problems with horizontal axis of timeline and vertical axis of frequency can be drawn.

Case Study

Extracting SAO Structures

We selected Derwent Innovations Index (DII) as data source and invited experts to determine the patent retrieval strategy for graphene sensor patents. After eliminating irrelevant patents, we got 196 patents. We extracted raw SAO from the "Title" and "Abstract" fields and got 4,823 raw SAO structures using an NLP tool named ReVerb (Anthony, Stephen & Oren, 2011).

Cleaning SAO Structures

We cleaned Subject and Object by using a commercial text mining tool, VantagePoint (Nils, 2011) and domain thesauri. We followed the term clumping framework to clean them, which includes general cleaning, terms pruning and terms

consolidating processes. After term clumping, we got 628 terms of Subject and Object. We normalized and categorized the verb phrases of Action based on a rule table made by experts. After the cleaning steps, we got 2250 SAO structures.

Tagging SAO Structures

We chose 167 SAO structures from 20 patents as a training set. We picked up Subject, Action as the classification features and C4.5 decision tree as the classifying algorithm to build a classification model which helps to categorize SAO to 4 classes of *problem*, *solution*, *function* and *effect*. Among the classified SAO structures, there are 208 tagged with *problem* label, 746 with *solution* label, 824 with *function* label and 472 with *effect* label. A sample of SAO is shown in table 1.

Clustering SAO of Solution

We clustered the SAO structures with *solution* label into *solution* groups using *k*-means algorithm. By comparing the cluster results, we set the *k*-value 20 and got 20 *solution* groups.

Drawing technology evolution map of problems

By calculating the distributions of problems over each *solution* group, a technology evolution map of problems in graphene sensor patents was drawn. A part of the map is shown in Figure 1.

Table 1. A sample of SAO after tagging.

Type	Subject	Action	Object
Problem	method	synthetize	graphene oxide
Solution	method	use	ultrasonic oscillation process
Solution	graphite powder	mixed with	sodium nitrate
Function	graphene oxide	used for	thin film transistor

Conclusions

The technologies in the upper left corner of Figure 1 appeared in many different *solution* groups and were applied for patents in earlier time, which can be considered as the basic problems in graphene sensor, such as ***producing carbon nanotube***, ***synthetizing graphene oxide***, etc. The technologies in the lower right corner of Figure 1 appeared in fewer *solution* groups and were applied for patents lately, which can be considered as the latest technologies or emerging technologies, such as ***manufacturing sensor array***, ***detecting nucleic acid***, etc**.**

We can draw a technology evolution map of solution, function or effect by following a similar process. The separate technology evolution maps of problem, solution, function and effect can be combined to a more comprehensive technology evolution map of graphene sensor. This study is ongoing.

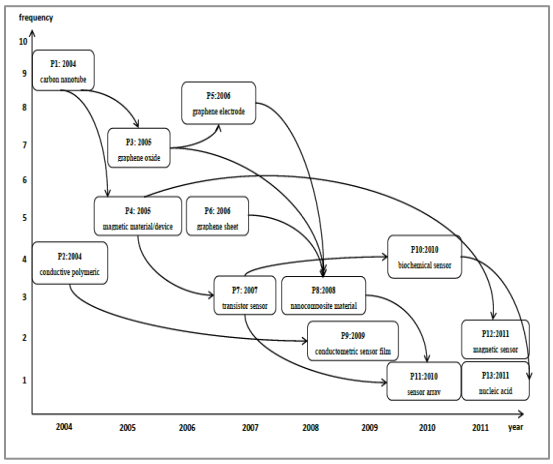


Figure 1. A part of technologies evolution map of problems in graphene sensor patents.

References

Anthony, F., Stephen, S., & Oren E. (2011). *Identifying Relations for Open Information Extraction*. Retrieved March 2, 2014 from: http://ai.cs.washington.edu/www/media/papers/reverb.pdf.

Cascini, G., Lucchhesi, D. & Rissone, P. (2001). Automatic patents functional analysis through semantic processing. *The 12th ADM International Conference*. Rimini, Italy.

Nils N. (2011). *VantagePoint*. Retrieved April 24, 2015 from: https://www.thevantagepoint.com/data/documents/VP%20INTRO%202011.pdf.

Sungchul, C., Hyunseok, P., Dongwoo, K., Lee, J.Y., & Kim, K. (2012). An SAO-based text mining approach to building a technology tree for technology planning. *Expert Systems with Application*, *39*, 11443-11455.

Kim, Y.G., Suh, J.H. & Park, S.C. (2008). Visualization of patent analysis for emerging technology. *Expert Systems with Applications*, *34*, 1804–1812.

Zhang, Y., Porter, A. L., Hu, Z., Guo, N., & Newman, N.C. (2014b). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells. *Technological Forecasting and Social Change*, *85*, 26-39.

Zhang, Y., Zhou, X., Porter, A. L., & Gomila, J. (2014a). How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: The semantic TRIZ tool and case study. *Scientometrics, 101*(2), 1375-1389.

Prediction of Potential Market Value Using Patent Citation Index

HeeChel Kim[1,2], Hong-Woo Chun[2], Byoung-Youl Coh[2]

{kim, hw.chun, cohby}@kisti.re.kr

[1]University of Science and Technology, 305-350, 217 Gajeong-ro, Yuseong-gu, Deajeon(South Korea)

[2]Korea Institute of Science and Technology Information, Dept. Of Technology Intelligence Research, 130-741, 66 Hoegiro, Dongdaemun-gu, Seoul (South Korea)

Introduction

Patent statistics have frequently been used as both technological and economic indicators, however, in order to fully utilize patent data in economic analyses, we must link patents to economic activity at a level of industry or product.

Many previous pieces of research showed the effectiveness of patents citation index (PCI), containing annual citation information, on economic indicators of respective firms. Hall et al. (2005) have studied the relation between a market value and PCI using the Tobin's q approach, and Patel and Ward (2011) have compared the stock market value of firms with the patent citation using the event study methodologies. Both studies showed that Patent statistics can be effectively used to micro-level economic analyses and the increase of PCI has the positive effect on the corresponding market value.

Meanwhile, our study aims to prove the effectiveness of PCI on the economic value of industry, so-called Meso-level study and, in this case, it is essential to develop technology-industry concordance method.

Method

The correlation analysis between Potential Market value (PMV) and PCI for the respective industry is carried out in three stages.

(1) Data concordance process. The market data was collected from Annual Survey of Manufactures (ASM) [1] in the US Census Bureau (http://www.census.gov) and PCI [2] data was collected from the patent set registered USPTO.

Next, we created an annual concordance matrix of IPC (international patent classification) 4-digit to NAICS (North American industry classification system) 6-digit (rev.2002, 2007, and 2012) by Algorithmic Links with Probabilities (ALP), ALP (Lybbert & Zolas, 2013), concordance method of the WIPO (http://www.wipo.int/). ALP is the most up-to-date method compared with those of YTC (Kortum & Putnam, 1997), OECD (Johnson, 2002) and DG (Schmoch et al., 2003).

Each IPC 4-digit is connected to multiple NAICS 6-digit probabilistically via a text mining-based matching rule.

PMV was calculated by model 1 as follows, and consequently, 593 annual pairs of PMV-PCI for each IPC were generated.

$$PMV_{ij} = \frac{\sum_{k=1}^{476} a_{ijk} \times b_{ik}}{\sum_{i=1}^{593} \sum_{k=1}^{476} a_{ijk} \times b_{ik}} \times \sum_{k=1}^{476} b_{ik} \quad \text{..... Model 1.}$$

a = Probability of IPC 4-digit to NAICS 6-digit
b = Value of shipment by NAICS in ASM
i = Year (2002 to 2013)
j = IPC 4-digit code (A01G, A01H, …, H05K)
k = NAICS 6-digit code (311111, 311119, …, 339999)

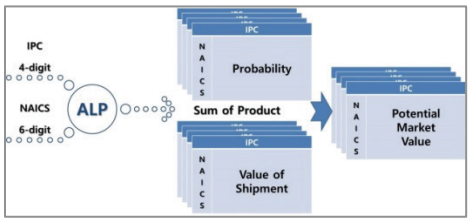


Figure 1. Process of IPC-NAICS Concordance and PMV Calculation.

(2) Statistical correlation analyses for all industry fields. We performed a statistical correlation analysis between the annual incremental of PMV and PCI. We used the Spearman's rho correlation analysis, a nonparametric correlation analysis algorithm, useful to calculate the correlation between the ranked variables (IBM, http://k:5172/help/index.jsp?topic=/com.ibm.spss.statistics.tut/introtut2.htm).

(3) Statistical correlation analyses for 4 major industry fields. The correlation analyses between the annual incremental of PMV and PCI for 4 major industry fields - electrical engineering, instruments, chemistry, and mechanical engineering – were also performed.

Result

Figure 2 shows annual trends of PMV, PCI, and Patent registered. All kinds of variables are trending upward in an accelerating degree.

[1]ASM is estimated sample statistics issued annually for more than one people employees firms in the manufacturing sector. ASM is classified industries by NAICS. In this study, using field of the value of shipment at the 2004 and 2006 edition of ASM that follow the revised NAICS 04 and 2008 to 2011 edition of ASM that follow the revised NAICS 07.

[2]PCI data was used granted patent of USPTO. During the year of from 2002 to 2013.

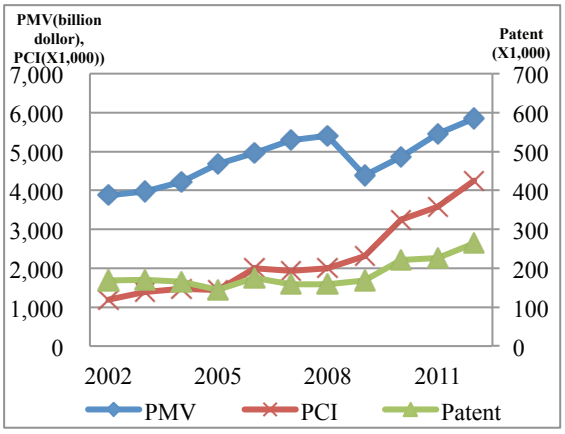


Figure 2. Structure of PMV, PCI and Patent.

PMV of each IPC

Table 1 shows the result of the PMV of each IPC calculated from model 1. It has a significant meaning that a set of patents can be expressed to market value.

Table 1. PMV (unit: million US$).

No.	IPC	2002	2003	…	2013
1	A01G	282	301	…	229
2	A01H	3,057	3,831	…	15,227
...	…	…	…	…	…
593	H05K	6,556	6,166	…	5,055

Correlation Analyses

In the analysis results over the entire industry fields (Table 2), we could find out that significance of correlation and direction varies depending on the Lagging time (differences in data collection year between PMV and PCI). It has a relatively weak positive correlation when the lagging time is 0, meanwhile, it showed relatively strong negative correlation when the lagging time is "PCI+1" – the data collection year for PCI is one year after to that of PMV - . And in case of the lagging time of "PCI-1", it has relatively strong positive correlation, which reveals patent citation activity's positive relation to the corresponding market value "one year later".

Table 2. Results of PMV-PCI rate's correlation analyses (all fields, **significance level 0.01).

Lagging time(year)	Correlation coefficient	p-value (two-tailed)	N
PCI-1	0.136**	0.000	5337
0	0.093**	0.000	5930
PCI+1	-0.323**	0.000	5337

The analyses results of 4 major industry fields showed similar tendencies to all-field-analysis except electrical engineering field.

Table 3. Results of PMV-PCI rate's correlation analyses (4 major fields, **significance level 0.01).

Field	Lagging time(year)	Correlation coefficient	p-value (two-tailed)
Electronic	PCI-1	-0.013	0.747
	0	0.143**	0.000
	PCI+1	-0.513**	0.000
Instrument	PCI-1	0.209**	0.000
	0	0.011	0.795
	PCI+1	-0.360**	0.000
Chemistry	PCI-1	0.180**	0.000
	0	0.022	0.434
	PCI+1	-0.265**	0.000
Mechanic	PCI-1	0.167**	0.000
	0	0.123**	0.000
	PCI+1	-0.266**	0.000

Conclusion

In this research, we made a systematic way for describing the technological impact on industry sector by using some indices, which has a significant meaning that a set of patents can be expressed to market value. We also had confirmed the potential of PCI to predict PMV of the industry. Experimental results showed that PMV in all industry fields was related by the corresponding field's patent-citation activity in one year before or after. After this work, we will deal with enhanced concordance approach to find out relationships between IPC 7-digit and NAICS 7-digit. Also, the self-citation ratio of patent-citation activity may affect economic activity at a level of industry or product, which is now on a study.

References

Hall, B. H., et al. (2005). Market value and patent citations. *RAND Journal of Economics*, 16-38.

Johnson, D., March (2002). The OECD Technology Concordance (OTC): Patents by Industry of Manufacture and Sector of Use, OECD Science, Technology and Industry Working Papers.

Kortum, S. & Putnam, J. (1997). Assigning patents to industries: tests of the Yale technology concordance. *Economic Systems Research*, *9*(2), 161-176.

Lybbert, T.J. & Zolas, N.J. (2014). Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity. *Research Policy*, *43*(3), 530-542.

Patel, D. & Ward, M.R. (2011). Using patent citation patterns to infer innovation market competition. *Research Policy*, *40*(6), 886-894.

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003). Linking technology areas to industrial sectors. *Final Report to the European Commission, DG Research*, *1*.

Knowledge Flows and Delays in the Pharmaceutical Innovation System

Mari Jibu[1], Yoshiyuki Osabe[2], and Katy Börner[3]
[1] OECD, Paris (France) and Japan Science and Technology Agency, Tokyo (Japan)
[2]Japan Patent Office, Ministry of Economy, Trade and Industry, Tokyo (Japan)

[3] *katy@indiana.edu,*
CNS, ILS, SOIC & IUNI, Indiana University, Bloomington, IN (USA)

Introduction

This paper presents an analysis of knowledge flows in the pharmaceutical innovation process. Backward citations, citations to non-patent literature (NPL), and forward citations that link patents, scientific publications, and pharmaceutical pipelines data on drug developments are analyzed and visualized to provide a more holistic understanding. Results show that patents linked to drugs tend to be technically specialized when compared to patents without linkages to drugs. Moreover, patents linked to drugs tend to cite older patents and scientific publications and impact wider technological and scientific fields than pharmaceutical patents not linked to drugs.

Diverse studies have been conducted to study the origin, trajectory, and destination of knowledge flows and the delays in the science and technology system. Patents and citations between patents and to non-patent literature (NPL) are analyzed to understand knowledge spillovers (Lukach & Plasmans, 2002) or to measure patent quality (Squicciarni et al., 2013). The OECD Science, Technology and Industry Scoreboard 2013 (OECD, 2013) uses comprehensive and up-to-date data to report on knowledge flows via collaboration networks (e.g., derived from co-authored publications and co-inventors on patents), international migration of researchers (e.g., estimated from changes in author's addresses on publications), but also flows of royalty and license fees for technologies. Recently, the OECD introduced a new indicator, called "Patent-Science Link," that aims to measure knowledge flows between the science base and the innovation system (OECD, 2013). According to this new indicator, patented pharmaceutical inventions account for the majority of citations made from patents to scientific publications. That is, the distance between the science base and the innovation system is much closer in pharmaceutical fields than it is in other technological fields. Pharmaceutical innovation is particularly important for drug discovery, as research and development (R&D) costs are huge and major challenges exist for arriving at cost-effective new drugs. In fact, there is a steady decrease in R&D productivity over the last number of years (Booth & Zemmel, 2004).

The structure of the paper is as follows: The next Section details data acquisition and preparation. This is followed by a description of the methodology and results. The paper concludes with a discussion of key insights and their comparison to prior work.

Data Acquisition and Preparation

Five datasets by Thomson Reuters covering 1981 to 2011 are used in this analysis. (1) Publication data from the *Web of Science (WoS)* database. (2) Patent data from the *Derwent World Patents Index (DWPI)* and associated citations from the (3) *Derwent Patents Citation Index (DPCI)*. (4) Linkages between publications and patents come from the *WoS-DPCI Linktable* computed by Thomson Reuters and JST that provides information on backward citations from patents and to the non-patent literature (NPL), i.e., scholarly publications, derived from the *DPCI*. (5) Drug pipeline data was retrieved from the *Cortellis for Competitive Intelligence* database including detailed information of exactly drugs a patent is associated with. Data was compiled on December 11, 2013.

Interested to identify patents and their linkages to the NPL in pharmaceutical fields, we extracted all 833,376 patents with the International Patent Classification (IPC) code "A61P: Specific therapeutic activity of chemical compounds or medicinal preparations" from the *DWPI* with their citations from *DPCI*, called "Pharma_Patents." Then, we extracted 57,800 patents linked to pipeline data from the *Cortellis for Competitive Intelligence* database, called "Drug_Patents." Next, the *Drug-Patents* were subtracted from the *A61P-Patents* resulting in a dataset of 325,576 "Non-Drug Pharma Patents" that have the A61P code but are not linked to drugs.

Finally, all 115,252 NPL for *Drug_Patents (DP)* and 718,269 *Non-Drug_Pharma_Patents (NDPP)* were retrieved using the *WoS-DPCI Linktable*.

Methodology

Four metrics were computed: (1) *citation lag*; (2) *generality index* computing the diversity of patents that are cited by a given focal patent as well as the diversity of patents that are citing the focal patent;

(3) *subject index*, a new indicator based on the generality index but computed for NPL; (4) *patent scope*, often associated with the technological and economic value of patents with broad scope patents having a higher value (Lerner, 1994).

Results

Using the four metrics, a number of novel results can be computed.

Technology Delays: Citation Lag

Comparing citation lag data for *DP* and *NDPP* reveals the temporal dynamics of knowledge flows. Table 1 shows that forward citations from *NDPP* come from patents that were published on average 2.17 years later while *DP* are cited faster—after 1.89 years on average. Backward citations from *NDPP* go to patents that were published on average 3.4 years earlier and they go to much more recent NPL—published only 1.69 years earlier on average. Interestingly, *DP* cite older works than *NDPP*: Cited patents are 5.64 years old and cited NPL are 2.5 years old on average. All values are statistically significant at the 1% level. In sum, they show that *DP* cover larger temporal ranges and are cited more quickly than *NDPP*.

Table 1. Forward and Backward Citation Lags.

	NDPP	DP
Forward Cites by Patents	2.17	1.89
Backward Cites to Patents	3.40	5.64
Backward Cites to NPL	1.69	2.50

Technology Diversity: Generality & Subject Index

The generality index was calculated for 4- and 6-digit IPCs for forward and backward citations for *NDPP* and *DP*, see Table 2. *DPs* have higher generality index and subject index than *NDPP*. That is, on average, *DP* draw on more diverse technology "base knowledge" and are cited by a more diverse set of patents that have more varied IPCs. All values are statistically significant at the 1% level.

Table 2. Generality Index for Forward Citations (FC) and Backward Citations (BC).

		NDPP	DP
Generality Index (4-Digits)	FC	0.36	0.37
	BC	0.40	0.54
Generality Index (6-Digits)	FC	0.46	0.50
	BC	0.52	0.73
Subject Index	BC to NPL	0.22	0.28

Technology Value: Scope

The patent scope was computed for *NDPP* and *DP*, see Table 3. The scope of *DP* is lower than that of *NDPP*. This is unexpected as patents linked to drugs are presumably more valuable than those not linked to drugs.

Table 3. Scope.

	NDPP	DP
Scope (4-Digits)	0.13	0.11
Scope (6 Digits)	0.16	0.15

Conclusions

This paper compared and contrasted patents that are linked or not linked to drugs to understand knowledge flows and delays in pharmaceutical innovation. The results indicate that *Drug_Patents* draw from a more diverse set of technologies and are cited more widely across the technology landscape. However, they tend to be more technically specialized (lower scope) than *Non-Drug_Pharma_Patents*. Concerning citation lag, *Drug_Patents* tend to refer to older patents and scientific publications and are cited faster than *Non-Drug_Pharma_Patents*.

In our prior work, we introduced new drug-patent indicators for identifying patents related with pharmaceutical entities' R&D progress (Jibu & Osabe, 2014) and that IPC count, forward citations, and citations to NPL are efficient drug-patent-indicators. The work presented here is novel is that it shows that citation lags and the generality of backward citations are statically significantly different for *Non-Drug_Pharma_Patents* and *Drug_Patents*.

Acknowledgments

We would like to thank Fernando Galindo-Rueda for his expert comments and support of this research. This work was partially funded by the National Institutes of Health under awards P01AG039347, U01GM098959, and U01CA198934.

References

Booth, B., & Zemmel. R. (2004). Prospects for Productivity. *Nature Reviews Drug Discovery 3*, 451-456.

Jibu, M. & Osabe, Y. (2014). Refined R&D Indicators for Pharmaceutical Industry. *Future Information Technology, Lecture Notes in Electrical Engineering*, *309*, 549-554.

Lerner, J. (1994). The Importance of Patent Scope: An Empirical Analysis. *The RAND Journal of Economics*. 25(2), 319-333.

Lukach, R. & Plasmans. J. (2002). Measuring knowledge spillovers using patent citations: evidence from the Belgian firm's data. CESifo Working Paper NO.754 Category 9: Industrial organization,

OECD, (2013). *OECD Science, Technology and Industry Scoreboard 2013: Innovation for Growth.* Paris, France: OECD Publishing.

Squicciarni, M., Dernis, H. & Criscuolo, C. (2013). Measuring Patent Quality: Indicators of Technological and Economic Value. *OECD/DSTI/DOC* 3.

2015

THEORY

METHODS AND TECHNIQUES

Can Numbers of Publications on a Specific Topic Observe the Research Trend of This Topic: A Case Study of the Biomarker HER-2?

Yuxian Liu[1,2,3], Michael Hopkins[2] and Yishan Wu[4]

yxliu@tongji.edu.cn, m.m.hopkins@sussex.ac.uk, wuyishan@istic.ac.cn

[1]Tongji University, Tongji University Library, Siping Street 1239, 200092 Shanghai (China);

[2]Tongji University Development & Planning Research Center, Siping Street 1239, 200092 (China);

[3]Univ. of Sussex, School of Business, Management and Economics, SPRU, Falmer BN1 9SL, Brighton (UK)

[4]Institute of Scientific and Technical Information of China, 15 Fuxinglu, Beijing 100038 (China)

Abstract

Using the accumulative publication data on HER-2 and its trend line, we draw the accumulative curve of the publication data. We discuss the characteristics of the accumulative publication curve, and how these characteristics change with respect to the different trend lines. We find that the points that regression line and the publication curve intersect with each other and the minimum points with respect to the trend lines do not change very much in both exponential trend line and linear trend line even if the exponential trend line raises itself much faster than the linear trend line. These data points are formed around the time when the significant discoveries are made and the related regulations are executed. These significant discoveries and regulations impact how and where the research should go and how the basic discoveries influence their application. The accumulative publication curve itself tells us very little about science. However the change of the accumulative publication curve with respect to the trend lines may tell us how science evolves. The content in the publications with significant scientific value may change the direction and trend of research, while research may change the publication trend the other way round. We may say that important scientific discoveries and regulations on clinical practice act as tipping points or act as drivers of change in the rates of scientific publications on the topic of HER-2. This induces us further to explore how scientific events drive the publication process. We may expect that through the publication process, we can monitor the scientific process.

Conference Topic

Theory

Introduction

The number of publications is widely used to measure the output or the productivity of researchers or their affiliated institutes. Hence, it is also used to compare the output of different countries (Bornmann & Marx, 2013; Zhu et al., 2004; Inglesi-Lotz & Pouris, 2011; Garfield, Pudovkin, & Paris, 2010). (China is ranked the second in terms of output of scientific research measured by the number of publications.) It is normally regarded as a quantitative indicator. The number of citations is supposed to measure the impact or the visibility of the researchers or their affiliated institutes that are investigated (Garfield, 1955). Sometimes it is even referred to as the indicator that measures the quality of the research in the cited article that a researcher has performed.

However, these measurements arouse a heated debate. In the December 16, 2012, the concerned scientists gathered in the Annual Meeting of the American Society for Cell Biology developed a set of recommendations referred to as the *San Francisco Declaration on Research Assessment* (DORA). DORA aimed to stop the use of the "journal impact factor" (JIF) in judging an individual scientist's work. They invited interested parties to indicate their support by adding their names to this declaration. Later the editor-in-chief of *Science* Bruce Alberts published an editorial to support this declaration. He thought the evaluation based on JIF was destructive and just encouraged "me-too science" and hence blocked innovation and created a strong disincentive to pursue risky and potentially groundbreaking work. Many leading scientists and scientific organization endorsed in this declaration (Alberts, 2013). JIF, a scientometric indicator based on the number of publications and the number of citations,

was originally created as a tool to help librarians to select journal to purchase, but later it is frequently used as a measure of the scientific quality of research in an article published in this journal and act as the primary parameter with which to compare the scientific output of individuals and institutions. Some academic institutes even use it to decide if a researcher should be funded or promoted as a tenure member (Garfield, 1999; Alberts, 2013). However, this practice arouses the fierce objection by scientists who are evaluated.

Bibliometricians also gave their voices to this phenomenon. Wouters, Glänzel, Gläser, & Rafols (2013) call for the urgent debate on the dilemmas of performance indicators of individual researchers. The Higher Education Funding Council for England (HEFCE), which distributes public money for higher education to universities and colleges in England and ensures that this money is used to deliver the greatest benefit to students and the wider public, carry out a work to review the role of metrics in the assessment and management of research. (http://www.hefce.ac.uk/whatwedo/rsrch/howfundr/metrics/). In the review, the working group launched a call for evidence to gather views and evidence relating to the use of metrics in research assessment and management. Elsevier and SPRU responded to the call. Ismael Rafols, Paul Wouters and Sarah de Rijcke organized a special session on the quality standards for evaluation indicators: Any chance for the dream to come true? (STI program). This session initiated to make the Leiden manifesto on the research assessment. van Raan, a scientometrics pioneer and gatekeeper (Garfield, Pudovkin, & Paris, 2010), will coordinate among different aspects so that this manifesto could be accepted widely. All these principles and responses, without exception, mention that quantitative information provided by metrics must be complemented by qualitative evidence to ensure the most complete and accurate input to answer a question. Even DORA recommended that the funding agencies should consider a broad range of impact measures including qualitative indicators of research impact, such as influence on policy and practice. DORA also recommended the publishers should make available a range of article-level metrics to encourage a shift toward assessment based on the scientific content of an article (DORA).

Garfield (1979, p. 62) illustrated that:

> "If the literature of science reflects the activities of science, a comprehensive, multidisciplinary citation index can provide an interesting view of these activities. This view can shed some useful light on both the structure of science and the process of scientific development."

However, can metrics drawn from publications and citations provide qualitative indicators that reveal the contents of the publications so that metrics can measure the way the contents of the publications influence policy and practice? Liu & Rousseau (2013, 2014) expounded that citation in essence is the interaction of the perspectives on a specific scientific phenomenon, hence can be used to reveal how the scientific phenomenon is understood. With the help of the regression line and a detrended curve, Liu & Rousseau (2012) show that the citation diffusion curve of an article containing a really original idea has an S-shape similar to the standard innovation diffusion curve. The convex part corresponds to the academic phase of the field that Kao's idea initiated, while the concave part corresponds to the technology dominated phase. The curve in the post-technology phase paralleled the regression line. The points of inflection correspond to the phase transition from academic to application research, while minima indicate a breakthrough in academic phase, and maxima indicate a breakthrough in the technology dominated phase. This implies that breakthroughs may directly influence the rate of change of the diffusion process while phase transfers may influence the rate of change implicitly. They claimed that the theory of diffusion process expounded in this article have the potential use of discerning breakthrough and turning points in an S & T area and finding social, technological, political and economic factors influencing the development of science. Can we use the number of publications on a specific topic to

observe the research trend? How the regression lines and the detrended forms of the publication curve tell us about the development of science? Can we discern the breakthrough and turning points between the academic phase and applied phase? Can we find social, technological, political and economic factors influencing the development of science? In this article, we will use the publications on Biomarker Her 2 to illustrate how the scientific activities on a specific research topic influence the publication process. With the help of the regression line and the detrended forms of the publication curve, we try to identify the breakthrough in this area and trajectory of translating research finding into diagnostic tools, medicines, procedures, policies and education. We will combine descriptive material on the development of the research domain with the publication growth - presents a model of interconnections of the publication and citation process, we analyze the cumulative publication curve and compare it to major events in the field. We will show that important scientific discoveries and regulation of clinical practice act as tipping points/ drivers of change in the rates of scientific publications on the topic of HER-2.

Data

After comprehensive literature research, we determined our search string:
TS=("CerbB2*" OR "CerbB-2*" OR "Cer-bB2*" OR "C-erbB2*" OR "Cer-bB-2*" OR "C-erbB-2*" OR "C-er-bB2*" OR "C-er-bB-2*" OR "Cerb B2*" OR "Cerb B 2*" OR "erbB2*" OR "erbB-2*" OR "er-bB2*" OR "er-bB-2*" OR "erb b2" OR "erb b 2" OR "HER2" OR "Epidermal growth factor receptor 2" OR "EGFR2" OR "CD340" OR "her 2")
These words include all the spelling variants related to the biomarker Human Epidermal Growth Factor 2. Among these words, “Her 2” is the only word that is not specific which may bring us some noising results because “her 2” can be used as in “her 2 children” which has nothing to do with Human Epidermal Growth Factor 2. Worse, children can be replaced by any nouns. Between her 2 and the nouns, any adjectives can be added in between. Even worse, since the Web of Science (WoS) ignores all punctuations, any punctuations can be added in between. Also one item that has “her 2 children” does not necessarily mean it is not what we need. Even the articles which deal with Human Epidermal Growth Factor 2 do not exclude the expression “her 2 children”. These situations make it very difficult for us to formulate an effective search string. However, we use the position information and its follow up to judge if these articles are related to the topic that we are searching by a program (Chavarro & Liu 2014, Lang, Liu & Chavarro, 2015), if it cannot be judged by a program, we judge it manually. We have got 98 articles that are not related to our topic. We downloaded all these data in 27 May 2014 and then excluded these 98 articles. Hence we get 30,056 articles. Since the gene of Her2/neu did not have a uniform name at the beginning when the scientists found this gene, we picked up some articles from the reference list of the early articles. And we exclude the articles published in 2014, and then we get 29,210 publications. Using these 29,210 records we do some bibliometric analysis.
The numbers of publications per year increase in roughly linearly. It is said that when a research topic turns to the application science, fewer and fewer publications will be published, instead, more and more patents will be approved. But in our case, it is the opposite, the research topic on HER-2 has already been applied in the diagnosis and therapy, the numbers of the publications on this topic do not decrease at all.

Table 1. Cumulative numbers of publications, the first and the second order differences.

Year	*cumulative numbers of publication*	*the first order difference*	*the second order difference*
1981	1	1	
1982	1	0	-1
1983	1	0	0
1984	3	2	2
1985	6	3	1
1986	12	6	3
1987	29	17	11
1988	68	39	22
1989	133	65	26
1990	261	128	63
1991	467	206	78
1992	763	296	90
1993	1126	363	67
1994	1581	455	92
1995	2046	465	10
1996	2530	484	19
1997	3048	518	34
1998	3624	576	58
1999	4312	688	112
2000	4996	684	-4
2001	5980	984	300
2002	7006	1026	42
2003	8141	1135	109
2004	9414	1273	138
2005	10922	1508	235
2006	12527	1605	97
2007	14196	1669	64
2008	16262	2066	397
2009	18633	2371	305
2010	21040	2407	36
2011	23500	2460	53
2012	26423	2923	463
2013	29210	2787	-136

Methodology: Regression Trend Lines and Detrended Curves of Time Series Data

Table 1 is a time series data. A time series is a sequence of data points, typically consisting of successive measurements made over a time interval. In informetrics, the time interval can be defined in different shift (Liu & Rousseau, 2008). Normally we make a scatter diagram to see whether data change linearly or nonlinearly. Then we make a regression analysis to find the best-fitting curve to see how the data change over time. We can get a regression equation to explain the

degree of association or the relationship between the data and time. Based on the equation that fits past data as well as possible, we can predict values of the variable at points other than the observation points.

The linear regression is the straight line. The curves of the nonlinear regression curves, depended on the regression equations, have different shapes. For example, the curve can be exponent curves if the regression equation is exponent function. The other possible curves can be logarithmic curve, power curve and multinomial curve. The straight line from the linear regression and the curve from the nonlinear regression are also called trend lines. Figure 1 show the exponential, multinomial, power, linear and logarithmic regression curves of data in Table 1.

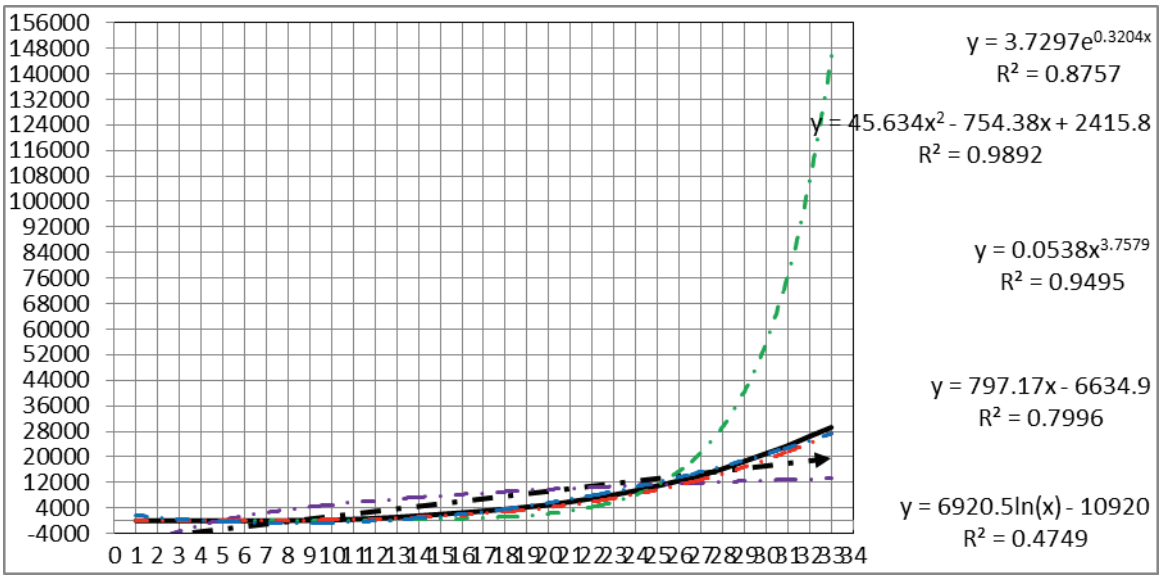


Figure 1.The exponential, multinomial, power, linear and logarithmic regression curves of data in Table 1.

Detrended curve (Shiavi, 1991) is a detrended description of the data. In order to draw a detrended curve, we find a trend line first and then calculating the difference between the overserved data and the trend line. It will give us a view on how the data change in terms of trend line. Peng et al. (1994, 1995) introduce the detrended fluctuation analysis. It is a scaling analysis method used to estimate long-range temporal correlations form. In other words, if a sequence of events has a non-random temporal structure with slowly decaying auto-correlations. It hence can eliminate the trend that self-affinity. By discerning long range correlation, it can help us understand what dominates the change of the data in the time series.

In this article, instead of calculating the difference between the observed data and the trend line, we will rotate abscissa to the paralleling line of the regression line and make the line touching the edge of the scatter diagram. The ordinate will pass through the first observation point so that all the numbers are positive. We then establish a new coordinate system. We will see how the data change with respect to the regression line

Results

We can choose different regression trend lines. In this article, we choose the best fitted straight line. Figure 2 shows the cumulative curve of the numbers of publications on her 2, its regression line and its minimum with respect to the regression line. We can see that the cumulative curve of the numbers of the publications on HER-2 is convex. The regression line intersects with the original data around 1987-1988 and 2007-2008. The minimum with respect to the regression line is around 1998-1999 (1 is the year 1981, 2 is 1982 and so on).

Now we know gene HER-2 was identified in 1981 by transfection studies with DNA from chemically induced rat neurogliobalstomas by Shih, Padhy, Murray and Weinberg (1981). From 1981 to 1987, several groups identified this gene independently (Schechter et al., 1985; Coussens et al., 1985; Semba, Kamata, Toyoshima, & Yamamoto, 1985; Fukushige et al., 1985). Slamon, Clark, Wong, Levin, Ullrich, and McGuire (1987) found correlation of relapse and survival with amplification of the HER-2 oncogene. HER-2 became a significant prognostic factor. Since then Slamon started to do research on binding to the HER-2 protein and prevents it from relaying a signal that stimulates the cancer cell to divide (Pioneers, 2007). In 1998, Herceptin was approved by FDA. Since then a revolutionary treatment started its journey in the history of human being to conquer the disease, based on the gene analysis, personalized treatment appear in the horizon that people can see. In 2007, American Society of Clinical Oncology (ASCO) and The College of American Pathologists (CAP) developed guidelines for when and how the status of HER-2 should be tested (Wolff et al., 2007). This guidelines were updated in 2013 (Wolff et al., 2013). Since then the test for the statues of HER-2 and clinical treatment with Herceptin become a standard test and treatment. However, as Herceptin did not take effect in some patients, the subpopulation remains to be defined, and side effects including cardiotoxicity need to be solved (Kumler, Tuxen, & Neilsen, 2014), HER-2 is still a topic that needs more investigations. We indicate these important events in Figure 2.

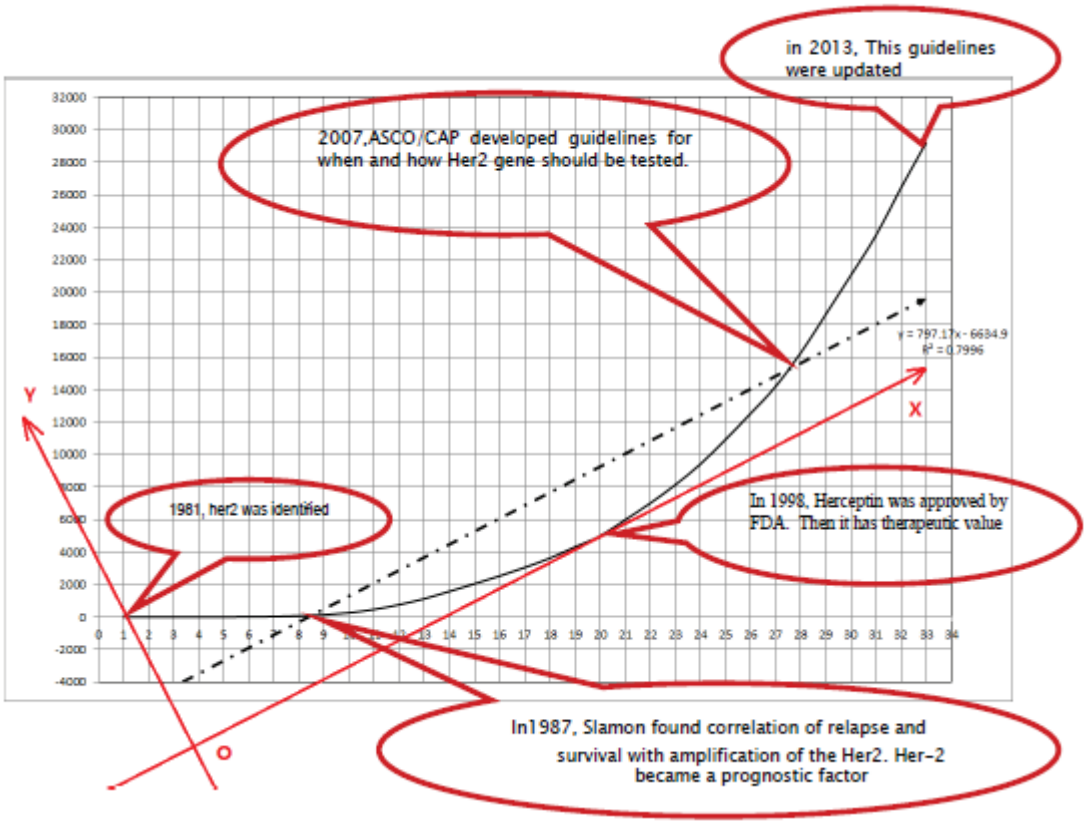


Figure 2. The cumulative curve of the numbers of publications on her 2, its regression line and its minimum with respect to the regression line.

We know that the minimum is a key point where the first-order derivative changes from negative to positive. If a curve shows the status of a thing that changes over time, we say it describes a kind of motion. The motion described by this curve changes from decreasing to increasing in the minimum point. A motion may have a different appearance as viewed from a different reference frame. If we choose the actual data as reference frame, to see how the trend changes, we can see that the discovery of the correlation of relapse and survival with amplification of her-2 oncogene in 1987 changes the trend reflected with publication data. This discovery made the amplification of her-2 a significant predictor and prognostic factor.

The numbers of publications start to increase significantly, the passion on this research topic is activated. Though with respect to the trend line, the original data curve decreases monotonically; the curve did not begin to increase until 1998 when Herceptin was approved by FDA. It is similar to the minimum point in cumulative number of citations curve of Kao when optical fiber was invented by Corning Glass Works in 1970. The crucial material problem, optical fiber, which Kao said in his conclusion "appears to be one, which is difficult but not impossible" was solved. This invention helped Kao realized his dream that no one believed it at the beginning (Liu & Rousseau, 2012). It is a coincidence that no one believed that the method Dr. Slamon used would work and the drug he created would be approved by FDA. On the contrary, everyone thought Dr. Slamon was crazy and he could not even find a student assistant majoring in science at the beginning (see the movie: Living proof and Bazell, 1998). In 2007, HER-2 test in breast cancer was recommended by ASCO-CAP, HER-2 research entered into another stage. The second order difference decreases after 2008. It dropped tremendously in 2010 and 2011. But in 2012, it went up tremendously which probably was caused by the fact that the recommendation guideline was challenged by the clinic practices and the new progresses. In 2013, ASCO/CAP convened an Update Committee that included coauthors of the 2007 guideline to conduct a systematic literature review and update recommendations for optimal HER-2 testing. In 2013, the second order difference become negative. Does the curve reach the point of inflection? We know the negative second order difference means the curve change from convex to concave. So far we cannot get to this conclusion. More observations are needed, at least we need to know how many publications on HER-2 will be published in 2014 so that we can judge whether it is an innate trend or just an occasional fluctuation. However, since major debate was settled down, though HER family oncogene (erbb1 erbb2, erbb3, erbb4) need to be dually blocked, and relative subpopulation needed to be defined and side effects refrain the use of some new developed medicine. For the moment there is an urgent need for prospective biomarker-driven trials to identify patients for whom dual targeting is cost effective (Kumler, Tuxen, & Neilsen 2014), we say it is not a major obstacle. We expect that the year when the breakthrough will make on these obstacles will appear in the maximum point on the curve drawn by the numbers of the publications on the HER-2. But it would depend on whether the research topic HER-2 gives rise to the other research topic.

The predictive, prognostic and therapeutic value of HER-2 are what changes the trend of research. The discoveries of these values of HER-2 influence the diffusion of the knowledge on HER-2 in the landscape of human intellectual space.

Selection of Trend Lines and the Different Implications that Detrended Line can Give Us

We can choose exponential, linear, logarithmic or power function as the trend lines to see what the data can tell us. Intuitively these trend lines are totally different, we hence imagine that the different trend lines can tell us totally different stories. But Figure 1 tells us the points that the different trend lines cross the data are slightly different, all around 2005-2008 even if the exponential trend is a much faster trend than the linear one. However, it is difficult to establish a new coordinate system to see clearly what the data tells us. Since the exponential curve is a straight line in semi-logarithmic system, we draw a scatter diagram in semi-logarithmic system (Figure 3). The data curve is concave upwards with respect to the exponential trend line. We can see the extremum with respect to the exponential trend line is around 1994, a little bit earlier than the time when the Herceptin was approved. However, it is in 1994 that Prof. Slamon finished phase 3 trial and was waiting for the decision of FDA. The first point that the trend line crossed the data is the same, but the second point is a little bit earlier. But the 2007 guideline was accepted for publication in September 27, 2006. The

expert panel was convened in 2005 and started to work on the guideline. It seems as if the shift of time is still in the acceptable region.

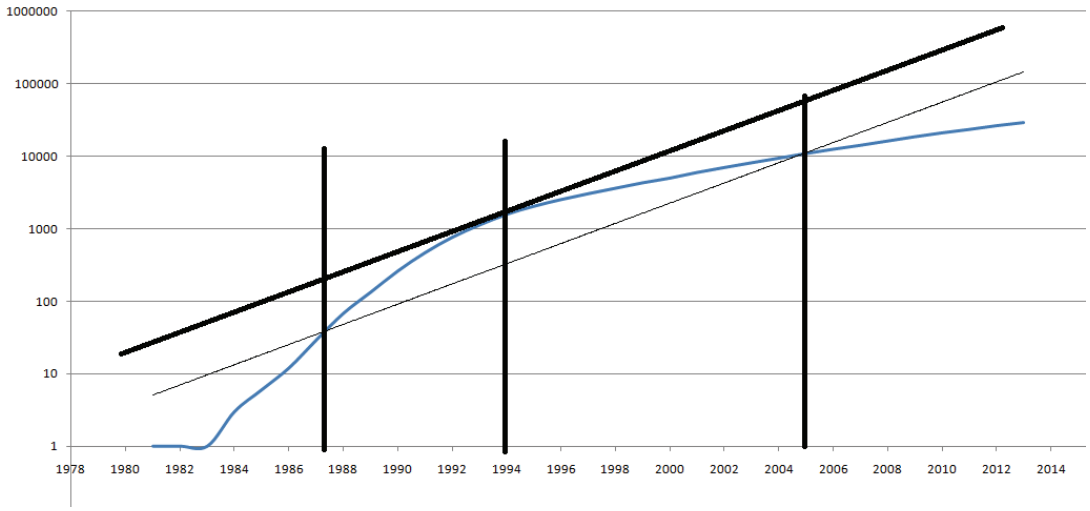


Figure 3. Cumulative publication data curve and its exponential trend line in the semi-logarithmic framework.

Publication and Citation Diffusion Process

Liu (2011) and Liu and Rousseau (2012, 2013, 2014) explore the determining factors that influence the citation process, and link the citation to the cognitive process of a scientific phenomenon under investigation. Through these articles, we illustrated the interaction of different perspectives on the phenomenon under investigation and how it is that the new ideas are accepted by academia determine the citation diffusion process.

In this investigation, we show that publication data curve with respect to the trend line can reflect how the important scientific events such as scientific discoveries and the release of government regulations in the clinical practice can change the trends of the publication process. Obviously, the primary knowledge creation process influences not only the citation process but also publication process. The change of research trends can show themselves in the publication data curve with respect to the trend line.

Liu and Rousseau (2010) studied two forms of diffusion, namely diffusion by publication and by citation. They tried to illustrate that publication diffusion is dominated by the internal diffusion mechanism that originates from the fact that a group of scientists expands their own (field) border. The citation diffusion is dominated by the external diffusion mechanism that the publication of the group of scientists, published in more and more fields, have potential to be applied in the other fields. Obviously, the publication diffusion process and citation diffusion process are interlinked with each other in that publication diffusion process determines the citation diffusion process.

As a matter of fact, publication process is entangled with citation process. Figure 4 shows how these two processes are entangled. Once the scientist(s) are interested in the scientific phenomenon, on the one hand they observe this phenomenon and get some preliminary impressions, and from these impressions they formulate some scientific ideas. On the other hand, they read the literature, which discusses this phenomenon and the perspectives to interact with the ideas that they formed by their observations to help them to get new insight into the phenomenon, and they then begin to make a thorough investigation. From these investigations scientists get new perspectives on the phenomenon. They articulate the new perspectives into a publication. When they write the manuscript they cite the old perspectives in the literature (perhaps they also read the other literatures for new evidence to convince the

readers). Publication and citation are thus born. In this process, scientific phenomenon is more and more clearly cognized.

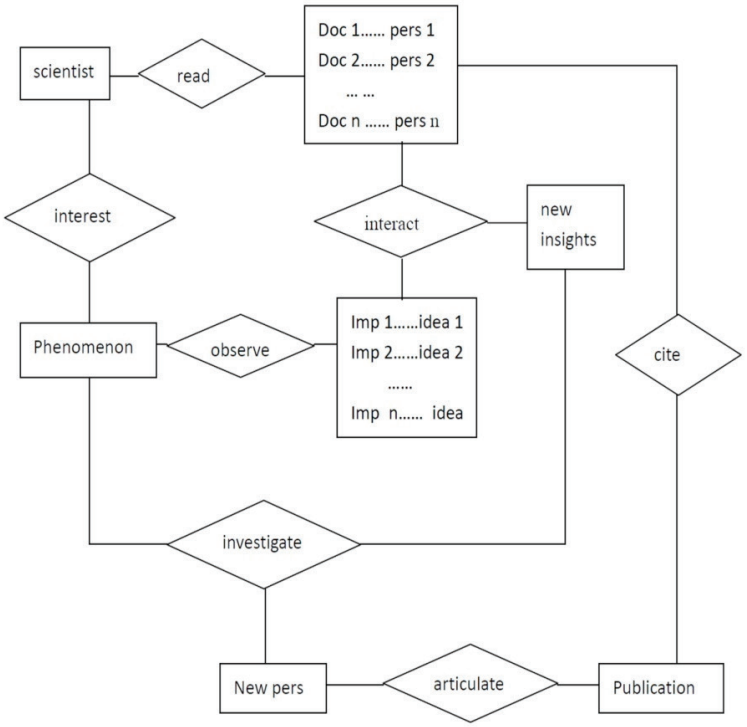


Figure 4. Entangled publication and citation diffusion process.

We can see from Figure 4 that the citation and publication processes are dynamic movement processes driven by the cognitive process of a phenomenon under investigation. The cognitive process is constituted (was led) by a series of scientific events. In this sense we may say that scientific events act as an engine to drive the evolution of science. Some events can lead the cognitive process to another direction. For example, the research in the publication Slamon, Clark, Wong, Levin, Ullrich and McGuire (1987) led HER-2 research from basic research to applied research. This event will change the research trend. Some events have no significant influence on the research trend.

Every scientific event could be represented by some publications on this event. In this sense, the scientific events drive the publication process, this process then drives the citation process. Scientific ideas that the publications convey are then diffused into the human intellectual landscape. So publication diffusion process may give us a deeper insight into the scientific events. The relationships between different publications are not as clear as that in citations, though through co-authorship or co-keywords we can establish different networks. But co-authorship or co-keywords did not reveal how the idea in one publication is diffused into the other publication. We cannot trace how the ideas in different publications interact with each other. Probably the mechanism of publication diffusion process needs to be explained via the citation diffusion process communicating different perspectives of the phenomenon under investigation. Therefore, citation and publication have a potential to reveal the cognitive process of the phenomenon under investigation.

However, we must understand how a scientific idea is diffused in the abstract intellectual landscape. This is the academic movement. In order to describe the academic movement we need to know where an idea comes from, where it will go, how fast the diffusion process is, how long is the distance from its start point to its destination. However, we face a lot challenges. First of all, we must mark the landscape with these scientific events. We have the

classification system such as the Library of Congress Classification System, Chinese Classification System, the WOS subject areas and the ESI fields. However, these systems alone cannot mark the scientific event. Because of the inaccuracy of this system, this kind of research does not give us more sense about the cognitive process of a research topic. Trochim and his colleagues (2011) proposed to identify "markers" in the translation process. They then assess the time that it takes for outputs to move across markers (Molas-Gallart, Este, Llopis & Rafols, 2014). Maybe this kind of mark system that embedded in a concrete scientific investigation will give us more information about the cognitive process of a scientific research.

Secondly, the distance in the human intellectual landscape may change over time and the destinations for the diffusion process are uncertain. These will make it very difficult to describe the scientific cognitive process via publication and citation diffusion process. These research questions deserve our effort. We would understand the scientific process more accurately if we could describe publication and citation diffusion processes more precisely. We can even anticipate what drives the evolution of science.

Conclusion

With the numbers of the publications on HER-2, we drew the accumulative curve of the publication data. We discuss the characteristics of the accumulative publication curve with respect to its trend lines and how its characteristics change in different trends. We find out the intersect points through regression line and the publication curve. These points are around the time when significant discoveries and regulations are made. These significant discoveries and regulations dominate how and where the research should go and how the basic discoveries influence their application. The accumulative publication curve itself tells us very little about how the science is evolving, but the change of the accumulative publication curve with respect to the trend lines may tell us more about the science. The content in the publication that has significant scientific value may change the direction and trend of research, hence change the publication trend reversely. We may say that important scientific discoveries and government regulations on clinical practice act as tipping points or act as drivers of change in the rates of scientific publications on the topic of HER-2. This makes us go further to explore how scientific events drive the publication process.

Acknowledgements

Yuxian Liu thanks Ismael Rafols, Raf Guns, Tim Engels, and Ronald Rousseau for the discussion in the early stage of this research. This work is supported by NSFC via 71173154. Yuxian Liu further acknowledges support from the China Scholarship Council.

References

Alberts, B. (2013). Impact Factor Distortions. *Science*, 340(6134), 787.

Bornmann, L. & Marx, W. (2013). Proposals of standards for the application of scientometrics in the evaluation of individual researchers working in the natural sciences. *Zeitschrift für Evaluation,* 12(1), 103-127.

Chavarro, D., & Liu, Y. (2014). How can a word be disambiguated in a set of documents: using recursive Lesk to select relevant records. http://www.gtmconference.org/pages/program.html

Coussens, L., Yang-Feng, T. L., Liao, Y. C., Chen, E., Gray, A., McGrath, J. et al. (1985). Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene. *Science*, 230, 1132-1139.

Elsevier. (2014).Response to HEFCE's call for evidence: independent review of the role of metrics in research assessment. http://www.elsevier.com/__data/assets/pdf_file/0015/210813/Elsevier-response-HEFCE-review-role-of-metrics.pdf.

Fukushige, S. I., Matsubara K.I, Yoshida, M., Sasaki, M., Suzuki, T., Semba, K. et al. (1986). Localization of a novel v-erbB-related gene, c-erbB-2, on human chromosome 17 and its amplification in a gastric cancer cell line. *Molecular and Cellular Biology,* 6, 955-958.

Garfield, E. (1999). Journal impact factor: a brief review. *Canadian Medical Association Journal*, 161(8), 979-980.

Garfield, E (1979). *Citation Indexing.* New York: Wiley.

Garfield, E. (1955). Citation index for science. *Science*, 122, 108-111.

Garfield, E., Pudovkin, A. I., & Paris, S. W. (2010). A bibliometric and historiographic analysis of the work of Tony van Raan: a tribute to a scientometrics pioneer and gatekeeper. *Research Evaluation*, 19(3), 161-172.

Inglesi-Lotz, R., & Pouris, A. (2011). Scientometric impact assessment of a research policy instrument: the case of rating researchers on scientific outputs in South Africa. *Scientometrics*, 88(3), 747-760.

Kumler, I., Tuxen, M.K., & Neilsen, D. L. (2014). Anti-Tumour Treatment, A systematic review of dual targeting in Her-2 positive breast cancer. *Cancer Treatment Reviews*, 40, 259-270.

Lang, F., Liu, Y., & Chavarro, D. (2015). Improving accuracy in data collection: Can machine learning classification help? (in press)

Liu, Y.X. (2011). The diffusion of scientific ideas in time and indicators for the description of this process. Unpublished Doctoral Thesis, University of Antwerp, Belgium.

Liu, Y. X. & Rousseau, R. (2008). Definitions of time series in citation analysis with special attention to the h-index. *Journal of Informetrics,* 2(3), 202-210.

Liu, Y.X., & Rousseau, R. (2010). Knowledge diffusion through publications and citations: A case study using ESI-fields as unit of diffusion. *Journal of the American Society for Information Science and Technology*, 61(2), 340-351.

Liu, Y. & Rousseau, R. (2012). Towards a representation of diffusion and interaction of scientific ideas: the case of fiber optics communication. *Information Processing and Management,* 48(4), 791-801.

Liu, Y. & Rousseau, R. (2013). Interestingness and the essence of citation. *Journal of Documentation*. 69(4), 580-589.

Liu, Y. & Rousseau, R. (2014). Citation analysis and the development of science: a case study using articles by some Nobel Prize winners. *Journal of the American Society for Information Science and Technology,* 65(2), 281–289.

Peng, C-K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., & Goldberger, A.L. (1994). Mosaic organization of DNA nucleotides. *Physical Review E,* 49:1685-1689.

Peng, C-K., Havlin, S., Stanley, H.E., & Goldberger, A.L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos, 5:82-87.

Dr. Dennis Slamon's development of Herceptin revolutionized breast cancer treatment and accelerated research into therapies customized for each individual patient: the personal approach. (2007, Fall-Winter). *Triumph*, 13-16.

Schechter, A. L., Stern, D. F., Vaidyanathan, L., Decker, S. J., Drebin, J. A., Greene, M. I. et al. (1984). The neu oncogene: an erb-B-related gene encoding a 185,000-Mr tumour antigene. *Nature*, 312(5994), 513-516.

Semba, K., Kamata, N., Toyoshima, K., & Yamamoto, T. (1985). A v-erbB-related protooncogene, c-erbB-2, is distinct from the c-erbB-1/epidermal growth factor-receptor gene and is amplified in a human salivary gland adenocarcinoma. *PNAS,* 82 (19), 6497-6501.

Shiavi, R. (1991). *Introduction to Applied Statistical Signal Analysis*. Homewood, IL: Irwin, Aksen.

Shih, C., Padhy, L., Murray, M., & Weinberg, R. A. (1981). Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature*, 290, 261-264.

Slamon, D. J., Clark, G. M., Wong, S. G. Levin, Ullrich, A., & McGuire, W. L. (1987). Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*, 235, 177-182.

Truex, D., Cuellar, M., & Takeda, H. (2009). Assessing scholarly influence: using the Hirsch indices to reframe the discourse. *Journal of the Association for Information Systems,* 10 (7), 560-594.

Wolff, A.C., Hammond, M. E. H., Schwartz, J.N.,Hagerty, K.L., Allred, D.C., Cote, R. J. et al. (2007). American Society of Clinical Oncology/College ofAmerican Pathologists Guideline Recommendations for Human Epidermal Growth Factor Receptor 2Testing in Breast Cancer. *Arch Pathol Lab Med*, 131, 18-43.

Wolff, A.C., Hammond, M. E. H., Hicks, D. G., Dowsett, M., McShane, L.M. et al. (2013, November 1). Recommendations for Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical practice Guideline Update. *Journal of Clinical Oncology*, 31(31), 399-4013.

Wouters, P., Glänzel, W., Gläser, J. & Rafols, I. (2013). The dilemmas of performance indicators of individual researchers: an urgent debate in bibliometrics. *ISSI Newsletter*, 9(3), 48-53.

Zhu, X., Wu, Q., Zheng, Y. Z., & Ma, X. (2004). Highly cited research papers and the evaluation of a research university: A case study: Peking University 1974-2003. *Scientometrics*, 60(2). 237-247.

Founding Concepts and Foundational Work: Establishing the Framework for the Use of Acknowledgments as Indicators

Nadine Desrochers[1], Adèle Paul-Hus[1] and Jen Pecoskie [2]

[1] *nadine.desrochers@umontreal.ca, adele.paul-hus@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, C.P. 6128, Succ. Centre-Ville, H3C 3J7 Montreal, QC (Canada)

[2] *jpecoskie@wayne.edu*
School of Library and Information Science, Wayne State University, Detroit, MI, 48202 (USA)

Abstract
Building on the concepts of the reward system of science and social capital, Blaise Cronin brought forth the idea that rewards in science are threefold, forming a triangle built from authorship, citations, and acknowledgements. Of these, acknowledgments are the hardest to grasp and evaluate. After nearly 45 years of multidisciplinary research on acknowledgments and a corpus of over 80 scientific contributions, there is still no consensus on the value of acknowledgments in scholarly communication. This study aims to further acknowledgments research with a meta-synthesis of the literature, establishing the theoretical framework for the use of acknowledgments as bibliometric indicators. Based on in-progress content analyses, broad categories emerge revealing contextual information crucial to the understanding of acknowledgments. Applying our framework on data from the Web of Science, further phases of this study will provide large-scale findings based on a multidisciplinary sample. From there, it will be possible to envision recommendations for the standardization and use of acknowledgments as indicators. However, grounding the study of acknowledgments in their underlying theoretical considerations and conceptual foundations will ensure these recommendations respect the diverse traditions of the scientific field.

Conference Topic
Theory

Introduction and background
It is a broadly recognized fact that the scientific field has a very "high degree of codification", to borrow the Bourdieusian phrase (Bourdieu, 1996, p. 226). How and when one is admitted into the academic community, how a researcher acquires credibility within the scientific realm, and what contributions turn a researcher into a renowned scholar are endlessly evaluated, measured, and scrutinized. This high degree of codification helps to both foster and assuage the paradox that underlies the use of empirical measures to define what remains an intrinsically nuanced and contextualized concept: scientific "success".

Merton (1973) presented the sociology of science with the reward system of science, its recognition paradigm, and the nepotistic undertones of the Matthew effect; Bourdieu reframed the concept of recognition to befit the concept of symbolic capital. Blaise Cronin brought forth the idea that these rewards are threefold, forming a triangle built from authorship, citations, and acknowledgements (Cronin, 1995; Cronin, 2005; Cronin & Weaver-Wozniak, 1993). These are all part of the *illusio*, which encompasses the stakes of the academic "game", its rules, and the very fact that its rewards are worth pursuing (Bourdieu, 1988, p. 56).

Of these rewards, acknowledgments are the hardest to grasp and evaluate; reasons range from lack of standardization to name-dropping and ambiguous wording (Cronin, 1995; Cronin, 2014), as well as the placement of acknowledgments, which can vary from in-text mentions to paratextual elements situated outside the body of the text (Genette, 1997). Researchers have also called for stricter policies to inform the use of acknowledgments, prescribe their form, offer conditions for inclusion, or establish their ethical ramifications (Brown, 2009; Chubin, 1975; Pontille, 2001). For example, while Cronin's research (Cronin, 1995) showed that in

most researchers' view, obtaining permission to thank is unnecessary, certain current editorial policies (e.g., PLOS ONE, 2015) require any acknowledging party to obtain the acknowledged party's permission. Extricating one aspect of acknowledgments is also not always straightforward. The "Funding Text" (FT) field of the Web of Science (WoS) database, indexed since 2008, is a telling example, since it often contains all things and people acknowledged, not just the agencies or institutions that provided funds to the project. That being said, the FT field of the WoS has opened new avenues for this research by making massive datasets available.

However, the literature heeds one important and overarching warning: after nearly 45 years of multidisciplinary study and a corpus of over 80 scientific contributions, there is still no consensus on the value of acknowledgments, no potential for meta-analysis within this corpus, and, despite common questions, no shared framework for further analysis, nor any clear recommendations for standardization. Given this situation, this study aims to further acknowledgments research with potential contributions to scientific policy guidelines (editorial and institutional) and research assessment (individual and disciplinary) in the scientometrics field, which has shown ongoing interest for acknowledgments as a potential indicator (Cronin & Weaver-Wozniak, 1992; Cronin, 2005; Díaz-Faes & Bordons, 2014).

In order to gain an understanding of where acknowledgments research had emanated from and where it is currently situated in the scientific ecology, an initial overview of the literature on acknowledgments was conducted, leading to the retrieval and document-level analysis of 115 scientific publications, which became the subject of a chapter submitted for inclusion in a book on theories in informetrics (Desrochers, Paul-Hus, & Larivière, in press).

This phase of the research established that the reward triangle can and should be studied, not only for its three constituting factors, but also for the relationships between them. It showed that the meeting point of citation and authorship is the apex of the reward triangle. Acknowledgements, however, are foundational in that they reveal the inner workings of the scientific *illusio* (Bourdieu, 1988) that support this apex and that have, historically, supported key conceptual frameworks: the "invisible college" (Crane, 1972), "trusted assessors," encountered before and during the peer review process (Mullins & Mullins, 1973), and the categorization of authors vs. acknowledged contributors (Patel, 1973).

Methodology

Following this initial review, it became clear that a meta-analysis of acknowledgments research would not be possible; however, the range of complex and varied approaches could form the basis for a meta-synthesis (Rousseau, Manning & Denyer, 2008) of the literature. This will: extract knowledge on the perceptions of acknowledgements across a variety of disciplines (e.g., Information Science, History, Astronomy, Literature, and Psychology); provide scientometricians with information pertaining to the nuances and contexts of research creation in various disciplines; and yield the conceptual framework necessary to undertake acknowledgements research on a larger scale using multidisciplinary datasets. The following research questions were thus devised:

1. What does "acknowledgment research" look like?
    a. Throughout history? (1970-present)
    b. What were its founding concepts and considerations?
    c. How are acknowledgments perceived and positioned in the acknowledgments literature itself?
2. Who is concerned with acknowledgment research?
    a. Scientists from what fields conduct acknowledgment research?
3. What aspects of acknowledgments are studied in acknowledgment research?

Using approaches based in the Social and Health Sciences (Rousseau et al., 2008; Dixon-Woods et al., 2005; Mays, Pope & Popay, 2005) and recommendations specific to the use of evidence-based literature in Information Science (Urquhart, 2010), a protocol for meta-synthesis was established using the PRISMA model for systematic literature reviews (Moher et al., 2009). The most recent searches place the corpus at 80 relevant documents. This paper presents preliminary findings and initial theoretical considerations.

Preliminary Findings and Discussion - Foundations for a theoretical framework

Based on in-progress content analyses, broad categories are emerging; they reveal contextual information crucial to the understanding of acknowledgments as potential bibliometric indicators.

Paratextual Status: Acknowledgements can be elusive, especially in structure-driven datasets. Standardized locations, conventions, separate paragraphs, in-text allusions, database fields defined as pertaining to one aspect but including others are all intrinsic to understanding their value.

Disciplinary Contexts: The literature stems from various disciplines, yielding a broad range of methods and reporting styles. It also approaches the topic from various angles: a discipline (e.g., Cronin, 2001), a culture or a group (e.g., Woolf, 1975), a linguistic community (e.g., Al-Ali, 2010), a specific journal or set of journals (e.g., Rattan, 2013), dissertations (e.g., Gesuato, 2004), or direct enquiry (e.g., Heffner, 1979), quantitative (e.g., Costas & van Leeuwen, 2012) or qualitative (e.g., Bashtomi, 2008). These differences do provide a spectrum of perspectives that need to be part of any standardization process of these scholarly rewards into contextualized indicators.

The Thankers and the Thanked: At its core, acknowledgments research is based on the basic questions of who or what gets thanked by whom and for what. From the expression of gratitude towards spouses to the mention of support from grant agencies, scientific acknowledgments reflect the same diversity as acknowledgments from other types of writers, such as literary writers (Desrochers & Pecoskie, 2014) and can be seen as a "'ledger' where debts are acknowledged" (Weber & Thomer, 2014, p. 84). Inconsistencies abound: people are thanked without specification of tasks, tasks are listed without names; financial capital is embedded with social capital and with messages of a highly personal nature (Coates, 1999).

Cloak and Dagger Reveals: The previous two categories show that scientific acknowledgments are sometimes as much a puzzle as they are clear; this in itself is information. Indeed, the last decades have shown interest in the fact that acknowledgments can expose the invisible college and pre-publication readers, including unknown reviewers, thereby setting boundaries between groups who know their identities and those who do not. This is obviously problematic in terms of using acknowledgments as indicators; yet abolishing this practice would mean revoking a practice that pays homage to the peer review process as it currently exists.

Language and Ethics: The acknowledgments genre has been studied in Linguistics and alluded to in other disciplines, including Information Science (Cronin, McKenzie & Stiffler, 1992). "How" entities are thanked is closely linked to prescribed funding-based requirements, cultural and disciplinary practices, and editorial guidelines, the latter being related to the ethics of thanks: securing permission to thank someone, paying 'lip service' to key players, and name-dropping (Cronin, 1995; Hollander, 2002)—angles reminiscent of the Matthew effect.

Value and Perception: Finally, acknowledgments research has the ingrained quality, seen elsewhere in science but perhaps rarely to this extent, to turn on itself. Numerous papers oscillate between two positions: perceiving acknowledgments as suitable for study and as potential indicators, true to the Merton-Bourdieu-Cronin theoretical continuum; and

criticizing them as problem-laden, lacking standardization, and fickle. Context and processes have come under scrutiny in the use of other indicators in research assessment; yet acknowledgment studies have a particular penchant for self-deprecation while relying on what is now four decades of research to insist upon the fact that there is something to this paratext.

Conclusion and Upcoming Phases

Quantitative content analysis will help weigh these concerns throughout the history of acknowledgments research. Qualitative analysis will help nuance these findings through context, history, and disciplinary boundaries. Together, these analyses will provide a meta-synthesis of the existing literature, from which the conceptual framework outlined here will be refined for use in further studies. The goal is to use this framework on data from the WoS and to provide large-scale findings based on a multidisciplinary sample. From there, it will be possible to envision recommendations for the standardization and use of acknowledgments as indicators.

However, since the literature provides many important warning signs, heeding them and grounding the study of acknowledgments in their underlying conceptual foundations will ensure these guidelines respect the multiple traditions of the scientific field and work within the boundaries of the evolving high stakes of codification. Furthermore, they will help take into account the fact that acknowledgments have long had a special standing in academia as the place where the *homo academicus* (Bourdieu, 1988) can make the invisible visible, but also vice-versa. This, in itself, is a stake of the *illusio* that deserves to be better understood.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council of Canada. The researchers further thank Vincent Larivière for his support and insight.

References

Al-Ali, M. N. (2010). Generic patterns and socio-cultural resources in acknowledgements accompanying Arabic Ph.D. dissertations. *Pragmatics*, *20*(1), 1–26.

Basthomi, Y. (2008). Interlanguage discourse of thesis acknowledgements section: Examining the terms of address. *Philippine Journal of Linguistics*, *39*(1), 55–66.

Bourdieu, P. (1975). The specificity of the scientific field and the social conditions of the progress of reason. *Social Science Information*, 14(6), 19–47.

Bourdieu, P. (1988). *Homo academicus*. Stanford, Calif.: Stanford University Press.

Bourdieu, P. (1996). *The Rules of Art: Genesis and Structure of the Literary Field*. Stanford, Calif.: Stanford University Press.

Brown, R. (2009). How scholars credit editors in their acknowledgements. *Journal of Scholarly Publishing*, *40*(4), 384–398.

Chubin, D. E. (1975). Trusted assessorship in science: A relation in need of data. *Social Studies of Science*, *5*(3), 362–367.

Coates, C. (1999). Interpreting academic acknowledgements in English studies: Professors, their partners, and peers. *English Studies in Canada*, *25*(3-4), 253–276.

Costas, R., & van Leeuwen, T. (2012). Approaching the "reward triangle": General analysis of the presence of funding acknowledgments and "peer interactive communication" in scientific publications. *Journal of the American Society for Information Science and Technology*, *63*(8), 1647–1661.

Crane, D. (1972). *Invisible Colleges: Diffusion of Knowledge in Scientific Communities*. Chicago, IL: University of Chicago Press.

Cronin, B. (2014). Foreword: The penumbral world of the paratext. In N. Desrochers & D. Apollon (Eds.), *Examining Paratextual Theory and its Applications in Digital Culture* (pp. xv-xix). Hershey, PA: IGI Global.

Cronin, B. (1995). *The Scholar's Courtesy: The Role of Acknowledgement in the Primary Communication Process*. London: Taylor Graham.

Cronin, B. (2001). Acknowledgement trends in the research literature of information science. *Journal of Documentation*, *57*(3), 427–433.

Cronin, B. (2005). *The Hand of Science: Academic Writing and its Rewards*. Lanham, Maryland: Scarecrow Press.

Cronin, B., McKenzie, G., & Stiffler, M. (1992). Patterns of acknowledgement. *Journal of Documentation*, *48*(2), 107–122.

Cronin, B., & Weaver-Wozniak, S. (1992). An online acknowledgment index: Rationale and feasibility. In D. Raitt (Ed.), *Online Information 92: Proceedings of the 16th International Online Information Meeting, London, 5-10 December 1992* (pp. 281–290). Oxford: Learned Information.

Cronin, B., & Weaver-Wozniak, S. (1993). Online access to acknowledgements. *Proceedings of the 14th National Online Meeting 1993* (pp. 93–98). New York: M.E. Williams.

Desrochers, N., Paul-Hus, A., & Larivière, V. (in press.) The angle sum theory: Exploring the literature on acknowledgments in scholarly communication. In C. R. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication.* Boston, MA: De Gruyter.

Desrochers, N., & Pecoskie, J. (2014). Inner circles and outer reaches: Local and global information-seeking habits of authors in acknowledgment paratext. *Information Research, 19*(1), paper 608. Retrieved from http://InformationR.net/ir/19-1/paper608.html

Díaz-Faes, A. A., & Bordons, M. (2014). Acknowledgments in scientific publications: Presence in Spanish science and text patterns across disciplines. *Journal of the Association for Information Science and Technology*, 65(9): 1834-1849.

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: A review of possible methods. *Journal of Health Services Research & Policy*, *10*(1), 45–53B.

Gesuato, S. (2004). Acknowledgments in PhD dissertations: The complexity of thanking. In C. Taylor Torsello, M. Grazia Bùsa, & S. Gesuato (Eds.), *Lingua inglese e mediazione linguistica. Ricerca e didattica con supporto telematico* (pp. 273–318). Padova: Unipress.

Heffner, A. G. (1979). Authorship recognition of subordinates in collaborative research. *Social Studies of Science*, *9*(3), 377–384.

Hollander, P. (2001). Acknowledgments: An academic ritual. *Academic Questions*, *15*(1), 63–76.

Mays, N., Pope, C., & Popay, J. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research & Policy*, *10*(suppl 1), 6–20.

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago: University of Chicago Press.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, *6*(7), e1000097.

Mullins, N. C., & Mullins, C. J. (1973). *Theories and Theory Groups in Contemporary American Sociology*. New York, NY: Harper and Row.

Patel, N. (1973). Collaboration in the professional growth of American sociology. *Social Science Information*, *12*(6), 77–92.

PLOS ONE. (2015). PLOS ONE manuscript guidelines: Acknowledgments. Retrieved from http://www.plosone.org/static/guidelines#acks

Pontille, D. (2001). L'auteur scientifique en question: Pratiques en psychologie et en sciences biomédicales. *Social Science Information*, *40*(3), 433–453.

Rattan, G. K. M. (2013). Acknowledgement patterns in annals of library and information studies 1999-2012. *Library Philosophy and Practice*, *e-journal* (paper 989). Retrieved from http://digitalcommons.unl.edu/libphilprac/989/

Rousseau, D. M., Manning, J., & Denyer, D. (2008). *Evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses* (SSRN scholarly paper 1309606). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=1309606

Urquhart, C. (2010). Systematic reviewing, meta-analysis and meta-synthesis for evidence-based library and information science. *Information Research, 15*(3), paper 708. Retrieved from http://www.informationr.net/ir/15-3/colis7/colis708.html

Weber, N. M. & Thomer, A. K. (2014). Paratexts and documentary practices: Text mining authorship and acknowledgment from a bioinformatics corpus. In N. Desrochers & D. Apollon (Eds.), *Examining Paratextual Theory and its Applications in Digital Culture* (pp. 84-109). Hershey, PA: IGI Global.

Woolf, P. (1975). The second messenger: Informal communication in cyclic AMP research. *Minerva*, *13*(3), 349–373.

Analysis on the Age Distribution of Scientific Elites' Productivity: A study on Academicians of the Chinese Academy of Science

Liu Jun-wan[1] and Zheng Xiao-min[2] and Feng Xiu-zhen[3] and Wang Fei-fei[4]

[1]*liujunwan@bjut.edu.cn,* [2]*xiaominzheng2014@sina.com,* [3]*xfeng@bjut.edu.cn,* [4]*feifeiwang@bjut.edu.cn*
School of Economics & Management, Beijing University of Technology, Beijing 100124, (China)

Introduction

Is there any regularity in scientists' research activities? For example, does there exist a period when a scientist makes his most contributions? If so, which period is the most productive period? To answer the questions above, many scholars have been contributed their efforts on studying the relationships between productivity and age, such as: (1) age distribution of scientists' creativity or productivity (Liming et al., 1996; Bonacarsi & Daraio, 2003; Jones 2010); (2) the relationship between the longevity and scientist's outputs (Levin & Stephan, 1991; Jonesa & Weinberg 2011; Todorovsky, 2014); (3) the effects of age on researcher's productivity (Bonacarsi & Daraio Costas & van Leeuwen, 2010). However, the previous research still leave some gaps need to be filled. One of them is what about the age distribution of an individual researcher's achievements in his research career. Our research efforts in this paper would contribute to this topic. Particularly, the object of our study is Academicians of the Chinese Academy of Science. And we explore the age distribution of publication by these academicians.

Data and Method

The website of Academic Divisions of the Chinese Academy of Sciences provides academicians' brief introduction and research experience, which including their birth day and affiliated institutions. We choose total 139 Academicians in field of *Mathematics & Physics*, and total 85 Academicians in field of *Information Technical Science* as our research data. *Mathematics & Physics* is an ancient and classical subject, and *Information Technical Science* is a rapid development subject. In order to analyze the age distribution of these academicians' publication, the academician's name and affiliation were used as joined retrieval terms to get their publications both in China National Knowledge Infrastructure (CNKI) and web of science (SCI) database. CNKI is the largest authoritative digital publishing platform and knowledge services platform in China. To get their whole publication output, the repetitive or mistaken publication data of these academicians were deleted.

The average age of 224 academicians is 74 years old, and all of these academicians are now alive until the retrieval day (11/2014). The number of the scientists' publications was selected as the scientific productivity indicator, but the co-author situation was equally considered. This paper considers age distribution of scientists' publication from the scientists' physiological age view.

Age distribution of academicians' publication

Firstly, we count the number of every individual academician's publication according to his physiological age. After that we sum the number of publication up according to the same physiological age of all academicians in the same field. So we can get the physiological age distribution of publication of total scientists in one field. We named papers indexed in SCI/CNKI as "SCI/CNKI" paper for short.

Age distribution of academicians' publication in Mathematics & Physics

The publication age distribution curve of CNKI paper and SCI paper of academicians in *Mathematics & Physics* are shown in Figure 1(a). The publication age distribution curve of total paper (sum of number of CNKI paper and SCI paper) is presented in Figure 1(b). Just as shown from the folder part of the two publication age distribution curves in Figure 1(a), we can see the period between the age of 50 and 65 is the same publication peak period of CNKI paper and SCI paper. Scientists published 61% of their total publications between the age of 50 and 71, the highest peak point is at the age of 68.

Age distribution of academicians' publication in Information Technical Sciences

The publication age distribution curve of CNKI paper and SCI paper of academicians in *Information Technical Sciences* are presented in Figure 2(a). The age period from 60 to 70 is the same publication peak period of CNKI paper and SCI paper. As Figure 2(b) is shown, scientists published 51% of their total publications between the age of 62 and 76, and the highest peak point is at the age of 67. In detail, there is a smaller publication peak period between the age of 45 to 51 before the higher one.

Significant differences test of academicians' productivity before and after tenure

Paired-Samples T Test was used to test if the scientists' productivity would be different before and after tenure. We sum up the number of publications for five years of every individual academician before and after tenure. Before testing, we assume that there is no significant difference of academicians' productivity before and after tenure, then we use the Paired-Samples T test to test the hypothesis. According to the analysis results, the assumption is rejected, which means that the number of publication is obviously different before and after tenure. After tenure, academicians are more productive than before in overall.

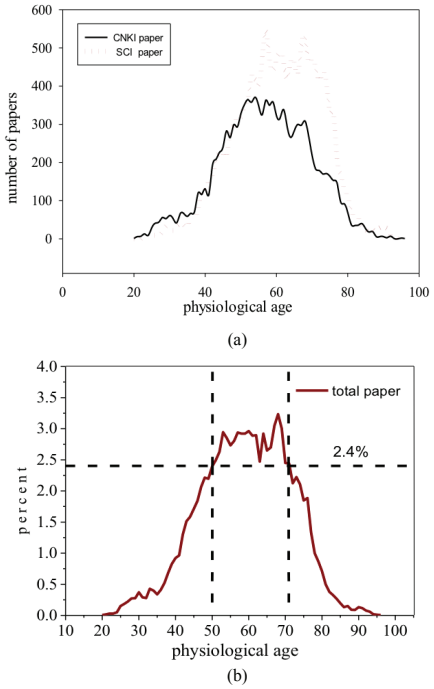


Figure 1. Publication age distribution of academicians in *Mathematics & Physics*.

Discussion and conclusion

The final results show that age distributions of academicians' publication have some regular features. The entire publication age curve of *Mathematics & Physics* shows a single peak distribution. The publication peak period is between the age of 50 and 71. However, publication peak period of academicians in *Information Technical Sciences* is between the age of 62 and 76. Moreover, it is different from *Mathematics& Physics,* which has a small publication peak period between 45 and 51 in publication age curve of *Information Technical Sciences'* academicians. Additionally, our results also reveal that there is significant difference of the scientists' productivity before and after tenure. The publication age distribution law on academicians of the Chinese Academy of Science brings us useful enlightenment. We should pay more attention to middle-aged scientists to improve their research input-output ratio.

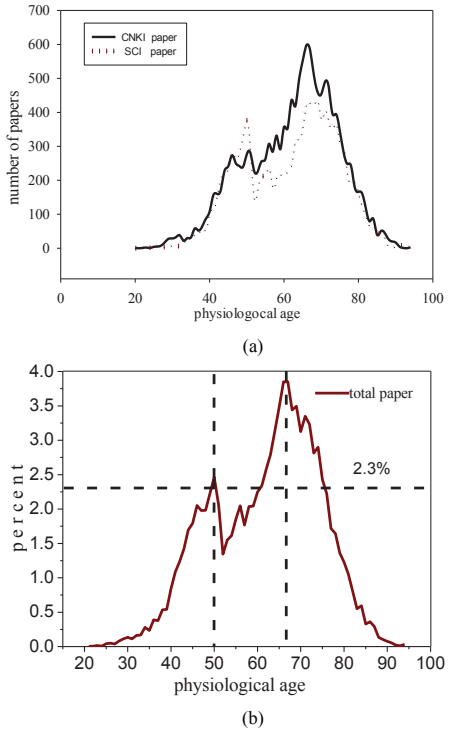


Figure 2. Publication age distribution of academicians in *Information Technical Sciences*.

References

Bonaccorsi, A., & Daraio, C. (2003). Age effects in scientific productivity. *Scientometrics*, *58*(1), 49-90.

Costas, R. & Van Leeuwen, T. N. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *JASIST*, *61*, 1564–1581.

Jones, B.F. (2010). Age and great invention. *Rev Econ Stat*, *92*, 1–14.

Jonesa, B. F. & Weinberg, B. A. (2011). Age dynamics in scientific creativity. *Proceedings of the national academy of sciences of the united states of America*, *108*, 18910-18914.

Levin, S.G. & Stephan, P. E. (1991). Research productivity over the life cycle: evidence for academic scientists. *American Economic Review*, *81*, 114–132.

Liming, L., Hongzhou, Z. & Yuan, W. (1996). Distribution of major scientific and technological achievements in terms of age group — Weibull distribution. *Scientometrics*, *36*, 3-18.

Todorovsky, D. (2014). Follow-up study: on the working time budget of a university teacher-45 years self-observation. *Scientometrics*, *3*, 2063-2070.

An Experimental Study on the Dynamic Evolution of Core Documents

Lin Zhang[1], Wolfgang Glänzel[2], Fred Y. Ye[3]

[1] *zhanglin_1117@126.com*
[1]Dept. Management and Economics, North China University of Water Conservancy and Electric Power, Zhengzhou (China)

[2] *Wolfgang.Glanzel@kuleuven.be*
[2]Centre for R&D Monitoring (ECOOM) and Dept. MSI, KU Leuven, Leuven (Belgium)
Dept. Science Policy & Scientometrics, Library of the Hungarian Academy of Sciences, Budapest (Hungary)

[3] *yye@nju.edu.cn*
[3]School of Information Management, Nanjing University, Nanjing 210023, (China)
Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023 (China)

Introduction

The concept of the core of documents had originally been introduced in connection of co-citation analysis (Small 1973). The term *core documents* has later been re-introduced in the context of bibliographic coupling (BC; see Glänzel & Czerwon, 1996) and hybrid BC and text based similarities (Glänzel & Thijs, 2011) in order to identify strongly interlinked papers that form important nodes in the network of scholarly communication. In order to study stability and dynamics of core-document sets we apply two different methods to h-index related literature in the period 2005–2013 for illustration.

Data Sources and Processing

Data were retrieved from Thomson Reuters Web of Science Core Collection (WoS) following the strategy of Zhang et al. (2011), with extension of the period 2005–2013. We also added citing papers but removed duplicates and papers with less than 5 references to avoid biases in BC similarities. We obtained a final set of 3,270 documents. Figure 1 shows the annual increment of papers in this set.

Research Questions, Methods and Results

In this study we apply two different methods to determine core documents, (Method I) the traditional one according to Glänzel & Czerwon (1996) with a fixed number of links ($n = 15$) and Method II using the h-core of the network (Glänzel, 2012). In both cases we applied a *hybrid approach*. We used link strengths of 0.5 and 0.4 according to Salton's cosine measure. Using these parameters, we analysed the dynamics of core documents along the following questions.

- How is evolution of core documents reflected by the two methods?
- Do the two methods provide stable results?
- Do core documents adequately represent the evolution of the topic?

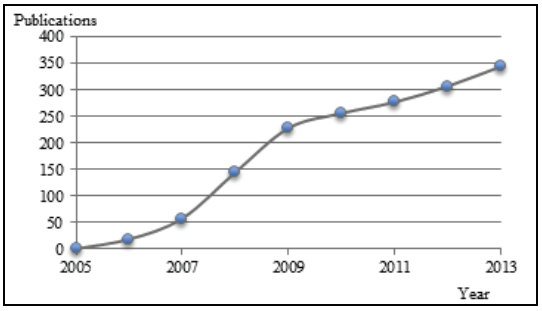


Figure 1. Distribution of h-related publications during 2005-2013.

Core document are by definition strongly interlinked with a large number of other documents in the set under study and thus represent the very core of the set. As expected, their number increases with expanding time spans, the average annual growth rate of the cumulative set amounted to 46% (Method I) and 25% (Method II), respectively. Not only the number of nodes in the network but also the number of their links is growing, however at a different pace. Indeed, we found that the complete h-related set increased at a large constant pace of 11% while the growth of the core sets was faster (see above), but its growth slowed down. This might in part be a consequence of the increasing age of references. In 2013 the core reached a representation of 2.0% and 2.4%, respectively. This characterizes the evolution of the core set with respect to the topic dynamics. The second question that arises from these figures is in how far do both methods mirror the same "core" of literature. In order to check the robustness of these methods, we compared the overlap of the sets of core documents obtained from the two methods. To this end we used BC with fixed number of links as reference standard. Concordance with Method I ranged between 83.8% and 95.2% with increasing trend from 2005–2007 to 2005–2013 and using Method II the shares ranged between 96.8% and 80.7%, however with decreasing trend.

In order to answer the third question, we analysed the core sets obtained from the two methods on the basis of authors and topics of the individual papers. The evolution of the core-document sets according to Method II is shown at three different stages in Figure 2 using Pajek with Kamada–Kawai layout (Batagelj & Mrvar 2003).

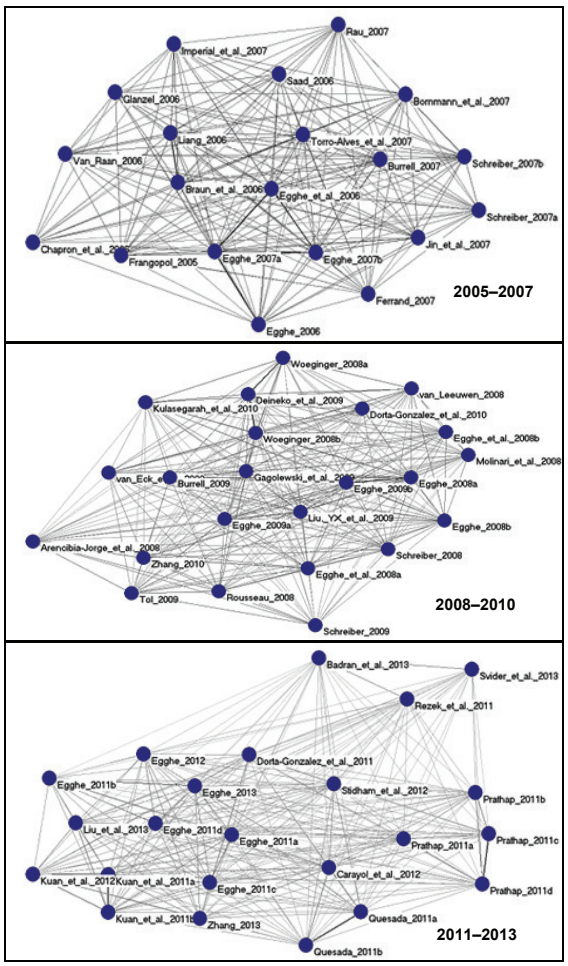


Figure 2. The evolution of the core-documents set (II).

Core nodes in Figure 2 are based on BC but hybrid similarities are used to measure the links between the nodes. This can be done because of the strong concordance between the sets obtained from the two methods. The links between core nodes in Figure 2 are denser and stronger than in the BC approach, which is due to the inclusion of textural information. The interpretation of Figure 2 is not straightforward, but the structural changes of the networks during different periods presented here are quite clear and noteworthy. The network in the first sub-period (2005–2007) comprises above all theoretical publications. The network of 2008–2010 already reflects a different picture. While most theoretical papers are still located in the centre of the network, also 'applied studies' started to appear in the core-documents set. These are distributed at the periphery of the network, which indicates that the topic starts to expand from pure theory to more application. The network of the last sub-period (2011–2013) reflects the clearest structure, where we could distinguish several sub-networks. As the most stable contributor, Egghe's six papers are found in one strongly interlinked sub-network, with the most theoretical roots. Unlike the network in 2008–2010, where some 'applied studies' were still scattered at the periphery of the network, we found more distinct sub-networks on 'applied' research in the network in 2011–2013. In this sense, core documents appear to follow the trend of the topic that is moving away from 'hard-core' informetrics towards research evaluation at different levels of aggregation and for various purposes.

Discussion and Conclusions

In the present study we focussed on 'core documents' with their evolution in publication networks using the example of a specific but nonetheless heterogeneous paper set. The two applied methods proved robust and representative. Their coverage amounted to about 2% of the topic literature, which is in line with the expectations (cf. Glänzel, 2012) but their links lead to related documents that represent a much broader coverage of the topic h-related literature.

The evolution of the core-document network represents the general tendency of shifts in topic, authors and application in an adequate manner. This gives also evidence that Hirsch-type indices have become a tool that is used also outside the informetric community.

Acknowledgments

Lin Zhang acknowledges the NFSC Grant No 71103064, Fred Ye acknowledges the NFSC Grant No 71173187 and Jiangsu Key Laboratory Fund for financial support.

References

Batagelj, V., & Mrvar, A. (2003). *Pajek–Analysis and visualization of large networks*. In M. Jünger & P. Mutzel (Eds.), Graph drawing software (pp. 77-103). Berlin: Springer.

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2), 195–221.

Glänzel, W. & Bart Thijs, B. (2011). Using 'core documents' for the representation of clusters and topics. *Scientometrics*, 88(1), 297-309.

Glänzel, W. (2012). The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics*, 93(1), 113–123.

Small, H. (1973). Cocitation in scientific literature – new measure of relationship between 2 documents. *JASIS*, 24(4), 265-269.

Zhang, L., Bart Thijs, B. & Glänzel, W. (2011). The diffusion of H-related literature. *Journal of Informetrics*, 5(3), 583-593.

How Related is Author Topical Similarity to Other Author Relatedness Measures?

Kun Lu[1], Yuehua Zhao[2], Isola Ajiferuke[3] and Dietmar Wolfram[2]

[1] *kunlu@ou.edu*
School of Library and Information Studies, University of Oklahoma, 401 West Brooks, Norman, OK 73019 (United States)

[2] *{yuehua,* dwolfram}*@uwm.edu*
School of Information Studies, University of Wisconsin-Milwaukee, P.O. Box 413, Milwaukee, WI 53201 (United States)

[3] *iajiferu@uwo.ca*
Faculty of Information and Media Studies, University of Western Ontario, London, ON N6A 5B7 (Canada)

Abstract
Using a dataset of 26,228 Psychology document surrogates from Elsevier databases, we compare author relatedness measure outcomes for 125 authors based on topic modelling to more traditional approaches that rely on direct citation, co-citation and collaboration. Outcomes for the author topical similarity measure are compared to existing co-authorships in the dataset using UCINET/NetDraw. We demonstrate how author topical similarity outcomes provide a similar, but more complete, picture of author relationships than the co-authorship network. Nonparametric correlation analysis results of author topical similarity, co-authorship, citation, and co-citation were also compared for thirty author pairs of differing author topical similarity values. There is a significant correlation between author topical similarity and co-authorship and direct citation-based measures for high similarity author pairs, but not with co-citation measures. The author topical similarity measure, therefore, may serve as a reasonable predictor of collaboration or direct citation for authors with high topical similarity. The measure may also identify potential collaborators based on high author pair similarity values, where there is a lack of existing collaboration, and serve as a complement to author relatedness based on co-citation analysis.

Conference Topic
Methods and techniques

Introduction
Understanding the relationships between authors is of great interest to researchers in scholarly communication and informetrics. Author relatedness can be revealing of the membership of research communities and potentially hidden similarities among authors that may not be readily apparent. The relatedness of authors is a multi-faceted concept that can be determined from different data sources, which include direct author citations, author co-citations (White & McCain, 1998), author bibliographic coupling (Zhao & Strotmann, 2014), author topical similarities (Lu & Wolfram, 2012), author collaborations (Glänzel & Schubert, 2005), and other derived measures (Jacobs & Wolfram, 2014; Jeong, Song, & Ding, 2014). These measures can be discursively categorized into three groups: citation-based (author citation or co-citation), content-based (author topical similarities), and collaboration-based measures (co-authorship). Among developed measures, citation-based measures, especially based on author co-citation analysis, are most influential and well-studied in the literature. The emergence of topic modelling techniques (Rosen-Zvi et al., 2010) has reheated the interest in content-based measures. Co-authorship has been widely used to understand scientific collaborations and reveal research communities. It is well understood that these measures focus on different aspects of author relationships and reveal different types of relatedness. However, the inter-relationships among the different measures have been rarely researched. Are authors with higher topical similarities more likely to collaborate with each other? Do they tend to cite

each other more often? Are they more likely to be co-cited by others? These questions are not adequately addressed in the literature. The purpose of this study is to examine the inter-relationships among several measures of author relationships, including citation-based measures, content-based measures, and collaboration-based measures. More specifically, the research aims to address the following questions:

1) Does author relatedness assessed by author topical similarity reveal similar relationships as a more traditional assessment approach based on co-authorship?
2) What is the relationship between author citation, author co-citation, author collaboration and author topical similarity?
3) Can author topical similarity be used as a predictor for other relatedness measures such as author collaboration, author direct citation or author co-citation?

Author topic modelling (Rosen-Zvi et al., 2010) will be used to determine author topical similarity using bibliographic records for the field of Psychology. Understanding the inter-relationships among the different author relatedness measures contributes to the better use of them in revealing scientific structures.

Literature Review

The present study builds on existing research examining author similarity comparison by employing topic modelling techniques and comparing outcomes to citation and co-citation-based measures.

Measuring the relatedness between scientific entities (e.g. articles, authors, and journals) has been studied for years. Typically, most similarity measures between units are based on quantifiable assessments arising from citation practices that link authors or through direct collaboration or other co-occurrence similarities (Börner, Chen, & Boyack, 2003). To date, the relatedness or similarity between authors has been investigated mainly through five perspectives: direct citation, bibliographic coupling analysis, co-citation analysis, co-authorship analysis, and co-word analysis. Direct citation relationships are built on citation behaviour when one author cites others' work (Boyack & Klavans, 2010). Bibliographic coupling relationships are measured by counting the same references two authors share in their publications and have been studied recently by Zhao and Strotmann (2008, 2014). Moreover, the most widely studied approach, co-citation analysis, assesses the association between two authors by the frequencies they were co-cited by others (White & McCain, 1998). A co-authorship relationship results from a direct collaboration (Glänzel & Schubert, 2005). Each of these methods relies on an explicit connection arising from citation or collaboration. Without these connections, no relationship can be identified. Implicit relationships can be revealed by comparing the content of documents authors have published. Until recently, this has taken the form of co-word analysis, where words or index terms from documents are used to determine how closely related entities of interest are (e.g., Law & Whittaker, 1992).

Although previous studies have used content-based methods to approach the relationships between authors, documents and disciplinary areas, topic-based methods have rarely been applied to date to capture the relationships between authors (Lu & Wolfram, 2012). Topic modelling seeks to automatically reveal the latent topics from a set of documents through machine learning. Hofmann (1999) first proposed a generative data model—called the Probabilistic Latent Semantic Indexing (PLSI)—that represented each document as a probability distribution over a set of topics. While Hofmann's work provided some advantages for document indexing, it may lead to serious problems of overfitting (Blei, Ng, & Jordan, 2003). To overcome the limitations of PLSI, Blei, Ng, and Jordan (2003) presented a three-level hierarchical Bayesian model, which is known as Latent Dirichlet Allocation (LDA). In the LDA model, each document is modelled as a finite mixture over an underlying

set of topics, where each topic is modelled as a mixture over an underlying set of terms (Blei et al., 2003). Follow-up efforts to extend content-level LDA modelling have been investigated using different approaches, such as the Author-Conference-Topic (ACT) model (Tang et al., 2008), correlated topic model (CTM) (Blei & Lafferty, 2006), interactive topic modelling (Hu, Boyd-Graber, Satinoff, & Smith, 2014), and supervised Latent Dirichlet Allocation (sLDA) (Mcauliffe & Blei, 2008). Most topic modelling studies explored the relationships between documents and topics. However, few studies have employed topic modelling methods to conduct author similarity comparison. The present study explores how topic modelling-based author relatedness assessment may complement existing methods based on citation and collaboration-based measures.

Method

Data collection

Elsevier, Inc. has provided a dataset consisting of selected data for 56,620 bibliographic records from 118 Elsevier Arts & Humanities journals. Initially, the authors explored the use of all the data, representing many disciplines within the humanities and social sciences. Outcomes using the author topic modelling approach outlined below resulted in inclusive topical assignments, likely due to the broad vocabulary represented that resulted in topical assignments that combined terms from different disciplines. The subset of the data assigned with Scopus subject classification code 3200, corresponding to "Psychology (all)", was used in this study. The Psychology subset represented the most frequent field appearing in the dataset.

The Psychology subset includes bibliographic records of 26,228 publications written by 63,695 different authors. The authors were identified using the *author_id* field included with the data. An Author-Topic LDA model (Rosen-Zvi et al., 2010) was trained on the title and abstract fields of the psychology subset. The number of topics (k) was set to 100 for exploratory purposes. Other parameters of the model were set as follows: alpha equals 0.5 ($50/k$), beta equals 0.01 and the number of iterations is 1000. All terms were normalized to lower case before processing. A standard list of English stop words were removed and Porter stemming was applied when processing the text. The descriptive statistics of the psychology subset are provided in Table 1. The document length is measured by the number of word tokens in the title and abstract after removing stop words (i.e. common words that were excluded). During the process, we found some authors were listed multiple times in an article because of their multiple affiliations. This was counted as one occurrence in the study.

Table 1. Descriptive statistics of the psychology subset (title and abstract fields).

Measure	*Frequency/Value*
# of documents	26,228
# of unique authors	63,695
Avg. document length	176.98
Title terms	349,410
Abstract terms	4,292,509

Author topical similarity measure

The author topical similarity measure is adopted from the topic-based author relatedness measure proposed by Lu and Wolfram (2012). The measure uses the cosine similarity between Author-Topic vectors from the training results of the Author-Topic modelling as the topical similarity between authors. Given the topic features of the Author-Topic modelling,

the author topical similarity measure is able to identify topical similarity even when the terms do not match. As is the case in any other probabilistic model, the Author-Topic modelling does not work well for authors with a limited number of publications. To ensure the quality of the topical similarity measure, we focused on authors with at least 10 publications in the Psychology subset. Higher cutoff values for the number of papers resulted in smaller numbers of authors for comparison. The cutoff of 10 papers resulted in 125 authors and 7750 author pairs. Table 2 provides descriptive statistics of the author topical similarities between the 125 prolific authors in the psychology subset.

Table 2. Descriptive statistics of the author topical similarity values (author publication count ≥ 10).

Measure	*Value*
# of author pairs	7,750
Mean	0.108
Standard deviation	0.166
Minimum	0.003
Maximum	0.997
Median	0.049

The similarity measures for the 7,750 author pairs were mapped using UCINET 6.0/NetDraw 2.1 network analysis software (https://sites.google.com/site/ucinetsoftware/home; Borgatti, Everett, & Freeman, 2002) and compared to a co-authorship map for the same authors using the data from the Elsevier Psychology dataset. The software allows the strength of ties between nodes to be represented by line thickness. Because each author topical similarity pair had essentially a non-zero similarity, the mapping of all possible author pairs resulted in an incomprehensible map filled with edges. Another advantageous feature of the software is that the display of edges may be controlled using a cutoff value. To allow the stronger relationships to be represented on the map, a similarity cutoff value of 0.5 was selected. The use of a cutoff value did not remove any data in the similarity calculation. It affected only the display of edges between author pairs by removing the edges for author similarity values below the cutoff value. Other cutoff values could have also been selected based on the strength of similarity sought. The 0.5 cutoff value resulted in 10 of the 125 authors not being included in the generated map. A co-authorship map of the Psychology authors was also generated from the Elsevier data and served as a comparison for similarity using a more commonly used measure of author similarity. There were far fewer co-authorship pairs generated from the dataset resulting in a much larger number of authors being excluded from the map because there were no collaborations present in the dataset to be represented in the map. Also, because the 0.5 cutoff value excluded 10 of the 125 authors, the same 115 authors were included in the co-authorship map. The map for the author topical similarity pairings and co-authorship relationships were compared visually for common groupings and differences.

Sampling and other data collection

To explore how the author topical similarity measure compares to other measures of similarity, a stratified random sample of 30 author pairs that spans the full range of author similarity measures was compared. Three pairs of authors were selected from each 0.1 similarity level stratum. The author topical similarity measure for each of these author pairs was compared to more commonly used similarity assessment measures including co-authorship, co-citation and mutual citations by the author pairs. The Elsevier dataset did not provide citation data and the authors did not have access to Elsevier Scopus. Citations between each author and co-citations were collected manually using Thomson Reuters Web of Science (WoS). Co-authorship data from WoS was also

incorporated because it included possible additional co-authored publications beyond those included in the Elsevier dataset. Nonparametric correlation outcomes were calculated for each measure due to the skewed distribution of the data.

Results

A histogram of the distribution of calculated author topical similarity values appears in Figure 1. Note that a logarithmic scale is used due to the large number of low similarity values. Approximately 25.7% of the similarity values exceed 0.1, and only 4.4% are above 0.5, indicating that high similarity measures may provide good discriminative capacity in distinguishing between author pairs with high and low levels of relatedness.

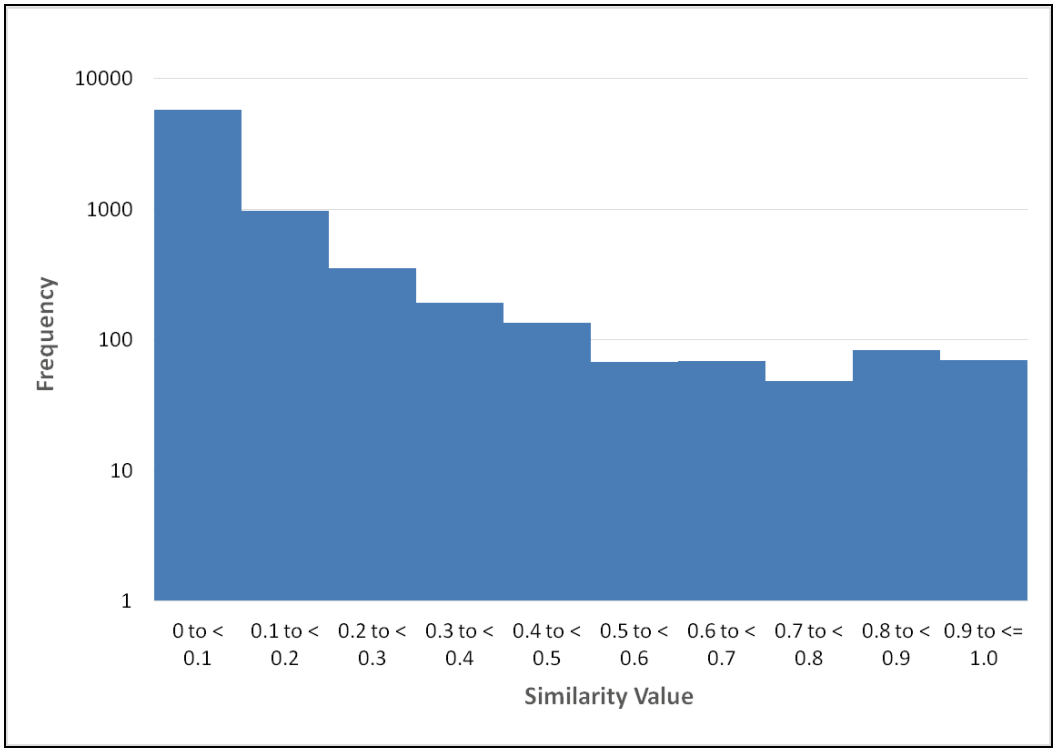


Figure 1. Histogram of calculated author topical similarity values

The UCINET map of the author topical similarity pairings with a 0.5 cutoff value appears in Figure 2. The node sizes and colours highlight comparable numbers of edges where the author similarity values are greater than 0.5. One can see that distinctive clusters of author groups are formed, with two relatively large clusters, a third mid-sized cluster and three smaller clusters with several authors. The large cluster on the right side of the map reveals an author, "Leino-Kilpi H.", who topically serves as a bridge between two parts of the cluster. The topical connection of this author to others in the cluster is missing in the co-authorship maps below due to a lack of collaboration evident in the dataset. The largest node with the greatest number of edges, "Keser H." near the centre of the large cluster to the left, indicates a high level of similarity with a large number of surrounding authors, which is also reflected in Figure 4 below.

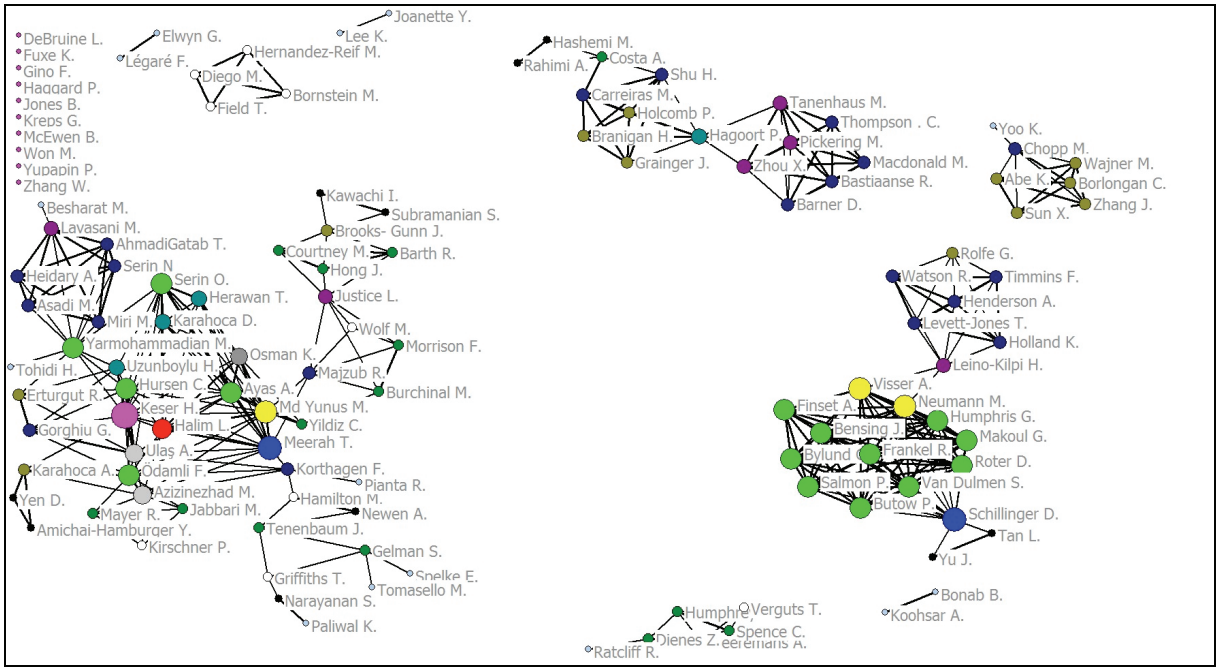


Figure 2. Author topical similarity map (similarity cutoff = 0.5).

Figure 3 summarizes the author collaboration map for the Psychology authors. One can immediately see one drawback of using co-authorship only to assess author relatedness. Fifty-six of the 125 authors were excluded because they did not collaborate with any of the other authors in the dataset. Those connections that do exist are much more limited than for the author topicality similarity outcomes, with two larger clusters and many smaller groups of two to six authors. The members of the two largest clusters in Figure 3 are almost identical to the two largest clusters in Figure 2, but represent only a fraction of the authors that appear in the Figure 2 clusters. Only three of the 10 authors excluded in Figure 2 are included in Figure 3, indicating that these three authors had no topical similarity values above 0.5 with the other authors.

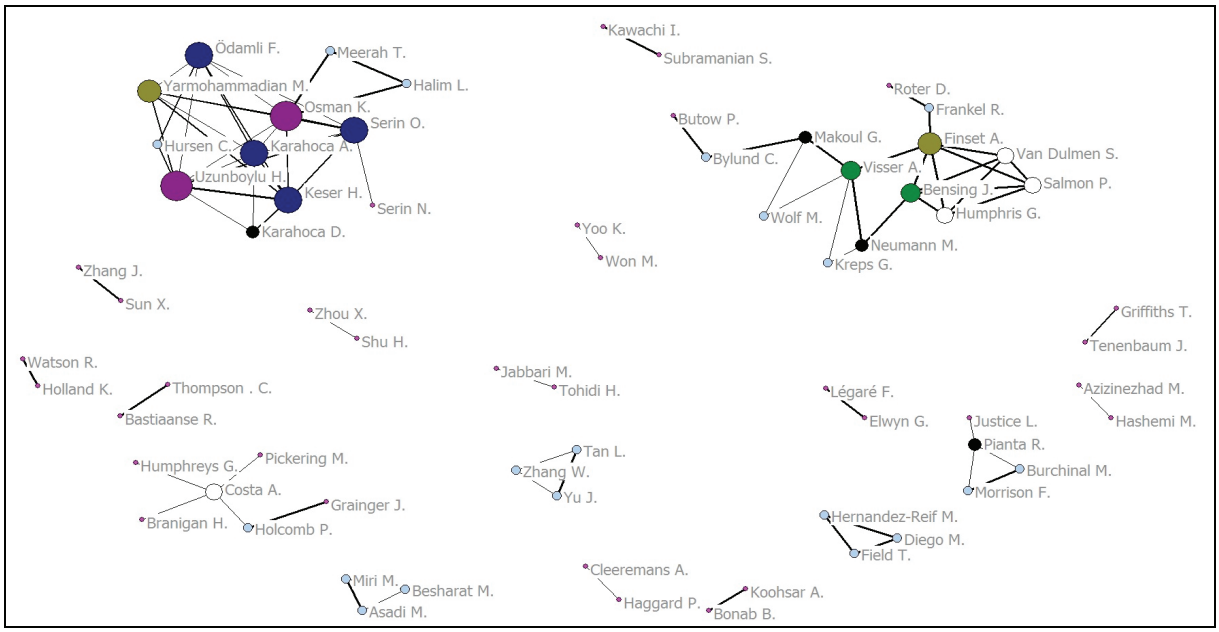


Figure 3. Co-authorship map for all author pairs

The map in Figure 2 shows relationships only for authors with a topical similarity of greater than 0.5. Figure 3 does not take into account the topical similarity of authors. Figure 4

provides the co-authorship map that includes only author pairs with a topical similarity of greater than 0.5. This eliminates a further 21 authors (77 total) from inclusion on the map. It is essentially the same map as Figure 3 flipped along the horizontal axis and, but with fewer edges arising from the removal of the additional authors.

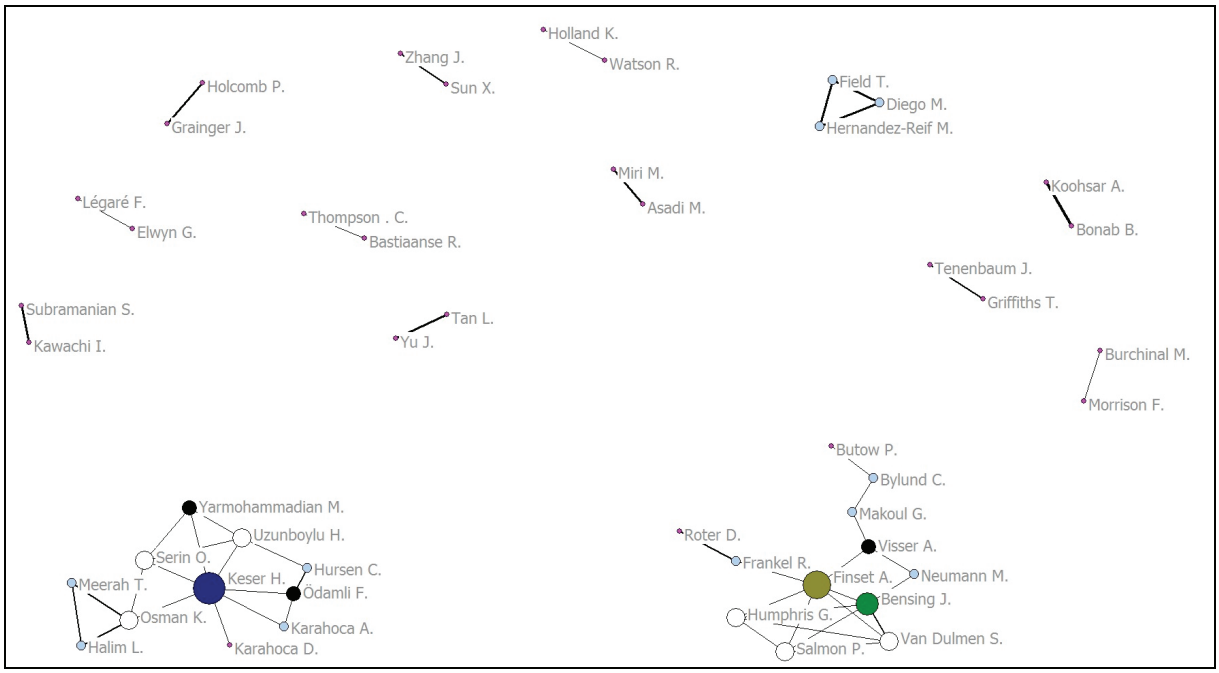


Figure 4. Co-authorship map for author pairs (Author topical similarity cutoff of 0.5)

To examine how the author topical similarity measure correlates to other author relatedness measures, 30 pairs of authors were randomly selected as described above. Outcomes for the author topical similarity, co-authorship, mutual author citing and co-citation values were compared using Spearman's rho nonparametric correlation coefficients (Table 3). There are significant, mid-level correlations observed between the author topical similarity measure and co-authorship for both the Elsevier and WoS data, as well as the mutual citing data for each author. However, there is not a significant correlation with the co-citation counts from WoS. Due to the lack of co-authorship observed for the selected author pairs for author topical similarity values below 0.5, the correlations were also run and included in Table 3 using the 15 similarity values above 0.5 (High) and the 15 values below 0.5 (Low). The positive correlations remained for high similarity author pairs but were not significant for low similarity author pairs.

Discussion

Outcomes of the author topical similarity measure provide a richer method by which author relationships may be mapped and assessed. Unlike co-authorship, direct citation and co-citation networks, where a linkage is created only through collaboration or citation behaviours. The lack of collaboration or citation does not indicate that there is no relationship between two authors; it may simply indicate that the research community has not yet recognized such a relationship. This is most evident when comparing the resulting edges based on author topical similarity and co-authorship. Even when limited to author topical similarity values of greater than 0.5, representing only 4.4% of all possible network connections, the resulting network is rich and demonstrates clusters of author relationships. The richness of the linkages in the resulting network may also be controlled by setting different cutoff values for the author topical similarity. The co-authorship map, conversely, is much sparser and only reveals explicit relationships. The relatively high correlation measure

implies that the author topical similarity measure may serve as a good predictor of existing collaboration. This is more evident for authors with higher topical similarities. Although one would expect there to be a high correlation between collaborating authors, in the Author-Topic model each word is generated from each author according to the author's profile, modelled as a distribution of topics. So, even though collaborating authors tend to be more similar, they may still be generating different words in the titles and abstracts. Excluding co-authored papers in these cases for topic modelling may be attempted, but this could result in less reliable outcomes if the majority of the text on which the models are based is removed.

Table 3. Spearman's rho correlation outcomes for author relatedness measures

		Author Topical Similarity	*Co-authorship Elsevier*	*Co-authorship WoS*	*A Cites B WoS*	*B Cites A WoS*	*Co-citation WoS*
Author Topical Similarity	All	1	.568**	.660**	.452*	.445*	.255
	High	1	.710**	.762**	.694**	.691**	.311
	Low	1	NA	NA	.141	.099	.373
Co-authorship Elsevier	All		1	.816**	.414*	.490**	.415*
	High		1	.812**	.531*	.607*	.573*
	Low		NA	NA	NA	NA	NA
Co-authorship WoS	All			1	.472**	.374*	.336
	High			1	.583*	.398	.492
	Low			NA	NA	NA	NA
A Cites B WoS	All				1	.669**	.587**
	High				1	.812**	.505
	Low				1	.492	.728**
B Cites A WoS	All					1	.536**
	High					1	.451
	Low					1	.650**
Co-citation WoS	All						1
	High						1
	Low						1

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

In the absence of existing collaborations, high author similarity values could serve as an indicator for possible future collaborations. We recognize that the motivations for collaboration are complex and go beyond authors having similar interests. Collaboration may also be prompted by the complementary areas of expertise collaborators bring, which would not be reflected using topic modelling techniques alone. Still, the similarity measure may be used to identify research "birds of a feather" that may not be evident using similarity measures based on collaboration or citation data.

In answer to the research questions posed at the beginning of this paper: 1) mapping of author relatedness based on author topical similarity can reveal a richer network of relationships between authors not evident through a more traditional relationship assessment based on co-authorship and can identify topical bridges; 2) co-authorship, co-citation and mutual citation between authors are significantly correlated, in particular for authors with high topical similarity, so authors with similar topical interests may be more likely to collaborate or cite each other; 3) high author topical similarity values can serve as a reasonably accurate predictor of co-authorship and mutual citation, but not of co-citation activity. The lack of a significant correlation between author topical similarity and co-citation provides evidence that

the topical similarity measure offers a different perspective on author relationships that complements the more traditional co-citation approach. The significant correlations observed between the author topicality similarity and other citation and co-authorship measures indicate that topicality may be a weak to moderately strong predictor of these other more traditional measures for authors with high topical similarity. This positive correlation between author topicality and co-authorship is not unexpected given that co-authored publications would result in more similar topical assignments.

The findings of this study have implications for author relatedness assessment. As the author topical similarity measure does not depend on collaboration or citation behaviour, it can serve as an alternative author relatedness measure where there is a lack of collaboration or citation connections. Even if the collaboration and citation connections exist, the topical similarity measure can provide complementary evidence of relatedness from the content perspective. In addition, the significant correlations between author topical similarity and collaboration shed light on recent developments in predicting and recommending collaborations. Most existing methods for predicting and recommending collaborations are based on the topological features of collaboration networks (Yan & Guns, 2014). The level of correlations between topical similarity and collaboration, particularly for authors with high similarity, provide strong evidence of including content-based predictors for this problem.

Topic modelling offers the ability to reveal relationships between authors that may not be evident through more traditional methods of similarity assessment, but it does have its limitations. The computational overhead associated with topic-based author relatedness modelling is more substantial than for citation and collaboration-based data. Also there must be a sufficient body of text to train the topic model and to accurately represent author relationships; therefore, this method may not be suitable for authors with a more modest publication record. In this case, analysis using citation-based methods may be more fruitful. Other limitations arise from the dataset itself. In identifying works attributable to an author, we have relied on the supplied Scopus author identifier. We recognize that author name disambiguation, regardless of the method used, may not be 100% accurate. In addition, the present study has limited itself to data from a single discipline. Furthermore, the dataset itself was not complete for the discipline of Psychology, but rather a subset. We cannot conclude that the outcomes for other disciplines will be similar. Outcomes would depend also on the collaboration traditions and citing behaviours of those disciplines. The computational overhead and limited ability for topic modelling to be able to produce meaningful topics with multidisciplinary datasets may limit the application of this approach beyond the disciplinary level.

Conclusions

Author topical similarity provides a novel way to assess author relatedness that complements existing methods based on co-authorship, direct citation or co-citation. While other methods require an existing form of connection based on collaboration or citations, author topical similarity assesses author relatedness based on the language used by the authors themselves. The small percentage of author pairs with high similarity values indicates that the measure is discriminating in the assessment of author relatedness. The present study has demonstrated how author relatedness based on topic modelling can provide a richer method to assess how closely related authors' research contributions are. Although significantly correlated with co-authorship and direct citation measures, author topical similarity between authors was not found to be significantly correlated with co-citations, which has been commonly used to assess author relatedness. Author topical similarity outcomes may serve as a reasonably accurate predictor of existing collaborations between authors, or an indicator of potential future collaborators in the absence of existing collaboration. Future research may investigate

how author topical similarity measures compare to other existing author relatedness measures for other disciplinary areas including the humanities and sciences, where collaboration and citation patterns may differ.

Acknowledgments

The authors would like to thank Elsevier, Inc. for providing access to the dataset used to conduct this study.

References

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Processing Systems*, *18*, 147.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). UCINET for Windows: Software for social network analysis. Retrieved from https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=6&ved=0CEMQFjAF&url=https%3A%2F%2Fwww.soc.umn.edu%2F~knoke%2Fpages%2FUCINET_6_User%2527s_Guide.doc&ei=WqatVLfOKtP3ggTgq4OIAg&usg=AFQjCNF_1umvC9bg07zqAb969AG7WJX5qw&cad=rja

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, *37*(1), 179–255.

Boyack, K.W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404.

Glänzel, W. & Schubert, A. (2005). Analysing scientific networks through co-authorship. In *Handbook of Quantitative Science and Technology Research* (pp. 257-276). Netherlands: Springer.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 50–57). New York, NY, USA: ACM.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning*, *95*(3), 423–469.

Jacobs, D., & Wolfram, D. (2014). Exploring author similarity using citing discipline analysis. In *Proceedings of the Annual Conference of CAIS/ Actes du congrès annuel de l'ACSI*. Retrieved from: http://www.cais-acsi.ca/ojs/index.php/cais/article/download/892/812.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics*, *8*(1), 197-211.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics*, *23*(3), 417–461.

Lu, K., & Wolfram, D. (2012). Measuring author research relatedness: A comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, *63*(10), 1973-1986.

Mcauliffe, J. D., & Blei, D. M. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems* (pp. 121–128). Retrieved from http://papers.nips.cc/paper/3328-supervised-topic-models

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, *28*(1), 1-38.Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 990–998). New York, NY, USA: ACM.

Tang, J., Jin, R., & Zhang, J. (2008). A topic modeling approach and its integration into the random walk framework for academic search. In F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan, & X. Wu (Eds.), *Eighth IEEE International Conference on Data Mining ICDM'08*. (pp. 1055-1060). IEEE.

White, H.D., & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4)*,* 327-355.

Yan, E., & Guns, R. (2014). Predicting and recommending collaborations: An author-, institution-, and country-level analysis. *Journal of Informetrics, 8*(2)*,* 295-309.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in Information Science 1996-2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070–2086.

Zhao, D., & Strotmann, A. (2014). The knowledge base and research front of information science 2006–2010: An author cocitation and bibliographic coupling analysis. *Journal of the Association for Information Science and Technology*, *65*(5), 995-1006.

Publication Rates in 192 Research Fields of the Hard Sciences

Ciriaco Andrea D'Angelo[1] and Giovanni Abramo[2]

[1] *dangelo@dii.uniroma2.it*
Department of Engineering and Management
University of Rome "Tor Vergata" – Italy, Via del Politecnico 1, 00133 Rome (Italy)

[2] *giovanni.abramo@uniroma2.it*
Laboratory for Studies of Research and Technology Transfer, Institute for System Analysis and Computer Science (IASI-CNR), National Research Council of Italy, Via dei Taurini 19, 00185 Rome (Italy)

Abstract
Bibliometricians are aware that the citation behavior of scientists varies across fields, and for this they carefully normalize citations by field. They are also aware of the different publication intensities across fields. This imposes that the research performance of a scientist must be compared with that of their colleagues in the same field. Every comparison of scientists in different fields should be preceded by the normalization of the performances, and the same holds for comparing multidisciplinary organizational units. If the Web of Science recognizes 251 subject categories, there should be a somewhat similar number of research fields for the classification of the scientists. The Italian academic system is quite unique in providing a classification of professors, into 370 fields, 192 of them in the hard sciences. In this work we measure the descriptive statistics on annual publication (full and fractional counting) by Italian academics in each of the 192 hard science fields. These statistics help recognize the extent of distortion from failing to normalize the research performance of scientists based in different fields. They could also serve as scaling factors for avoiding distortion in rankings, including in other nations.

Conference Topic
Methods and techniques

Introduction
The purpose of bibliometrics is to provide continuously better support for the policy-makers and administrators of research institutions, in the achievement of their specific objectives, through the provision of methods and indicators for the evaluation of performance that are themselves always more accurate, robust, reliable and functional. The principle obstacle to bibliometrics is the insufficiency of the data to meet such high standards. The practitioner is thus forced to resort to proxies in measurement, which cause varying degrees of distortion in the results.

Research organizations are likened to other productive organizations, but where the product is new knowledge, rather than some other good or service. An organization's performance is then better than that of another one if, at parity of resources, it produces more knowledge or if, at parity of output, it consumes less resources. It is the shortage of information on inputs (production factors) that presents the greatest problem to bibliometricians. The production factors are labor and capital. Capital embeds all those resources other than labor (facilities, technical instruments, materials, databases, etc.). When we wish to measure labor productivity we must thus normalize for capital. But who can really know the financial and technical resources available to all the different institutions, departments, and then individual researchers? The bibliometrician also frequently lacks information on the realities of labor, due to the absence of databases on the researchers, and on their institutional, discipline and field affiliations.

Given these obstacles, practitioners often use indicators that do not relate output to input. This means they produce ranking lists that are highly size-dependent. At that point we cannot

know what part of an organization's or nation's rank arises from its performance or is due to size. Examples of this are the CWTS Leiden[1] and SCImago[2] lists, which rank universities by publications and fractional publications. Others have proposed indicators that attempt to get around the problems by relating the impact or excellence of research not to input, but rather to the output itself. Examples of this are the "new crown indicator" (Waltman et al., 2011), which measures the average impact per publication, or the "proportion of highly-cited articles to total publications" (Waltman et al., 2012). However, with this type of indicator, even when the output of the scientist increases, other factors remaining equal, his or her performance could still decrease: a paradox and a violation of the fundamental principle of the measure of efficiency.

In those cases where an indicator does relate output to input, it is still often applied at levels of organizational aggregation that are too high, ignoring the differing intensity of publication across fields. Bibliometricians have been aware of this problem for many years (Butler, 2007; Moed et al., 1985; Garfield, 1979), and are also aware of the distortion that afflicts the resulting aggregate rankings (Abramo, D'Angelo, & Di Costa, 2008). However the task of finer aggregation is difficult to solve without a database that classifies the researchers by field of research. Where they exist, such databases are maintained at central levels. Apart from the Italian one[3], maintained by the Italian Ministry of Education, Universities and Research (MIUR), the only other large-scale one we are aware of is the Norwegian Research Personnel Register[4] compiled by the Nordic Institute for Studies in Innovation, Research and Education (NIFU).

The NIFU system classifies scientists in 58 scientific fields grouped in five main domains. Perhaps the lower number of scientists in Norway works against finer classification: in fact comparing the performance of small numbers of researchers per field creates serious problems of significance. However, on the other hand, the Web of Science (WoS) identifies a full 251 subject categories for the classification of journals. And if there are this many fields for classifying scientific journals, there must be at least that many fields for classifying scientific work, and the scientists. In smaller nations or emerging economies we could expect to see fewer number of these fields present, since research structures will be unable to deal with all the areas, and we would expect to see research in more concentrated fields. However, in larger, developed countries we can expect to see the full spectrum of research fields. In fact in Italy the MIUR manages a system for the classification of all professors into a total of 370 "scientific disciplinary sectors" (SDSs).[5] Each professor belongs to one and only one of the SDSs, which are grouped into 14 university disciplinary areas (UDAs). Further, 192 of the SDSs from 9 of the UDAs fall in the so-called hard sciences. In the following we refer to these SDS by their code or acronym.[6] These 192 SDSs compare to the 176 WoS subject categories identified in the JCR-Science Citation Index (see the Annex 1[7] for a conversion of SDSs to WoS subject categories.

As noted above, the lack of field classification of scientists means that measures of research performance will inevitably be affected by distortions in rankings, due to the different intensity of publication across fields. The higher the level of aggregation, the stronger these distortions become. The corollary is that, rising to international levels, it has been impossible

[1] http://www.leidenranking.com/ranking/2014, last accessed on April 8, 2015.
[2] http://www.scimagoir.com/research.php, last accessed on April 8, 2015.
[3] http://cercauniversita.cineca.it/php5/docenti/cerca.php, last accessed on April 8, 2015.
[4] http://www.nifu.no/en/statistikk/databaser-og-registre/4897-2/ last accessed on April 8, 2015.
[5] The complete list is accessible on attiministeriali.miur.it/UserFiles/115.htm, last accessed April 8, 2015.
[6] The full names can be found in www.iasi.cnr.it/laboratoriortt/TESTI/Altro/ISSI-ANNEX%202_P.pdf, last accessed on April 8, 2015
[7] www.iasi.cnr.it/laboratoriortt/TESTI/Altro/ISSI-ANNEX1.pdf, last accessed on April 8, 2015

to correctly compare institutional or national research performance.

To date, in fact there is no international standard for the classification of scientists. Thus in this work we provide our colleagues and practitioners with descriptive statistics on yearly publications (both full and fractional counting) of Italian academics in each of the 192 hard science SDSs. Our intention is that these statistics might first permit recognition of the extent of distortions that occur when evaluations compare the research performance of scientists within the same discipline, but in different fields. For those nations lacking databases of researchers by field, our statistics could also serve as normalization factors, serving to reduce the distortions when comparing research performance of individuals, groups or entire research organizations.

Data and Methods

In the study we measure "publication rates" in 192 SDSs, meaning average yearly publications of individual scientists, over the period 2009-2013.[8] Data on Italian academics are extracted from the official database maintained by the MIUR. The database indexes the name, academic rank, affiliation, and SDS of all academics in Italian universities. At 31/12/2013 the entire Italian university population consisted of 56,600 professors employed in 96 universities, which are authorized by the MIUR to grant legally recognized degrees. It has been shown (Moed, 2005) that in the so-called hard sciences, the prevalent form of codification for research output is publication in scientific journals. For reasons of robustness, we thus examine only the nine UDAs that deal with the hard sciences,[9] including a total of 192 SDSs. Furthermore, again for reasons of robustness, we calculate the yearly average publication rates only of those professors who have been on staff for at least three years over the observed period.

Table 1. Dataset for the analysis: number of fields (SDSs), universities, research staff and WoS publications in each UDA under investigation

UDA	SDS	Universities	Research staff	Publications*
Mathematics and computer science	10	72	2,930	16,262
Physics	8	65	2,003	22,597
Chemistry	12	60	2,701	26,054
Earth sciences	12	49	974	6,066
Biology	19	67	4,423	34,406
Medicine	50	65	8,998	72,661
Agricultural and veterinary sciences	30	56	2,820	14,951
Civil engineering	9	54	1,394	7,462
Industrial and information engineering	42	73	4,791	40,572
Total	192	86	31,034	207,132†

* Figures refer to publications authored by at least one professor pertaining to the UDA.

† Total is less than the sum of the column data due to double counts of publications co-authored by researchers pertaining to SDSs of more than one UDA.

Publication data are drawn from the Italian Observatory of Public Research (ORP), a database developed and maintained by the authors and derived under license from the WoS. Beginning from the raw data of Italian publications[10] indexed in WoS-ORP, we apply a complex

[8] For the most appropriate publication period to be observed see Abramo et al. (2012b).

[9] Mathematics and computer sciences; Physics; Chemistry; Earth sciences; Biology; Medicine; Agricultural and veterinary sciences; Civil engineering; Industrial and information engineering.

[10] We exclude those document types that cannot be strictly considered as true research products, such as editorial material, meeting abstracts, replies to letters, etc.

algorithm for disambiguation of the true identity of the authors and their institutional affiliations (for details see D'Angelo et al., 2011). Each publication is attributed to the university professors that authored it, with a harmonic average of precision and recall (F-measure) equal to 96 (error of 4%). We further reduce this error by manual disambiguation. Because each professor belongs to one and only one SDS, we can then calculate the distribution of annual publication rates and the relevant descriptive statistics in each SDS.

The dataset for the analysis includes 31,034 professors, employed in 86 universities, authoring over 200,000 WoS publications, sorted in the UDAs as shown in Table 1.

Research projects frequently involve a team of researchers, a fact revealed in the co-authorship of publications. Various performance measures account for the fractional contributions of single co-authors to outputs. The contributions of the individual co-authors to the achievement of the publication are not necessarily equal, and in some fields the authors signal the different contributions through the ordering of the byline. The conventions on the order of authors for scientific papers differ across fields (Pontille, 2004; RIN, 2009), thus in the current study, the fractional contribution of the individuals is weighted accordingly.

Fractional contribution equals the inverse of the number of authors, in those fields where the practice is to place the authors in simple alphabetical order but assumes different weights in other cases, particularly in the life sciences. For these disciplines, we give different weights to each co-author according to their order in the byline and the character of the co-authorship (intra-mural or extra-mural). If first and last authors belong to the same university, 40% of citations are attributed to each of them; the remaining 20% are divided among all other authors. If the first two and last two authors belong to different universities, 30% of citations are attributed to first and last authors; 15% of citations are attributed to second and last author but one; the remaining 10% are divided among all others.[11] Failure to account for the number and position of authors in the byline would result in notable ranking differences, both at the individual level (Abramo, D'Angelo & Rosati, 2013a) and at the institution level (Abramo, D'Angelo & Rosati, 2013b).

Applying the above conventions, for each of the 192 SDS we will provide descriptive statistics on the intensity of annual publication: referred to as P for full counting and FP for fractional counting. We then examine further statistics on P and FP for the SDSs included in each UDA.

Results

Publication rates of professors in a specific field

The publication intensity of professors in a given field is known to be particularly skewed, with a small percentage of individuals authoring a large share of the total papers, and the others authoring a small share (Egghe, 2005; Kyvik, 1989; Lotka, 1926). Figure 1 provides the example of the field of Organic chemistry (SDS CHIM/06), showing the distribution of the average number of publications per year over the period under examination, for each of the 554 professors in the SDS. The distribution fits quite well a logarithmic curve, as indicated by the particularly high value of R^2 (0.974). Here, 10% of the professors have produced on average less than one publication per year, and six were totally unproductive. On the opposite front, we find 20 professors with over 10 publications per year, and one absolute outlier with 25.

The box plot (right side of Figure 1) refers to the same distribution. It shows a median of 3 publications per year and an interquartile range (difference between third and first quartile) of

[11] The weighting values were assigned following advice from senior Italian professors in the life sciences. The values could be changed to suit different practices in other national contexts.

2.6. It also brings out the presence of 30 outliers: hyper-productive professors with a performance that exceeds that of the third quartile by over 1.5 times the interquartile difference.

The distribution of frequencies by class of publication rates (Figure 2) shows a mode between 2 and 3 publications annually and a particularly long right tail, with a final peak for the hyper-productive professors.

The distribution of the average yearly publications measured by fractional counting (FP) shows a very similar situation: in Figure 3 the right tail is actually longer than that for only full counting (Figure 2).

The distributions seen for SDS CHIM/06 show structural elements that recur in the analyses of the other 191 SDSs. Most obvious is the skewness, although there are some interesting exceptions, for example as in VET/04 (Inspection of food products of animal origin). The 77 professors of this SDS have a publication rate that is almost uniform, as illustrated in Figure 4.

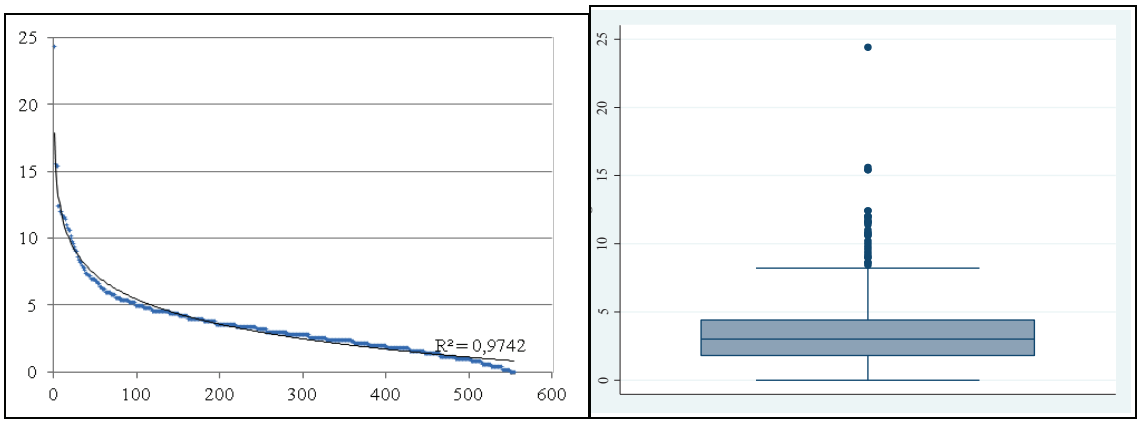


Figure 1. Distribution and box plot of annual publication rate P (full counting, 2009-2013) for 554 Italian professors in Organic chemistry (CHIM/06).

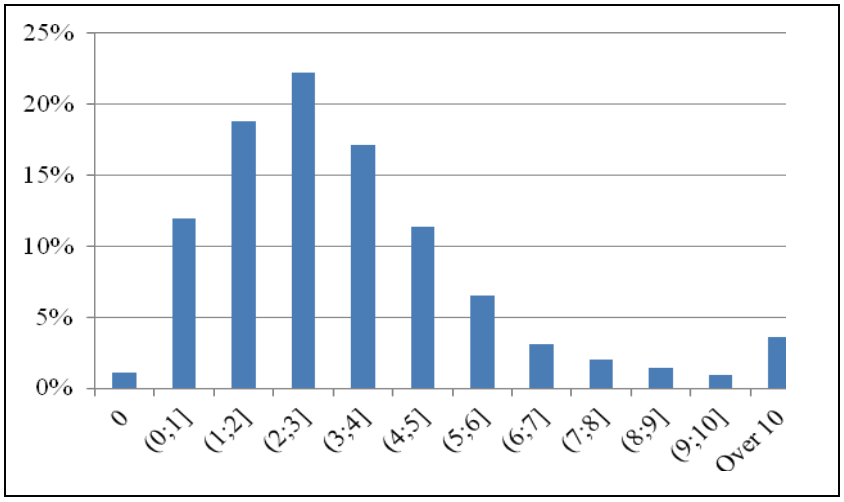


Figure 2. Frequency distribution for classes of annual publication rate P (2009-2013) for the 554 Italian professors in CHIM/06.

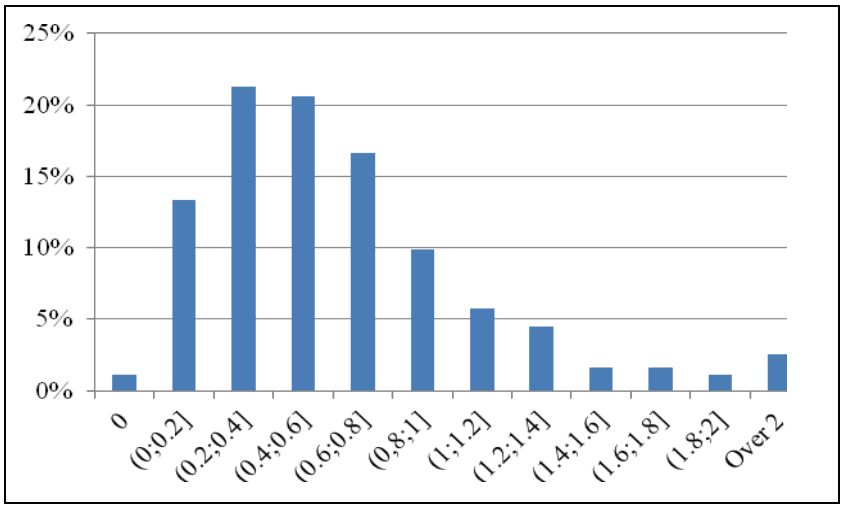


Figure 3. Frequency distribution for classes of annual publication rate FP (fractional counting, 2009-2013) for the 554 Italian professors in CHIM/06.

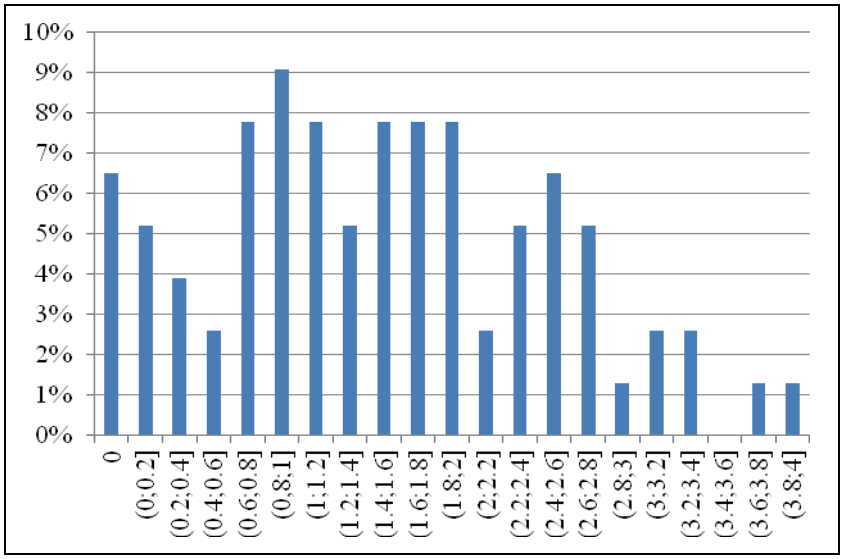


Figure 4. Frequency distribution for classes of annual publication rate P (2009-2013) for Italian professors in Inspection of food products of animal origin (VET/04).

Publication rates of fields within a discipline

As with the two examples above (CHIM/06 and VET/04), the publication rates in the various SDSs are never superimposable. Thus the calculation of the descriptive statistics for the SDSs provides useful benchmarks for the professors that work in them. Table 2 provides the statistics for all the SDSs in the Earth sciences discipline.

This UDA consists of a total of 12 SDSs with very different sizes in terms of national research staff, from a minimum of 17 professors in Applied geophysics (GEO/12) to a maximum of 137, in Palaeontology and palaeoecology (GEO/02). The intensity of publication is structurally very different. In Stratigraphic and sedimentological geology (GEO/03) only 2.2% of the professors (2 of 92) did not produce any publications over the five-year period under examination. On the opposite front there are 19 unproductive professors among the 121 of Physical geography and geomorphology (GEO/05), or 15.7% of the total. This SDS also registers the lowest average annual rate of publication, at 1.12 per year, followed by Structural geology (GEO/04), GEO/02 and Geophysics of solid earth GEO/11 (1.44, 1.48 and 1.49, respectively). In half the SDSs there is an average intensity of publication of 2 per year,

with a peak in Applied geology GEO/06 (3.09). Clearly, among all those of the UDA, this SDS has the greatest publication rate: the distribution of the performances shows all values in the highest quartiles. The top 25% of professors (3rd quartile) produce on average more than 4 publications per year, with the absolute record being a professor who produces almost 18. The dispersion of the performances in all the SDSs, indicated by the variation coefficients in the last column of Table 2, results as greatest in GEO/03 and GEO/05, where the coefficient is above 1.

The analyses of the distributions for fractional counting of the publication rate (FP) (Table 3) provide a picture similar to that for full counting. The average intensity of collaboration evidently does not vary in a substantial way between the SDSs, and thus the differential of publication rates between the SDSs does not vary in going from a full counting approach to fractional counting.

Table 2. Descriptive statistics for intensity of annual publication rate P (2009-2013) for the SDSs of Earth sciences.

SDS	Research staff	Unproductive	I quartile	Median	III quartile	Max	Average	Std dev.	Variat. coeff.
GEO/01	93	3.2%	0.8	1.6	2.2	8	1.76	1.40	0.80
GEO/02	137	7.3%	0.6	1	2.20	6.4	1.48	1.25	0.84
GEO/03	92	2.2%	1	1.8	2.8	22	2.40	2.69	1.12
GEO/04	116	6.9%	0.6	1	2	4.8	1.44	1.21	0.84
GEO/05	121	15.7%	0.2	0.8	1.4	8.2	1.12	1.22	1.09
GEO/06	76	1.3%	1.55	2.6	4.05	17.8	3.09	2.51	0.81
GEO/07	82	2.4%	1	1.8	2.75	8.2	1.99	1.46	0.73
GEO/08	67	3.0%	1.3	2.4	3.5	10.6	2.69	2.03	0.75
GEO/09	63	6.3%	0.8	1.8	2.9	11.4	2.21	2.04	0.92
GEO/10	69	4.3%	1.2	1.8	2.4	10.2	2.14	1.82	0.85
GEO/11	41	2.4%	0.6	1.2	2	5.6	1.49	1.12	0.75
GEO/12	17	5.9%	0.8	1.6	2	4.6	1.75	1.34	0.77

Table 3. Descriptive statistics for intensity of annual publication rate FP (2009-2013) for the SDSs of Earth sciences

SDS	I quartile	Median	III quartile	Max	Average	Std dev.	Variat. coeff.
GEO/01	0.20	0.33	0.53	2.61	0.45	0.45	1.00
GEO/02	0.14	0.28	0.45	1.47	0.34	0.29	0.85
GEO/03	0.26	0.43	0.65	2.64	0.53	0.42	0.79
GEO/04	0.14	0.25	0.47	1.81	0.33	0.30	0.91
GEO/05	0.07	0.24	0.42	1.81	0.29	0.31	1.07
GEO/06	0.32	0.56	0.87	3.52	0.71	0.61	0.86
GEO/07	0.20	0.37	0.59	1.62	0.44	0.31	0.70
GEO/08	0.29	0.53	0.72	1.61	0.56	0.39	0.70
GEO/09	0.13	0.39	0.65	3.06	0.48	0.48	1.00
GEO/10	0.29	0.45	0.74	2.44	0.56	0.46	0.82
GEO/11	0.19	0.31	0.61	1.50	0.45	0.38	0.84
GEO/12	0.19	0.32	0.60	0.90	0.38	0.28	0.74

For the descriptive statistics of the full 192 SDSs investigated, we refer the reader to Annex 2[12] for the full counting, and to Annex 3[13] for fractional counting. Below, in Table 4, we show for each UDA the SDSs with minimum and maximum values of some of the above statistics

[12] www.iasi.cnr.it/laboratoriortt/TESTI/Altro/ISSI-ANNEX%202_P.pdf, last accessed on April 8, 2015
[13] www.iasi.cnr.it/laboratoriortt/TESTI/Altro/ISSI-ANNEX%203_FP.pdf, last accessed on April 8, 2015

of P (full counting). The data indicate substantial variability in the intensity of publication between the SDSs in all the UDAs. In Mathematics the percentage of unproductive professors varies from a minimum of 3.9% in MAT/09 (Operations research) and a maximum of 43.2% in MAT/04 (Complementary mathematics). Such substantial variations also occur in Medicine, with 1.1% unproductive professors in MED/08 (Pathological anatomy) and 45.5% in MED/02 (History of medicine). In Agricultural and veterinary sciences, VET/02 (Veterinary physiology) does not have any unproductive professors, while AGR/01 (Rural economics and valuation) registers a share of 45.5%. More contained heterogeneity in unproductive professors is seen in some other UDAs: certainly in Earth sciences, which we have already examined, but also in Biology. In this UDA the maximum incidence of unproductive professors (11.8% of the total professors) is seen in BIO/08 (Anthropology) and the minimum (1.2%) in BIO/15 (Pharmaceutical biology). The median intensity of annual publication also presents high variability between the SDSs of a UDA. In Mathematics the median ranges from 0.2 publications per year in MAT/04 (Complementary mathematics) to 1.8 in MAT/09 (Operations research). In effect the interval of variation of the median values is very substantial in almost all the UDAs. Within Industrial and information engineering, the median intensity of publication registered in ING-INF/06 (Electronic and information bioengineering) and in ING-INF/02 (Electromagnetic fields) is more than 40 times that registered in ING-IND/01 (Naval architecture). In Medicine the two extreme situations concern MED/02 (History of medicine) and MED/16 (Rheumatology): the median intensity of publication registered in the first SDS (0.2) is $1/25^{th}$ of that for the second (5.0). The differences are more contained in Chemistry (2.0 vs. 3.4), Earth sciences (0.8 vs. 2.6) and Biology (1.1 vs. 3.3). The consistency of the outliers is also significantly different between the SDSs of a given discipline. In the Mathematics UDA, the most productive professor in absolute terms is one in INF/01 (Computer science), with an average of 28.6 publications per year, against the 3.6 of the most productive professor in MAT/04 (Complementary mathematics). In Medicine, a professor in MED/24 (Urology) registers a median of 76 publications per year over the five years examined; the most prolific in MED/47 (Nursing and midwifery) has barely 1.4 publications. In Industrial and information engineering the most prolific professor of ING-IND/01 (Naval architecture) authors an average of 1.4 publications annually, against the 33.2 of the most productive in ING-IND/34 (Industrial bioengineering). Finally, Physics FIS/01 (Experimental physics) includes a professor with an average of over 100 publications per year. In effect, this SDS consists of a range of subfields, including "high energy physics", where scientists regularly author hundreds of publications together with hundreds of co-authors. In this case (but not only in this case) a more opportune benchmark could be the distribution of the publication rate under the fractional counting method. Table 5 shows, for every UDA, the SDS with minimum and maximum values of the main statistics[14] of the fractional counting distributions. We see a level of superimposability with the data of Table 4, both in terms of the SDSs featured and for the intervals of variation in the main statistics of the SDSs, for each UDA.

[14] To avoid pointless duplication, the table does not show the incidence of unproductive professors, and instead provides statistics on average publication rate.

Table 4. SDSs with Min and Max values of descriptive statistics of intensity of annual publication P (2009-2013), for all UDAs.

	Unproductive (%)		Median		Max	
UDA*	Min	Max	Min	Max	Min	Max
1	3.9 (MAT/09)	43.2 (MAT/04)	0.2 (MAT/04)	1.8 (MAT/09)	3.6 (MAT/04)	28.6 (INF/01)
2	2.1 (FIS/04)	37.5 (FIS/08)	0.2 (FIS/08)	5.6 (FIS/01)	4.4 (FIS/08)	102.2 (FIS/01)
3	0.0 (CHIM/04)	8.6 (CHIM/11)	2.0 (CHIM/11)	3.4 (CHIM/02)	7.6 (CHIM/12)	66.2 (CHIM/08)
4	1.3 (GEO/06)	15.7 (GEO/05)	0.8 (GEO/05)	2.6 (GEO/06)	4.6 (GEO/12)	22 (GEO/03)
5	1.2 (BIO/15)	11.8 (BIO/08)	1.1 (BIO/02)	3.3 (BIO/15)	6.4 (BIO/08)	37.6 (BIO/12)
6	1.1 (MED/08)	45.5 (MED/02)	0.2 (MED/02)	5.0 (MED/16)	1.4 (MED/47)	76 (MED/24)
7	0.0 (VET/02)	42.0 (AGR/01)	0.2 (AGR/01)	2.8 (VET/06)	3.2 (AGR/06)	32.6 (VET/06)
8	5.8 (ICAR/03)	29.9 (ICAR/06)	0.2 (ICAR/06)	1.6 (ICAR/03)	2.8 (ICAR/05)	21.2 (ICAR/08)
9	0.0 (ING-IND/18)	50.0 (ING-IND/01)	0.1 (ING-IND/01)	4.4 (ING-INF/02 and ING-INF/06)	1.4 (ING-IND/01)	33.2 (ING-IND/34)

* 1 = Mathematics and computer sciences; 2 = Physics; 3 = Chemistry; 4 = Earth sciences; 5 = Biology; 6 = Medicine; 7 = Agricultural and veterinary sciences; 8 = Civil engineering; 9 = Industrial and information engineering

Table 5. SDSs with Min and Max values of descriptive statistics of intensity of annual publication FP (2009-2013), for all UDAs.

	Median		Average		Max	
UDA*	Min	Max	Min	Max	Min	Max
1	0.10 (MAT/04)	0.55 (MAT/09)	0.16 (MAT/04)	0.70 (MAT/07)	1.00 (MAT/04)	6.47 (MAT/02)
2	0.07 (FIS/08)	0.74 (FIS/03)	0.20 (FIS/08)	0.96 (FIS/03)	0.80 (FIS/08)	13.74 (FIS/03)
3	0.35 (CHIM/12)	0.70 (CHIM/02)	0.58 (CHIM/12)	0.83 (CHIM/02)	2.38 (CHIM/12)	17.60 (CHIM/08)
4	0.24 (GEO/05)	0.56 (GEO/06)	0.29 (GEO/05)	0.71 (GEO/06)	0.90 (GEO/12)	3.52 (GEO/06)
5	0.24 (BIO/08)	0.58 (BIO/15)	0.32 (BIO/08)	0.85 (BIO/15)	1.04 (BIO/08)	10.50 (BIO/12)
6	0.01 (MED/02)	0.84 (MED/16)	0.08 (MED/47)	1.18 (MED/16)	0.19 (MED/47)	13.28 (MED/11)
7	0.04 (AGR/01)	0.60 (AGR/15)	0.14 (AGR/01)	0.78 (VET/06)	0.65 (AGR/06)	9.14 (VET/06)
8	0.10 (ICAR/06)	0.48 (ICAR/08)	0.17 (ICAR/06)	0.73 (ICAR/08)	1.27 (ICAR/05)	6.85 (ICAR/08)
9	0.03 (ING-IND/01)	1.08 (ING-INF/02)	0.10 (ING-IND/01)	1.28 (ING-INF/02)	0.54 (ING-IND/02)	9.18 (ING-IND/19)

* 1 = Mathematics and computer sciences; 2 = Physics; 3 = Chemistry; 4 = Earth sciences; 5 = Biology; 6 = Medicine; 7 = Agricultural and veterinary sciences; 8 = Civil engineering; 9 = Industrial and information engineering

Conclusions

The great majority of the bibliometric indicators and the relative rankings lack fine-grained normalization of performance to the field to which the scientists belong. While bibliometricians intelligently field-normalize citations to account for the different citation behaviors across fields, they often close an eye when it comes to accounting for the different intensity of publication. At most they distinguish scientists as belonging to a few large disciplines, which cannot be sufficient if we accept the WoS as a true characterization, where scientific work is distinguished in 251 subject categories. Why would we normalize the citations for these 251 subject categories but then the scientists' performance for only a few disciplines? The answer is simple: in most cases the bibliometricians lack information about the field of research of each scientist under observation. Even at the national level the challenge of identifying the scientist's field is daunting, let alone for the task of international comparison.

Taking advantage of a particular feature of the Italian academic system, in this work we have provided descriptive statistics on the yearly publication rates of all Italian professors (over 30,000) in each of the 192 hard sciences fields, with both full and fractional counting method. Although the dataset refers to a specific nation, the very substantial size and the fine-grained field stratification certainly make it a useful reference system for the comparative evaluation of scientists in all the world. The only condition is that scholars recognize in which field of the Italian system the core of their scientific production falls. To this aim, in the Appendix, we have provided the reader with a conversion table, which establishes a link between SDSs and WoS subject categories, based on incidence of publications authored by Italian academics. Through this link, scientists outside Italy, knowing the distribution of their scientific production in the subject categories, can identify the corresponding SDS and select relevant statistic parameters as benchmark for comparative evaluation of their publication rates.

The statistics from the current analyses very clearly demonstrate the heterogeneity of publication rates even in the fields belonging to a single discipline. They help recognize the extent of distortions that occur when comparing the research performance of scientists from different fields, and could then serve as normalization factors to reduce such distortions when comparing the research performance of individuals, groups, or entire research organizations.

In future extensions of this work we could envisage a longitudinal analysis to assess the trends in publication intensity by field. We also know that publication rates of full, associate and assistant professors are different (Abramo, D'Angelo, & Di Costa, 2011). Gender differences in productivity have been demonstrated as well (Abramo, D'Angelo, & Caprasecca, 2009; Leahey, 2006; Fox, 2005; Pripić, 2002; Long, 1992). Because the composition of research staff by academic rank and gender varies across fields, a further extension of the analysis may then entail examining the differing publication intensity across fields by academic rank and gender.

References

Abramo, G., D'Angelo, C.A., & Caprasecca, A. (2009). Gender differences in research productivity: a bibliometric analysis of the Italian academic system. *Scientometrics*, *79*(3), 517-539.

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2008). Assessment of sectoral aggregation distortion in research productivity measurements. *Research Evaluation*, *17*(2), 111-121.

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2011). Research productivity: are higher academic ranks more productive than lower ones? *Scientometrics*, *88*(3), 915-928.

Abramo, G., D'Angelo, C.A., & Rosati, F. (2013a). The importance of accounting for the number of co-authors and their order when assessing research performance at the individual level in the life sciences. *Journal of Informetrics, 7*(1), 198–208.

Abramo, G., D'Angelo, C.A., & Rosati, F. (2013b). Measuring institutional research productivity for the life sciences: the importance of accounting for the order of authors in the byline. *Scientometrics*, *97*(3), 779-795.

Butler, L. (2007). Assessing university research: A plea for a balanced approach. *Science and Public Policy*, *34*(8), 565-574.

D'Angelo, C.A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in large-scale bibliometric databases. *Journal of the American Society for Information Science and Technology*, *62*(2), 257–269.

Egghe, L. (2005). Relations between the continuous and the discrete Lotka power function. *Journal of the American Society for Information Science and Technology*, *56*(7), 664–668.

Fox, M.F. (2005). Gender, family characteristics, and publication productivity among scientists, *Social Studies of Science, 35*(1), 131–150.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics, 1*(4), 359-375.

Kyvik, S. (1989). Productivity differences, fields of learning, and Lotka's law. *Scientometrics*, *15*(3-4), 205-214.

Leahey, E. (2006), Gender differences in productivity: research specialization as a missing link, *Gender and Society, 20*(6), 754-780.

Long, J.S. (1992), Measure of sex differences in scientific productivity, *Social Forces, 71*(1), 159-178.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16 (12), 317–324.

Moed, H.F. (2005). *Citation Analysis in Research Evaluation*. Dordrecht: Springer.

Moed, H.F., Burger, W.J M., Frankfort, J.G. & Van Raan, A.F.J. (1985). The application of bibliometric indicators: Important field- and time-dependent factors to be considered. *Scientometrics*, *8*(3-4), 177-203.

Pontille, D. (2004). La Signature Scientifique: Une Sociologie Pragmatique de l'Attribution. Paris: CNRS Éditions.

Pripič, K. (2002). Gender and productivity differentials in science, *Scientometrics, 55*(1), 27-58.

RIN (Research Information Network) (2009). *Communicating Knowledge: How and Why Researchers Publish and Disseminate Their Findings*. London, UK: RIN. Retrieved April 8, 2015 from www.jisc.ac.uk/publications/research/2009/communicatingknowledgereport.aspx.

Waltman, L., Van Eck, N.J., Van Leeuwen, T.N., Visser, M.S.& Van Raan, A.F.J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E.C.M., Tijssen, R.J.W., Van Eck, N.J., Van Leeuwen, T.N., Van Raan, A.F.J., Visser, M.S., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419-2432.

A Technology Foresight Model: Used for Foreseeing Impelling Technology in Life Science

Yunwei Chen*, Yong Deng, Fang Chen, Chenjun Ding, Ying Zheng and Shu Fang

* *chenyw@clas.ac.cn*
Chengdu Library of the Chinese Academy of Sciences, Chengdu, 610041 (China)

Abstract
This paper constructs an Impelling Technology Foresight Model (ITFM) for foreseeing impelling technology in the field of life science, which is a comprehensive model consisting of four class indicators: international scientific environment, evolving of papers and patents, collaboration features of patent assignees' collaboration networks, and impacts. A case study was carried out in the field of life science. Recombinant DNA (RbDNA) and Monoclonal Antibody (mAb) were selected as impelling technologies to carry out the case study. ELISA Diagnosis (ELISA) and Fermentation Technology (FT) were defined as non-impelling technologies to be control group. Results revealed that impelling technologies have higher evolving rates from the stage of growth to maturity. Significant policies or programs usually boost the rapid progress of impelling technologies. Impelling technologies have much higher impact than non-impelling ones. Collaboration behaviour is much more broad and general for impelling technologies. To our knowledge, this is the first study carried out to date to foreseeing impelling technologies at this way.

Conference Topic
Methods and techniques

Introduction
Technology has made enormous contributions to modern society and many future social developments can be realized only through better technical developments and better management (Compton, 1939). Nevertheless, not all technical progress makes substantial contributions to social development. Only a few techniques brought revolutionary change to the human society, such as Transistor Technology and Recombinant DNA technique, which belong to the field of information technology and biotechnology, respectively. Information technology and biotechnology are also regarded as dominant technologies and will essentially impel the social development in the 21st century (Das, 2001).

Thus, it is an attractive topic all the time for scientists from many scientific fields to foresee what kind of technologies can become such impelling technologies, especially in the field of biotechnology. Impelling technology is defined in this paper as technologies that can bolster, lead and push the scientific development and technology progress in given fields, and that can drive the industry fast development and breed emerging industry. Transistor Technology and Recombinant DNA technique are just such technologies. However, people desire to know which technologies can become impelling technologies in the near future, especially for new technologies. For example, synthetic biology, which uses unnatural molecules to reproduce emergent behaviours from natural biology with the goal of creating artificial life (Benner & Sismour, 2005), is recognized as a powerful technique that can produce re-engineered organisms that will change our lives over the coming years, leading to cheaper drugs, green fuel and targeted therapies for diseases. The de novo engineering of genetic circuits, biological modules and synthetic pathways is beginning to address these crucial problems (Khalil & Collins, 2010). If that is true, synthetic biology could be regarded as an impelling technology. However, except for synthetic biology, there are still a large number of techniques emerging in the field of biology. Which can become impelling technology in the near future? Foresight analysis provides the idea of solutions.

Technology foresight, like technology forecasting, is the generation of reasoned statements about the future, the interpretation of such statements in terms of informed action, and the collective learning processes that are involved in responding to challenges of the future (Salo & Cuhls, 2003). Amanatidou (2014) pointed out that the major impacts of foresight belong to knowledge, network creation and promoting public engagement in policy-making. The scope of technology foresight comprises not only technologies and their applications but also public policies and societal challenges (Salo & Cuhls, 2003). UNIDO defined technology foresight as the most upstream element of the technology development process. It provides inputs for the formulation of technology policies and strategies that guide the development of the technological infrastructure. In addition, technology foresight provides support to innovation, and incentives and assistance to enterprises in the domain of technology management and technology transfer, leading to enhanced competitiveness and growth (UNIDO, 2014).

Indeed, similar forms of foresight technology also include technology intelligence, technology forecasting, road mapping and assessment (Firat, 2008). Many of these forms use similar tools and get similar results. Particularly forecasting and foresight are often confused in practice. According to the interpretation from the Technology Futures Analysis Methods Working Group 1 (TFAMWG), all these similar methods could be used in technology futures analysis (TFA). Technology foresight is used to analyse the effecting development strategy, often involving participatory mechanisms. Technology forecasting is to anticipate the direction and pace of changes. But there is a general tendency that forecasting usually focuses on specific technologies. Foresight studies usually bring together people with different expertise and interests, and use instruments and procedures that allow participants to simultaneously adopt a micro view of their own disciplines and a systems view of overriding or shared objectives (Firat, 2008). Some foresight related studies are introduced below and their findings contributed partly to the theoretical and technical basis of this study.

Based on the below related works analysis, we found that although many techniques have been used to answer many kinds of questions, impelling technology foresight works were lacking, especially by the method of model construction. Therefore, this study advanced the existing works by constructing an ITFM model to carry out impelling technology foresight analysis. ITFM model can be used for impelling technology foresight. To our knowledge, none of the existing studies has done such work as ever. The significance of this work is that if an impelling technology could be known before it becomes impelling technology or at the earlier stage of its life cycle, that would be very valuable for many kinds of scientists, policy makers and stakeholders to deal with it.

Related works

The term “Technology Foresight” was introduced by Irvine and Martin and took off in the 1990s as European, and then other countries (Miles, 2010). Until now, a lot of studies have been carried out to do such analysis in recent years, which could be divided into four aspects: function, subject areas of use, features of products and results, and techniques. Related works are discussed below.

Function

The focuses of technology foresight studies have been often motivated by the desire to shape S&T policies and analyse the challenges of education, services, health, and environment, etc. (Salo & Cuhls, 2003). For example, Carlson (2004) discussed the using of technology foresight to create business value. Sanz-Menendez (2001) made technology foresight as a useful tool for policy making. Havas (2010) analysed the impact of foresight on innovation policy-making. Weigand et al. (2014) studied collaborative foresight method to complement long-horizon strategic planning.

Subject areas of use

Based on the fields of science and technology, Linstone (2011) discussed the unique impacts of technology foresight on nanotechnology, biotechnology and materials science. Weinberger, Jorissen and Schippl (2012) carried out a study about technology foresight analysis in the field of environmental technologies with the purpose of supporting the process of identifying and recommending options for the prioritisation of future research funding. Furthermore, foresight has also been used in the field of education studies (Goldbeck & Waters, 2014; King, 2014), drugs discovery (Lintonen et al., 2014).

Features of products and results

From the aspect of products and results of foresight, the works of technology foresight usually have the following products: Strategic advice or guidance, particular technologies or their consequences, price or trends of markets, and production. For example, Cook, Inayatullah and Burgman (2014) concluded that foresight could play a more significant role in environmental decisions by the following ways: monitoring existing problems, highlighting emerging threats, identifying promising new opportunities, testing the resilience of policies, and defining a research agenda. Markus and Mentzer (2014) discussed the future consequences of ICT. Weinberger, Jorissen and Schippl (2012) used foresight methods to support the process of identifying and recommending options for the prioritisation of future research funding among the wide range of environmental technologies available that can contribute to progress in the field of environment.

Techniques

At the angle of techniques used for foresight, many kinds of methods have been used to carry out technology foresight analysis. One typical technique is bibliometric methods. Van Raan (1996) overviewed the potentials and limitations of bibliometric methods for the assessment of strengths and weaknesses in research performance, and for monitoring scientific developments. The study suggested that research performance assessment is based on advanced analysis of publication and citation data. While for monitoring scientific developments, bibliometric mapping techniques are essential. Actually, mapping has been widely used for technology foresight. For example, Yoon, Lee and Lee (2010) developed a keyword-based knowledge map to use to establish a policy to support promising R&D areas and devise a long-term research plan. Another typical method is modelling and system. For instance, Shiue and Lin (2011) developed a foresight MASA model for future technology evaluation in electric vehicle industry, which integrated the concept of vision, linking analysis planning, Markov chain, and Scenario analysis (SA). Chen (2012) proposed a structural variation model for answering what kinds of information may serve as early signs of potentially valuable ideas. Peer review and Delphi have also been used in foresight as in forecasting. For example, Lintonen et al. (2014) had done a drugs foresight analysis in 2020 through the method of Delphi expert panel study. Forster & Gracht (2014) had also assessed Delphi panel composition for strategic foresight based on company-internal and external participants.

Model of Impelling Technology Foresight Model (ITFM)

Definition and Hypothesis

As is stated above, impelling technologies are such technologies that could bolster, lead and push the scientific development and technology progress and drive the existing industry fast develop and bread emerging industry in given fields. However, this definition explains only the functional feature reflecting the results generated by impelling technologies, and lacks the

description of its inherent features, especially the features at the early stage of technology lifetime, which are much more important to foresee whether a technology at the early stage could become impelling technologies. Therefore, the inherent features of impelling technologies especially the features at the early stage could be used as indicators for reflecting impelling technologies. Thus, some hypothesises had been proposed as the theoretical base for constructing an Impelling Technology Foresight Model (ITFM) for foreseeing impelling technologies, particularly in the field of life science.

Hypothesis 1. Viewed by the concept of technology life cycle, technologies' development process can be divided into four stages (Little, 1981) of emerging, growth, maturity and saturation. Impelling technologies grow rapidly to the stage of maturity after short growth stage. Impelling technologies seldom show signs of turning to saturation stage for their competitive impact could remain much longer than non-impelling technologies. In order to evaluate the current stages of a technology, patents have been widespread used to do such analysis. For example, Patent analysis was applied by Zhou et al. (2014) to monitor the developmental stage of a particular New and Emerging Science & Technologies, dye-sensitized solar cells (DSSCs), and traced its potential evolutionary pathways. Some other related works have high impacts include Haupt, Kloyer & Lange (2007), Trappey & Wu (2011), Jarvenpaa, Makinen & Seppanen (2011), etc. This paper uses patent data to disclose the different/given features at the different stages of impelling technologies.

Hypothesis 2. During the development process of an impelling technology, pushing policies or programs usually would like to be attracted to boost the progress of impelling technology. For example, Human Genome Project has been the first major foray of the biological and medical research communities and it boosted the development of an array of new technologies (Collins, Morgan & Patrinos, 2003), among which Recombinant DNA technique have achieved considerable development and have also been generally recognized as an impelling technology in the field of life science.

Hypothesis 3. Impelling technologies have higher level of collaboration, especially in patent assignees' collaboration. A lot of studies have shown that there is a positive correlation between collaboration and better production of science. For instance, Guimerà, et al. (2005) pointed out that collaboration could spur creativity, solving old problems and inspiring fresh thinking. In the field of scientific researches, Whitfield (2008) pointed out that there is a picture of science's increasingly collaborative nature and which determine a team's success. Wuchty, Jones and Uzzi (2007) found that there's something about between-school collaboration that's associated with the production of better science. Kato & Ando (2013) found a positive correlation between their research performance and degree of internationalization.

Hypothesis 4. Impelling technologies have higher level of impacts. Citation-based analysis is the most frequently used method to carry impact analysis. The original use of citation for evaluation is Journal Citation Reports from Thomson Reuters to evaluate journals impact factors. Garfield (1979) pointed out that citation analysis could introduce a useful measure of objectivity into the evaluation process at relatively low financial cost. Numerous approaches have been devised to assess future technological impacts based on patent citation information with the core purpose of identifying the current technologies that will drive technological changes over the coming few years (Lee et al., 2012). There are also some network-based method were used to do technology impact analysis. For example, Ko et al. (2014) presented a combined approach for constructing a technology impact network basing on patent co-classification and identifying the impact and intermediating capability of technology areas from the perspective of a national technology system. This paper uses paper citations to compare the difference of impacts between impelling technologies and non-impelling technologies.

ITFM frame

A few factors from four aspects were introduced to validate the above hypothesis.

Technology life cycle - Evolving of patents and paper were introduced to disclose the evolving features of impelling technologies during the four stages of emerging, growth, maturity and saturation.

International environment - The ITFM model took only policy, plan or program as indicators to reflect the international scientific environment although the related factors are more.

Collaboration - The following network statistics of patent assignees collaboration networks were used to represent the collaboration features of impelling technology.

- Ratio of isolates, which have no collaborators in the assignees collaboration networks G. Counted as n (isolates)/n.
- Ratio of nodes in the largest cluster, counted as n (largest cluster)/n.
- Ratio of clusters compare to nodes, counted as #clusters/n.
- Average degree, let N(i) be the set of assignees collaborating with assignee i. The total number of collaboration assignees with assignee i is the degree of assignee i and is defined as $\eta(i) = |N(i)|$. The average degree of a network G is defined by $\eta(G)=\Sigma_{i \square N}\eta(i)/n$.
- Diameter, which is measured by shortest-path length, has been used to estimate the stage of development through documentation data (Chen, Borner & Fang, 2013, Bettencourt, Kaiser & Kaur, 2009) or patent data (Chen & Fang, 2014). There is a theory that collaboration graph that densify with constant or decreasing diameters. All these studies have showed that collaboration graphs in several scientific and technological fields exhibit initial rapid growth in their diameter, which then tends to stabilize and stay approximately constant at 12~14 (Bettencourt, Kaiser & Kaur, 2009). The assignees collaboration network diameters seem to stabilize at about 12 when a technology come into the stage of maturity (Chen & Fang, 2014).

Note that n is the total number of nodes in the network.

Impact - Two factors of times cited per paper and times cited per patent were used for expressing the technology impacts.

The ITFM frame is listed in Table 1, which is the origin of the following case study.

Table 1. Factors contributing to the ITFM.

Factors		*For validating hypothesis (purpose)*
Technology life cycle	evolving of papers	hypothesis 1
	evolving of patents	
International scientific environment	policy	hypothesis 2
	plan or program	
Collaboration-patent assignees collaboration networks	ratio of isolates	hypothesis 3
	ratio of nodes in the largest cluster	
	ratio of clusters compare to nodes	
	average degree	
	diameter	
Impact	times cited per paper	hypothesis 4
	times cited per patents	

Data and methods

According to the opinions of thirty experts in the field of life science through email consultation, Recombinant DNA (RbDNA) and Monoclonal Antibody (mAb) were selected as impelling technologies to carry out case study. ELISA Diagnosis (ELISA) and Fermentation Technology (FT) were defined as non-impelling technologies to be control group.

Publications in Web of Science^{TM} from 1960s to 2012 (publication year) and US patents in Derwent Innovations Index^{SM} from 1970s to 2012 (basic patent year, defined by DII based on the earliest year of all the publication dates of all members of a patent family) were chosen as quantitative data of case study. Data was acquired from the Web of Science in May 2013. Thomson Data Analyzer (TDA) and Science of Science (Sci^2) Tool (http://cns.iu.indiana.edu) were used to extract the statistic and network information.

Search terms to retrieval papers and patents are listed in Table 2.

Table 2. Search terms used for this study.

	Papers	*Patents*
RbDNA	TS=("DNA recombination" or "recombinant DNA" or "DNA cloning" or "molecular cloning" or "gene cloning")	IPCs: from C12N-015/09 to C12N-015/90
mAb	TS=(("monoclon* antibod*") OR (monoclon* same antibod*))	IP=C12P-021/08
FT	TS=ferment*	IP=(C12C-011/* OR C12G* OR C12P* OR C12J*) AND TS=ferment*
ELISA	TS=elisa, removed the papers in WC class of Spectroscopy, Optics, Physics Condensed Matter, Nuclear Science Technology, Behavioral Sciences, Astronomy Astrophysics and Microscopy.	TS=Elisa

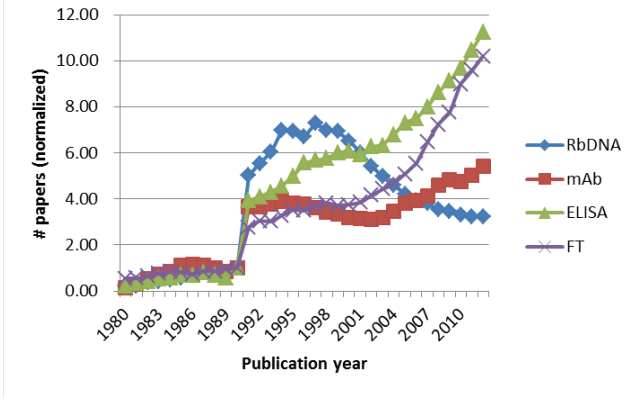
Results and Analysis

Evolving of papers and patents

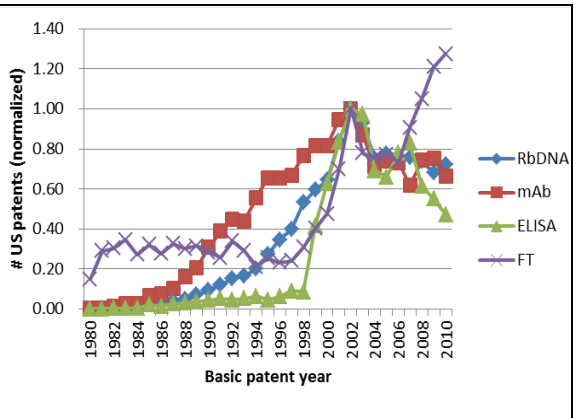
Papers and patents are two external indicators for reflecting the evolving of technologies. The output of papers and patents of the two impelling technologies and two non-impelling technologies were normalized to 1 by their numbers of papers in 1990 and numbers of patents in 2002 separately. The reason of choosing 1990 was that the year 1990 was a jumping-off year, after when the number of papers jumped at least more than three times in 1991. The reason of choosing 2002 was that the year 2002 was a dividing crest, which year had the maximum number of patents, except for FT. Fig. 1 illustrates that the number of papers of both the two impelling technologies stabled at a certain range after three or four years development following the jumping-off from 1990 to 1991. The patents trends show that the number of patents of impelling technologies stabled at a certain level after two years of the patent outputs peak. However, both the papers trends and patent trends of non-impelling technologies had no stable signal no matter which way they go, increase or decrease constantly.

In order to compare the features of impelling technologies at different stages of life cycle, time were sliced into four sections, -1986 (emerging stage), 1987-1993 (growth stage), 1994- (maturity and saturation stages). This division mainly depended on the evolving histories of the two impelling technologies. Although it was not adaptive for on-impelling technologies it

had also been used for distinguishing non-impelling technologies' life cycles with the purpose of comparison.



papers-normalized by the number of papers in 1990



patents-normalized by the number of patents in 2002

Figure 1.Growth of papers and US patents.

International scientific environment

Through watching the histories of the two impelling technologies, we found that Human Genome Project, the first major foray of the biological and medical research communities launched in 1990, boosted the two impelling technologies fast into maturity stage, which could be reflected by the jump of the number of papers. Nevertheless, although the two non-impelling technologies, ELISA and FT, had also been boosted by the Human genome Project, these two technologies had not entered into maturity stage throughout. Actually, beside for the Human Genome Project, there were still more crucial policies had been drawn and put into effect. For example, USA had announced the first Recombinant DNA research Guidelines for normalizing such researches. Even till now, government still made positive policies to maintain the driving functions of impelling technologies. For instance, US Federal Court ruled that synthetic DNA could be patented, which might become a new pushing for the development of RbDNA.

In the aspect of industry, at the stage of growth there were one or a few professional companies born and the number of companies rose sharply at the stage of development and the early maturity stage. For instance, benefited from the development of RbDNA, the first biotechnology company Genentech had been established in 1976. When an impelling technology is mature, the relevant industry would expand rapidly. For example, mAb had brought a rapid growth market of 26 billion USD in 2006 while it was only 4 billion in 2002.

Patent assignees collaboration networks

Figure 2 and Figure 3 illustrate the network features of patent assignees collaboration networks. It is clearly showed in Figure 2c that as time gone on, the ratio of isolates (assignees have no collaborators) decreased year by year and seemed to stabled at a certain level. However, the ratios of isolates of the impelling technologies were much lower all along than that of the non-impelling technologies. The values of the latter were more than twice of the former. The gap was enlarged to more than three times at the stage of development. As a result of the reduction of isolates, the clusters increased and there were many a big cluster became bigger and bigger. It has to be noted that an isolate was also regarded as a cluster. Therefore, a network with high level of collaborative behaviours must has less clusters because of much more isolates and small clusters tend to merge to bigger clusters. Thus the

excellent performance of collaboration leads to generate a super big cluster and less ration of clusters (see Fig. 2a). Figure 2b shows that the biggest cluster of impelling technologies gathered about more than half of the total number of assignees particularly after the stage of development, which was much higher than that of the non-impelling technologies.

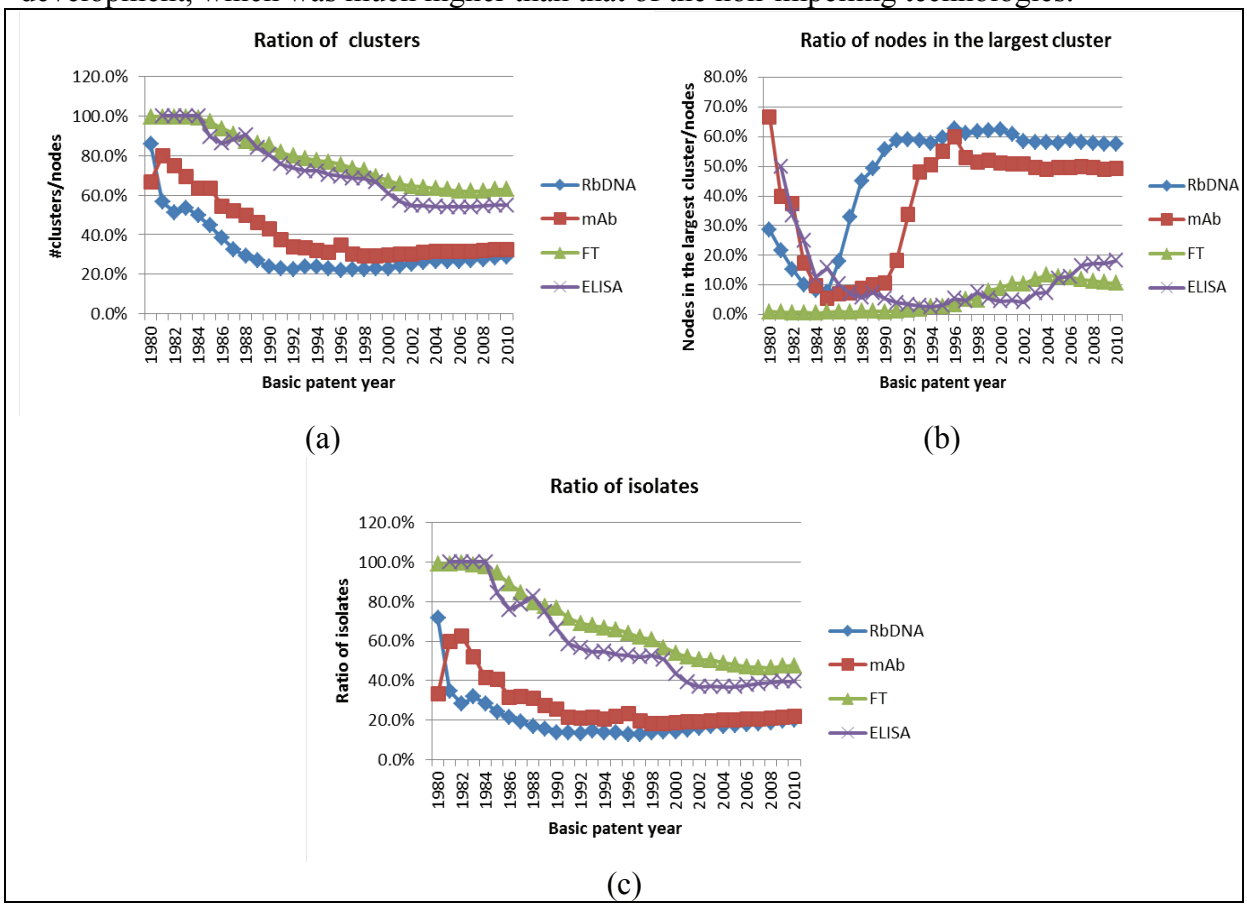


Figure 2. Network features of US patents' assignees' collaboration.

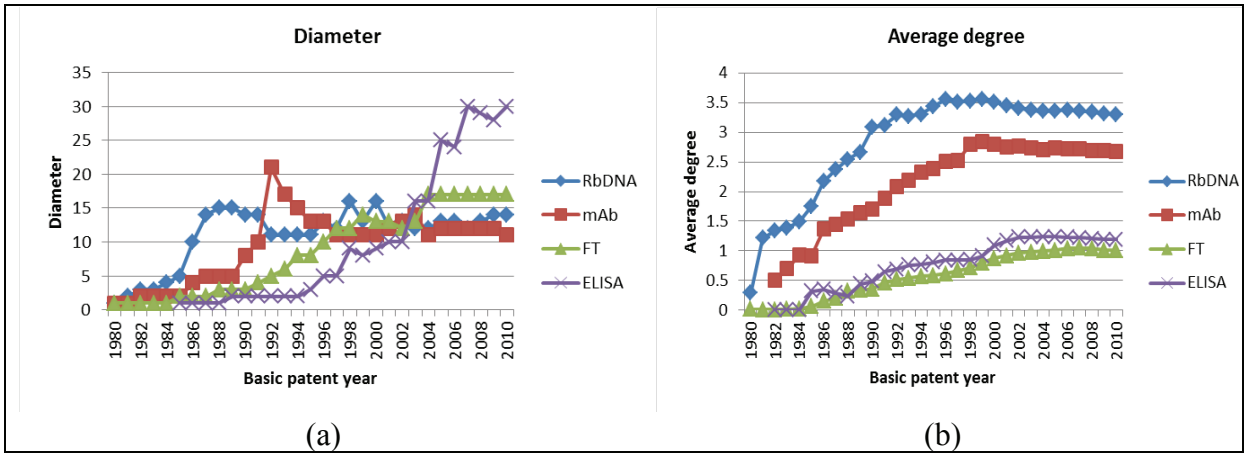


Figure 3. Diameters and average degrees of assignees collaboration network.

Benefited from the good network performance, the impelling technologies had higher average degree all the time. It was about three times higher than that of the non-impelling technologies at the stage of maturity, ten and four times during the period of growth and development respectively.

Impact

The average times cited of papers and patents of the two impelling technologies and two non-impelling technologies were illustrated in Figure 4.

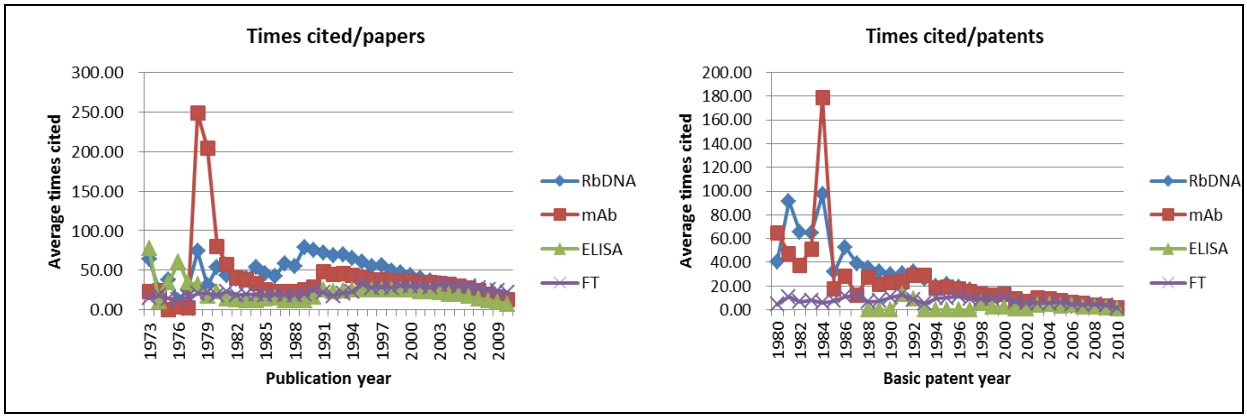


Figure 4. Average times cited of papers and US patents.

The results showed that the average times cited of papers of impelling technologies was two times higher than non-impelling technologies during the whole period of this analysis. The value of impelling technologies was 30 and 50-80 compared to 18 and 20 of non-impelling technologies at the stages of growth (before 1986) and development (1987-1993). For patents, the average times cited of impelling technologies and non-impelling technologies were 66 and 7 in stage of growth, 24 and 9 in stage of development correspondingly. However, the advantages of impelling technologies were eroded as time goes on in the stage of maturity.

Quantitative ITFM

Through the above case study we might conclude some unique features of impelling technologies in the field of life science.

First, impelling technologies had higher rates of evolution from the stage of growth to maturity, which could be illustrated particularly by the papers evolving patterns. When it comes to the technologies of RbDNA and mAb, it took only about one year that both of the two impelling technologies had finished their transform. At the same time, impelling technologies represented distinct feature at the stage of maturity. Nevertheless, the two non-impelling technologies represented no obvious such transformation. It seemed like that both the two non-impelling technologies were still at the stage of growth. However, the fermentation technology had a much longer history than both the two impelling technologies. The reason of it represented such evolutionary feature might just due to the position as a non-impelling technology, which contributes more and more to the society development, but always is a applied technology and will not play more impelling functions.

Second, significant policies or programs boosted the rapid progress of impelling technologies. Although non-impelling technologies had also been pushed by specific policies or plans, the range was lower than that of impelling technologies. When the impelling technologies switched into maturity stage, they usually drove the explosive increase of industry.

Third, impelling technologies had much higher impact than non-impelling technologies, which could be reflected by the times cited per paper/patent. The value of times cited per paper/patent of impelling technologies was two to three times higher than non-impelling technologies. It was highlighted during the process of involving from the stage of development to maturity. In the case of life science, for papers, the value of impelling technologies was 50-80 compared to 20 of non-impelling technologies, for patents, the values were 24 and 9 correspondingly.

Last, collaboration behaviour measured by the collaborations of patent assignees was much more broad and general for impelling technologies. Assignees collaboration networks of impelling technologies had fewer isolates, and there were only about 20% assignees were isolates at the stages of development and maturity. Much more assignees had collaborated with others and become much bigger clusters with a result of the number of clusters decreased. The biggest cluster (principal component) gathered a large number of assignees that took up more than half of the total number of all nodes in the networks at the stage of maturity. As a result, the average degree of impelling technologies reached to 3 which were three times to that of non-impelling technologies at the stage of maturity. The diameters of impelling technologies stabilized at 12 at the stage of maturity. Non-impelling technologies had no such features of stable diameters.

The results indicate that hypothesises listed above were answered by the case study. Based on the results of the comparison of impelling technologies and non-impelling technologies in the field of life science, a quantitative model is induced in table 3. The model can be used for foreseeing any new impelling technologies that have just born or at different stages, especially at the stages of development and maturity.

Table 3 Quantitative ITFM.

	Indicators	***Features***		
		Growth (- 1986)	***Development (1987-1993)***	***maturity (1994-)***
International scientific environment	**Policies, plans &projects**	New incentive, convenient policies enacted	Pushed significantly by major project	Still focus of policies, plans & projects
	Industry	Start-up companies	Number of companies would rise sharply	Industry expand rapidly
Evolving of papers and patents	**Papers evolution**	/	Evolved into maturity stage in few years	Stable (no sign of stable)
	Patents evolution	Steady increase	Steady increase	Stable(no sign of stable)
Collaboration-Features of patent assignees collaboration networks	**Ratio of isolates**	40% (*95%*)	20% (*70%*)	20% (*50%*)
	Nodes in the largest cluster/nodes	20% (*10%*)	35% (*3%*)	55% (*10%*)
	#clusters/#nodes	60% (*97%*)	35% (*80%*)	30% (*65%*)
	Average degree	1 (*0.1*)	2 (*0.5*)	3 (*1*)
	diameters	3 (*1*)	12 (*3*)	Stable at 11-14 (no sign of stable)
Impacts	**average times cited of papers**	30 (*18*)	50～80 (*20*)	Decreased yearly
	average times cited of patents	66 (*7*)	24 (*9*)	No difference between impelling and no-impelling technologies

Notes. The values of non-impelling technologies were listed in brackets.

Discussions

This paper defines impelling technologies and constructs an ITFM model for foreseeing technologies that have potential to become impelling technologies. There is no doubting that this is an attractive topic all the time for many kinds of scientists, policy makers and stakeholders. The theoretical basis of this study is the positive correlation between the four hypothesises and the performance of an impelling technology. Four classes of indicators were introduced into the ITFM model and demonstrated on two impelling technologies and two contrasted non-impelling technologies in the field of life science. Indeed, this work is the first study about impelling technologies foresight and got some valuable results which could be

used for many new technologies foresight, such as synthetic biology. Such application study would be carried out in the near future.

Nevertheless, there are still some shortages of this study. First, the ITFM model can be used only for evaluating existed technologies and not for future technologies that have not born yet. Indeed, this topic is also interesting and important. Second, the values in the ITFM were concluded from the four technologies from life science, which might volatile when used in other fields. Actually, different impelling technologies even in the field of life science might get different values. Therefore, the values in ITFM model are referenced values. The relative performance of impelling technologies is more important when the model is used for evaluating other technologies. Third, impelling technologies foresight is a complex question, which is hard to be identified easily through one or two models or methods. There must be many other indicators that could reflect the unique features of impelling technologies. Therefore, this work is just a beginning of such efforts for foreseeing impelling technologies.

Acknowledgements

We would like to thank many fellows in the field of life science from CAS, who gave a lot of advises in choosing the technologies used for case study. The paper did benefit greatly from detailed comments by an anonymous reviewer. This work is funded by the Documentation and Information Special Project of Chinese Academy of Sciences (2013). This work is funded in part by the National High Technology Research and Development Program of China (863 Program) under grant no. 2014AA021503. This work is supported in part by the West Light Foundation of the Chinese Academy of Sciences, China under grant no. [2013]165(3-6). This work is funded in part by the Main Direction Program of Knowledge Innovation of Chinese Academy of Sciences (KSCX2-EW-G-9).

References

Amanatidou, E. (2014). Beyond the veil - The real value of Foresight. *Technological Forecasting and Social Change*, *87*, 274-291.

Benner, S. A. & Sismour, A. M. (2005). Synthetic biology. *Nature Reviews Genetics*, *6*(7), 533-543.

Bettencourt L. M. A., Kaiser D. I. & Kaur J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, *3*(3), 210-221.

Carlson, L. W. (2004). Using technology foresight to create business value. *Research-Technology Management*, *47*(5), 51-60.

Chen, C. M. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, *63*(3), 431-449.

Chen, Y. W., Börner, K. & Fang, S. (2013). Evolving collaboration networks in Scientometrics in 1978-2010: a micro-macro analysis. *Scientometrics*, *95*(3), 1051-1070.

Chen, Y. W. & Fang, S. (2014). Mapping the evolving patterns of patent assignees' collaboration networks and identifying the collaboration potential. *Scientometrics*, *101*(2), 1215-1231.

Collins, F. S., Morgan, M. & Patrinos, A. (2003). The human genome project: Lessons from large-scale biology. *Science*, *300*(5617), 286-290.

Compton, K. T. (1939). Technical progress and social development. *Electrical Engineering*, *58*(1), 12-15.

Cook, C. N., Inayatullah, S., Burgman, M. A., et al. (2014). Strategic foresight: how planning for the unpredictable can improve environmental decision-making. *Trends in Ecology & Evolution*, *29*(9), 531-541.

Das, M. R. (2001). Biotechnology in the 21(st) Century. *Defence Science Journal*, *51*(4), 327-332.

Firat, A. K. (2008). Technological Forecasting – A Review. Working Paper CISL# 2008-15.

Forster B. & Gracht H. (2014). Assessing Delphi panel composition for strategic foresight - A comparison of panels based on company-internal and external participants. *Technological Forecasting and Social Change*, *84*, 215-229.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool. *Scientometrics*, *1*(4), 359-375.

Goldbeck, W. & Waters, L. H. (2014). Foresight education: When students meet the future(s). *Futurist*, *48*(5), 30.

Guimerà, R., Uzzi, B., Spira, J. & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*. *308*, 697–702.

Haupt, R., Kloyer, M. & Lange, M. (2007). Patent indicators for the technology life cycle development. *Research Policy*, *36*(3), 387-398.

Havas, A., Schartinger, D. & Weber, M. (2010). The impact of foresight on innovation policy-making: recent experiences and future perspectives. *Research Evaluation*, *19*(2), 91-104.

Jarvenpaa, H. M., Makinen, S. J. & Seppanen, M. (2011). Patent and publishing activity sequence over a technology's life cycle. *Technological Forecasting and Social Change*, *78*(2), 283-293.

Kato, M. & Ando, A. (2013). The relationship between research performance and international collaboration in chemistry. *Scientometrics*, *97*(3), 535–553.

Khalil, A. S. & Collins, J. J. (2010). Synthetic biology: applications come of age. *Nature Reviews Genetics*, *11*(5), 367-379.

King, K. (2014). Foresight in middle school: Teaching the future for the future. *Futurist*, *48*(5), 41-42.

Ko, S. S., Ko, N., Kim, D., et al. (2014). Analyzing technology impact networks for R&D planning using patents: combined application of network approaches. *Scientometrics*, *101*(1), 917-936.

Lee, C., Cho, Y., Seol, H., et al. (2012). A stochastic patent citation analysis approach to assessing future technological impacts. *Technological Forecasting and Social Change*, *79*(1), 16-29.

Lintonen, T., Konu, A., Ronka, S., et al. (2014). Drugs foresight 2020: a Delphi expert panel study. *Substance Abuse Treatment Prevention and Policy*, 9.

Little, A. D. (1981). *The Strategic Management of Technology*. Cambridge, Mass.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica*, *29*(4), 741-766.

Markus, M. L. & Mentzer, K. (2014). Foresight for a responsible future with ICT. *Information Systems Frontiers*, *16*(3), 353-368.

Miles, I. (2010). The development of technology foresight: A review. *Technological Forecasting and Social Change*, *77*(9), 1448-1456.

Salo, A. & Cuhls, K. (2003). Technology foresight—past and future. *Journal of Forecasting*, *22*(2-3), 79-82.

Sanz-Menendez, L., Cabello, C. & Garcia, C. E. (2001). Understanding technology foresight: the relevance of its S & T policy context. *International Journal of Technology Management*, *21*(7-8), 661-679.

Shiue, Y. C. & Lin, C. Y. (2011). Developing a new foresight model for future technology evaluation in electric vehicle industry. *Journal of Testing and Evaluation*, *39*(2), 119-125.

Trappey, C. V., Wu, H. Y., Taghaboni-Dutta, F., et al. (2011). Using patent data for technology forecasting: China RFID patent analysis. *Advanced Engineering Informatics*, *25*(1), 53-64.

UNIDO. (2014). *Technology Foresight*. Retrieved May 10, 2014 from: http://www.unido.org/foresight.html.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*(3): 397-420.

Weigand, K., Flanagan, T., Dye, K., & Jones, P. (2014). Collaborative foresight: Complementing long-horizon strategic planning. *Technological Forecasting and Social Change*, *85*, 134-152.

Weinberger, N., Jorissen, J. & Schippl, J. (2012). Foresight on environmental technologies: options for the prioritisation of future research funding - lessons learned from the project "Roadmap Environmental Technologies 2020+". *Journal of Cleaner Production*, *27*, 32-41.

Whitfield, J. (2008). Collaboration: Group theory. *Nature*. *455*, 720-723.

Wuchty, S., Jones, B. F. & Uzzi, B. (2007). The Increasing dominance of teams in production of knowledge. *Science*. *316*, 1036–1039.

Yoon, B., Lee, S. and Lee, G. (2010). Development and application of a keyword-based knowledge map for effective R&D planning. *Scientometrics*, *85*(3), 803-820.

Zhang, J. X., Zhang, H. S., de Pablos, P. O., et al. (2014). Challenges and foresights of global virtual worlds markets. *Journal of Global Information Technology Management*, *17*(2), 69-73.

Zhou, X., Zhang, Y., Porter, A. L., et al. (2014). A patent analysis method to trace technology evolutionary pathways. *Scientometrics*, *100*(3), 705-721.

Lung Cancer Researchers, 2008-2013: Their Sex and Ethnicity

Grant Lewison[1], Philip Roe[2] and Richard Webber[3]

[1] *grant.lewison@kcl.ac.uk*
King's College London, Guy's Hospital, Great Maze Pond, London SE1 6RT (UK)

[2] *philip@evaluametrics.co.uk*
157 Verulam Road, St Albans, AL3 4DW (UK)

[3] *richardwebber@blueyonder.co.uk*
Kings's College London, Department of Geography, Strand, London WC2R 2LS (UK)

Abstract
This paper describes the process by which almost all authors of papers in the Web of Science (WoS) can be characterised by their sex and ethnicity or national background, based on their names. These are compared with two large databases of surnames and given names to determine to which of some 160 different ethnic groups they are most likely to belong. Since 2008 the authors of WoS papers are tagged with their addresses, and many have their given names if they appear on the paper, so the workforce composition of each country can be determined. Conversely, the current location of members of particular ethnic groups can be found. This will show the extent of a country's "brain drain", if any. Key results are shown for one subject area, and *inter alia* it appears that the majority of researchers of Indian origin who are active in lung cancer research are working in the USA. But East Asians (Chinese, Japanese and Koreans) tend to stay in their country of birth.

Conference Topic
Methods and techniques

Introduction
There is continuing research interest in the sex and ethnic composition of research personnel. A brief survey of the literature in 2013-2014 indicates that there is a widespread interest in the problems faced by female researchers (no fewer than 24 countries were involved in such research, and there were 71 papers in the two years, including several exploring the problems in countries outwith North America and western Europe (*e.g.*, Gonenc et al., 2013; Homma, Motohashi, & Ohtsubo, 2013; Bettachy et al., 2013; Isfandyari-Moghaddam & Hasanzadeh, 2013; Garg & Kumar, 2014). However there is much less interest in the situation of ethnic groups, and that only in the USA (Griffin, Bennett & Harris, 2013; Pololi et al., 2013; Campbell et al., 2013; Hassouneh et al., 2014), with one exception (Johansson & Sliwa, 2014; Sliwa & Johansson, 2014), which concerned foreign women in a UK business school. Attention in the USA is focussed almost entirely on under-represented minorities (African-Americans, Hispanics, and in some cases Native Americans), and hardly at all on the problems that may be encountered by researchers of Asian origins, notably Chinese and Indians, who may have to cope with difficult immigration (Teich, 2014), integration and living experiences when they move to the USA. In fact, as we shall see, they are hardly "under-represented minorities" but rather over-represented compared with their presence in the population. (A fuller survey of the relevant prior literature was given in Roe et al., 2014.)
This paper provides a method whereby the researchers in a given scientific subject area can be characterised by their ethnicity or national background and their sex. This is important for science policy, including the monitoring of the changing roles and positions of women in research and the extent to which a country is welcoming to researchers from abroad and helps them to integrate. It builds on the methods described earlier (e.g., Roe et al., 2014) but now allows all the authors on multi-national papers to be classified, and is applicable to all the countries represented in the subject area. Conversely, it can reveal the location of researchers

of any particular ethnicity or national origin. The methods have been applied to the subject area of lung cancer research, and results for this area are given in some detail, but they can equally be applied to any other research area.

Attention was focussed on 24 leading countries, responsible for the large majority of global lung cancer research output, as shown in Table 1 with their digraph ISO codes. However, some results are also given for others, because the database listed all countries contributing to lung cancer research, and researchers with names characteristic of 90 different countries.

Table 1. List of 24 leading countries in lung cancer research, 2004-13.

Countries	*ISO*	*Countries*	*ISO*	*Countries*	*ISO*	*Countries*	*ISO*
Australia	AU	Denmark	DK	Japan	JP	Sweden	SE
Austria	AT	France	FR	Netherlands	NL	Switzerland	CH
Belgium	BE	Germany	DE	Norway	NO	Taiwan	TW
Brazil	BR	Greece	GR	Poland	PL	Turkey	TR
Canada	CA	India	IN	South Korea	KR	United Kingdom	UK
China (PR of)	CN	Italy	IT	Spain	ES	USA	US

Methodology

The file of lung cancer papers (articles and reviews) was obtained from the Web of Science (WoS) for the six years, 2008-2013, from the intersection of two "filters". One was for cancer, and was based on journal names and title words. These included the names of many individual cancers, genes known to pre-dispose people to an enhanced (or reduced) risk of cancer, and specialist drugs and other treatments such as radiotherapy. The other was for lung disease, and consisted of a number of specialist respiratory journals, such as *Experimental Lung Research, Jornal Brasileiro de Pneumologia, Lung* and *Respiration*, and two title words *lung* and *trachea**. In addition, all the papers in the journals *Lung Cancer* and *Clinical Lung Cancer* were retained, together with papers with *SCLC* or *NSCLC* in their titles. The file contained details of 22,433 papers.

The analysis of the researchers was based on their names, both surnames and given names. The surnames were compared with our listing of 2.6 million family names which is based on records of the majority of the adult population in the following countries: Australia, Brazil, Denmark, Germany, Ireland, Italy, Netherlands, Norway, South Africa, Spain, Sweden, the UK and the USA as well as surname frequency distributions for Austria, Belgium, France, India and Japan. For some countries in Eastern Europe and the Middle East, the files were supplemented by data on the names of scientists from these countries found in the WoS. We were able to classify names into over 160 different ethnicities, nationalities and regions within countries, but in this study the classification was simplified to include own country and eight main groups:

- own country (OWN) – this also included representatives of countries who have been the main sources of immigrants, such as France and the UK in Canada;
- other European country (EUR: Albania, Balkan, Belgium, Bosnia, Britain, Bulgaria, Croatia, Cyprus, Czech, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Malta, Montenegro, Netherlands, Nordic, Norway, Poland, Portugal, Romania, Russia, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland);
- Latin America (LAT: including Brazil, Guyana and Mexico);
- Levant and Mediterranean (LEV: Algeria, Egypt, Israel, Lebanon, Libya, Morocco, Saudi Arabia, Tunisia, Turkey, Ukraine);

- Africa (AFR: Afrikaaner, Angola, Cameroon, Congo, Eritrea, Ethiopia, Ghana, Ivory Coast, Kenya, Malawi, Mauritius, Nigeria, Sierra Leone, Somalia, Sudan, Uganda);
- South Asia (SAS: Bangladesh, Burma, India, Pakistan, Sri Lanka);
- China (CHI);
- other Asia (ASI: Afghanistan, Armenia, Azerbaijan, Cambodia, Georgia, Iran, Iraq, Japan, Korea, Laos, Malaysia, Mongolia, Nepal, Philippines, Singapore, Thailand, Vietnam);
- other non-European and Oceanic (OCE: Australia, Caribbean, Fiji, Indonesia, New Zealand).

The methodology is more fully described in a recent paper by Roe et al. (2014).

Given names often (but not always) connote the sex of the person, and we have compiled a list of some 0.7 million such names, including some misspellings and phonetic misrepresentations. This has recently been complemented with the given names of all UK doctors on the Medical Register – over 328,000 individuals, many of whom come from other countries. Some given names connote a different sex in different countries – for example, Andrea is female in the UK but male in Italy. A few countries (in the present study, only Poland) have surnames with gender endings and this can also be used to determine the sex of an author.

In (Roe et al., 2014), attention was confined to papers from a single country, but we were now able to identify the names of the authors from each of the countries in a multi-national paper because the WoS lists them with their addresses in the following format:

> *[Scagliotti, Giorgio V.] Univ Torino, Thorac Oncol Unit, Dept Clin & Biol Sci, S Luigi Hosp, I-10043 Turin, Italy; [Germonpre, Paul] Univ Ziekenhuis Antwerpen, Edegem, Belgium; [Planchard, David] CHU Poitiers, Poitiers, France; [Reck, Martin] Krankenhaus Grosshansdorf, Grosshansdorf, Germany; [Lee, Jin Soo] Natl Canc Ctr Korea, Goyang, South Korea; [Biesma, Bonne] Jeroen Bosch Ziekenhuis, Shertogenbosch, Netherlands; [Szczesna, Aleusandra] Mazowieckie Ctr Leczenia Chorob Pluc & Gruzlicy, Otwock, Poland; [Morgan, Bruno] Leicester Royal Infirm, Dept Radiol, Leicester, Leics, England*

although not all the authors have given names that would allow their sex to be determined.

A special macro was written to enable the names of all authors from each of the countries to be listed in appropriate columns of a spreadsheet for each paper. These were then each classified by national group and sex, where available, so that the contributions of each of the national groups and sexes could be determined. However, the main analysis was performed on the long list of 84,533 different names, each of which was associated with a country and had its frequency of occurrence listed. For each of the 24 selected countries, and for the rest of the world (RoW), the composition of the lung cancer research workforce and the contributions (sums of the numbers of papers) from researchers from each ethnic group (or world region) were determined.

However, we found during our analysis that some East Asian names belonging to researchers working in China, Japan or South Korea, had been misclassified as European as they were ambiguous, such as Jung, Lee and Park. It was obvious from the given names of these researchers if they were Orientals or Europeans. Thus Jung, Andreas working in Germany was clearly German, but Jung, Deuk-Kju working in South Korea was Korean. Likewise, Park, Bernard J. working in the USA was considered to be of European origin, but Park, Byung-Joo in Korea was taken as Korean. These were manually corrected, and some other adjustments to ethnicity were made.

It also became apparent that some names with different given names or initials actually referred to the same person. Thus there were only two Aaronsons in our list of researchers,

one was Neil and the other Stuart A. Both could be classed as male. Another Aaronson, S.A. was clearly the same as Aaronson, Stuart A, and so could be counted as male. We were able to sex quite a lot of researchers without given names in this way.

Results

The data on the national origins and on the sex of the lung cancer researchers in the 24 selected countries, plus the Rest of the World, were obtained from a large file that looked like this:

Table 2. Small excerpt from the file listing the names of all lung cancer researchers.

Name	*Country*	*ISO*	*Count*	*Ethnic*	*Sex*	*Region*
Aakre, J.	USA	US	1	NO	M	EUR
Aakre, Jeremiah	China	CN	1	NO	M	EUR
Aakre, Jeremiah A.	USA	US	4	NO	M	EUR
Aamini, Mahnaz	Iran	IR	1	IR	F	ASI
Aapro, M.	Switzerland	CH	1	FI	X	EUR
Aarab-Terrisse, S.	France	FR	1	MA	X	LEV
Aarndal, Steinar	Norway	NO	2	NO	M	EUR
Aaron, Jesse	USA	US	1	UK	M	EUR
Aarons, Y.	Australia	AU	1	ES	F	EUR
Aarons, Yolanda	Australia	AU	1	ES	F	EUR

The top person in this list evidently worked both in China and the USA, and the first and ninth names were sexed by comparison with the row(s) below.

For the analysis by sex, all 24 countries, plus the RoW, have been included in Table 3. The table shows the percentages of names that could be sexed, and the percentage of such names that were female. The calculation was made both for the number of researchers (this will be an over-estimate, as in Table 2 there are only 7 people, not 10) and for their total contributions.

The high percentage of females in China is clearly anomalous as fewer than half the names could be sexed – this was also the case for Taiwan and Korea. Among European countries, Canada and the USA, on average just over 80% of names could be sexed, and the female percentages are therefore more reliable. Austria, Belgium, Germany and the Netherlands score noticeably low on female participation. On the other hand Poland, a former Communist country where females were strongly encouraged to work (Webster, 2001), ranked highly, and the 10 other eastern European countries (the new "accession Member States" of the European Union) as a group ranked more highly still, with an actual majority of female researchers (51.5%) though their collective contribution was only 46.6%.

Table 3. Analysis of lung cancer researchers in different countries by sex. P = number of people; C = number of contributions (integer count). F = number of females; M = number of males. Countries are ranked by percentage of female researchers.

	Total			*Males*	*Females*	*Unknown*	*Sexed, %*		*F/(M+F), %*	
ISO	*P*	*C*	*C/P*	*P*	*P*	*P*	*P*	*C*	*P*	*C*
CN	13500	29897	2.21	2241	3918	7341	46	42	63.6	63.9
RoW	5226	8475	1.62	1920	1733	1573	70	74	47.4	45.8
PL	842	1643	1.95	396	348	98	88	91	46.8	43.2
IT	4647	9220	1.98	2060	1802	785	83	87	46.7	39.6
BR	721	911	1.26	338	282	101	86	86	45.5	43.9
ES	2300	4376	1.90	983	808	509	78	81	45.1	42.2
KR	3990	10533	2.64	938	754	2298	42	43	44.6	44.7
TR	1827	2747	1.50	819	648	360	80	83	44.2	39.0
SE	560	1159	2.07	268	205	93	84	86	43.3	39.7
TW	2867	8243	2.88	508	378	1981	31	34	42.7	38.5
Wld	36480	77204	2.12	10471	10876	15139	59	56	50.9	48.5
FR	3319	7976	2.40	1346	946	1027	69	80	41.3	38.2
DK	502	965	1.92	257	179	66	87	90	41.1	44.0
UK	2908	4782	1.64	1403	914	591	80	84	39.4	35.1
US	19962	44423	2.23	9854	6416	3692	82	84	39.4	34.9
AU	1101	2336	2.12	531	343	227	79	84	39.2	38.6
GR	1247	2194	1.76	620	369	258	79	85	37.3	31.1
CA	1933	4585	2.37	940	551	442	77	79	37.0	37.1
IN	940	1339	1.42	363	212	365	61	62	36.9	34.3
NO	300	923	3.08	172	95	33	89	93	35.6	26.2
NL	1638	3738	2.28	865	462	311	81	86	34.8	31.1
CH	756	1293	1.71	417	212	127	83	87	33.7	29.6
BE	606	1186	1.96	287	143	176	71	72	33.3	28.9
AT	412	851	2.07	242	105	65	84	89	30.3	23.1
DE	3523	6935	1.97	2083	841	599	83	88	28.8	23.9
JP	8900	24503	2.75	4260	1703	2937	67	68	28.6	22.1

The five South American countries (Argentina, Brazil, Chile, Colombia and Venezuela) also scored well for female participation with nearly 46% of researchers and 44% of contributions, slightly higher than the values for Brazil alone. The three Mediterranean Latin countries (Italy, Portugal and Spain) also scored well, and Portugal had the highest female participation, with over 61% of female researchers, whose contribution was 58%.

The correlation of the percentage of females in the above table (for the 11 countries for which a comparison could be made) with that obtained from another (unpublished) study on cancer screening where a similar methodology was used is quite high ($r^2 = 0.63$). However lung cancer averaged only 39% compared to 46% for cancer screening. Sweden was an exception, with a higher female percentage in lung cancer (43%) compared with 40% for cancer screening.

For the analysis of ethnicity/national origins of the researchers, we first determined the percentage of researchers with "own country" ethnicity. Table 4 shows, for each country, the national background(s) of the names that were selected and the corresponding percentages of their numbers and contributions.

Table 4. Numbers and percentages of "own country" researchers

Country	*Own CU*	*P, %*	*C, %*		*Country*	*Own CU*	*P, %*	*C, %*
BR	BR	26.4	27.1		NL	NL	62.9	63.8
DK	DK,SC	41.0	41.8		IN	IN	67.8	68.3
CA	FR,UK	42.0	42.9		ES	ES	68.3	67.3
SE	SC,SE	48.2	50.7		DE	DE	70.3	71.2
AU	UK	51.9	55.7		BE	BE,FR,NL	76.2	72.2
NO	NO,SC	55.3	58.8		TW	CN	78.9	74.5
FR	FR,UK	58.5	60.6		PL	PL	80.0	76.7
UK	UK	59.8	60.1		CN	CN	83.7	85.3
US	EUR	60.1	61.4		TR	TR	85.6	86.6
GR	GR	60.5	64.0		IT	IT	90.5	91.2
AT	DE	61.9	59.5		KR	KR	92.4	92.9
CH	DE,FR,IT	62.0	64.9		JP	JP	95.3	96.3

The result for Brazil is anomalous, as most of its researchers are descended from Europeans and would have European or Latin American names. (A scientific conference in Caxambu of the Brazilian Biochemical Society, which one of us attended in 1994, was almost entirely populated by Brazilians who appeared to be of European origin.) If these are allowed as "own country" names, then they would represent 90% of Brazilian researchers with a contribution of 91%.

The countries with the greatest fraction of their lung cancer workforce of non-native origin appeared to be the Nordic ones (Denmark, Sweden and Norway), and Canada. The UK also had a high proportion of its lung cancer researchers with non-national ethnic backgrounds (40%) and the same percentage of contributions. On the other hand, Italy had only 10% of non-Italians, and Korea and Japan even fewer foreigners (8% and 5% respectively) though there were rather more in Taiwan (21%) and in China (16%). This feature of Italian research was found in a previous study (Roe et al., 2014).

We now consider the contribution of other European researchers to the lung cancer research of the 14 selected European countries. This is shown in Table 5.

Table 5. Contributions of researchers from other European countries to the lung cancer research of 14 selected European countries. P = people; C = contributions (integer count).

	Other EUR, %				*Other EUR, %*				*Other EUR, %*	
Country	*P*	*C*		*Country*	*P*	*C*		*Country*	*P*	*C*
DK	52.4	53.8		FR	28.7	29.5		ES	17.3	19.9
NO	36.3	27.1		CH	27.4	25.4		BE	16.7	22.1
SE	35.7	36.1		NL	27.0	27.1		PL	16.4	19.5
GR	33.9	32.2		DE	21.5	21.2		IT	6.6	6.0
AT	33.7	37.4		UK	21.3	21.3				

The results are similar to those of Table 4, except that the UK dropped from fifth to tenth place with its proportion of other European nationals among its lung cancer researchers. Its acceptance of non-Europeans was therefore correspondingly greater. There were 7.0% with a South Asian background, three fifths of them Indian, 3.1% Chinese and 4.0% from other Asian countries. These percentages are much higher in Europe except that Sweden had a slightly greater percentage of researchers of Chinese origin. The UK also had 2.2% of lung cancer researchers with North African or Levantine names (third highest in Europe), 0.8% with African names (second to the Netherlands) and 0.7% with names from Latin America (highest in Europe). Altogether, its lung cancer research population with non-European names amounted to 19% of the total.

These percentages can be compared with census data for England and Wales in 2011 (ONS, 2012). There were about 5.3% of "other White" including Irish (corresponding approximately to "other Europeans" in the above table), 2.5% of Indian origin, 4.2% of other Asians, and 0.7% of Chinese. So the Chinese were over-represented among lung cancer researchers by 3.1/0.7 = 4.4, the Indians by 4.2/2.5 = 1.7 and other Asians were slightly under-represented by 4.0/4.2 = 0.95. The other Europeans were also over-represented by 21.3/5.3 = 4.0. Many of the Chinese would have been graduate students and would probably have returned to China or gone elsewhere after obtaining their doctorates or other degrees.

Canada and the USA were even more accepting of non-Europeans, and their percentages of the different groups are shown in Table 6. Almost 40% of US lung cancer researchers were of non-European ethnicity or national background, of whom by far the largest group were Chinese (13.8% of the total), followed by Indians (5.8%) and Koreans (3.5%). Despite the large numbers of Latin Americans now in the population, they represent only 4.3% of American lung cancer researchers, even when people with Brazilian, Portuguese and Spanish names are included. US Census data for 2010 show that "Latinos" accounted for well over one third of those living in the USA but born abroad, compared with the Chinese (5%) and Indians (4%). However, only 5% of them had university degrees, compared with 50% of the Chinese and 74% of the Indians (US Census Bureau, 2012).

Table 6. Percentages of non-European lung cancer researchers in Canada and the USA.

	CHI	*ASI*	*SAS*	*LEV*	*LAT*	*AFR*	*Other*	*Total*
CA	11.0	9.6	5.6	4.2	0.9	0.4	2.7	34.4
US	13.8	9.6	7.7	4.5	1.4	1.0	1.8	39.8

The file also allows us to determine where lung cancer researchers with given ethnicities are now based and how much they are contributing to either their countries of origin or their new host countries. We previously found (Basu, Roe & Lewison, 2012) that the output of cancer research papers by people of Indian origin now living in Canada and the USA was greater than that of Indians remaining in India. In lung cancer research, of the 2,233 researchers with Indian names, over half (1,164 or 52%) are working in the USA and only 637 (28.5%) in India. There are 124 in the UK, 80 in other European countries, 73 in Canada and 155 elsewhere. The situation is very different for the Chinese, Japanese and Koreans, see Table 7.

Table 7. Current locations of lung cancer researchers from China, Japan and Korea (S).

Ethnicity \ Workplace	*China*	*Europe*	*Japan*	*Korea*	*USA*	*Other*	*Total*
CN	11301	220	124	178	2762	2725	17310
JP	18	27	8485	9	341	90	8970
KR	1151	40	51	3688	702	443	6075
CN, %	65.3	1.3	0.7	1.0	16.0	15.7	
JP, %	0.2	0.3	94.6	0.1	3.8	1.0	
KR, %	18.9	0.7	0.8	60.7	11.6	7.3	

Clearly, most of these East Asians remain in their own country, although the Chinese travel abroad the most, and the Japanese the least, and hardly at all to China or Korea. There is also very little movement to Japan by Chinese and Koreans, and some of the 51 Koreans working in Japan may be ones whose families have been there for several generations. In 2005, there were some 901,000 people of Korean ancestry living in Japan (out of a population of 128 million) or 0.7%. The percentage of the lung cancer researchers in Japan with Korean names was 0.6%, which is slightly less.

We can also see where the lung cancer researchers with various "European" names are now - some will have stayed in their own country, some have gone to the United States, and some have gone elsewhere. The two figures below show the situation. The five largest countries (in terms of numbers of named researchers) are on the left chart and the next nine are on the right chart. However, many of those with British, German, Polish and Irish names will have been resident in the USA for several generations rather than being recent immigrants.

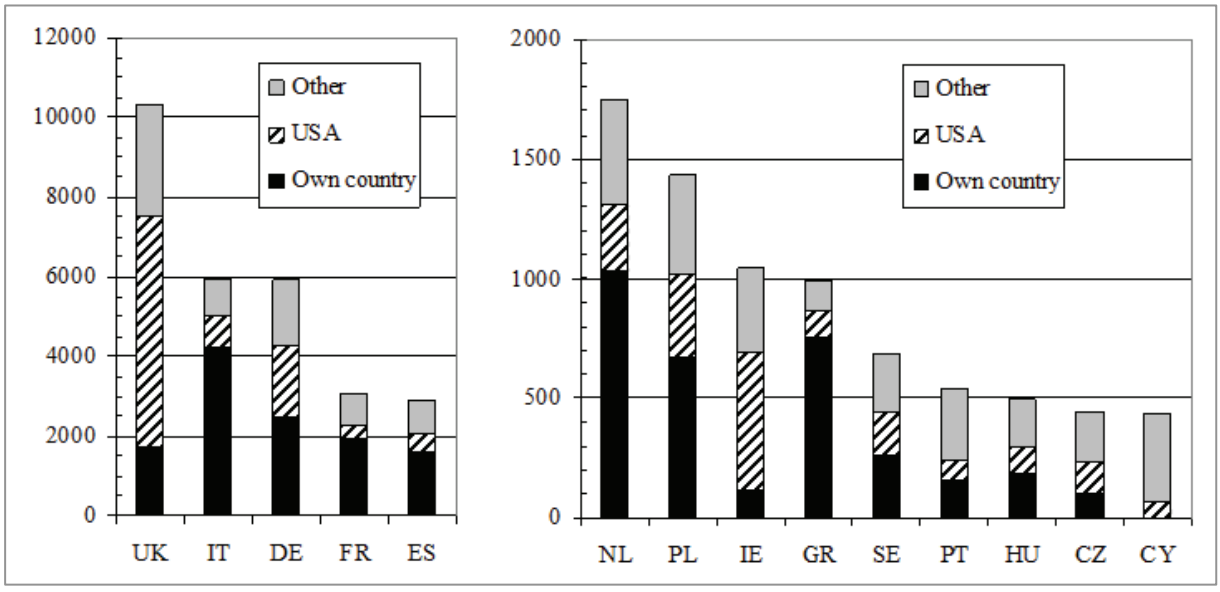


Figure 1. Locations of lung cancer researchers with names characteristic of different European countries - in own country, in the USA, and in other countries.

The file of lung cancer researchers also enables us to investigate whether there is a difference between men and women in the numbers of papers that they write. Figure 2 shows the sex ratio F/(M+F) for groups of authors who publish sufficient papers to put them in a given centile. Thus of the 84,533 authors, the top 1% (n = 845) each wrote at least 17 papers, and the figure shows that just under 26% of those whose sex could be determined were female. By contrast, the 53,143 authors with but a single paper (probably mainly graduate students) were nearly 44% female. This shows clearly that the percentage of females falls off with production, which is probably strongly correlated with seniority. A similar graph could be

produced for individual countries, or ethnic groups, provided that there are enough people in the group or country to make the analysis worth-while.

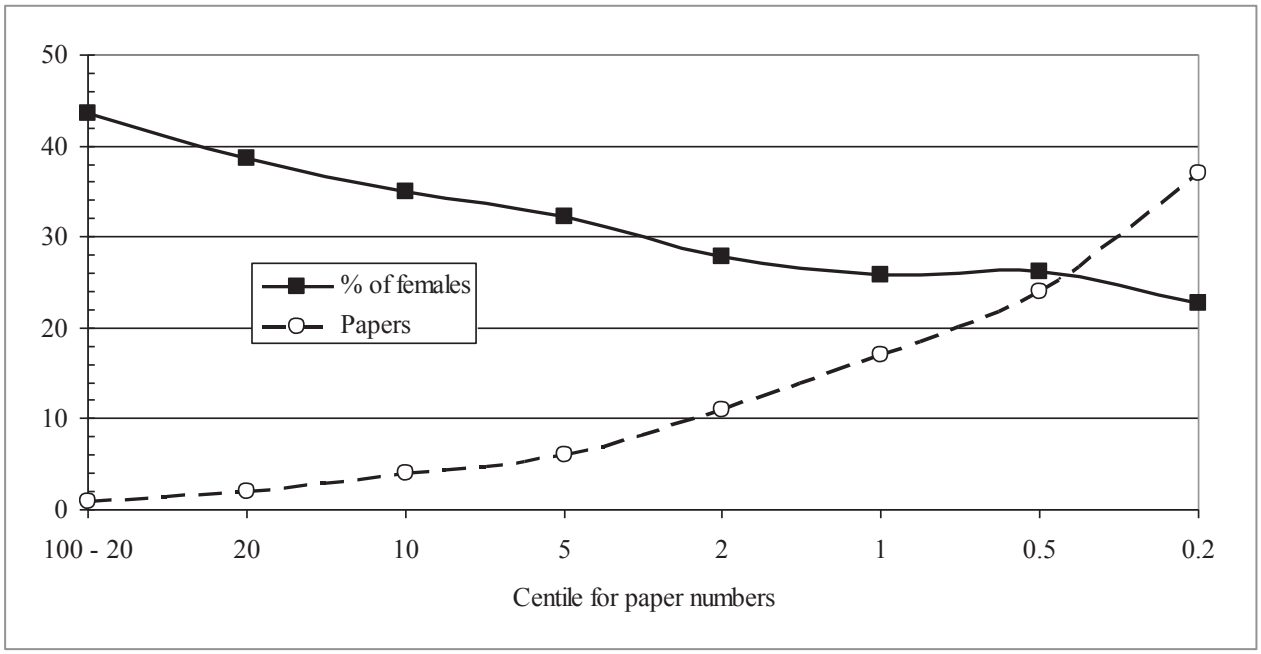


Figure 2. Percentage of female authors whose number of lung cancer papers put them in given centiles of the population of 84,533 authors.

Discussion

This paper greatly extends the methodology used in Roe *et al*., 2014 by its application to all the papers in a subject area, including multi-national ones, and by the provision of a file of all the named researchers, classified by their ethnicity and sex, and the country or countries in which they were working. This allows many research questions to be addressed, and some of them have been in this paper.

However, the methodology still has some limitations, and these are currently being tackled. The first is that, although Aakre, J. can be identified as the same as Aakre, Jeremiah and so classed as male, the file contains two separate entries (actually three in this case because he also published a paper with a Chinese address), which should be amalgamated. The second limitation is that the number of each researcher's papers is given only as an integer count, and for many purposes it would be more useful to have a fractional count, based on the number of different authors of each paper. This is sometimes problematic, as quite a lot of papers list individuals with more than one affiliation. This would not matter if these are all in the same country, as is usual, but increasingly nowadays senior researchers have appointments in more than one country. We would need to fractionate these people's contributions by country in order to make the sum of the individual contributions equal the number of papers (less those with anonymous authors).

A further problem is that, although most names can be classed by country or region within it, some can not be, at present. (The lung cancer database only has 392 names not classified by ethnicity, less than 0.5% of the total.) This is well within the margin of error for most bibliometric studies. However, there is a bigger problem with ambiguous family names where the given names are not on the paper. We have approached this on the basis that most East Asians stay in their own country (see Table 7). However this method would not apply so strongly to Europeans, and as movement and marriage between EU Member States becomes increasingly common, there will be more errors in attribution of researchers to countries.

We have also found that the percentage of names that cannot be sexed is quite high, so that the results for some countries are not at all representative – notably for China. Clearly, we need to acquire more information on the sex associated with particular Chinese, Japanese and Korean names, although some names may not be strictly unisexual. (This occurs also with some European and some British given names, such as Hilary and Robin, where a minority of holders are respectively male and female.) We previously took a ratio of at least 10:1 as indicative of the association of a given name with just one sex, but there may be some errors, though these could be reduced if a researcher has two given names and one can be sexed definitively. This again will need improvements to the software.

Acknowledgments

This study was funded by King's Health Partners and the Global Lung Cancer Coalition as part of their evaluation of lung cancer research world-wide.

References

Basu, A., Roe, P. & Lewison, G. (2012) The Indian diaspora in cancer research: a bibliometric assessment for Canada and the USA. *Proceedings of the 17th International Conference on Science and Technology Indicators* (eds. Éric Archambault, Yves Gingras and Vincent Larivière), Montréal: Science-Metrix and OST; 110-120.

Bettachy, A., Baitoul, M., Benelmostafa, M. & Mimouni, Z. (2013) Women in scientific research in physics in Morocco. *Women in Physics*, *1517*, 128-129.

Campbell, A.G., Leibowitz, M.J. Murray, S.A., Burgess, D., Denetclaw, W.F., Carrero-Martinez, F.A. & Asai, D.J. (2013) Partnered research experiences for junior faculty at minority-serving institutions enhance professional success *CBE-Life Sciences Education*, *12*, 394-402.

Garg, K.C. & Kumar, S. (2014) Scientometric profile of Indian scientific output in life sciences with a focus on the contributions of women scientists. *Scientometrics*, *98*, 1771-1783.

Gonenc, I.M., Akgun, S., Bahar Ozvaris, S. & Emin Tunc, T. (2013). An analysis of the relationship between academic career and sex at Hacettepe University. *Egitim ve Bilim-Education and Science*, *38*, 166-178.

Griffin, K.A., Bennett, J.C. & Harris, J. (2013). Marginalizing merit? Gender differences in black faculty discourses on tenure, advancement, and professional success. *Review of Higher Education*, *36*, 489-512.

Hassouneh, D., Lutz, K.F., Beckett, A.K., Junkins, E.P. & Horton, L.L. (2014). The experiences of underrepresented minority faculty in schools of medicine. *Medical Education Online*, *19*:24768. doi: 10.3402/meo.v19.24768.

Homma, M.K., Motohashi, R. & Ohtsubo, H. (2013). Maximizing the potential of scientists in Japan: promoting equal participation for women scientists through leadership development. *Genes to Cells*, *18*, 529-532.

Isfandyari-Moghaddam, A. & Hasanzadeh, M. (2013). A study of factors inhibiting research productivity of Iranian women in ISI. *Scientometrics*, *95*, 797-815.

ONS. (2012). http://www.ons.gov.uk/ons/dcp171776_290558.pdf

Pololi, L.H., Evans, A.T., Gibbs, B.K., Krupat, E., Brennan, R.T. & Civian, J.T. (2013). The experience of minority faculty who are underrepresented in medicine, at 26 representative US medical schools. *Academic Medicine*, *88*, 1308-1314.

Roe, P., Lewison, G. & Webber, R. (2014). The sex and ethnicity or national origins of researchers in astronomy and oncology in four countries, 2006-2007 and 2011-2012. *Scientometrics*, *100*, 287-296.

US Census Bureau. (2012). https://www.census.gov/newsroom/pdf/cspan_fb_slides.pdf.

Webster, B.M. (2001). Polish women in science: a bibliometric analysis of Polish science and its publications, 1980-1999. *Research Evaluation*, *10* (3), 185-194.

A Model for Publication and Citation Statistics of Individual Authors

Wolfgang Glänzel[1, 2], Sarah Heeffer[1], and Bart Thijs[1]

[1] *{wolfgang.glanzel, sarah.heeffer, bart.thijs}@kuleuven.be*
[1]KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
[2] Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
One of the most important requirements of building applicable models and meaningful indicators for the use of scientometrics at the micro and meso level is the correct identification and disambiguation of authors and institutes. Platforms like ResearcherID or ORCID with author registration providing high reliability but lower coverage now provide appropriate data sets for the development and testing of stochastic models describing the publication activity and citation impact of individual authors. This paper proposes a triangular model incorporating papers, citations and authors analogously to the dichotomous model used at higher levels of aggregation like countries or fields. This model is applied to a set of authors in any field of science identified by their ResearcherID. However, the main advantage of classical citation indicators to study citation impact under conditional productivity turned out to be the main problem in this triangle: the possible heterogeneity of the collaborating authors results in low robustness. A mere technical solution to this problem would be fractional counting at three levels, but the conceptual issue, the different roles of co-authors causing this heterogeneity, will never be solved by any algorithm.

Conference Topics
Methods and techniques; Data Accuracy and disambiguation

Introduction
Spectacular progress has been made in author identification, the disambiguation of names and their institutional assignment on the basis of correct affiliation and cleaned address data extracted from bibliographic databases. In particular, this is one of the most important and basic requirements of building applicable models and meaningful indicators for the use of scientometrics at the micro and meso level. Correct author identification is not only indispensable in studies of academic careers, researchers' mobility, authors' publication and collaboration patterns (Braun et al., 2001) but also in monitoring constitution and performance of research teams (Strotman & Zhao, 2012). The task outlined here is practically twofold: On the one hand, the large-scale disambiguation and assignment of authors forms still one of the big challenges in scientometrics. Although the quality of disambiguation and assignments of authors has considerably improved due to sophisticated algorithms and scientometric techniques, e.g., using "bibliometric fingerprints" (Tang & Walsh, 2010) and similarity patterns (cf. Caron & van Eck, 2014), automated processes proved not sufficient to provide reliable reference standards even if optional interaction of individual authors has been made possible. In this context author identification of the Mathematical Reviews and Elsevier's Scopus databases might just serve as examples. Mathematical Reviews was one of the first databases that applied automated processes (since 1985) for author identification. Challenges are, among others, mobility, topic shifts, career breaks, occasional and infrequent publication activity, e.g., so-called transients (Price & Gürsey, 1976). Incorrect institutional assignment, multiple identities as well as unresolved homonyms are still frequently observed errors. This is contrasted by the possibly higher reliability but lower coverage of identifiers that are based on author registration as, for instance, the ResearcherID of the Web of Science database (Thomson Reuters) and the Open Researcher and Contributor ID (ORCID). The latter IDs are sensitive to human errors and their willingness to regularly update and maintain publication assignment to their IDs. A previous study has pointed to the representativeness bias in favour of more prolific authors (Heeffer et al., 2013).

The second issue is partially related to methodology but also of conceptual nature. The methodological issues arise from the superposition of multiple assignments of publication to subjects, on the one hand, and to co-authors and their particular profiles, on the other hand.
Stochastic models for publication activity and citation impact of authors, however, require partitions, which can only partially approximated by corresponding fractionation procedures (cf. Glänzel et al., 2014 for multiple subject assignment in the context of Characteristic Scores at Scales at different levels of aggregation). A further issue arises from the different stages of the individual careers of authors at the same time; while the same publication year ensures the same age of papers in a given citation window, a pre-set publication year collects papers of scientists who are situated in completely different stages of their careers at the same time. The fact that a PhD student or post-doctoral fellow might collaborate with a senior scientist makes the situation even more complex. Thus the question arises whether the same reference standard derived from the data set should apply to the junior as to the senior co-author. And this leads us directly to the conceptual problem: What is the weight of co-authors and their profiles in determining standards for possible benchmarking exercises? This implies that large-scale statistics calculated on the basis of given publication periods and selected subject fields will not be appropriate as reference standards at the micro level but might indeed mirror the profiles of larger institutions and countries adequately and thus serve as general model at these levels of aggregation.
In this paper a triangular stochastic model analogously to the models used at higher levels of aggregation will be described and opportunities and limitations of such a model will be discussed. In the following we will mainly focus on the following questions.

1. What is the relationship between authors' productivity and their citation impact?
2. How can the relationship between the authors' citation impact and the impact of their publications be described?
3. What is the possible effect of co-authorship on these patterns?
4. Can any reference standard for evaluative studies be derived from the model and the empirical data?

This short introduction already adumbrates the possibilities but also the limitations of scientometric models that are created on the basis of the identification and assignment of individual authors. We optionally attempted to use Thomson Reuters' *Distinct Author Identification System* (DAIS), which is based on clustering author names, institution names, and citing and cited author relationships (Thomson Reuters, 2012). As all automated processes, this results in a broader coverage, but suffers from false positives. We have found nearly 30 authors with more than 300 papers each in 2011 according to the DAIS and the most productive author had 1272 WoS indexed papers. However, a simple manual check of names and profiles of authors associated with the same DAIS code revealed different persons with the same family name and first initial but partially different given names and different research profiles. In order to reduce uncertainty we decided therefore to use Thomson Reuters' ResearcherID in conjunction with journal articles published in the same year hazarding the consequences of representativeness bias. From the viewpoint of the model and the analysis this restriction is, however, immaterial. In this context we would like to stress again that the possible biases in representativeness of author selection is insignificant from the viewpoint of the creation and applicability of the model. More important in this context is the reliability of identification of the authors and their affiliations. Nevertheless, we will first have a look at representativeness of author selection on the basis of Thomson Reuters' ResearcherID (RID). This first part of the analysis forms a straight continuation of a previous study on productivity of registered authors by Heeffer et al. (2013).

Data sources and data processing

All papers indexed as articles, proceedings papers, reviews and letters in the 2011 volume of Thomson Reuters Science Citation Index Expanded (SCIE) have been selected. The reason for this choice of a single year publication window, which results from structural properties of author representation and productivity reflected by annual document indexation in bibliographic databases, is as follows. We have already mentioned at the outset that citation processes of scientific papers published in the same year have the following properties: Within a given citation window, all documents in the set have the same age at any particular time and the citation process is not *homogeneous*, that is, citation frequencies at the initial period differ from those at later stages. Paradigmatically this phenomenon has been characterised as a combination of phases of maturing and decline in citation processes (Glänzel & Schoepflin, 1995; Moed et al., 1998). As a consequence, enlarging the citation window will not simply result in a multiplication of citations by a factor proportional to the length of the window. The situation is completely different when a population with heterogeneous age structure is underlying the process and authors are constantly entering and leaving the system. While the citation process of a fixed document set can be described, for instance, by a simple birth process (e.g., Glänzel & Schoepflin, 1994), the publication distribution of an author set, which is subject to changes and interacts with the "environment", requires a different model taking also the effect of immigration and emigration into account. Such model has been proposed by Schubert and Glänzel (1984). This is the situation we find in any publication period in a bibliographic database: Newcomers are entering the author population, terminators are leaving the system and continuants are members of the population for a longer time including the complete period under study (cf. Price & Gürsey, 1976). As a consequence, publication activity in a longer time period can be simulated by multiplying productivity by a proportionality factor according to the length of the period. Therefore it is initially sufficient to select a shorter period of, e.g., one year as the basis of the analysis.

The reason why we have chosen the year 2011 was that in this particular year the share of papers with registered RID was the largest. We expected, of course, that this share will increase and that more authors will be registered in more recent years but the fact that this share decreases beyond 2011 is probably caused by the attitude of authors to update registration and register newly indexed papers not always immediately and regularly but rather intermittently. The choice of 2011 was also convenient because it allows the observation of citations in an appropriate time span. In addition to this publication year we could therefore choose the three-year citation window 2011-2013.

Methods and results

Theoretical considerations

As already mentioned in the previous section, the inclusion of productivity patterns in citation statistics permits insight into a complex system with the provision of a whole set of benchmarks and reference values. From the mathematical viewpoint, we deal with two basic variables that can stochastically be considered random variables, ζ expressing publication activity and ξ standing for citation rates. Yet the two variables are not assumed to be independent and it is commonly known that more prolific authors tend to be more cited as well. Therefore $P(\xi=i|\zeta=j)$ does not necessarily equal $P(\xi=i)$ for all $i, j \geq 0$ and the conditional expectation $E(\xi|\zeta=j)$, being a function of ζ and taking its values with probability $P(\zeta=j)$ is not necessarily constant. In our case, the following measurable variables occur: The publication activity of a (randomly chosen) author in the mirror of the SCIE database in 2011, the citation impact of a (randomly chosen) paper indexed in the 2011 volume of the SCIE and the citation impact of an author with one or more papers in 2011 with the intermediate conditional

measure of citation impact, provided the author has a given number of publications $j \geq 0$ in 2011.

The following mathematical description, which is indeed necessary to avoid confusions, will, however, be restricted to the absolute necessary. The first question formulated in the introduction relates to the relationship between authors' productivity and their citation impact. This can be formulated as follows. Since citation impact is always measured through the citation rates of individual publications, an author's citation impact can theoretically be obtained as

$$P(\xi=i) = \Sigma_j P(\xi=i|\zeta=j)\cdot P(\zeta=j) \text{ for all } i \geq 0,$$

with the corresponding expectation

$$E(\xi) = \Sigma_j E(\xi|\zeta=j)\cdot P(\zeta=j).$$

Index j is assumed to be positive because the trivial case $P(\xi=i|\zeta=0) = 1$, if $i = 0$ and $P(\xi=i|\zeta=0) = 0$, otherwise, can be excluded (no citations without publications). The corresponding statistics are then denoted as $f_i|j$ and $x|j$. Both statistics (conditional empirical distribution and mean value) refer to the citation impact of authors. Furthermore, the corresponding conditional mean citation rate of an author's *papers* can be obtained by dividing $x|j$ by the number of papers j, that is, $(x|j)/j$ with $j > 0$ is an estimator of the expected citation rate of the individual papers of an author with j papers in the given publication year.

In order to tackle the second problem, we have to introduce a third variable, which will complete the triangular model. Using the notation η for the citation impact of a single paper by an individual author, we obtain a more complex formula than above for the conditional probabilities taking all possible combinatorial combinations concerning number of publications and their citations into account but the relationship of their expectations simply reduced to $E(\xi) = E(\eta)\cdot E(\zeta)$. Under the simple assumption that the likelihood not to be cited is the same for all papers of the author, i.e., $q = P(\eta=0)$ for all $j > 0$, we can approximate the probability of author uncitedness and citedness as $P(\xi=0) = \Sigma_j q^j \cdot P(\zeta=j) = P(\eta=0)^j$ and $P(\xi>0) = 1–P(\xi=0)$, respectively. The reason for the relative simplicity of this expression is that uncitedness of an author in a given period implies that none of his/her papers is cited. The extreme cases $P(\xi=0) = 0$ and $P(\xi=0) = 1$ are obviously equivalent with $q = 0$ and $q = 1$, respectively. We will denote the empirical value of q by g_0. Using the mean values x, z and y as estimators of expected citation rate of an author, the expected publication activity of an author and the expected impact of the author's papers, respectively, we obtain the simple relationship $x = y\cdot z$. From the elementary considerations we can conclude that at least basic statistics can be readily expressed with the aid of two variables.

Finally, it might be worth mentioning in this context that the above random variables and the corresponding statistics also form the groundwork for modelling Hirsch-type indices, notably their cumulative versions such as the successive h-index (e.g., Schubert, 2007).

The sample

The sample of RID authors does – as already observed by Heeffer et al. (2013) – not form a *random sample* of the complete author population in the database as RID authors are less frequent at the low end (particularly among single-paper authors), and are more productive at the high end of the productivity distribution.

Table 1. Share of papers with RID authors and their relative citation impact by countries [Data sourced from Thomson Reuters Web of Science Core Collection].

Country	*Papers*	*RCR*	*NMCR*	*%HC*	*RCR*	*NMCR*	*%HC*	*%RID*
Argentina	7702	1.03	0.98	1.3%	1.44	1.91	4.5%	14.7%
Australia	40979	1.16	1.36	2.1%	1.22	1.63	2.9%	42.4%
Austria	12274	1.22	1.45	2.6%	1.39	2.01	4.7%	29.7%
Belgium	17598	1.22	1.51	2.5%	1.34	1.96	4.1%	32.6%
Brazil	33940	0.99	0.72	0.7%	1.02	0.88	1.0%	45.2%
Canada	54511	1.14	1.38	2.1%	1.38	2.08	4.3%	21.0%
Chile	5073	1.15	1.08	1.3%	1.31	1.49	2.6%	31.8%
Czech Rep.	9350	1.18	1.09	1.5%	1.27	1.40	2.4%	40.4%
Denmark	12772	1.30	1.62	3.1%	1.41	2.03	4.4%	36.2%
Egypt	6251	1.02	0.75	0.6%	1.41	1.52	2.9%	15.1%
Finland	9945	1.20	1.42	2.2%	1.35	1.91	3.8%	34.7%
France	65238	1.09	1.29	1.8%	1.20	1.71	3.0%	28.4%
Germany	91263	1.14	1.39	2.1%	1.23	1.81	3.4%	30.7%
Greece	10647	1.13	1.12	1.6%	1.45	1.91	4.1%	22.2%
Hungary	5763	1.15	1.16	1.8%	1.36	1.63	3.4%	36.2%
India	46532	0.98	0.68	0.7%	1.20	1.26	1.8%	13.0%
Iran	20234	1.15	0.71	0.8%	1.55	1.36	2.8%	9.1%
Ireland	6833	1.18	1.42	2.3%	1.34	1.85	3.5%	35.5%
Israel	11558	1.06	1.34	2.1%	1.28	1.97	4.2%	21.4%
Italy	53919	1.10	1.22	1.7%	1.19	1.52	2.6%	32.8%
Japan	76799	0.94	0.96	1.1%	1.13	1.52	2.5%	20.9%
Malaysia	7325	1.12	0.71	0.7%	1.15	0.84	0.9%	41.1%
Mexico	9830	1.02	0.89	1.2%	1.40	1.69	3.4%	21.0%
Netherlands	31883	1.21	1.60	2.8%	1.28	1.90	3.8%	36.8%
New Zealand	7186	1.17	1.33	2.1%	1.45	1.98	4.0%	30.5%
Norway	9694	1.23	1.43	2.4%	1.43	2.07	4.8%	26.7%
Pakistan	5371	1.18	0.69	1.1%	1.52	1.58	3.3%	16.0%
China PR	156403	1.04	0.91	1.1%	1.24	1.53	2.9%	20.2%
Poland	20261	1.08	0.82	0.9%	1.30	1.41	2.3%	20.1%
Portugal	9844	1.14	1.19	1.6%	1.17	1.29	1.9%	63.9%
Romania	6618	1.26	0.71	1.2%	1.30	0.97	1.9%	40.0%
Russia	27853	1.03	0.55	0.7%	1.12	0.94	1.5%	26.5%
Saudi Arabia	5417	1.15	0.92	1.3%	1.35	1.42	2.4%	31.4%
Singapore	9458	1.17	1.53	2.8%	1.29	1.91	4.1%	47.0%
South Africa	7787	1.26	1.19	2.2%	1.50	1.73	4.4%	25.3%
South Korea	44228	0.97	0.89	1.0%	1.13	1.44	2.4%	22.2%
Spain	47885	1.10	1.24	1.7%	1.19	1.56	2.6%	35.9%
Sweden	19923	1.18	1.44	2.4%	1.31	1.90	3.8%	30.8%
Switzerland	23582	1.29	1.73	3.3%	1.38	2.16	5.0%	34.6%
Taiwan	25550	0.92	0.93	1.1%	1.19	1.55	2.9%	17.0%
Thailand	5819	1.08	0.89	1.0%	1.32	1.48	2.6%	16.9%
Turkey	22571	1.02	0.63	0.8%	1.39	1.34	2.7%	12.8%
UK	91438	1.16	1.46	2.4%	1.28	1.90	3.9%	31.6%
USA	333610	1.09	1.40	2.2%	1.25	1.95	3.9%	20.0%
World total	1229248	1.00	1.00	1.2%	1.13	1.42	2.2%	21.1%

Nevertheless, from the viewpoint of the objectives of this study, this bias is primarily insignificant. In total we have 1,229,248 documents among which 259,341, that is, 21.1% had

at least one registered (RID) author. This share considerably varies among countries. The share ranges between about 10% in Africa, Arabic countries and India till about 50% and even more in Brazil, Singapore and Portugal.

Table 1 displays statistics of countries with at least 5,000 publications in 2011. In particular, the variable RCR represents the relation of observed citation impact and the corresponding journal-based expectation, NMCR stands for corresponding relation between observation and discipline-based expectation and %HC is the share of highly cited papers, that is, of papers that have received at least seven times as many citations as the standard of their discipline (see Glänzel et al., 2009 for exact definitions). The last variable %RID, finally, expresses the share of papers with (at least one) author with registered RID. The comparison of relative citation rates and the share of highly cited papers provides empirical evidence that papers by registered authors exhibit distinctly higher citation impact than the corresponding national standards. We would also like to mention that only very few exceptions have been found in smaller countries not displayed here, e.g., Jordan and Latvia, where the share of highly cited papers and the RCR values did not reach their national standards created by all authors.

Representativeness of publications by authors with RID in individual subject fields is in line with our intuitive expectations: The share of papers by RID authors is the lowest in Mathematics (13.0%), clinical and experimental medicine (14.2% for general & internal medicine and 14.2% for non-internal specialties) and engineering (18.7%). This is contrasted by the corresponding shares in physics, chemistry and biosciences (29.8%, 28.5% and 25.0%, respectively).

Productivity and impact of RID authors

The bias in publication-activity statistics of registered authors has already been stressed (cf. Figure 3 in Heeffer et al, 2013). In particular, RID authors are less frequent at the low end, and more productive at the high end of the productivity scale. Figure 1 shows the distribution of papers over RID authors in 2011. The underlying data are based on the short period of only one year so that the share of single-paper authors is consequently large. Nevertheless, the productivity distribution has the expected long tail: 87 authors have (co-)authored more than 50 papers each. We just mention in passing that the maximum count amounted to 296. This almost incredibly large annual publication output of publishing almost one paper a day is, however, formally correct. The author with an affiliation at the University Sains in Malaysia and a second, more recent one at the King Saud University in Saudi Arabia is active in crystallography. In this context we have to notice that the number of his co-authors per paper is rather low, so that even fractionation would not essentially decrease this author's publication count. This example also illustrates that conceptual issues might have more weight than the number or seniority of co-authors. Before we discuss field-specific aspects of authorship statistics, we still have a look at general citation patterns.

In Figure 2, the citation distribution over authors is compared with the corresponding distribution by papers. In addition to the two series of bars expressing the frequency of citations by RID authors and their papers, respectively, a solid line displays the citation distribution of all papers indexed in the SCIE database to illustrate the bias of the sample. The more moderate skewness and greater expectation of the distribution of citations over authors are plausible and in line with the theoretical rudiments described in the previous subsection since usually we have $z > 1$ and $g_0 \in (0, 1)$.

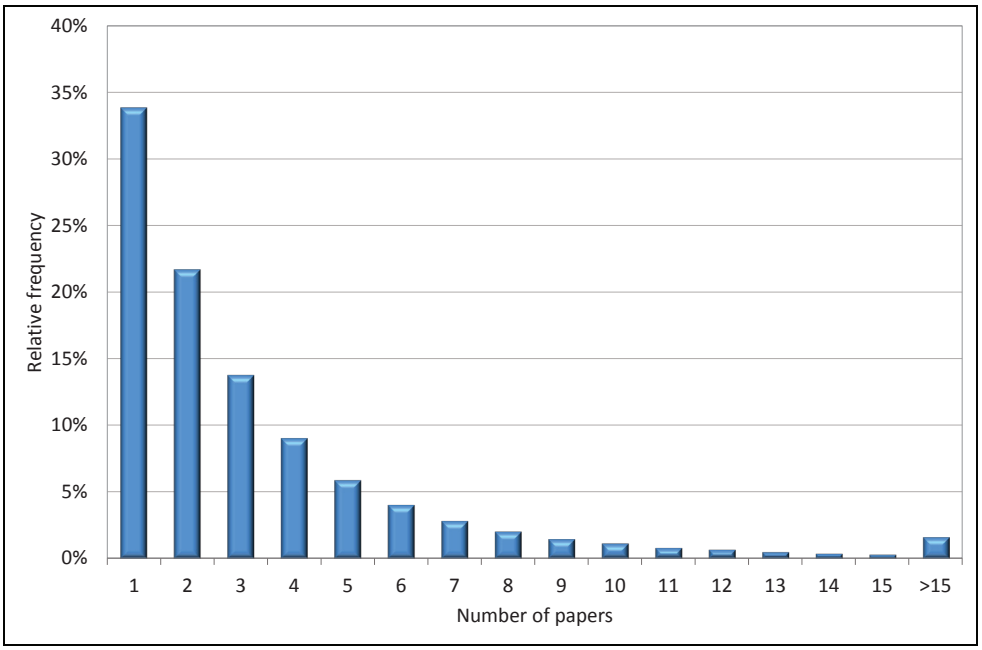


Figure 1. Relative frequency of publication activity of RID authors in 2011. [Data sourced from Thomson Reuters Web of Science Core Collection].

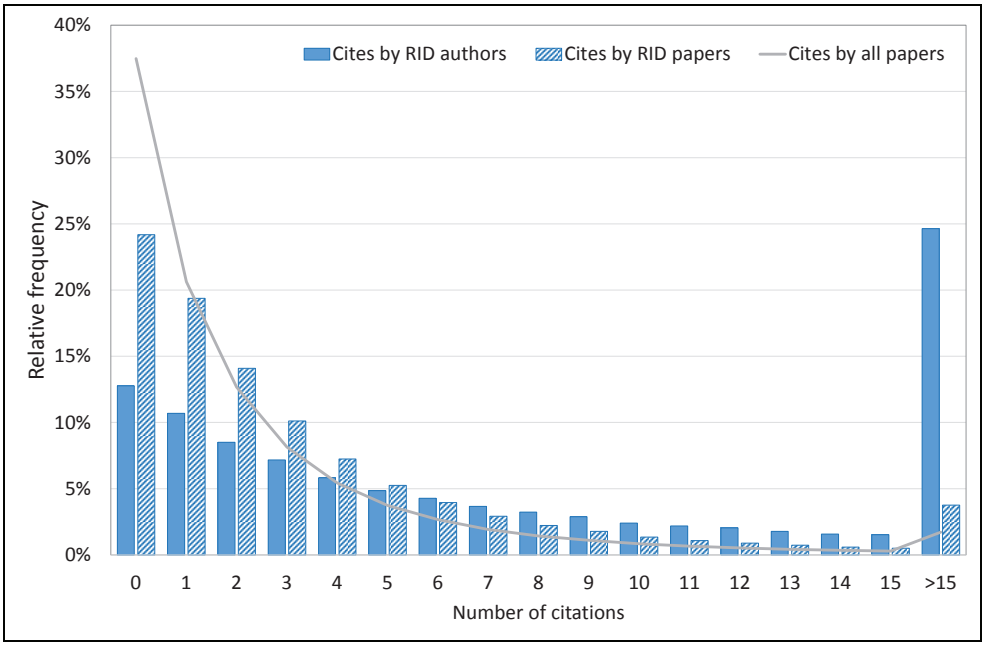


Figure 2. Empirical citation distribution related to RID authors in 2011 in a 3-year citation window. [Data sourced from Thomson Reuters Web of Science Core Collection].

A simple regression analysis aims at studying the relationship of productivity and citation impact of authors, on the one hand, and his/her publications, on the other hand. Conditional mean citation rates in the citation window 2011–2013 received by papers published in 2011 by registered authors have been plotted against their productivity (see Figure 3). Productivity higher than 32 papers has been omitted because of low frequency and considerably fluctuations beyond this level. A power-law model for author citations reflects a very strong correlation, whereas the regression for article citations by authors proved to be linear with somewhat weaker correlation.

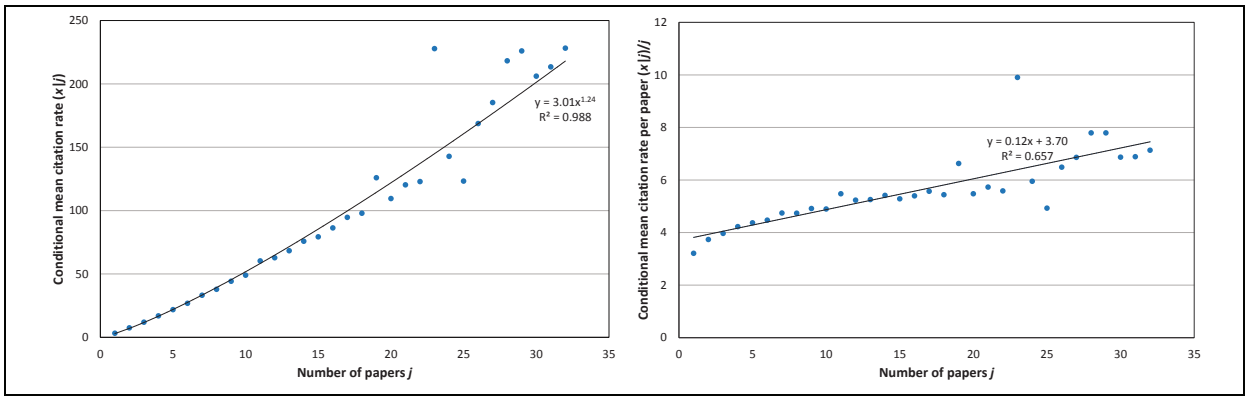


Figure 3. Plot of conditional citation impact of RID authors (left-hand side) and RID papers (right-hand side) based on a 3-year citation window vs. productivity in 2011 [Data sourced from Thomson Reuters Web of Science Core Collection].

While a positive effect of productivity on the expected citation impact of authors was, of course, expected (an increase of papers cannot result in less citations), the positive correlation between number of papers and the mean citation rate of *those* papers is as such not necessarily an inherent property of the model and as we described in the first subsection, the three variables ξ, ζ and η are not assumed to be independent. This indeed substantiates that the publication output of more productive authors exhibit also higher mean citation rates of their output. We have to emphasise that this holds at least for registered authors.

Table 2. Indicators of productivity and citation impact of RID authors and their papers by major science fields. [Data sourced from Thomson Reuters Web of Science Core Collection].

Field	*y*	*z*	*x*	f_0	g_0
A	2.76	1.32	3.65	17.9%	26.4%
Z	3.26	1.40	4.57	14.6%	22.8%
B	4.85	1.19	5.78	11.6%	15.9%
R	3.56	1.15	4.10	16.7%	22.0%
I	4.83	1.58	7.62	13.0%	20.9%
M	3.01	1.76	5.29	16.8%	27.8%
N	3.75	1.54	5.78	14.0%	20.8%
C	4.27	1.89	8.08	12.9%	20.8%
P	3.56	1.66	5.91	14.7%	25.1%
G	3.85	1.40	5.39	15.1%	20.9%
E	2.19	1.35	2.96	26.0%	36.4%
H	1.52	1.47	2.23	35.5%	44.6%

* Legend: A: agriculture & environment; B: biosciences (general, cellular & subcellular biology; genetics); C: chemistry; E: engineering; G: geosciences & space sciences; H: mathematics, I: clinical and experimental medicine I (general & internal medicine); M: clinical and experimental medicine II (non-internal medicine specialties); N: neuroscience & behavior; P: physics; R: biomedical research; Z: biology (organismic & supraorganismic level)

In order to conclude the analysis, we have calculated the mean values of the basic statistics *x*, *y* and *z* as well as the shares of cited authors and papers f_0 and g_0 by subject fields (see previous subsection for description). Table 2 shows these indicators for the 12 major fields in the sciences according to the Leuven–Budapest classification scheme (see Glänzel & Schubert, 2003). As explained in the theoretical part $x = y \cdot z$, $x \geq y$ and $f_0 \leq g_0$ is to be observed. Also subject-specific peculiarities are expected. The *y* and g_0 values concerning the citation impact of papers are by and large in line with the expectations: high impact and low

share of uncited papers in the biomedical sciences and the opposite situation in engineering and mathematics. Nevertheless, the very high impact of chemistry (with low uncitedness) was somewhat surprising and somewhat deviates from the general citation patterns of the fields. Chemistry seems also to be somewhat overrepresented in terms of author registration; 33.5% of all RID authors are active in this field. This is followed by physics with 27.4% and biosciences with 20.8%. All other fields have shares of registered authors below 20% with neuroscience and mathematics having the lowest ones (7.6 % and 4.4%, respectively). In this context we have to mention that the distribution of shares of RID authors over fields is rather strongly correlated with the corresponding distributions of their papers ($r = 0.928$). Hence the question arises whether statistics as presented in Table 2 could be used as reference standards for publication activity and citation impact of authors at the national or institutional level. It has already be stressed in the introduction that an application at the individual level is not recommended because of the heterogeneous age and profile structure of the underlying reference data. Other details regarding this question will be tackled in the following subsection.

Limitations

After the methodological groundwork has been laid for capturing and describing the relationship between productivity and citation impact of authors and their papers, we have also to look at considerable limitations of possible applications of the indicators derived from this model. The low variation of average productivity over subject fields gives already a first hint of possible issues. As already observed by Heeffer et al. (2013) on the basis of the three-year publication period 2009–2011 and RID authors from eight selected countries, the distribution of average productivity was rather flat and ranged – except for physics – roughly between 2 and 3 papers by RID author. Only the average activity in physics with 5 papers per author was distinctly higher. The accustomed and specific inequality of citation impact of papers in different subject areas is almost missing in the productivity statistics what surprises since it is known that scientists in mathematics and engineering are usually less productive – at least as reflected by journal literature – than their colleagues in most fields of the natural and above all in the life sciences. The reason for the observed phenomenon is quite complex but readily explicable. In order to discuss this in detail we have first to refer to the corresponding statistics on citation rates of given paper sets. Provided that the publication year or period as well as the citation window is properly defined and chosen and the subject classification is appropriate, multiple subject assignment of individual papers is then the only severe issue to cope with. Various fractional counting and weighting models have been developed to overcome this problem and to build suitable reference standards for benchmark analysis. Even for more complex statistics than simple shares and means, fractionation by subject can still yield extremely robust statistics as the methods of characteristic scores and scales has shown for various citation windows and aggregation levels (cf. Glänzel, 2007; Glänzel et al., 2014). The question of co-authorship, in general, and how the individual co-authors' actual contribution to a paper should be credited, in particular, is at least in the context of paper-based citation indicators a secondary issue and not primarily related to the definition of citation indicators. The situation becomes completely different, whenever author productivity is directly included in indicator building as, for instance, in our "triangle model" based on the author-paper-citation relationship. The different (academic) age and the different profiles of authors have already been mentioned as possible sources of bias or even distortion, notably in the context of creating benchmarks for individual-author statistics. The most serious issues are related to co-authorship and cannot be simply solved by fractionation by co-authors and/or subjects. Collaboration of senior with junior co-authors, that is, of authors with strong publication record and less active authors, independently of their actual contribution to

the paper in question and their function in preparing it, might have quite strong effect on the resulting indicators at the author level but also at higher level of aggregations. Here we would also like to point to two further issues, firstly the fact that a prolific author in one subject might only play a marginal part as researcher in a different subject in which he/she is collaborating with a possibly less prolific author, who, however, takes the part of the senior co-author of the paper(s) in this topic. Secondly, when it comes to measuring citation impact, an uncited author might be a co-author of a frequently cited author but the joint publications are not cited. This also implies that a mere author–citation analysis in conjunction with productivity studies does not yet suffice; an additional paper–citation analysis is needed for an adequate interpretation. And it becomes clear that a simple fractionation algorithm will not be able to solve these problems. A superposition of fractional counting at three levels (co-author credit, assignment by author profile and subject of publications) is required to solve at least the technical part of this problem: the large overlap by multiple assignments (authors, papers, subjects) could, of course, be resolved and indicators could then be additive over these actors and units at the price of very low robustness. Finally, the most important conceptual issue described in this subjection, the different roles of authors in different environments, will never be solved by using any algorithm.

Concluding discussion

Elementary statistics including relative frequencies and (conditional) mean values have been used to illustrate a simple model of the author-paper-citation relationship. Both opportunities and limitations have been sketched. The use of a joint model for studies of author productivity and impact at higher levels of aggregation is a topical issue in scientometrics: Hitherto the celebrated but also disputed h-index (Hirsch, 2005), originally proposed for the assessment of research performance at the micro level, was the only one that has combined these two aspects, and afterwards been extended for the use at higher aggregation level in the context of institutional and journal evaluation as well.

For illustration purposes, we have selected authors with ResearcherID and active in 2011 in order to exclude errors in author identification as far as possible. Of course, we have to mention that homonyms and synonyms still occur in RIDs too (cf. Heeffer et al., 2013) but the weight of errors is reasonably small. The main advantage of this model is the possibility of studying citation impact under the condition of the author's productivity, and the identification of high performance in terms of both productivity and impact. However, the same precision as experienced with "classical" citation indicators defined on paper sets could not be reached. The main problem is of conceptual nature: Authors and their papers might hold a different position in various environments created by co-authorship of subject-related issues. This has already induced Hirsch to revise his index in terms of co-authorship (Hirsch, 2010). His new indicator also substantiated that complex constellations cannot be described by separately fractionated parts of the model.

The conclusions drawn from this study are two-fold: On the one hand, author-identification systems need to extended in a reliable way to reach a nearly complete coverage of the author population in the database so that indicators based on author IDs can be considered representative enough to be used as reference standards. The limited discriminative power of author-based indicators and the heterogeneity of the underlying author population, on the other hand, prevents the use of the indicators for the analysis of individual research performance as well as in the context of fine-grained benchmark studies at higher levels of aggregations.

Finally, we would like to emphasise again the necessity and general use of the model introduced in this study, which is formally independent of any author-identification system. The model makes is possible to formalise and describe the relationship between authors, their

publications and the citations those publications receive. The neglect of the structural properties and peculiarities of this "triangle relationship" might result in misinterpretation or even miscalculation of statistics and indicators at this level. The use of author identification in this context is an important means of demonstrating the measurement of this relationship for at least a considerable share of active authors.

References

Braun, T., Glänzel, W., & Schubert, A. (2001), Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics*, *51*(3), 499–510.

Caron, E. & van Eck, N.J. (2014), *Large scale author name disambiguation using rule-based scoring and clustering.* In: E. Noyons (Ed.), *"Context Counts: Pathways to Master Big and Little Data". Proceedings of the STI Conference 2014, Leiden University, 2014*, 79–86.

Glänzel, W. & Schoepflin, U. (1994), A stochastic model for the ageing analyses of scientific literature. *Scientometrics*, *30*(1), 49–64.

Glänzel, W. & Schoepflin, U. (1995), A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, *21*(1), 37–53.

Glänzel, W. & Schubert, A. (2003), A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Glänzel, W. (2007), Characteristic scores and scales. A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics*, *1*(1), 92–102.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2009), Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, *78*(1), 165–188.

Glänzel, W., Thijs, B., & Debackere, K. (2014), The application of citation-based performance classes to the disciplinary and multidisciplinary assessment in national comparison and institutional research assessment. *Scientometrics*, *101*(2), 939–952.

Heeffer, S., Thijs, B., & Glänzel, W. (2013), Are registered authors more productive? *ISSI Newsletter*, *9*(2), 29–32.

Hirsch, J.E. (2005), An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Hirsch, J.E. (2010), An index to quantify an individual's scientific research output that takes into account the effect of multiple co-authorship. *Scientometrics*, *85*(3), 741–754.

Moed, H.F., van Leeuwen, T.N., & Reedijk, J. (1998), A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation*, *54*(4), 387–419.

Price, D.D. & Gürsey, S. (1976), Studies in scientometrics. Part 1. Transience and continuance in scientific authorship. *International Forum on Information and Documentation*. *1*, 17–24.

Schubert, A. & Glänzel, W. (1984), A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, *6*(3), 149–167.

Schubert, A. (2007), Successive h-indices. *Scientometrics*, *70* (1), 201–205.

Strotman, A. & Zhao, D. (2012), Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820–1833.

Tang, L. & Walsh, J.P. (2010), Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics*, 84(3), 763–784.

Thomson Reuters (2012), Web of Science® Help. Accessible at: http://images.webofknowledge.com/WOKRS58B4/help/WOS/hp_das1.html. Last modified on 09/18/2012, accessed on 28/12/2014.

A Delineating Procedure to Retrieve Relevant Research Areas on Nanocellulose

Douglas H. Milanez[1] and Ed C. M. Noyons[2]

[1] *douglas@nit.ufscar.br*
Federal University of Sao Carlos, Centre for Information Technology in Materials (NIT/Materiais), Washington Luis Highway, km 235, São Carlos – SP (Brazil)

[2] *noyons@cwts.leidenuniv.nl*
Leiden University, Centre for Science and Technology Studies (CWTS), PO Box 905, 2300 AX Leiden (The Netherlands)

Abstract
Advances concerning publication-level classification system have been demonstrated striking results by dealing properly with emergent, complex and interdisciplinary research areas, such as nanotechnology and nanocellulose. However, less attention has been paid to propose a delineation procedure using specific subjects and understand how it could provide interesting regards about it. This study aimed at proposing a delineation procedure to retrieve relevant research areas addressed to nanocellulose using the research areas clustered by the CWTS Web of Science Publication-level Classification System. The procedure involved an iterative process, which includes developing and cleaning set of core publication regarding the subject and analysis of which cluster they might be associated. Nanocellulose was selected as the subject of study. A discussion about each step of the procedure was also provided. The proposed delineation procedure enabled to retrieve relevant publications from research areas involving nanocellulose. Twelve research topics were identified, mapped and associated with current research challenges on nanocellulose.

Conference Topic
Methods and techniques

Introduction
In recent years, bibliometrics has been used often to monitor and quantitatively assess scientific fields within the context of science policy and research management (Moed, Glänzel, & Schmoch, 2004; Okubo, 1997; Raan, 2014). Partly, it is a consequence of the increased use of Internet since the early 1990s and the development of information technologies. Together, they made a huge volume of scientific databases available. Meanwhile, scientific studies have become more complex and interdisciplinary, involving the exchange of knowledge between scientists from different disciplines. Nanotechnology-focused research is a good example. Bibliometric indicators and tools are useful instruments to study and gain insight in science and, in particular, complex fields or research areas, c.f., van Raan (2004). Therefore, many studies on nanotechnology relied on bibliometric approaches (Hullmann & Meyer, 2003; Igami, 2008; Kostoff, Koytcheff, & Lau, 2009; Milanez, Faria, Amaral, Leiva, & Gregolin, 2014; Mogoutov & Kahane, 2007; Wang, Notten, & Surpatean, 2012). The problems often are: how to delineate a field or research area, how to retrieve the relevant data, and which publications to include and which not.

In this sense, classification systems have been used as an indispensable tool to study the structure and dynamics of scientific fields (Boyack, Klavans, & Börner, 2005; Glanzel & Schubert, 2003; Leydesdorff, Carley, & Rafols, 2013; Waltman & van Eck, 2012). They can simplify literature search and retrieving procedures (Glanzel & Schubert, 2003; Waltman & van Eck, 2012). According to Glanzel and Schubert (2003), classification of science into a disciplinary structure can be as old as science and, currently, most of them are based on journal assignment, such as the Web of Science and Scopus systems. The drawback of these

journal-based classification systems is the fact they do not deal properly with multidisciplinary journals or interdisciplinary research (Waltman, van Eck, & Noyons, 2010).

The development of publication-level classification systems has been a current subject of research (Boyack et al., 2011; Waltman & van Eck, 2012). Boyack et al. (2011) clustered a corpus of 2.15 million biomedical publications from Medline database (2004-2008) which generated coherent and concentrated cluster solution of text-based similarity approaches based on keywords extracted from titles and abstracts. They found their approach more precise than the Medical Subject Headings. Waltman and van Eck (2012) proposed a methodology to clustering a large-scale set of scientific publication indexed on Thomson Reuters' Web of Science database. Each publication was assigned to a single research area, which was organized in a three-level hierarchical structure. Their methodology took into account direct citation to cluster the publication. They labelled each research area with discriminative keywords extracted from titles and abstracts. Such publication-level classification systems may be used to gain insights on research areas involved in specific subjects.

In the present study, we intended to map relevant research areas associated with nanocellulose, which is a sustainable nanomaterial that has a great potential for innovation (Isogai, 2013; Mariano, Kissi, & Dufresne, 2014; Milanez, Amaral, Faria, & Gregolin, 2013; Moon, Martini, Nairn, Simonsen, & Youngblood, 2011). Nanocellulose has been a research area for many countries, including the major producers of cellulose worldwide, such as the USA, Canada, Finland, Sweden and Brazil (Milanez et al., 2013). Different disciplines are involved with nanocellulose research since its properties and behaviour have allowed applications as reinforcement agent in composite materials, packing material, optically transparent paper for electronic devices, texturizing agent in cosmetics and food, bio-artificial implants and bandages (Isogai, 2013; Klemm et al., 2011; Mariano et al., 2014; Moon et al., 2011; Siqueira, Bras, & Dufresne, 2010).

Nanocellulose is a generic term referring to cellulose nanofibrils on the one hand and cellulose nanocrystals on the other (Dufresne, 2013; Klemm et al., 2011; Moon et al., 2011; Siqueira et al., 2010; TAPPI, 2011). Cellulose nanocrystals are basically shorter and rod-like crystalline cellulose, whereas cellulose nanofibrils are long chains of alternate amorphous and crystalline cellulose. Consequently, they differ on their mechanical and functional properties (Eichhorn et al., 2010; Mariano et al., 2014; Moon et al., 2011). Both types of nanocellulose can be obtained from renewable sources, including natural fibres, plants, pulp and forest and agricultural residues. Moreover, cellulose nanocrystals can be biosynthesized by bacteria, resulting in the also called bacterial cellulose (Klemm et al., 2011; Milanez et al., 2013; Moon et al., 2011).

Checking the research topics associated with nanocelluloses will provide insights into current technical challenges concerning this nanomaterial, such as increasing the scale of production minimizing costs, characterization of sources and mechanical properties. Surface modifications to reduce moisture adsorption and improve the adhesion between the nanomaterial and the polymeric matrix, thermal degradation, and biocompatibility with living tissues has also been target of research (Gardner, Opo, Oporto, Mills, & Samir, 2008; Isogai, 2013; Klemm et al., 2011; Mariano et al., 2014; Milanez et al., 2013; Moon et al., 2011; Siqueira et al., 2010).

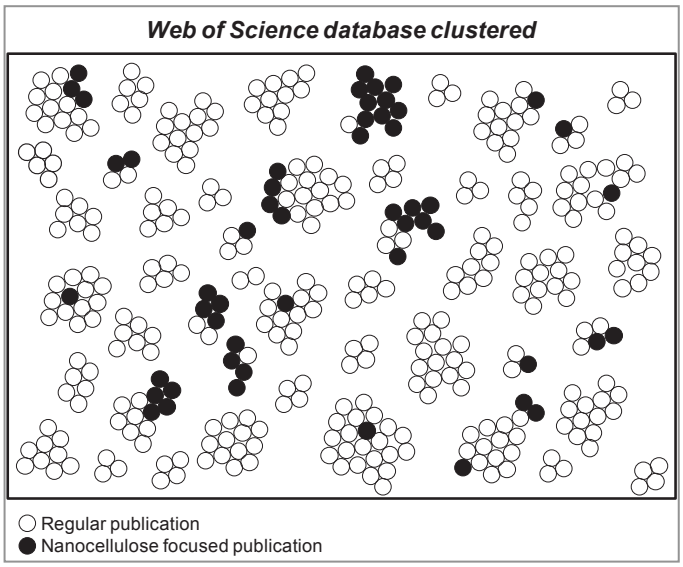
This study aims at proposing a delineation procedure to retrieve relevant research areas addressed to a specific topic. Nanocellulose was selected as a case, but it may be used for other subjects, of course. The approach involves research areas identified in the CWTS Web of Science Publication-level Classification System, a 2014 update of the version introduced by Waltman & van Eck (2012). This paper is structured as follows. In the next section, we describe the overall delineating procedure and its general issues. Next, we discuss details

concerning specific parts and tasks. We present and discuss results in Section 3 and finally in Section 4 we draw our conclusions.

Methodology

Overall delineation procedure

To delineate the field, i.e., to collect a relevant set of publications to represent it, we will select clusters from the CWTS publication level classification system. By this method we will identify papers that will not easily be picked up by keyword or journal based search strategies. Figure 1 presents a schematic representation of the distribution of the clustered Web of Science publications according to CWTS Publication-level classification system (Waltman & van Eck, 2012). Predefined nanocellulose publications are indicated as black circles and the first step is retrieving all research area that contains at least one of them.



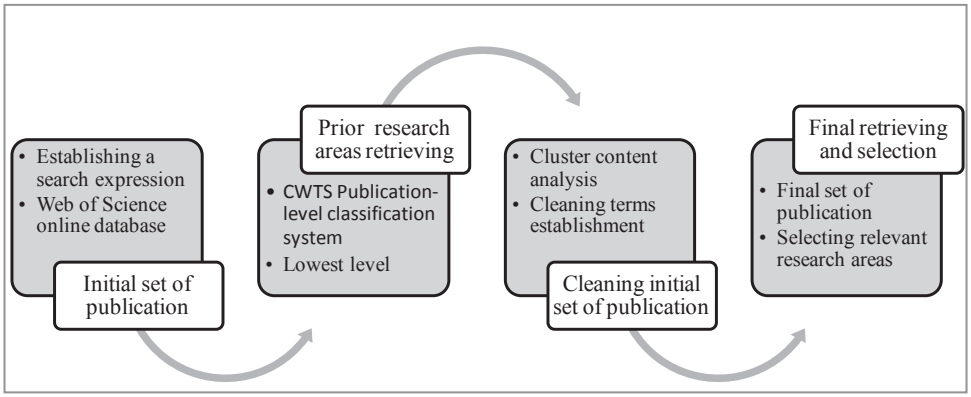
Source: authors.

Figure 1. Schematic representation of Web of Science publications clustered according to the CWTS Publication-level Classification System. The black nodes represent the publications focused on nanocellulose.

Figure 2 depicts the proposed procedure as an iterative process which can be described in four main steps:

1. *Determine an initial set of publication concerning the theme of interest.* In this first step, a set of publication which well represents the theme of interest (nanocellulose) is retrieved via the online Web of Science database, using a straightforward search strategy. This set of publication is a starting set and will be refined as well as expanded through the next steps;
2. *Prior retrieval of nanocellulose research areas*. The second step involves locating the research areas (publication clusters) with at least one publication from the initial set of nanocellulose. The bottom level of the classification scheme was used in this study (Waltman & van Eck, 2012);
3. *Analysis of retrieved research area and cleaning of the initial set*. The content of each research area was analysed pragmatically. A cleaning task was developed by selecting terms to eliminate part of the initial set of nanocellulose publication. This step provided a final set of nanocellulose publication clusters and enhanced the precision of research area assigned to nanocellulose;

4. *Final retrieval and selection of relevant nanocellulose research areas*. After cleaning the initial set of nanocellulose publication, the research areas (publication clusters) were retrieved again. Finally, as the number of topics retrieved was high, a selection that relies on the 80/20 rule was conducted reaching the final research areas associated with nanocelulose.



Source: authors.

Figure 2. Iterative process of the overall procedure proposed.

Determine an initial set of publication on nanocellulose

A search expression was developed considering several terms and synonyms recommended by experts and found in nanocellulose literature (Klemm et al., 2011; Milanez et al., 2013; Siqueira et al., 2010; Siró & Plackett, 2010), as can be seen from Table 1. The search expression encompassed different words that refer to cellulose nanocrystals, cellulose nanofibrils, and bacterial cellulose as well as other generic forms, such as nanocellulose, cellulose nanoparticles, and cellulose nanofiller. The search was conducted in March 31th 2014 in the online Web of Science database (topic search). Only articles that attended the CWTS Web of Science publication-level classification system criteria[1] were used, though.

Table 1. Boolean search expression to retrieve the initial set of nanocellulose publications.

("bacterial cellulos*") OR ("cellulos* crystal*") OR ("cellulos* nanocrystal*") OR ("cellulos* whisker*") OR ("cellulos* microcrystal*") OR ("cellulos* nanowhisker*") OR ("nanocrystal* cellulos*") OR ("cellulos* nano-whisker*") OR ("cellulos* nano-crystal*") OR ("nano-crystal cellulos*") OR ("cellulos* micro-crystal*") OR ("cellulos* microfibril*") OR ("microfibril* cellulos*") OR ("cellulos* nanofibril*") OR ("nanofibril* cellulos*") OR ("micro-fibril* cellulos*") OR ("nano-fibril* cellulos*") OR ("cellulos* micro-fibril*") OR ("cellulos* nano-fibril*") OR ("cellulos* nanofiber*") OR ("nanocellulos*") OR ("cellulos* nanoparticle*") OR ("nano-cellulos*") OR ("nanoparticl* cellulos*") OR ("nanosiz* cellulos*") OR ("cellulos* nanofill*") OR ("nano-siz* cellulos*") OR ("cellulos* nano-fiber*") OR ("cellulos* nano-particle*") OR ("cellulos* nano-fill*") OR ("nano-particl* cellulos*"))

Source: Developed considering nanocellulose-focused terms found in the literature (Klemm et al., 2011; Milanez, Amaral, Faria, & Gregolin, 2013; Siqueira, Bras, & Dufresne, 2010; Siró & Plackett, 2010) and expert opinions.

Prior retrieval of nanocellulose research areas

Research areas that contained at least one publication from the nanocelulose set were retrieved from the CWTS Web of Science Publication-level database. In total, 533 research

[1] The classification system takes into account only *article*, *letter* and *review* published from 2000 to 2013 and indexed in the Science Citation Index Expanded and the Social Science Citation Index. Moreover, to be part of one research area, a publication must be related, either directly or indirectly, to at least 49 other publications in terms of citation (Waltman & van Eck, 2012).

topics were found. These clusters showed large differences in terms of volume (number of publications included). The largest cluster contains 2,751 publications whereas the smallest one covers only 50 publications. Almost 80% of these clusters contained less than three publications from the initial set.

Interestingly, we found that two research areas (clusters) included 56.3% of the initial nanocellulose set of publications. Moreover, in these two clusters, more than 80% overlapped with the initial set. Their descriptive labels also pointed towards nanocellulose research. Therefore, they were considered as nuclei of research in nanocellulose. Other clusters in which the representation of the initial set was much lower, were considered peripheral research areas and their relevance to nanocellulose research was evaluated (see next section).

Analysis of retrieved research area and cleaning of the initial set

An analysis of the content of publications in the peripheral research areas was conducted. We wanted to check whether these articles focused on the nanomaterial as an object of research. If not they were considered noise. Because an evaluation of all research area retrieved would be too labour intensive, we made a selection. The checking task was performed only on those clusters that matched one of the following criteria:

- Research topics that contained at least 20 publications from initial dataset;
- Research topics of which at least 5% overlapped (percentage proportion) with the initial set.

A total of 20 (peripheral) clusters were evaluated. The analysis regarded only articles from the initial dataset. The task involved reading each title to decide whether the article was a study focused upon nanocellulose or not. When the title was not clear, the abstract was also consulted.

Once the checking process was completed, specific terms were identified to clean the initial set of nanocelulose publications. Only research topics with high percentage of "noise publication" were used[2]. Noun-phrases were obtained with support of VOSviewer corpus map analysis applied to titles and abstracts from publications belonging to these clusters. Table 2 present the terms used to clean the nanocellulose-focused publications retrieved using the search expression from Table 1. They were applied on the title, abstract, author's keyword and keyword plus search field. The effect of this cleaning task on the nuclei clusters and the peripheral clusters we used will be discussed in the results.

Table 2. Boolean expression of terms used to clean the nanocellulose-focused publications.

"gene" OR "xyloglucan" OR "microtubule" OR "*cyto*" OR "kinesi" OR "tubulin" OR "*cell wall*" OR "spindle" OR "phragmoplast" OR "mitosis" OR "preprophase" OR "phenotype" OR "*plant growth*" OR "meiosi" OR "*lignin distribution*" OR "delignification" OR "hemicellulose" OR "saccharification" OR "ethanol yield" OR "lignocellulos*" OR "glucosidase" OR "xylanase"

Source: Authors.

Final retrieving and selection of relevant research areas

The final set of nanocellulose publication comprised 2,600 nanocellulose publications (named now as core-nanocellulose) and they were assigned to 428 research areas, which still would be a highly number of cluster to be evaluated. Furthermore, 81.0% of these clusters included only one or two publications from the core-nanocellulose publication, which questions their actual relevance to the advances on nanocelulose studies. Therefore, a selecting step was introduced.

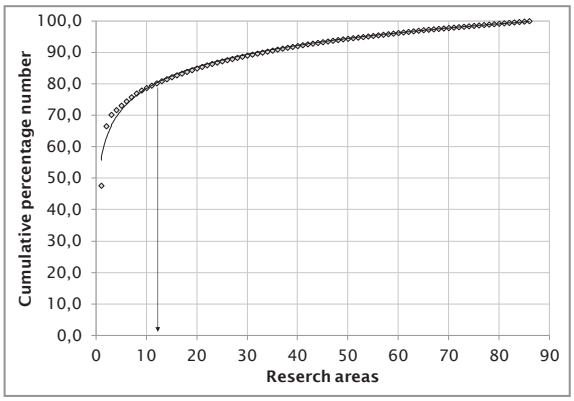
We introduce here the *Pareto Principle* (or 80/20 rule). This principle states that "roughly 80% of the effects come from 20% of the causes" (Juran & Godfrey, 1998) and is found in

[2] The presence of "noise publications" is usual in bibliometric analysis because there is no exhaustive search.

bibliometric and library studies (Gupta, 1989; Kao, 2009; Stephens, Hubbard, Pickett, & Kimball, 2013). We hypothesize that 80% of the core set will be assigned to 20% of the areas. To reach these relevant research areas, the steps below were carried out:

1. The research areas were listed in descending order of the total number of publications from the core-nanocellulose;
2. Research topics with one or two publications from the core-nanocellulose were excluded[3]. This yields 85 research areas remaining;
3. The representativeness of each research area was calculated by the number of publication of the core-nanocellulose of that cluster divided by 2,200 (which is the total of publication found in the 85 remaining research areas);
4. The cumulative percentage number of publications from the core-nanocellulose was obtained summing the values from the step before, as can be seen from Figure 3. The number of research to be assessed was those where the cumulative percentage number of publication reach approximately 80%.

We found that twelve research areas covered the required 80%, which means 14.1% of the total of 85 research topics. We do not claim that our selecting procedure was perfect, but a quick analysis of the chosen research topics showed themes currently found in nanocellulose literature.



Source: CWTS Web of Science Publication-level database.

Figure 3. Cumulative percentage number to research areas with six or more publications from the core-nanocellulose.

Independency test

An independency test was conducted to evaluate the effectiveness of the procedure proposed. The test involved retrieving the number of publication from the top five authors before and after cleaning and selecting the relevant research areas. The percentage decreases of their overall number of publication and from their main cluster were verified.

Results and discussion

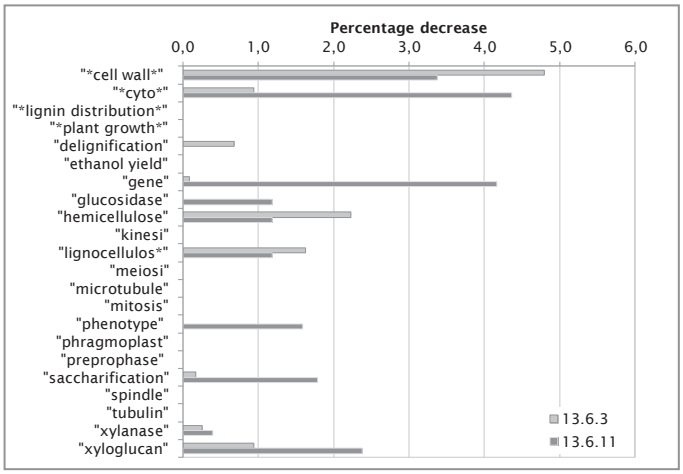
In this section we discuss the effect of cleaning up the core set of publications by using 'cleaning terms', i.e., terms to increase the accuracy of our initial set. Moreover, we present a basic structure of the field on the basis of the delineation we developed.

Effect of cleaning the initial set of nanocellulose publications

Half of the 22 terms we used to clean the nanocellulose search strategy did not affect the coverage of core-nanocellulose publications in the nuclei research areas, as depicted in Figure

[3] According to Waltman and van Eck (2012), the lowest research area contains 50 publications, consequently, clusters with less than 1% of proportion were not accounted for.

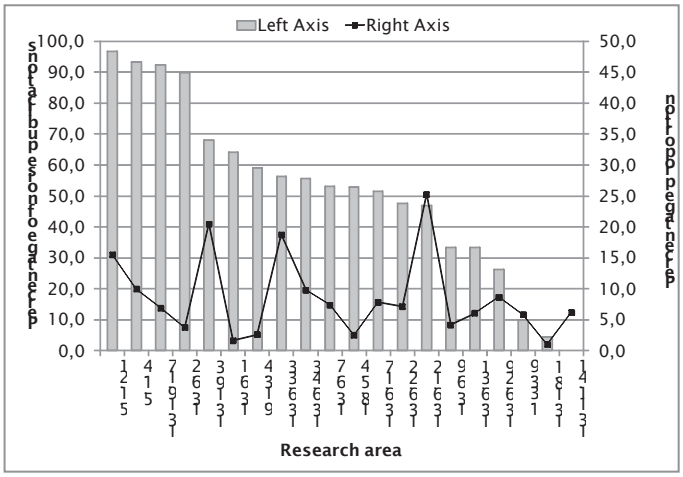
4. To the other half, none term could reduce the coverage in more than 5%. The terms that influenced research area 13.6.4 the most were "*cell wall*" and "hemicelluloses" while "*cyto*", "gene" and "*cell wall*" were the ones that decreased the most core-nanocellulose coverage in cluster 13.6.11. Overall, research topic 13.6.11 had its core-nanocellulose publication reduced in 17.5% while the decrease to cluster 13.6.3 was 10.2%. Nonetheless, both clusters still concentrated publication from the core-nanocellulose after the cleaning tasks (the proportion was 74.0% to research area 13.6.3 and 72.1% to 13.6.11). Therefore, they still had the status of nuclei research areas.



Source: CWTS Web of Science Publication-level database.

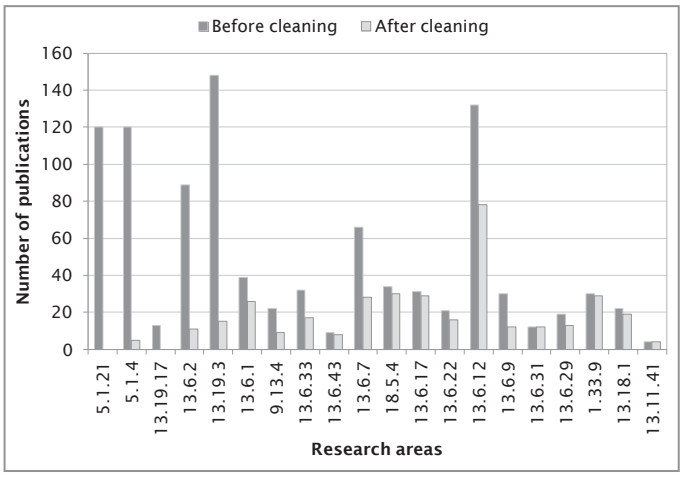
Figure 4. Effect of cleaning terms on the number of publication from nuclei research areas.

As to the 20 peripheral research topics whose nanocellulose set of publication were evaluated, no direct correlation was observed between the proportional relevance of each clusters and the percentage of noise, according to Figure 5. Four research topics had a high percentage (>70%) of 'noisy' publications mainly focusing on biological issues of plants, ethanol production, and enzymes aspects, not having the nanomaterial as a final object of research. Since these four were used to select the cleaning terms, the cleaning affected them highly. Two of them were even eliminated. Furthermore, other peripheral clusters had their nanocellulose publication coverage diminished, as shown on Figure 6.



Source: CWTS Web of Science Publication-level database.

Figure 5. Percentage of noise of core-nanocellulose publications and proportion between core-nanocellulose publications and total number of publications over research area.



Source: CWTS Web of Science Publication-level database.

Figure 6. Effect of cleaning terms on the number of publication from selected peripheral research areas.

Effect of cleaning procedure on top five authors (independency test)

A second test verified the effect of the cleaning process on the coverage of key-authors (top 5). The decrease in the number of publication is presented in Table 3. All authors concentrated their publications on nuclei research topics, mainly on 13.6.3. Only author E focuses primarily on research area 13.6.11. Although the result shows that the overall number of publication diminished in more than 10%, their position as the top authors did not changed but for author E, who went down to the seventh position. It should be noted, however, that research area 13.6.11 was affected more by the cleaning procedure than 13.6.3.

Table 3. Effect of on main authors publications.

Author	*Number of publication*		*Decrease (%)*	
	*Before**	*After**	*Overall*	*Nuclei*
A	87	78	-10,3	-6,33
B	51	40	-21,6	-14,9
C	50	43	-14	0
D	50	39	-22	-18,2
E	48	29	-39,6	-26,5

* Before and after the cleaning step.

Source: CWTS Web of Science Publication-level database.

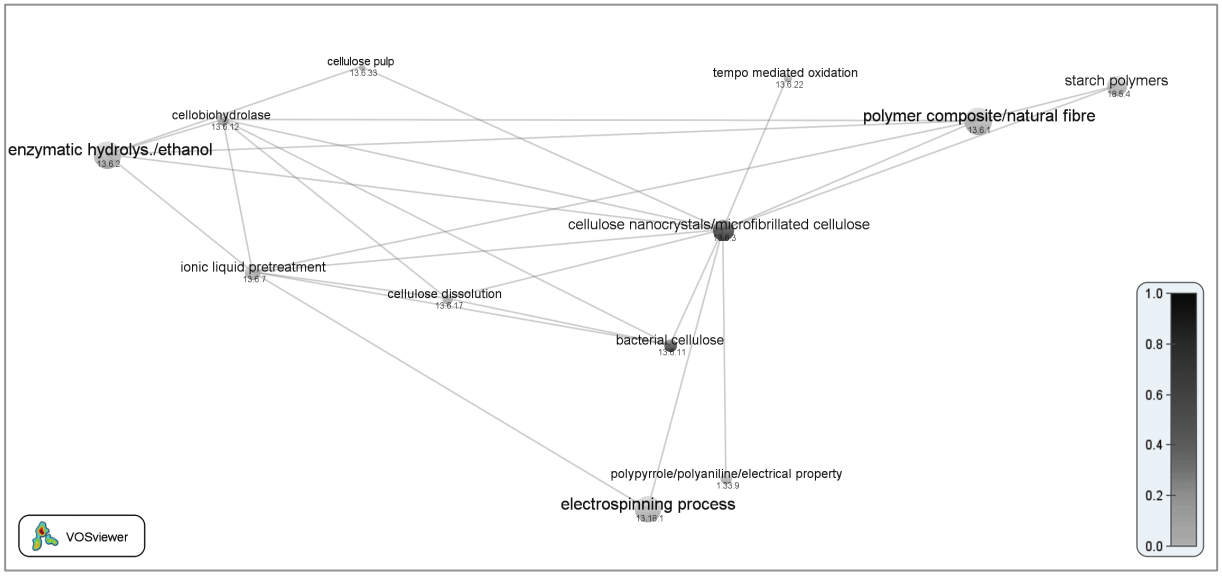
Map of the Nanocellulose research topics

The delineating approach was able to retrieve two nuclei research areas, one associated with cellulose nanocrystals and nanofibrils and other to bacterial cellulose. The peripheral research topics regards biodegradable polysaccharides (starch polymers), polymer composites based on natural fibres, and intrinsically conducting polymers. Other peripheral research areas included enzymatic hydrolyses and ethanol production, cellobiohydrolyse, cellulose pulp and cellulose dissolution, and ionic liquid pre-treatment. Electrospinning process and tempo mediated oxidation, which is an treatment that uses the chemical compound (2,2,6,6-Tetramethylpiperidin-1-yl)oxy (TEMPO), were also part of the final selection. These themes appears frequently in nanocellulose-focused studies (Azizi Samir, Alloin, & Dufresne, 2005; Charreau, Foresti, & Vazquez, 2013; Chirayil, Mathew, & Thomas, 2014; Dai et al., 2014; Domingues, Gomes, & Reis, 2014; Durán, Lemes, & Seabra, 2012; Eichhorn et al., 2010; Isogai, 2013; Klemm et al., 2011; Moon et al., 2011; Orts et al., 2005; Pääkkö et al., 2007; Siqueira et al., 2010; Siró & Plackett, 2010)

Figure 6 presents a map with these research topics (nodes). The map positions the topics on the basis of their citation relations. The closer two topics, the more frequent the citation traffic between them. The node labels match the main content of the clusters. Moreover, all selected clusters had their set of nanocellulose publication evaluated in the cleaning task.

The nuclei research areas are darker and positioned in the centre of the map. Research area 13.6.3 (cellulose nanocrystals/microfibrillated cellulose) has citation connections to all clusters. On the other hand, research topic 13.6.11 is connected only with four other clusters, which might indicate its lower relevance than the other nucleus research area. At the top right of the map are located two research areas addressed to starch polymers and polymer composites based on natural fibres. These research topics regard the development of sustainable materials (Durán et al., 2012; Moon et al., 2011; Siqueira et al., 2010; Isogai, 2013).

Research area concerning enzymatic hydrolysis is highly close to the research topic cellobiohydrolase, i.e., enzymes that perform the process of hydrolyse, and ionic liquid pre-treatment, which also relies on enzymatic approaches. However, they were located further than the nuclei clusters. Indeed, one of them was considered as highly noisy (13.6.2), but we should take into account that nanocellulose obtainment has been also studied as a secondary product of bio-ethanol production (Beecher, 2007; Zhu, Sabo, & Luo, 2011). Moreover, enzymatic pre-treatment has been researched to improve nanocellulose defibrillation (Pääkköet al. 2007; Moon et al., 2011; Klemm et al., 2011; Siqueira et al., 2010; Isogai, 2013).



Source: CWTS Web of Science Publication-level database.

Figure 6. Selected research area according to the procedure proposed.

At the bottom of the map, electrospinning process and conductive polymers were positioned closely, but there is no citation connection between them. Electrospinning is a technique used to produce micro- and nano-sized polymer-based fibres, and nanocellulose has been studied to improve the mechanical property of the final fibre (Dai et al., 2014). Nanocellulose electrical and magnetic properties have also been explored to be used with conductive polymers (Moon et al., 2011; Klemm et al., 2011). The other three research areas (cellulose pulp, cellulose dissolution and tempo mediated oxidation) are the smallest ones, and probably the publications that belong to them might be associated with other clusters on new updates performed using the classification system (Waltman & van Eck, 2012). Tempo mediated

oxidation is a current technique to perform pre-treatment of nanocellulose (Klemm et al., 2011; Isogai, 2013).

Conclusion

The proposed delineation procedure enabled us to retrieve relevant publications from research areas involving nanocellulose. Twelve research topics were identified, mapped and associated with current research challenges on nanocellulose. Two of them were highlighted as nuclei since they contain most part of the initial set of publications. The effect of the cleaning step on nuclei and peripheral clusters provided valuable feedback and demonstrated its importance to establishing relevant clusters afterwards. The independency test showed that the cleaning procedure could have been too rigorous and further research should be carried out to understand how it affected core authors' publication.

Delineating scientific fields is a complex task as boundaries are not frequently well established since scientific studies have become more complex and interdisciplinary. More and more exchange of knowledge between scientists from different disciplines is involved. Our approach retrieves and delineates the real nuclei and the peripheral research areas concerning nanocellulose studies. This clear separation provides suggestions for further research, putting the nuclei research in context. One of the ideas involves the knowledge flow from peripheral research topics to the nuclei areas. We intend to map how they provide the necessary knowledge to face nanocellulose current challenges and how country and scientific institutions are contributing to this evolution.

Acknowledgments

The authors are grateful to the São Paulo Research Foundation (process number 2012/16573-7) and comments from researchers of CWTS and NIT/Materiais. We are also thankful to the Graduate Program in Materials Science and Engineering at the Federal University of São Carlos for supporting this work.

References

Azizi Samir, M. A. S., Alloin, F., & Dufresne, A. (2005). Review of recent research into cellulosic whiskers, their properties and their application in nanocomposite field. *Biomacromolecules*, *6*(2), 612–26. doi:10.1021/bm0493685

Beecher, J. (2007). Wood , trees and nanotechnology. *Nature Nanotechnology*, *2*(August), 466–467.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., … Börner, K. (2011). Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches. *PloS One*, *6*(3), e18029. doi:10.1371/journal.pone.0018029

Charreau, H., Foresti, M. L., & Vazquez, A. (2013). Nanocellulose patents trends: a comprehensive review on patents on cellulose nanocrystals, microfibrillated and bacterial cellulose. *Recent Patents on Nanotechnology*, *7*(1), 56–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22747719

Chirayil, C. J., Mathew, L., & Thomas, S. (2014). Review of recent researhc in nanocellulose preparation from different lignocellulosic fibers. *Review of Advanced Materials Science*, *37*, 20–28.

Dai, L., Long, Z., Ren, X., Deng, H., He, H., & Liu, W. (2014). Electrospun polyvinyl alcohol/waterborne polyurethane composite nanofibers involving cellulose nanofibers. *Journal of Applied Polymer*, *41051*, 1–6. doi:10.1002/app.41051

Domingues, R. M. A., Gomes, M. E., & Reis, R. L. (2014). The potential of cellulose nanocrystals in tissue engineering strategies. *Biomacromolecules*, *15*, 2327–2346.

Dufresne, A. (2013). Nanocellulose: A new ageless bionanomaterial. *Materials Today*, *16*(6), 220–227.

Durán, N., Lemes, A. P., & Seabra, A. B. (2012). Review of cellulose nanocrystals patents: preparation, composites and general applications. *Recent Patents on Nanotechnology*, *6*(1), 16–28. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21875405

Eichhorn, S. J., Dufresne, A., Aranguren, M., Marcovich, N. E., Capadona, J. R., Rowan, S. J., … Peijs, T. (2010). Review: current international research into cellulose nanofibres and nanocomposites. *Journal of Materials Science*, *45*(1), 1–33. doi:10.1007/s10853-009-3874-0

Gardner, D. J., Opo, Oporto, G. S., Mills, R., & Samir, M. A. S. A. (2008). Adhesion and Surface Issues in Cellulose and Nanocellulose. *Journal of Adhesion Science and Technology*, *22*(5-6), 545–567. doi:10.1163/156856108X295509

Glanzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Gupta, D. . (1989). Scientometric study of biochemical literature of Nigeria, 1970-1984: application of Lotka's Law and the 80/20-rule. *Scientometrics*, *15*(3-4), 171–179.

Hullmann, A., & Meyer, M. (2003). Publications and patents in nanotechnology An overview of previous studies and the state of the art. *Scientometrics*, *58*(3), 507–527.

Igami, M. (2008). Exploration of the evolution of nanotechnology via mapping of patent applications. *Scientometrics*, *77*(2), 289–308. doi:10.1007/s11192-007-1973-8

Isogai, A. (2013). Wood nanocelluloses: fundamentals and applications as new bio-based nanomaterials. *Journal of Wood Science*, *59*(6), 449–459. doi:10.1007/s10086-013-1365-z

Juran, J. M., & Godfrey, A. B. (Eds.). (1998). *Juran's quality handbook* (5th ed., p. 1730). New York: McGraw-Hill.

Kao, C. (2009). The authorship and internationality of industrial engineering journals. *Scientometrics*, *81*(1), 123–136. doi:10.1007/s11192-009-2093-4

Klemm, D., Kramer, F., Moritz, S., Lindström, T., Ankerfors, M., Gray, D., & Dorris, A. (2011). Nanocelluloses: a new family of nature-based materials. *Angewandte Chemie (International Ed. in English)*, *50*(24), 5438–66. doi:10.1002/anie.201001273

Kostoff, R. N., Koytcheff, R. G., & Lau, C. G. Y. (2009). Seminal Nanotechnology Literature: A Review. *Journal of Nanoscience and Nanotechnology*, *9*(11), 6239–6270. doi:10.1166/jnn.2009.1465

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, *94*(2), 589–593. doi:10.1007/s11192-012-0784-8

Mariano, M., Kissi, N. El, & Dufresne, A. (2014). Cellulose nanocrystals and related nanocomposites: review of some properties and challenges. *Journal of Polymer Science*, *52*, 791–806. doi:10.1002/polb.23490

Milanez, D. H., Amaral, R. M. Do, Faria, L. I. L. De, & Gregolin, J. A. R. (2013). Assessing nanocellulose developments using science and technology indicators. *Materials Research*, *16*(3), 635–641. doi:10.1590/S1516-14392013005000033

Milanez, D. H., Faria, L. I. L., Amaral, R. M., Leiva, D. R., & Gregolin, J. A. R. (2014). Patents in nanotechnology: an analysis using macro-indicators and forecasting curves. *Scientometrics*. doi:10.1007/s11192-014-1244-4

Moed, H. F., Glänzel, W., & Schmoch, U. (2004). *Handbook of quantitative science and technology research: the use of publication and patent statistics in studies of S&T systems*. (H. F. Moed, W. Glanzel, & U. Schmoch, Eds.) (Kluwer Aca., p. 785). New York: Kluwer Academic Publishers.

Mogoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. *Research Policy*, *36*(6), 893–903. doi:10.1016/j.respol.2007.02.005

Moon, R. J., Martini, A., Nairn, J., Simonsen, J., & Youngblood, J. (2011). Cellulose nanomaterials review: structure, properties and nanocomposites. *Chemical Society Reviews*, *40*(7), 3941–94. doi:10.1039/c0cs00108b

Okubo, Y. (1997). Bibliometric Indicators and Analysis of Research Systems: Methods and Examples (No. 01). doi:101787/208277770603

Orts, W. J., Shey, J., Imam, S. H., Glenn, G. M., Guttman, M. E., & Revol, J.-F. (2005). Application of Cellulose Microfibrils in Polymer Nanocomposites. *Journal of Polymers and the Environment*, *13*(4), 301–306. doi:10.1007/s10924-005-5514-3

Pääkkö, M. et al. (2007). Enzymatic hydrolysis combined with mechanical shearing and high-pressure homogenization for nanoscale cellulose fibrils and strong gels. *Biomacromolecules*, *8*(6), 1934–41. doi:10.1021/bm061215p

Raan, A. F. J. Van. (2014). Advances in bibliometric analysis : research performance assessment and science mapping. In W. Blockmans, L. Engwall, & D. Weaire (Eds.), *Bibliometrics: Use and abuse in the review of research performance* (Vol. c, pp. 17–28). Portland.

Reuters, T. (2014). Web of Science. Retrieved March 03, 2014, from http://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&SID=4CGNLFV3uyF6vP2Mtvi&search_mode=GeneralSearch

Siqueira, G., Bras, J., & Dufresne, A. (2010). Cellulosic bionanocomposites: a review of preparation, properties and applications. *Polymers*, *2*(4), 728–765. doi:10.3390/polym2040728

Siró, I., & Plackett, D. (2010). Microfibrillated cellulose and new nanocomposite materials: a review. *Cellulose*, *17*(3), 459–494. doi:10.1007/s10570-010-9405-y

Stephens, J., Hubbard, D. E., Pickett, C., & Kimball, R. (2013). Citation Behavior of Aerospace Engineering Faculty. *The Journal of Academic Librarianship*, *39*(6), 451–457. doi:10.1016/j.acalib.2013.09.007

TAPPI. (2011). Roadmap for the development of international standards for nanocellulose (p. 36).

Van Raan, A. F. J. (2004). *Science meansuring. In Handbook of quantitative science and technology research: the use of publication and patent statistics in studies of S&T systems* (pp. 19–50). New York: Kluwer Academic Publishers.

Waltman, L. & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392. doi:10.1002/asi.22748

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629–635. doi:10.1016/j.joi.2010.07.002

Wang, L., Notten, A., & Surpatean, A. (2012). Interdisciplinarity of nano research fields: a keyword mining approach. *Scientometrics*, *94*(3), 877–892. doi:10.1007/s11192-012-0856-9

Zhu, J. Y., Sabo, R., & Luo, X. (2011). Integrated production of nano-fibrillated cellulose and cellulosic biofuel (ethanol) by enzymatic fractionation of wood fibers. *Green Chemistry*, *13*(5), 1339. doi:10.1039/c1gc15103g

Sapientia: the Ontology of Multi-dimensional Research Assessment

Cinzia Daraio[1], Maurizio Lenzerini[1], Claudio Leporelli[1], Henk F. Moed[1], Paolo Naggar[2], Andrea Bonaccorsi[3], Alessandro Bartolucci[2]

[1] *daraio@dis.uniroma1.it; lenzerini@dis.uniroma1.it, leporelli@dis.uniroma1.it; henk.moed@uniroma1.it;*
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

[2] *paolo.naggar@gmail.com; alessandro_bartolucci@fastwebnet.it*
Studiare Ltd., Rome (Italy)

[3] *a.bonaccorsi@gmail.com*
DISTEC, University of Pisa, Pisa (Italy)

Abstract
This paper proposes an Ontology-Based Data Management (OBDM) approach to a multi-dimensional research assessment. It is shown that an OBDM approach is able to take into account the recent trends in quantitative studies of Science, Technology and Innovation, including computerization of bibliometrics, multidimensionality of research assessment, altmetrics, and, more generally, the generation of new indicators with higher granularity and cross-referencing specificities according to increasingly demanding policy needs. The main features of *Sapientia* are presented, the Ontology of Multi-dimensional Research Assessment, developed within a project funded by the University of Rome La Sapienza. Illustrative examples are given of its usefulness for the specification of well known as well as recently developed indicators of research assessment.

Conference Topics
Methods and techniques; Indicators; Science policy and research assessment

Introduction: An Ontology-Based-Data-Management Approach to Multi-Dimensional Research Assessment

The quantitative analysis of Science and Technology is becoming a "big data" science, with an increasing level of "computerization", in which large and heterogeneous datasets on various aspects of Science, Technology and Innovation (STI) are combined. Within this framework, optimistic views, supporting "the end of theory" in favour of data-driven science (Kitchin, 2014), have been opposed to more critical positions in favour of theory-driven scientific discoveries (Frické, 2014) while a more balanced view emerged from a critical analysis of the current existing literature (Ekbia et al., 2015), leading the information systems community to further deeply analyse the critical challenges posed by the big data development (Agarwal, 2014). It has been rightly highlighted that "Data are not simply addenda or second-order artifacts; rather, they are the heart of much of the narrative literature, the protean stuff that allows for inference, interpretation, theory building, innovation, and invention" (Cronin, 2013, p. 435). Moreover, the need for accountability of STI activities to sustain their funding in the current difficult economic and financial situation is increasingly asking for rigorous empirical evidence to support informed policy making. Indeed, the needs to overcome the logic of rankings and the new trends in indicators development, including granularity and cross-referencing, can be explored and exploited in open data platforms with a clear description of the main concepts of the domain (Daraio & Bonaccorsi, 2015). The multidimensionality of research assessment and scholarly impact (Moed & Halevi, 2015), and the recent altmetrics movements (Cronin & Sugimoto, 2014), are questioning the traditional approach in indicators development.

Research assessment, indeed, is becoming increasingly complex due to its multi-dimensionality nature. A Report published in 2010 by the Expert Group on the Assessment of University-Based Research, installed by the European Commission proposed "a consolidated multidimensional methodological approach addressing the various user needs, interests and purposes, and identifying data and indicator requirements" (AUBR, 2010, p. 10). A key notion holds that "indicators designed to meet a particular objective or inform one target group may not be adequate for other purposes or target groups". Diverse institutional missions, and different policy environments and objectives require different assessment processes and indicators. In addition, the range of people and organizations requiring information about university-based research is growing. Each group has specific but also overlapping requirements (AUBR, 2010, p. 51).

Table 1. Main types of research outputs.

Printed outputs (texts)	**Non-printed outputs (non-text)**	**Main type of impact**
Scientific journal paper; book chapter; scholarly monograph	Research data file; video of experiment; software	Scientific-scholarly
Patent; commissioned research report;	New product or process; material; device; design; image; spin off	Economic or technological
Professional guidelines; newspaper article; communication submitted to social media, including blogs, tweets.	Interview; event; art performance; exhibit; artwork; scientific-scholarly advise;	Social or cultural

A research assessment has to take into account a range of different types of research output and impact. As regards output forms, one important distinction is between text-based and non-text based output forms. The main types are presented in Table 1. This table is not fully comprehensive. The specifications of the Panel Criteria in the Research Excellence Framework in the UK (REF, 2012, page 51 a.f.) provide more detailed lists of possible output forms arranged by major research discipline. Table 1 includes forms that are becoming increasingly important such as research data files, and communications submitted to social media and scholarly blogs. A framework for the assessment of these forms is being developed in the field of altmetrics (e.g., Taylor, 2013). The last column indicates the main types of impact a particular output may have. A distinction is made between scientific-scholarly impact, and wider impact outside the domain of science and scholarship, denoted as "societal", a concept that embraces technological, economic, social and cultural impact. A comprehensive overview of the types of impact, and the most frequently used impact indicators is presented in Table 2. The reader is referred to AUBR (2010 and Moed & Halevi (2015) for a further discussion of this table.

It is also important to include the inputs in the analysis; they should be jointly analysed with the outputs to assess the overall impact of the process (see e.g. Daraio et al., 2014, for a conditional multidimensional approach to rank higher education institutions). To meet all these new trends and policy needs a shift in the paradigm of the data integration for research assessment is needed. In this paper we advocate an OBDM approach to research assessment. This new approach radically changes the traditional paradigm of construction of STI indicators and offers a flexible and powerful tool for designing new indicators and develop rigorous policy making. The confidence in this new approach comes from three directions: (i) recent efforts from policy makers to support the creation of new datasets on S&T; (ii) bottom up standardization initiatives; (iii) development of almetrics and web-based indicators. To start with, in the last few years, several initiatives at European level have been based on an intense production and use of new data.

Table 2. Types of Research Impact and Indicators.

Type of impact	Short Description; Typical examples	Indicators (examples)
Scientific-scholarly or academic		
Knowledge growth	Contribution to scientific-scholarly progress: creation of new scientific knowledge	Indicators based on publications and citations in peer-reviewed journals and books
Research networks	Integration in (inter)national scientific-scholarly networks and research teams	(inter)national collaborations including co-authorships; participation in emerging topics
Publication outlets	Effectiveness of publication strategies; visibility and quality of used publication outlets	Journal impact factors and other journal metrics; diversity of used outlets;
Societal		
Social	Stimulating new approaches to social issues; informing public debate and improve policy-making; informing practitioners and improving professional practices; providing external users with useful knowledge; Improving people's health and quality of life; Improvements in environment and lifestyle;	▪ Citations in medical guidelines or policy documents to research articles ▪ Funding received from end-users ▪ End-user esteem (e.g., appointments in (inter)national organizations, advisory committees) ▪ Juried selection of artworks for exhibitions ▪ Mentions of research work in social media
Technological	Creation of new technologies (products and services) or enhancement of existing ones based on scientific research	Citations in patents to the scientific literature (journal articles)
Economic	Improved productivity; adding to economic growth and wealth creation; enhancing the skills base; increased innovation capability and global competitiveness; uptake of recycling techniques;	▪ Revenues created from the commercialization of research generated intellectual property (IP) ▪ Number patents, licenses, spin-offs ▪ Number of PhD and equivalent research doctorates ▪ Employability of PhD graduates
Cultural	Supporting greater understanding of where we have come from, and who and what we are; bringing new ideas and new modes of experience to the nation.	▪ Media (e.g. TV) performances ▪ Essays on scientific achievements in newspapers and weeklies ▪ Mentions of research work in social media

Legend to Table 2: Partly based on AUBR (2010) and Moed & Halevi (2015)

In the field of data on universities, the pioneering efforts of Aquameth (Daraio et al., 2011; Bonaccorsi & Daraio, 2007) and subsequently of Eumida (Bonaccorsi, 2014) have been transformed in an institutional initiative called ETER (European Tertiary Education Register), which will make publicly available microdata on universities in 2015. In the same field, the mapping of diversity of European institutions (Huisman, Meek & Wood, 2007; van Vught, 2009) led to the experimental project U-Map, after which there has been an institutional effort towards a multidimensional ranking exercise, called U-Multiranking (van Vught & Westerheijden, 2010). In the field of Public Research Organisations, there has been an effort to build up a comprehensive list of institutions and to survey their activities within the European Research Area (ERA) context. The results of the large ERA surveys, run in 2013 and 2014, will be made available in 2015. These efforts from Europe have a major counterpart on the other side of the Atlantic, where the STAR Metrics initiative (see https://www.starmetrics.nih.gov/) has promoted a federal and research institution collaboration to create a repository of data and tools that is producing extremely interesting results. All these efforts, however, are based on the construction of new datasets, or the integration of existing datasets into new ones. They do not solve the issue of comparability

and standardization of information and of inter-operability, updating and scalability of databases. It is interesting to observe that, in parallel to these efforts put in place by public institutions and policy makers, there have also been massive bottom up efforts aimed at standardizing the elementary pieces of information. Moreover, these efforts have been based on the construction of partial ontologies. Consider the following.

- ORCID (http://orcid.org/) is a non-profit organization, supported by research organizations, agencies, providers of publication management systems, and publishers, aiming at giving all researchers a unique identifier (ORCID_id number) and keeping it persistent over time. Established at the end of 2009, but operational since end 2012, it has almost reached one million researchers worldwide. Most of the increase has been achieved in a very short time frame: from 100,000 in March 2013 to almost 970,000 as of October 2014 (with 35% from European, Middle East and Asian countries);
- CERIF is a Europe-based initiative aiming at standardizing the operations of funding agencies, with the help of a full-scale ontology of almost all research products (http://www.eurocris.org);
- CASRAI (www.casrai.org) is a Canada-US initiative for the standardization of data on research institutions and funders (also supported by a committee of Science Europe; http://www.scienceeurope.org/scientific-committees/Life-sciences/life-sciences-committee);
- ISNI (www.isni.org) provides lists and metadata on higher education, research, funding and many other types of organizations, while Ringgold (www.ringgold.com) does the same in the world of publishers and intermediaries.

These initiatives are strongly supported by international scientific associations (see for example CODATA http://www.codata.org and the VIVO network of scientists: http://www.vivoweb.org/).

Finally, the rapid growth of alternative metrics and web-based metrics has also created a large space for the production of data from publicly available and other sources (Cronin & Sugimoto, 2014). Summing up, there are powerful trends that point to the need to change the overall philosophy of the production of S&T indicators. Instead of an environment in which indicators are produced in close circles, by constructing ad hoc databases, with no built-in interoperability, updating and scalability features, we have to move towards an environment in which elementary pieces of information are fully standardized, micro-data consistent with standardized definitions are (mostly) publicly available, and indicators are constructed following the policy demands on the basis of stable platforms constantly integrated and updated, instead of starting from scratch each time a new indicator is needed.

Main advantages of an OBDM approach compared to conventional data-base integration approaches

While the amount of data stored in current information systems and the processes making use of such data continuously grow, turning these data into information, and governing both data and processes are still tremendously challenging tasks for Information Technology. The problem is complicated due to the proliferation of data sources and services both within a single organization, and in cooperating environments. The following factors explain why such a proliferation constitutes a major problem with respect to the goal of carrying out effective data governance tasks:

- Although the initial design of a collection of data sources and services might be adequate, corrective maintenance actions tend to re-shape them into a form that often diverges from the original conceptual structure.
- It is common practice to change a data source (e.g., a database) so as to adapt it both to specific application-dependent needs, and to new requirements. The result is that

data sources often become data structures coupled to a specific application (or, a class of applications), rather than application-independent databases.
- The data stored in different sources and the processes operating over them tend to be redundant, and mutually inconsistent, mainly because of the lack of central, coherent and unified coordination of data management tasks.

The result is that information systems of medium and large organizations are typically structured according to a "sylos"-based architecture, constituted by several, independent, and distributed data sources, each one serving a specific application. This poses great difficulties with respect to the goal of accessing data in a unified and coherent way. Analogously, processes relevant to the organizations are often hidden in software applications, and a formal, up-to-date description of what they do on the data and how they are related with other processes is often missing. The introduction of service-oriented architectures is not a solution to this problem per se, because the fact that data and processes are packed into services is not sufficient for making the meaning of data and processes explicit. Indeed, services become other artifacts to document and maintain, adding complexity to the governance problem. Analogously, data warehousing techniques and the separation they advocate between the management of data for the operation level, and data for the decision level, do not provide solutions to this challenge. On the contrary, they also add complexity to the system, by replicating data in different layers of the system, and introducing synchronization processes across layers. All the above observations show that a unified access to data and an effective governance of processes and services are extremely difficult goals to achieve in modern information systems. Yet, both are crucial objectives for getting useful information out of the information system, as well as for taking decisions based on them. This explains why organizations spend a great deal of time and money for the understanding, the governance, the curation, and the integration of data stored in different sources, and of the processes/services that operate on them, and why this problem is often cited as a key and costly Information Technology challenge faced by medium and large organizations today (Bernstein & Haas, 2008).

We argue that ontology-based data management (OBDM, Lenzerini, 2011) is a promising direction for addressing the above challenges. The key idea of OBDM is to resort to a three-level architecture, constituted by the ontology, the sources, and the mapping between the two. The ontology is a conceptual, formal description of the domain of interest to a given organization (or, a community of users), expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge. The data sources are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others. The mapping is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology.

The main purpose of an OBDM system is to allow information consumers to query the data using the elements in the ontology as predicates. In this sense, OBDM can be seen as a form of information integration, where the usual global schema is replaced by the conceptual model of the application domain, formulated as an ontology expressed in a logic-based language. With this approach, the integrated view that the system provides to information consumers is not merely a data structure accommodating the various data at the sources, but a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships between such concepts. The distinction between the ontology and the data sources reflects the separation between the conceptual level, the one presented to the client, and the logical/physical level of the information system, the one stored in the sources, with the

mapping acting as the reconciling structure between the two levels. This separation brings several potential advantages:

- The ontology layer in the architecture is the obvious mean for pursuing a declarative approach to information integration, and, more generally, to data governance. By making the representation of the domain explicit, we gain re-usability of the acquired knowledge, which is not achieved when the global schema is simply a unified description of the underlying data sources.
- The mapping layer explicitly specifies the relationships between the domain concepts on the one hand and the data sources on the other hand. Such a mapping is not only used for the operation of the information system, but also for documentation purposes. The importance of this aspect clearly emerges when looking at large organisations where the information about data is widespread into separate pieces of documentation that are often difficult to access and rarely conforming to common standards. The ontology and the corresponding mappings to the data sources provide a common ground for the documentation of all the data in the organisation, with obvious advantages for the governance and the management of the information system.
- A third advantage has to do with the extensibility of the system. One criticism that is often raised to data integration is that it requires merging and integrating the source data in advance, and this merging process can be very costly. However, the ontology-based approach we advocate does not impose to fully integrate the data sources at once. Rather, after building even a rough skeleton of the domain model, one can incrementally add new data sources or new elements therein, when they become available, or when needed, thus amortising the cost of integration. Therefore, the overall design can be regarded as the incremental process of understanding and representing the domain, the available data sources, and the relationships between them. The goal is to support the evolution of both the ontology and the mappings in such a way that the system continues to operate while evolving, along the lines of "pay-as-you-go" data integration pursed in the research on data-spaces (Sarma et al., 2008).

The notions of ODBM were introduced in (Calvanese et al. 2007; Poggi et al. 2008), and originated from several disciplines, in particular, Information Integration, Knowledge Representation and Reasoning, and Incomplete and Deductive Databases. The central notion of OBDM is therefore the ontology, and reasoning over the ontology is at the basis of all the tasks that an OBDM system has to carry out. In particular, the axioms of the ontology allow one to derive new facts from the source data, and these inferred facts greatly influence the set of answers that the system should compute during query processing. In the last decades, research on ontology languages and ontology inferencing has been very active in the area of Knowledge Representation and Reasoning. Description Logics (DLs, Baader et al., 2007) are widely recognized as appropriate logics for expressing ontologies, and are at the basis of the W3C standard ontology language OWL. These logics permit the specification of a domain by providing the definition of classes and by structuring the knowledge about the classes using a rich set of logical operators. They are decidable fragments of mathematical logic, resulting from extensive investigations on the trade-off between expressive power of Knowledge Representation languages, and computational complexity of reasoning tasks. Indeed, the constructs appearing in the DLs used in OBDI are carefully chosen taking into account such a trade-off (Calvanese et al., 2007).

As indicated above, the axioms in the ontology can be seen as semantic rules that are used to complete the knowledge given by the raw facts determined by the data in the sources. In this sense, the source data of an OBDI system can be seen as an incomplete database, and query answering can be seen as the process of computing the answers logically deriving from the

combination of such incomplete knowledge and the ontology axioms. Therefore, at least conceptually, there is a connection between OBDM and the two areas of incomplete information (Imielinski & Lipski, 1984) and deductive databases (Ceri et al., 1990).

Sapientia at a glance

The main objective of *Sapientia* is to model all the activities relevant for the evaluation of research and for assessing its impact. For impact, in a broad sense, we mean any effect, change or benefit, to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia.

Table 3. Modules of the Sapientia Ontology.

N.	Module Name	Module Description
1	Overview	presents the terminological inventory needed to define the ontology domain: what is to be known to assess research activities and their impact on human knowledge and the economic system
2	Agent	models the individuals involved in the world of research, carrying out knowledge-related activities
3	Activity	models the main knowledge related activities matching them with public and relevant commitments of the agents involved in the domain (each module from 4 to 11 is devoted to a kind of knowledge-related activity - the module name corresponds to the appropriate specialization of the concept *Activity*)
4	*Research activity*	models, among the knowledge-related activities, those that allow the scientific community to advance the state of the art of knowledge
5	*Educational_activity*	models, among the knowledge-related activities, those that allow people to improve their knowledge
6	*Conferring degrees activity*	models, among the knowledge-related activities, those that grant degrees allowing people to widely qualify themselves
7	*Publishing activity*	models, among the knowledge-related activities, those that allow people to know the results of research activities
8	*Preservation activity*	models, among knowledge-related activities, those that permit the preservation of the value of things (related to research activities)
9	*Funding activity*	models, among the knowledge-related activities, those that assign and distribute the funds needed to carry out research, educational and service activities
10	*Inspecting activity*	models, among the knowledge-related activities, those that control and assess research, educational and service activities
11	*Producing activity*	models, among the knowledge-related activities, those that produce economic, society and cultural value
12	Space	models the space and its roles
13	Taxonomy	models the relevant taxonomies that classify the elements of the domain
14	Time	models the depth of time of the domain (this module is spread through the others)

Hence, *Sapientia* covers what is to be known about assess research activities and their impact on human knowledge and the economic system. For this purpose the ontology embraces:

- the inter-relationships between research activities (Modules Research_activity, Publishing_activity);

• the relationships between research activities and people's personal knowledge (Modules Teaching_activity, Conferring_degrees_activity, Publishing_activity, Producing_activity);
• the relationships between research activities and other missions of individuals and institutions (Modules Inspect_activity, Producing_activity);
• the relationship between research activities and the knowledge locally available to the companies in the economic system, enabling their innovative behavior (Module Producing_activity).

The *Sapientia* ontology includes also the activities that are needed for fostering these relationships (Modules Preservation_activity, Inspecting_activity and Funding_activities). The 14 modules that compose *Sapientia* are listed in Table 3.

Modelling choices

We pursued a modelling approach based on processes, which were conceived as collections of activities. A process is composed by inputs and outputs. Individuals and activities are the main pillars of the ontology.

We consider the building of descriptive, interpretative, and policy models of our domain as a distinct step with respect to the building of the domain ontology. However, the ontology will intermediate the use of data in the modelling step, and should be rich enough to allow the analyst the freedom to define any model she considers useful to pursue her analytic goal.

Obviously, the actual availability of relevant data will constrain both the mapping of data sources on the ontology, and the actual computation of model variables and indicators of the conceptual model. However, the analyst should not refrain from proposing the models that she considers the best suited for her purposes, and to express, using the ontology, the quality requirements, the logical, and the functional specification for her ideal model variables and indicators. This approach has many merits, and in particular:

- it allows the use of a common and stable ontology as a platform for different models;
- it addresses the efforts to enrich data sources, and verify their quality;
- it makes transparent and traceable the process of approximation of variables and models when the available data are less than ideal;
- it makes use of every source at the best level of aggregation, usually the atomic one.

More generally, this approach is consistent with the effort of avoiding "the harm caused by the blind symbolism that generally characterizes a hasty mathematization" put forward by Georgescu Roegen in his seminal work on production models and on methods in economic science (Georgescu-Roegen, 1970, 1971, 1979). In fact, one can verify the logical consistency of the ontology and compute answers to unambiguous logical queries.

Moreover, the proposed ontology allows us to follow the Georgescu-Roegen approach also in the use of the concept of process. We can analyze the knowledge production activities, at an atomic level, considering their *time* dimension and such *funds* as the cumulated results of previous research activities, both those available in relevant publications, and those embodied in the authors' competences and potential, the infrastructure assets, and the time devoted by the group of authors to current research projects. Similarly, we can analyze the output of teaching activities, considering the joint effect of *funds* such as the competence of teachers, the skills and the initial education of students, and educational infrastructures and resources. Thirdly, service activities of research and teaching institutions provide infrastructural and knowledge assets that act as a *fund* in the assessment of the impact of those institutions on the innovation of the economic system. The perimeter of our domain should allow us to consider the different channels of transmission of that impact: mobility of researchers, career of alumni, applied research contracts, joint use of infrastructures, and so on. In this context,

different theories and models of the system of knowledge production could be developed and tested (Etzkowitz & Leydesdorff, 2000).

Table 4. Indicators considered for the test of the completeness of Sapientia.

#	Indicator (I)	*Sapientia's Modules*												
		2	3	4	5	6	7	8	9	10	11	12	13	14
I1	Number of published articles	A					F					F,D1	D1	D2
I2	Number of citations	A					F					F,D1	D1	D2
I3	Citations per article	A					F					F,D1	D1	D2
I4	Normalized citation rate	A					F					F,D1	D1	D2
I5	Highly cited publications	A					F					D1	D1	D2
I6	Journal Impact Factor	A					F						F	D2
I7	Subject Normalized Impact Factor	A					F						F	D2
I8	Scimago Journal Ranking Impact Factor	A					F						F	D2
I9	H-index	F					F					A	F	D2
I10	E-index	F					F					A	F	D2
I11	Number of patents	A									F	F,D1	D1	D2
I12	Full text article download count	A					F					F,D1	D1	D2
I13	Mentions in social media	A					F					F,D1	D1	D2
I14	Research output per academic staff	A					F					F,D1	D1	D2
I15	Percentage of Highly Cited Publications	A					F					D1	D1	D2
I16	Number of keynote addresses at conferences	A					F					F,D1	D1	D2
I17	Number of prestigious awards and prizes	A								F		F,D1	D1	D2
I18	Number of visiting research appointments	F,A										D1	D1	D2
I19	Member of editorial board	A					F					D1	D1	D2
I20	Refereeing activity for journals	A	F							F		D1	D1	D2
I21	External research income	A							F		F	D1	D1	D2
I22	Number of competitive grants won	A							F		F	D1	D1	D2
I23	Percentage of competitive grants	A							F		F	D1	D1	D2
I24	External research income per academic staff	A							F		F	F,D1	D1	D2
I25	Employability of PhD graduates	A				F						D1	F,D1	D2
I26	Commerc. of research generated intellectual property	A									F	D1	D1	D2
I27	End-users esteem	A								F		D1	D1	D2
I28	Number of funding from end-users	A,D1							F			D1	D1	D2
I29	Percentage of funding from end-users	A,D1							F			D1	D1	D2
I30	Post-graduate research student load	A		F	F							D1	D1	D2
I31	Involvement of early career researchers in teams	A,F		F								D1	D1	D2
I32	Number of collaborations and partnerships	F		A								F,D1	D1	D2
I33	Doctoral completions	A				F						D1	D1	D2
I34	Research active academics	F,A					F					D1	D1	F,D2
I35	Percentage of research active per total academic staff	F,A					F					D1	D1	F,D2
I36	Total R&D investment	A							F			D1	D1	D2
I37	Research Infrastructures and facilities	A					F	F				D1	D1	D2
I38	Research ethics						F					A,F	F,D1	D2

Testing the Ontology: analysis of the competency questions

One way to check if the ontology contains all the relevant information and/or details to represent the domain of interest, currently used in knowledge representation, is based on the specification of competency questions (Gruninger & Fox 1995). These questions correspond to check whether the ontology contains enough information to answer these types of questions or whether the answers require a particular level of detail or representation of a particular module of the ontology that needs to be further developed. The analysis of the competency questions of *Sapientia* has been carried out on the indicators contained in the paper by Moed and Halevi (2015), integrated with the additional indicators reported in the AUBR (2010) document. In addition, other key references of the ontological commitments have been Moed, Glanzel and Schmock (2004), Moed (2005) and Cronin and Sugimoto (2014), together with the knowledge background of the team of the project.

Table 4 contains the list of indicators considered for the verification of the competency questions. Associated to each indicator are reported the following pieces of information:

- Facts (F) are the content of the data, the relevant information about atomic events relevant for the construction of the indicator;
- Aggregation level (A) is the minimal aggregation level: the concept which classifies the objects included in the indicator;
- Dimensions of the analysis (D), are descriptive properties which are relevant to access higher level of aggregation. They are evaluated by the dimension of taxonomy (D1) and that of time (D2).

Table 5 summarizes the number of facts (F), aggregations (A) and dimensions (D) by module, as reported in Table 4, to check the comprehensiveness of *Sapientia* with respect to the indicators listed therein. Put it in another way, we checked whether our ontology was able to include all the relevant conceptual information requested by the specification of the listed indicators in Table 4. The answer to this question is indeed positive.

Table 5. Some statistics on the "usage" of the Ontology modules.

	2	3	4	5	6	7	8	9	10	11	12	13	14
F	7	1	2	1	2	20	1	7	3	6	13	5	2
A	34	0	1	0	0	0	0	0	0	0	2	1	0
D	2	0	0	0	0	0	0	0	0	0	31	31	38

By inspecting Table 5 it clearly appears that only a few modules are used for the specification of the indicators reported in Table 4. This means that our ontology covers a much broader conceptual domain with respect to the one underlying (even if not formally specified) by the indicators reported in Table 4. The most frequently used module is the Publishing module (7), followed by Space (12) and Funding (9). We note that the modules 12 (Space), 13 (Taxonomy) and 14 (Time) are used in the majority of the cases to further characterize the dimensions of the considered indicators.

A new way to conceive and specify STI indicators

By adopting an OBDM perspective a new approach to designing indicators can be implemented. This new approach aligns very well with the recent trends described in the introduction.

The traditional approach to indicators' design is based on informal definitions expressed in a natural language (English, typically). An indicator is defined as a relationship between variables, e.g. a ratio between number of publications per academic staff, chosen among a predefined set of data collected and aggregated ad hoc, by a private or a public entity, according to the user needs, and hence not re-usable for future assessment and use.

The OBDM approach we pursue in this paper permits a *more advanced specification* of an indicator according to the following dimensions:

- the *ontological dimension*. It represents the domain (portion) of the reality to be measured by the indicator (obviously, in the scope of this paper, all indicators will share the *Sapientia* ontology as their ontological part);
- the *logical dimension*. It denotes the question that has to be asked to the ontological portion in order to retrieve all the information (data) needed for calculating the indicator value. In this case the data are extracted from the sources through the mapping considering the logical specification of the query;
- the *functional dimension*. It indicates the mathematical expression that has to be applied on the result of the logical extraction of data carried out in the previous point in order to calculate the indicator value;
- the *qualitative dimension*. It specifies the questions that have to be asked to the ontological part in order to generate the list of problems affecting the meaningfulness of the calculated indicator. An indicator will be considered meaningful if the list of its problems is empty.

In addition to the advantages of the OBDM recalled in previous sections above, the main specific benefits of this approach for designing indicators are the following:

1. It offers a space to *freely* explore the *generation of new indicators,* not previously specified by users, thanks to the *multiple inheritance* in the hierarchy of the concepts (a concept can be subsumed in several concepts).
2. For standard indicators specified by the users it can be seen immediately what is *missing* or which *problems* exist to calculate them;
3. It provides more alternatives and diagnostic ways to check the *robustness* of indicators with respect to opportunistic behaviour and the general goals of the assessment;
4. The formal specification of the indicators is made *independently* of the data. In this way, when applied to heterogeneous data sources, OBDM offers the opportunity to compute "comparable" indicator values at different level of aggregation. Moreover, it offers a reference system to *check the comparability* level among the heterogeneous sources of data and to identify where to invest in order to overcome the remaining existing comparability problems.
5. This approach permits an *unambiguous* way to define and compute the indicators. The indicator is calculated always in the same way.

Conclusions and further developments

In this paper we advocated the use of an OBDM approach to research assessment. We explained the reasons why a paradigm shift in research assessment is needed and outlined the main advantages of an OBDM approach over traditional databases integration approaches. We described the main objectives and structure of *Sapientia* the Ontology of Multi-dimensional Research Assessment. Finally, we illustrate the new indicator design methodology implicitly provided by an OBDM approach.

Sapientia 1.0 has been closed on the 22nd December 2014 and consisted of around 350 symbols (including concepts, relations and attributes). The full documentation of the Ontology is under way together with the mapping with several sources of data. Due to the works on the documentation and the mapping with the data in progress, as well as the limited number of pages available, we concentrated our presentation on the methodological aspects related to the development of the Sapientia.

We believe in fact that it will open a new stream of studies to further explore and exploit the OBDM approach for STI indicator designers and policy makers.

Acknowledgments

The financial support of the "Progetto di Ateneo 2013", University of Rome La Sapienza, is gratefully acknowledged.

References

Agarwal, R., & Dhar, V. (2014). Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. I*nformation Systems Research, 25*(3), 443-448.

AUBR Expert Group (2010). Expert Group on the Assessment of University-Based Research. Assessing Europe's University-Based Research. European Commission – DG Research. EUR 24187 EN

Baader F., D. Calvanese, D. McGuinness, D. Nardi, & P. F. Patel-Schneider, (eds) (2007). *The Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, 2nd edition.

Bernstein P. A. & Haas L.(2008). Information integration in the enterprise. *Communication of the ACM*, *51*(9), 72–79.

Bonaccorsi A., & Daraio C. (eds.) (2007) *Universities and strategic knowledge creation. Specialization and performance in Europe*. Cheltenham, Edward Elgar.

Bonaccorsi, A. (ed.) (2014) *Knowledge, diversity and performance in European higher education*. Cheltenham, Edward Elgar.

Calvanese D., G. De Giacomo, D. Lembo, M. Lenzerini, & R. Rosati (2007), Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated Reasoning*, *39*(3), 385–429.

Ceri S., G. Gottlob, & L. Tanca (1990). *Logic Programming and Databases*. Springer, Berlin (Germany).

Console M., Lembo D., Santarelli V. & Savo D.F. (2014a). Graphol: Ontology Representation Through Diagrams. *Proc. of the 27th Int. Workshop on Description Logic*.

Console M., Lembo D., Santarelli V., & Savo D.F. (2014b). Graphical Representation of OWL 2 Ontologies through Graphol. *Proc. of the 13th International Semantic Web Conference Posters & Demos*.

Cronin B. & Sugimoto C. (ed) (2014). Beyond bibliometrics. Harnessing multidimensional indicators of scholarly impact. MIT Press, Cambridge Mass.

Cronin, B. (2013). Thinking about data. *Journal of the American Society for Information Science and Technology*, *64*(3), 435–436.

Daraio, C. et al. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy 40*, 148–164.

Daraio C. & Bonaccorsi A. (2015). Beyond university rankings? Generating new indicators on universities by linking data in open platforms. under review for *JASIST*.

Daraio, C., Bonaccorsi A., & Simar L. (2015). Rankings and university performance: a conditional multidimensional approach. *European Journal of Operational Research*, *244*, 918–930.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., ... & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology*.

Etzkowitz H., & L. Leydesdorff. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations, *Research Policy*, *29*(2), 109-123.

Frické, M. (2014). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4), 651-661.

Georgescu-Roegen, N. (1970). The economics of production. *The American Economic Review, 60*(2), 1-9.

Georgescu-Roegen, N. (1972). Process analysis and the neoclassical theory of production, *American Journal of Agricultural Economics,* 279-294.

Georgescu-Roegen, N. (1979). Methods in economic science, *Journal of Economic Issues,* 317-328.

Gruninger, M. & Fox, M.S. (1995). Methodology for the Design and Evaluation of Ontologies. In: *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing*, IJCAI-95, Montreal.

Huisman, J., Meek, V.L. & Wood, F.Q. (2007). Institutional diversity in higher education: a cross-national and longitudinal analysis, *Higher Education Quarterly,* 61/4: 563-577.

Imielinski T. & W. Lipski, Jr. (1984) Incomplete information in relational databases. *Journal of the ACM, 31*(4), 761–791.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 1-12.

Lenzerini M. (2011). Ontology-based data management, *CIKM* 2011, 5-6.

Moed, H.F. (2005). *Citation Analysis in Research Evaluation*, Springer NY.

Moed, W. Glanzel & U. Schmoch (ed.) (2004), *Handbook of Quantitative Science and Technology Research,* Kluwer Academic Publishers, 51-74.

Moed, H. F., & Halevi, G. (2015). The Multidimensional Assessment of Scholarly Research Impact, *Journal of the American Society for Information Science and Technology*, forthcoming.

Parnas, D. L. (1972). On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, *15*(12), 1053-1058.

Poggi A. , D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, & R. Rosati (2008). Linking data to ontologies. *Journal on Data Semantics*, *10*,133–173.

REF (Research Excellence Framework) (2012). Panel Criteria and Working Methods. Retrieved January 7, 2015 from: http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf.

Sarma A. D., Dong X., & Alon Y (2008). Bootstrapping pay-as-you-go data integration systems. *Proc. of ACM SIGMOD*, 861–874.

Taylor, M. (2013). Exploring the Boundaries: How Altmetrics Can Expand Our Vision of Scholarly Communication and Social Impact. *Information Standards Quarterly*, *25*, 27-32.

Van Vught, F. (ed.) (2009). Mapping the higher education landscape: Towards a European classification of higher education. Dordrecht: Kluwer.

Van Vught, F., & Westerheijden, D.F. (2010). Multidimensional ranking: a new transparency tool for higher education and research, *Higher Education Management and Policy 22*/3, 1-26.

The Research Purpose, Methods and Results of the "Annual Report for International Citations of China's Academic Journals"

Junhong Wu, Hong Xiao[1,*], Shuhong Sheng[2,*], Yan Zhang, Xiukun Sun, Yichuan Zhang

[1, 2] *xh6613@cnki.net; SSH7600@cnki.net*

[1,2]Research Center of the Evaluation of the Scientific & Technical Literature of China, China National Knowledge Infrastructure (CNKI), 100192 Beijing (China)

Abstract

Before 2012, it was hard to come to a comprehensive evaluation of academic journals in China. For this reason the international influence of journals published in China hadn't been paid enough attention, leading to a bias in the Chinese research assessment system. Since 2012, China National Knowledge Infrastructure (CNKI) invested and carried out the project of the development of the "Annual Report for International Citations of Chinese Academic Journals ". In the same year, CNKI made a comprehensive study on the international citations of more than 6000 journals in China, and found that some journals had a certain international influence. In order to make a comprehensive assessment of the international influence of those journals, CNKI has developed a comprehensive indicator, named the CI index (clout index), combining the effect of both the impact factor and the citation counts. This article describes the purpose, methods and results of part of this project, providing a fresh idea for a comprehensive evaluation of the influence of Chinese academic journals.

Conference Topic

Methods and techniques

Background

In the era of big data and we-media as shown by Bowman & Willis (2003), direct publication and free access are all around, leading to the question: "how can academic journals survive"? It is known that journals, in particular journals sharing a scientific community compete in one market, but journals will survive as long as they have a function for a specific academic community. The main problem that Chinese journals, especially academic journals, are faced with, is the competition with huge international publishing companies. It has been a common knowledge that it is hard for the domestic journals to compete with those international academic journals.

According to Thomson Reuters' SCI data, as shown in Fig. 1, Chinese scholars published 114,130 papers in international journals in 2008. This number has greatly increased to 232.000 in 2013, which is a doubling of that in 2008. While Fig. 2 shows a comparison of the papers Chinese scholars published in the journals covered by the SCI with the papers Chinese scholars published in domestic academic journals in 2013. It can be seen in Fig. 2 that 1.035 million papers have been published in 3569 domestic academic journals in 2013. Compared to the 1,035,142 domestic papers, 206,598 academic papers were published in journals covered by the SCI. This means that one sixth of the Chinese academic papers had flowed overseas. There is also a rapid increase in quantity for the papers in the field of social sciences. The number of Chinese SSCI papers had increased from 4,430 in 2008 to 9,722 in 2013, which means a doubling over five years.

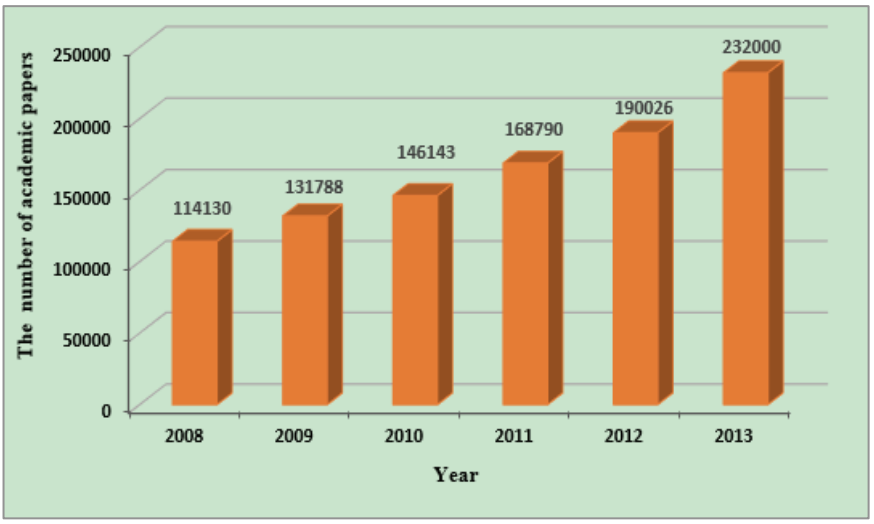


Figure 1. Evolution of the number of academic papers Chinese scholars published in international journals covered by the SCI during 2008-2013.

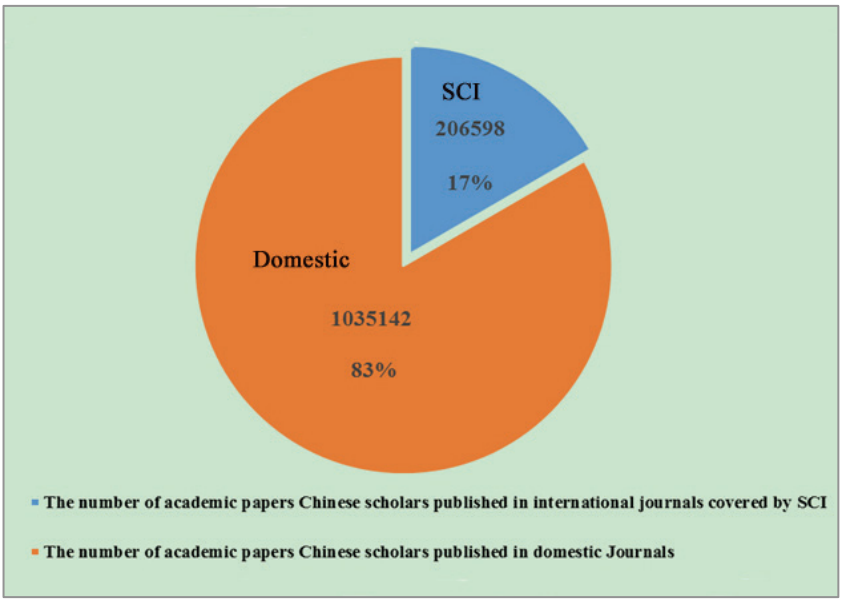


Figure 2. A comparison between Chinese papers published in journals covered by the SCI with papers published by Chinese scholars in domestic academic journals in 2013.

It is seen that so many China's qualified academic papers have flowed overseas and published in international journals, especially SCI journals. While the impact of China's academic journals on international audience was rarely revealed before. We think that, in the information era, with an increasing quantity of journals, the full importance of academic journals should be revealed through an objective evaluation based on a large amount of data. Journal management departments often use an "index set" of journal characteristics. This index set is used for the quantitative assessment of journal's quality. It has become a social consensus that "Scientific decision-making needs the support based on the big data". China's publishing management system, in particular, urgently needs a comprehensive, objective and impartial data set for the allocation of journal publishing resources.

Most scholars agree that publishing academic papers in the journals of a high academic standing is a means of academic communication and the success of a scholar in this can be used as a factor in evaluation exercises. The problem is where to draw the line between journals of high standing and journals of lower standing. In the past, data regarding

international journals were not taken into account for various evaluations of domestic journals. Therefore, it was hard for the domestic journals to compete with those international academic journals.

It is common that the reputation of Chinese scientific journals in the international community is measured by journal Impact factor (IF). Ren (1999) proposed the challenge for Chinese scientific journals using this indicator for the evaluation of Chinese journals. Based on the fact that a journal's international impact had not been adequately considered in the past, research management department such as that of CNKI could only take the SCI as evaluation standard. The Science Citation Index (SCI) initiated by Eugene Garfield is a unique retrieval and evaluation tool (Garfield, 1955). Yet it is known that it is not adequate for the local evaluations of less developed non-English speaking countries, or for the retrieval of these countries' publications (Ren & Rousseau, 2002). Since English is the most widely used language in science, journal publishers prefer to publishing in English to attract a larger reader base, resulting in more visibility, increased citations and higher IF, as shown in the study by Ren & Rousseau (2004). According to the above discussion, it is necessary for us to take an international perspective and a domestic view to evaluate the influence of China's academic journals, i.e. consider both domestic and international journals' citing citations to Chinese academic journals, in order to conduct scientific and reasonable evaluations of Chinese academic journals. For the citations by domestic journals, CNKI has developed "Annual Report for Chinese Academic Journal Impact Factors" since 2009. In this study, we focus on the citations by international journals, introducing the research purpose, methods and results of the "Annual Report for International Citations of Chinese Academic Journals".

Purpose

In this study, we conduct a quantitative assessment of academic journals published in Mainland China, either in Chinese or in English, in order to make an evaluation of their quality. Moreover, we analyze their world-wide influence by mining their cited records of citations by international journals. In the following we first give our understanding of an academic journal of high quality.

A journal of high quality provides products and services meeting or exceeding its readers' expectations. As such the quality of an academic journal is a comprehensive reflection of its publication level as manifested through the importance of its articles for the advancement of science. Following national and international norms, timeliness of publication and a large reader base also contribute to a journal's quality.

Influence of an academic journal refers to the ability of the journal to arouse its readers' attention and thinking, obtain their recognition and even alter their thoughts, opinions and behavior. A high-level journal influences academic development, by the ideas, concepts, theories, methods, findings, inventions and facts it introduces to the scientific community. Besides these objective aspects, a high-level journal has also an emotional influence, associated with its brand name, on its reader community.

Influence is not only a reflection of quality, but also a function of time. High quality papers, including editorials, show their influence gradually over time. Dissemination of journals can be judged scientifically and objectively by the frequency of being cited in domestic and foreign academic literature.

Method

Index system

Some scholars have suggested that data including downloading and online comments should be considered. These ideas are related to the altmetrics, or social influmetrics, movement

(Rousseau & Ye, 2013). Even the new Nature Index includes altmetric data (refer to the website: www.natureindex.com). However, downloading is a complex issue. It ranges from results of web crawling to students' learning, or providing intelligence services, and does not only include use for academic research. Moreover, based on current data analysis technology, it is still a challenge to judge if online comments are scientific or rigorous. In contrast, citation is a reflection of academic norms. Each author is required to respect the intellectual property rights of the literature he or she cites. Otherwise his/her behavior might be considered as misconduct. Therefore, statistical analysis of citations is considered as a relatively reliable and quantifiable technique.

Citation and publication statistics may include the following items:

(1) Statistics related to received citations such as the total cites in a year. Citations may include mutual journal citations and a history of received citations over a period of several years.

(2) Quantity of published literature such as the amount of published papers (further subdivided into types such as 'normal' articles, reviews, editorials, etc.), proportion of funded papers and proportion of articles with foreign collaboration.

(3) "Calculated indicators for evaluation: Indicators related with cited frequency such as immediacy index, the 2-year impact factor, 3-, 4-, or 5-year impact factor, etc. Indicator related with mutual citations: mutual citation index. Indicators related to the life cycle of literature: citing half-life, cited half-life, etc."

(4) The composition of the editorial board and the prestige of the editor-in-chief.

Selection of statistical sources for the international citations report

Statistical sources for the international citation report must be journals selected according to the standard for the evaluation of the international influence. Besides the journals from American and European countries, representative journals from other countries should also be included. The list of source journals should be based on suggestions from domestic and foreign experts. It is known that SCI database includes the most representative journals from the American and European countries and, as such, may be acceptable for reflecting the international influence of Chinese academic journals. Hence, at least for the current year, we still use the SCI database as the statistical source to evaluate academic journals. This means that we consider 8,621 academic journals covered in this database.

The case for humanities and social science fields is more complicated. It is not enough to merely use the 6,429 journals of SSCI and the A&HCI to evaluate the humanities and social sciences journals. For a more comprehensive statistic of the international influence of China's humanities and social sciences journals, we add well-known databases as a supplement, including those of leading international publishing groups such as Elsevier, Springer, Wiley, and Emerald. In this way, 1483 source journals (non-WOS humanities or social science journals) are included, which are good supplements for the source journals. According to experts' recommendations, we have also supplied 441 journals in minor languages, which pay attention to Chinese issues. These journals have not been included in the worldwide major databases, but they are indispensable for the research of local social science experts (Ossenblok, Engels & Sivertsen, 2012).

Data standards

In order to ensure the accuracy of the statistical data, we have established data processing standards, procedures, as well as quality requirements. Accordingly, we normalized and standardized the raw data and set up a series of databases as follows:

(1) The document database of norms for titles of more than 7,000 Chinese and English journals in China.
(2) The bibliographic database of China's academic journals, a collection of about 8,000 domestic academic journals and more than 42 million publications, used for citation links.
(3) Set up "the Statistical Standards of the Published Paper Amount". According the norm, make the statistics on the amount of published papers, as well as the cited papers published in the recent six years.

The development process

(1) Collection of data: including data retrieval in the WOS database, and the processing of data from supplementary journals.
(2) Standardized data processing: automatic processing of data such as citation links, and fuzzy title matching. If necessary these techniques were augmented by manual inspection to improve efficiency and accuracy.
(3) Detection: verification of data integrity and accuracy checks to ensure that the data meets the quality standards, plus an annual appraisal of a group of experts.
(4) Trial calculation, validation, and sample verification: indicator calculation must be double checked by several persons, and those journals with large inter-annual variations in one or more indices are the target of special attention.

Results

According to the method mentioned above, we developed the "International citation annual report" providing evaluation data of Chinese academic journals, first released in 2012. In December 2014, the 2014 Annual Report (Xiao & Du, 2014) was published and the evaluation data for more than 6000 academic journals were provided. These results were released in the "Database of Statistical Analysis of Individual Journal's Impact" available on the website of the CNKI (www.cnki.net).

Selection of highlights

Definition of a new international impact index: the clout index, denoted as CI

Several indicators like the journal impact factor and total cites are commonly used for the evaluation of a journal. Before continuing we first provide a short review of those indicators. The idea of a journal IF was first propagated by Eugene Garfield in the journal Science in 1955 (Garfield, 1955, 2006). Currently, the journal IF is generally regarded as representing the quality of academic journals in terms of citations received by its published articles. It is usually assumed that journals with a high IF carry meaningful, prominent, and quality research (Saxena, 2013). However, this single parameter is clearly not sufficient (Vanclay, 2012, Glaenzel, 2009). First of all, according to its definition, the IF reflects the performance of a journal in the most recent two years. The cited half-life for an academic journal is about 4 to 12 years, while the period of the most recent two years is merely the peak time for citation (and even that depends on the field), accounting for about 20% of the total citation amount, as shown in Fig.3. Second, the journal IF is independent of factors like the history and scale of journals and may reflect the popularity of published topics (Rousseau et al., 2013). Besides, journal IF depends on the research field: high journal IF is likely achieved for journals covering large areas of basic research with a rapidly expanding but short lived literature that use many references per article (Seglen, 1997). Using the IF as the single measure would lead to artificial constraints on the quantity of published work as well as the tendency to publish a large number of papers in line with the current popular trends without solving any

fundamental problem. Journals that act like this would lose their basic function as real academic communication platforms.

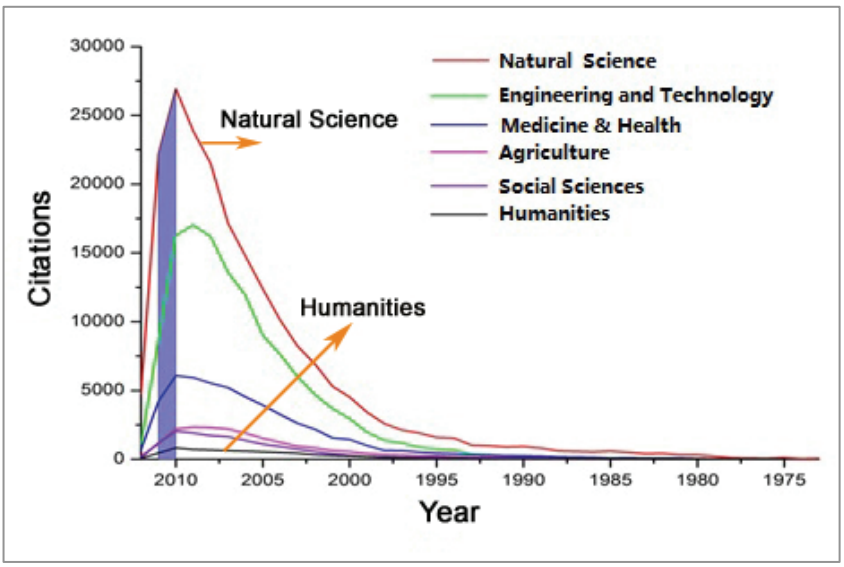


Figure 3. Cited half-life for different subject categories.

Total cites is directly related to journals' publishing history and scale. As shown in Fig. 3, we should also mention that total cites, which include the other 80% of citations that are excluded from the calculation of the IF, reflect the total impact power. However, it is also unreasonable to only consider total cites as the single evaluation factor. It might encourage researchers to aggressively increase their publishing quantity at the expense of academic quality.

Here, we should mention the topic of self citations. Whenever citations are used as indicators to evaluate scientific research, self-citations are often considered controversial. Many scholars have studied self-citation and some suggest that self-citations should be removed from citation counts, at least at micro and meso levels (e.g. analyses of persons, research groups, departments, and institutions) (Aksnes, 2003, Thijs & Glaenzel, 2005). Today the indicator of self-citations has been widely used in the evaluation of scientific journals.

We all know that IF and citations are field dependent, and therefore, indicators which compare expections to observed values are also interesting, see the work of Glaenzel, Schubert and Braun to MOCR (The Mean Observed Citation Rate) and MECR (The Mean Excepted Citation Rate) (Braun et al., 1985, Schubert et al., 1989). While Bonitz et al. (1997) stuied the Matthew effect of countries and Matthew citation journals, and presented the established characteristic of the so-called Matthew Effect for countries: field-dependency, time-stability and order of magnitude. Boyack & Klavans (2014) made the analysis on non-source publications in a different context, including non-source items in a large-scale map of science. These studies have inspired our work on exploring a comprehensive indicator for the evaluation of non-WOS-source domestic academic journals.

Thus, we have developed a comprehensive indicator, named as the Clout Index (CI), which takes both the IF and total cites into account. To be precise, we replace the WOS IF and total citations with the non-self-cited IF and total non-self-cited citations in the calculation of the CI values, taking into account that most of Chinese journals are not covered by the WOS.

First, we normalize the non-self-cited IF and total non-self-cited cites by a linear normalization method (the same for the two indices) shown in Equation (1), where V represents the parameter that has to be normalized, while N represents the normalized value. In this way the two values are in the range [0, 1].

$$N_i = \frac{V_i - V_{\min}}{V_{\max} - V_{\min}} \quad (1)$$

For the next step, we apply Eq. 2 to calculate the CI value:

$$CI = \sqrt{2} - C = \sqrt{2} - \sqrt{(1-x)^2 + (1-y)^2} \quad (2)$$

Fig. 4 shows a schematic distribution of CI values calculated by Eq.2. The points scattered in Fig. 4 represent the CI values of the selected journals. It should be noted that, in Eq. 2 and in Fig. 4, *x* and *y* stand for the normalized non-self-cited IF and total non-self-cited cites, respectively. From Fig.4, it can been seen that the origin coincides with non-self-cited IF = 0 and total non-self-cited cites = 0, while the point (1, 1) represents that the journal has reached the maximum value in both the non-self-cited IF and total non-self-cited cites. If we take the point (1, 1) as the center and CI value as radius to draw circles, then points on the same circle have the same CI value. The points located in the bottom-left area have lower CI values while the ones in the up-right area have higher values.

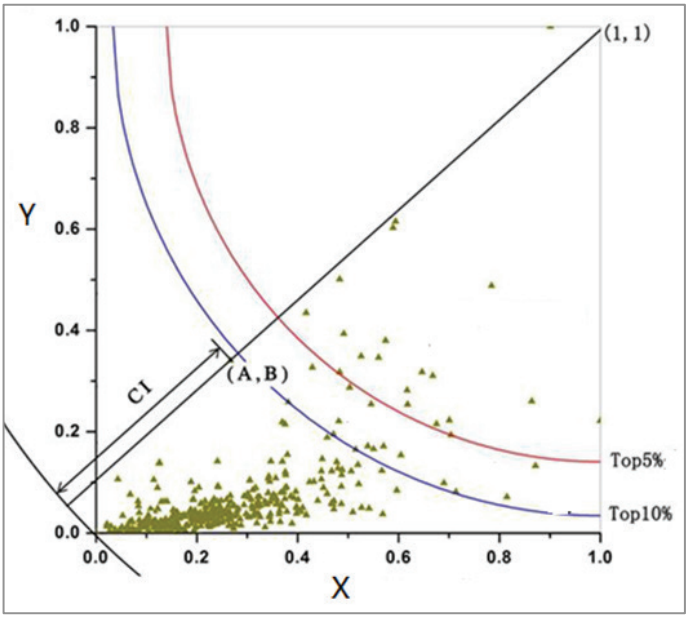


Figure 4. Schematic view of the Clout Index (CI) and selection of top journals in China.

In Figure 4, we drew two curves with the point (1,1) as their center, and CI(1) and CI(2) as radius, respectively. Here, CI(1) and CI(2), represent the critical CI values of selected Top 5% and Top 10% journals, respectively. We consider journals with CI values above the Top 5% as "The Highest International Impact Academic Journals of China", while CI values between CI(1) and CI(2) as "The Excellent International Impact Academic Journals of China". Here, China's academic journals published either in English or in Chinese are both considered.

At this point, we would like to explain why we choose a vector sum method for the calculation of the CI value instead of using a simple linear sum. Figure 5 shows a comparison of the linear sum and the vector sum method. The scatterplot itself in Figure 5 is the same as in Figure 4. First, we consider that IF and cites have the same weight as evaluation indicators. If we take the linear sum method, i.e. CI=*x*+*y*, we obtain oblique straight lines with different lengths to show equal CI values. Here, *x* and *y* have the same definition as those for Figure 5. Compared to the result obtained from the linear sum method, the CI value is smaller by the vector sum method, when the points closed to x or y axes. In this system, journals with higher single index, either IF or cites are easily excluded. Thus, our algorithm gives a better way to match the evaluation principle: "not only quality but also quantity matters".

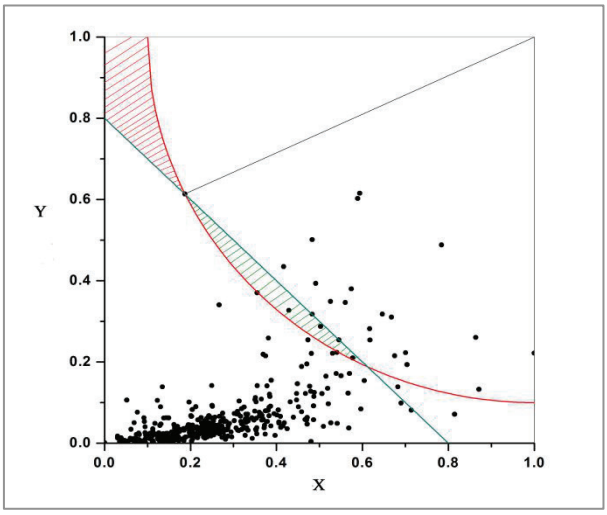


Figure 5. A comparison of the linear sum and the vector sum method.

In our system, we consider that journals with both higher IF and higher cites are journals of high-quality. Those journals commonly have higher influences in their scientific field. Fig. 6 shows the schematic view of the scatterplot of the CI value for both the SCI journals as well as Chinese domestic journals. We use double logarithmic coord. system in Fig. 6, where our developed vector sum method is applied to calculated the CI values for SCI journals and China's domestic journals. Both of the statistical sources are WOS database. The green triangles represent the CI values for the international SCI journals, while the black ones show the domestic journals covered by SCI. Those related to the Top 5% journals are shown in red, while for Top 5-10% journals in dark blue. Lower CI values for other domestic journal are shown in orange. It is clear that the majority of international journals covered by the SCI is situated in the area with both higher IF and higher citations. Situations are similar for the Top 5% and Top 5-10% domestic journals. Thus justifies that our vector sum method is a good method for the evaluation of the journals, and the CI index can be considered as an effective and reasonable indicator for the quantitative assessment of journal impact.

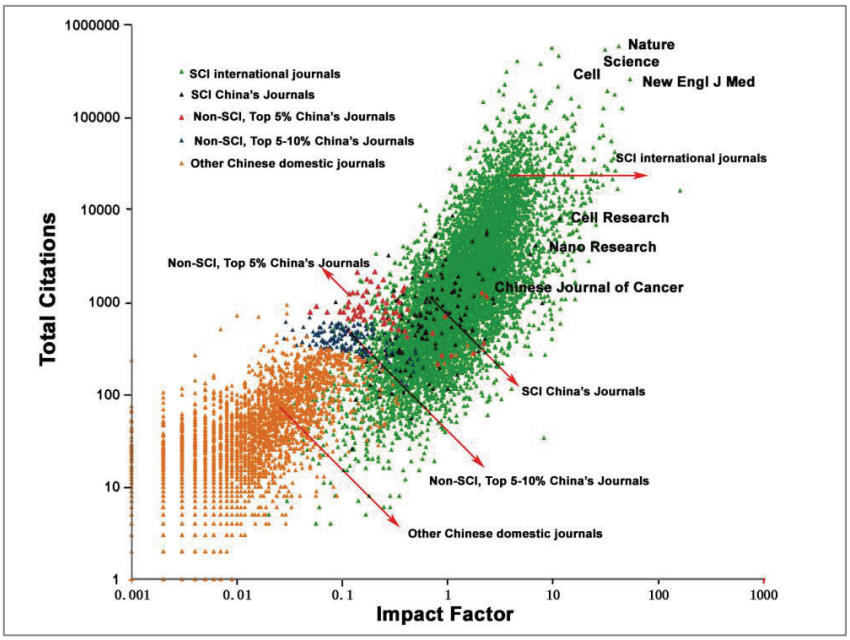


Figure 6. A schematic view of the scatterplot of the CI values for both the SCI journals and China's domestic journals.

Problems in the selection of top journals

Division of subjects

There is, among domestic journals in various disciplines, a large imbalance in the level of international influence and the extent of "going global". If top journals were selected per discipline, excellent journals in highly visible disciplines may be excluded, while journals with less international influence in disadvantaged disciplines may come into the list. This is not the way we want to highlight the most influential journals. As time goes by, we expect that the international influence of various disciplines will be improved and developed in a balanced way. At that time we will be able to perform sub-discipline rankings.

Comprehensive consideration of domestic and international influence

In the evaluation of international influence we should consider domestic and international influences as two sides of the same coin. However, first, a separate evaluation is more helpful for understanding the situation of the journals in both domestic and international market. Secondly, there is no recognized method showing how to merge these two reports, however, see Jin & Rousseau's study for an earlier merging attempts (Jin et al., 1999, Jin & Rousseau, 2004). The main difficulty lies in the point that there are different opinions on the issue of whether "a domestic citation is equal to an international citation." Considering the fact that "the annual report of the impact factors of China's academic journals" has been well developed for years, in this article, we mainly discuss the research method of the annual report of the international citations.

The selection process and resulting top list

To highlight the most influential journals, this year we continue selecting "The Highest International Impact Academic Journals of China" and "The Excellent International Impact Academic Journals of China". By ranking the journals of STM (Science/ Technology/ Medicine) and AH&SS (Humanities and Social Sciences) according to the CI values, we selected the Top 5% and the Top 5-10% journals; then sent the selection method, data of indicators and the primary list to more than 70 experts for peer review.

Some journals were removed from the list for their bad reputation evaluated by peer reviewers, while other ones were supplemented in sequence. This was done in such a way as to make sure that the total number of selected journals stayed the same. Finally we determined 176 STM journals and 61 AH&SS journals as "The Highest International Impact Academic Journals of China", and 174 STM journals and 60 AH&SS journals as "The Excellent International Impact Academic Journals of China" Among these 471 Top 5-10% journals, 458 are core journals selected by various domestic institutions, and most of the other 13 are journals in English or newly created ones.

Summary and outlook

(1) An international report has been issued for three successive years and approved by government departments in charge of journals, editorial department of journals and academic circles. This encourages us to keep this work going.

(2) With the accumulation of data, many meaningful conclusions can be drawn from the analysis of inter-annual variations in data.

(3) There is certainly room to improve the evaluation methods, including the selection criteria of the international source journals, possible improvement of the determination of the CI indicator and the integration of domestic and international lists.

(4) The main points of this article have been written in a manuscript submitted to Acta Editologica in Chinese language (Wu et al., 2014).

Acknowledgments

This study is sponsored by the National Social Science Project of "Evaluation and development strategy of the international influence of Chinese academic journals in English" (No.14BTQ055). The authors would like to express the deepest appreciation to Prof. Ronald Rousseau for his very careful corrections and valuable comments on this manuscript. We are very grateful to Prof. Jerold Mathews in Iowa State University for his careful corrections and good suggestions. And we thank Dr. Bo Liu in CNKI for his valuable discussion on the manuscript.

References

Aksnes, D.W. (2003). A macro study of self-citation. *Scientometrics*, *56*, 235–246.

Bonitz, M., Bruckner, E. & Scharnhorst, A. (1997). Characteristics and impact of the Matthew effect for countries, *Scientometrics 40*(3), 407-422.

Bowman, S. & Willis, C. (2003). We media-hypergene. Retrieved July 2003 from: http://www.hypergene.net/wemedia/download/we_media.pdf.

Boyack, K. W. & Klavans, R. (2014). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics 8*(3), 569-580.

Braun, T., Glaenzel, W. & Schubert, A. (1985). *Scientometric indicators: A 32 country comparison of publication productivity and citation impact.* (pp. 424). Singapore: World Scientific Publishing Co.

Garfield, E. (1955). Citation indexes for science: a new dimension in documentation through association of ideas. *Science*, *122*, 108-111.

Garfield, E. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*, 90-93.

Glaenzel, W. (2009). The multi-dimensionality of journal impact. *Scientometrics*, *78*(2), 355-374.

Jin, B., Wang, S., Wang, B. et al. (1999). A unified method of counting international and domestic articles. J*ournal of Management Sciences in China*, *2*(3), 59–65.

Jin, B.H. & Rousseau, R. (2004). Evaluation of Research Performance and Scientometric Indicators in China. In H. F. Moed, W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 497-514). Dordrecht, etc.: Kluwer Academic Publishers.

Ossenblok, Truyken L.B., Engels, Tim C. E. & Sivertsen, G. (2012). The representation of the social science and humanities in the Web of Science: a comparison of publication patterns and incentive structures in Flanders and Norway (2005-9). *Research Evaluation*, 21(4), 280-290.

Ren, S.L., Liang, P. & Zu, G. (1999). The challenge for Chinese scientific journals. *Science*, *286*, 1683.

Ren, S.L. & Rousseau, R. (2002). International Visibility of Chinese scientific journals. *Scientometrics*, *53*, 389-405.

Ren, S.L. & Rousseau, R. (2004). The role of China's English-language scientific journals in scientific communication. *Learned Publishing*, *17*, 99-104.

Rousseau, R., Garcia-Zorita, C. & Sanz-Casado, E. (2013). The h-bubble. *Journal of Informetrics*, *7*(2), 294-300.

Rousseau, R. & Ye, Fred Y. (2013). A multi-metric approach for research evaluation. *Chinese Science Bulletin*, *58*, 3288-3290.

Saxena, A., Thawani, V., Chakrabarty, M. & Gharpure, K. (2013). Scientific evaluation of the scholarly publications. *Journal of Pharmacology and Pharmacotherapeutics*, *4*(2), 125-129.

Schubert, A., Glaenzel, W. & Braun, T. (1989). World flash on basic research: scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields, 1981–1985. *Scientometrics*, *16*(1–6), 3–478.

Seglen, Per O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314, 497-502.

Thijs, B. & Glaenzel, W. (2005). The influence of author self-citations on bibliometric meso-indicators. The case of European universities. *Scientometrics*, *66*, 71–80.

Vanclay, J.K. (2012). Impact factor: outdated artefact or stepping-stone to journal certification? *Scientometrics*, *92*(2), 211-238.

Wu, J.H., Xiao, H., Zhang, Y., et al. (2014). A comprehensive index algorithm of the international influence evaluation of Chinese academic journals, submitted to *Acta Editologica*.

Xiao, H. & Du, W. T. (Eds.) (2014). Annual Report for International Citation of Chinese Academic Journals. Beijing: "China Academic Journals (CD-ROM Version)" Electronic Publishing House.

Is the Year of First Publication a Good Proxy of Scholars' Academic Age?

Rodrigo Costas[1], Tina Nane[2] and Vincent Larivière[3]

[1] *rcostas@cwts.leidenuniv.nl;* [2] *g.f.nane@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX, Leiden (the Netherlands)

[3] *vincent.lariviere@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, C.P. 6128, Station Centre-Ville Montreal, Quebec (Canada)

Abstract
Individual scholars are the central unit of the research system and are increasingly the focus of bibliometric studies. An important aspect in the study of individual scholars is their academic age, which allows for the comparison of scholars that have been academically active in a similar period of time. Based on a sample of Quebec researchers for whom their year of birth, PhD year as well as the year of their first publication are known, we study the relationships among these ages with the aim of determining how their year of first publication can be used to estimate their 'real' age. Moderate correlations have been found among the ages, and the first publication year has a higher correlation with the PhD year than with the birth year. However, an important dispersion of scholars across the different ages is observed; thus, the year of first publication can only be taken as proxy of the real age of scholars. Alternatively, the consideration of cohorts of around 5 years seems to be a reasonable approach. Further research will focus on the exploration of other bibliometric indicators in order to refine the preliminary developments discussed here.

Conference Topic
Methods and techniques

Introduction
In individual-level bibliometric studies, the socio-demographic characteristics of scholars are of central importance to understand and better frame the results obtained (Costas & Bordons, 2011; Gingras, Larivière, Macaluso, & Robitaille, 2008; Mauleón & Bordons, 2006). Among these socio-demographic characteristics we can mention gender (Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013; Mauleón & Bordons, 2006), mobility (Canibano, Otamendy, & Solis, 2011; Franzoni, Scellato, & Stephan, 2012), and nationality (Moed & Halevi, 2014), among others. The development of large-scale author-name disambiguation algorithms (Caron & Van Eck, 2014) as well as the increasing quantity of papers' metadata indexed (e.g. author names and surnames, affiliations, e-mail data, etc.) have allowed the study of the socio-demographic characteristics of scholars at a larger scale. For example, the analysis of the first author names of authors (Larivière et al., 2013) allowed the macro analysis of gender disparities worldwide. The large-scale analysis of the relationship between author names, affiliations and countries collected from scientific publications has open the possibility of studying academic mobility at the world level (Moed, Aisati, & Plume, 2013), as well as the nationality (Costas & Noyons, 2013), migrations (Moed & Halevi, 2014) or even the ethnic origin (Freeman, 2014) of scholars.

A critical element for individual-level bibliometrics is the age of the researchers (Costas & Bordons, 2011; Larivière, Archambault, & Gingras, 2008; Levin & Stephan, 1989), especially from the point of view of its relationship with productivity (Falagas, Ierodiakonou, & Alexiou, 2008; Levin & Stephan, 1989). Age is also a common point of debate in science policy, as it aims to compare scholars of the same 'academic age' (Bornmann & Leydesdorff,

2014). However, one of the main reasons that hinders the development of bibliometric studies at the individual level is the lack of systematic data on the age of scholars, as this information is not systematically collected in bibliographic databases. A commonly used proxy for the study of the age of scholars has been the so-called 'scientific (or academic) age', often defined as the publication year of the first paper of a scholar (Radicchi & Castellano, 2013).[1] The use of this age is very convenient, as it is possible to directly extract it from bibliometric data. However, so far there has not been any analysis on the relationship between this proxy and the real age of scholars. This paper is intended to fill this gap and shed some light on the relationship between the 'bibliometric' age of scholars that can be calculated based on bibliographic information and the 'real' age(s) of individual scholars, namely their birth age and their PhD age. In other words, we aim to infer the birth year and PhD year of scholars based on models that are exclusively based on bibliometric indicators[2] (e.g. first publication year, position of signature, co-authors, etc.). Thus, the main research question can be operationalized as follows: *could the year of first publication (YFP) of a scholar (as recorded in the Web of Science) be considered as a relevant proxy of the birth and/or PhD ages of scholars?*

Methodology

In order to answer the research questions it is necessary to have a dataset of scholars for whom their real ages are certainly known as well as the publication years of their scientific publications. Thus, as our golden set, in this study we have considered one of the (possibly) largest datasets of individual scholars for whom real individual characteristics are known (this dataset has been used in some other studies, e.g. Gingras et al., 2008; Larivière et al., 2011). This dataset is composed by 13,626 university professors from Quebec who have published at least one article during the 1980-2012 period. For every scholar in the dataset, the following individual elements have been codified:

- Year of birth [*Birth year*]
- Year of PhD (year when the scholar has obtained her (first) PhD) [*PhD year*].
- Publication year of their first publication in the Web of Science (WoS) [*YFP*]
- [*Birth year to YFP*], which is calculated as [*YFP*]-[*Birth year*]
- [*PhD year to YFP*] which is calculated as [*YFP*]-[*PhD year*]
- Domain (nine disciplinary fields of activity of the scholar, which is based on the 2000 revision of the U.S. Classification of Instructional Programs (CIP)[3] developed by the U.S. Department of Education's National Center for Education Statistics (NCES).

Complementary, we have also calculated the total number of publications of the scholars in the period 1980-2012 [*P*].

A technical limitation of the dataset is that the WoS publication data starts in 1980, thus meaning that for very old individuals it is not possible to know with certainty if the first publication recorded in the WoS during the period 1980-2012 truly corresponds with their actual first publication. To reduce the effect of this issue, we decided to focus only on those individuals that have a birth year later than 1959 (i.e. we don't expect that many scholars would have a publication before their 20's) and a PhD year also later than 1980 (same criteria

[1] Although this term has also been proposed for the time since the PhD has been awarded (Bar-Ilan, 2014). Some other studies have also focused on the starting year of publication of individuals as proxies of age (Fronczak, Fronczak & Holyst, 2006).

[2] Due to space restrictions, in this paper we focus only on the first publication year as a proxy, and leave for a further version of this paper the consideration of other bibliometric variables.

[3] The Classification of Instructional Programs (CIP) is developed by the U.S. Department of Education's National Center for Education Statistics (NCES). More details can be found at: http://nces.ed.gov/pubs2002/cip2000/

as before). As a result of this filtering we ended up with 3,596 scholars that are the final dataset of our analysis.

Main results

This section presents the main results of the analysis. In Appendix 1 the descriptive scores are presented. Results show that there are differences in individual productivity by domain, which is of course not a surprise. For instance scholars from the Basic Medical Sciences and Health sciences exhibit the highest number of WoS papers, while Humanities the lowest. Similarly, the median birth year of the whole sample is 1965, although there are small differences by domain, with Basic Medical Sciences with the oldest individuals (median=1964) and Social Sciences the youngest (median=1967). The median PhD year of the whole sample is 1998, with the Basic Medical Sciences as the oldest median (1994) and domains such as Business & Management, Education, Non-health professionals getting their PhD on median in 1998.

Regarding the time between the birth of the scholars and the time of their first publication, scholars from Basic Medical Sciences, Engineering, Health Sciences and Science are on median the fastest (32 years) while scholars from Business & Management, Education or Humanities are slower (35 years). From the PhD to the first publication, the fastest are the scholars in Health Sciences (1 year) and the slowest the Humanities (4 years). It is important to keep in mind that here we also have cases with negative values, which means that researchers publish publications before their PhD date; a finding coherent with Larivière (2012).

Relationship between the different ages

In Appendix 2 we present the main correlations between the different ages of the scholars. In Figure 1 a summary of the correlations is presented. In general, there is a reasonably good correlation between birth year and PhD year, and the two real ages of the scholars have moderate correlations with YFP, although the PhD year has a generally better correlation with YFP than the Birth Year. These results suggest that it is reasonable to consider the YFP as a proxy of the scientific age of the researchers.

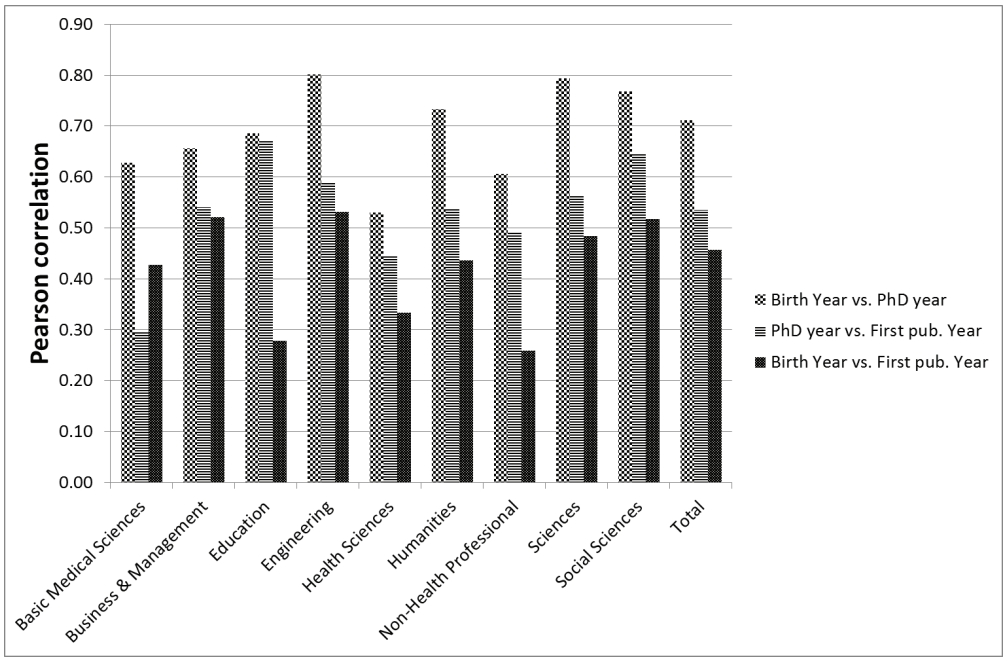


Figure 1. Pearson correlation values of the different ages – by disciplines and all disciplines combined.

YFP as a proxy of the age of researchers

Considering the moderate correlations between the YFP and the real ages of the researchers, we explore the dispersion of the scholars by the different ages. In Figure 2 box plots of each of the three variables (YFP, Birth year and PhD year) grouped by the combination of the same variables are presented. Thus it is possible to understand how scholars distribute across the different ages.

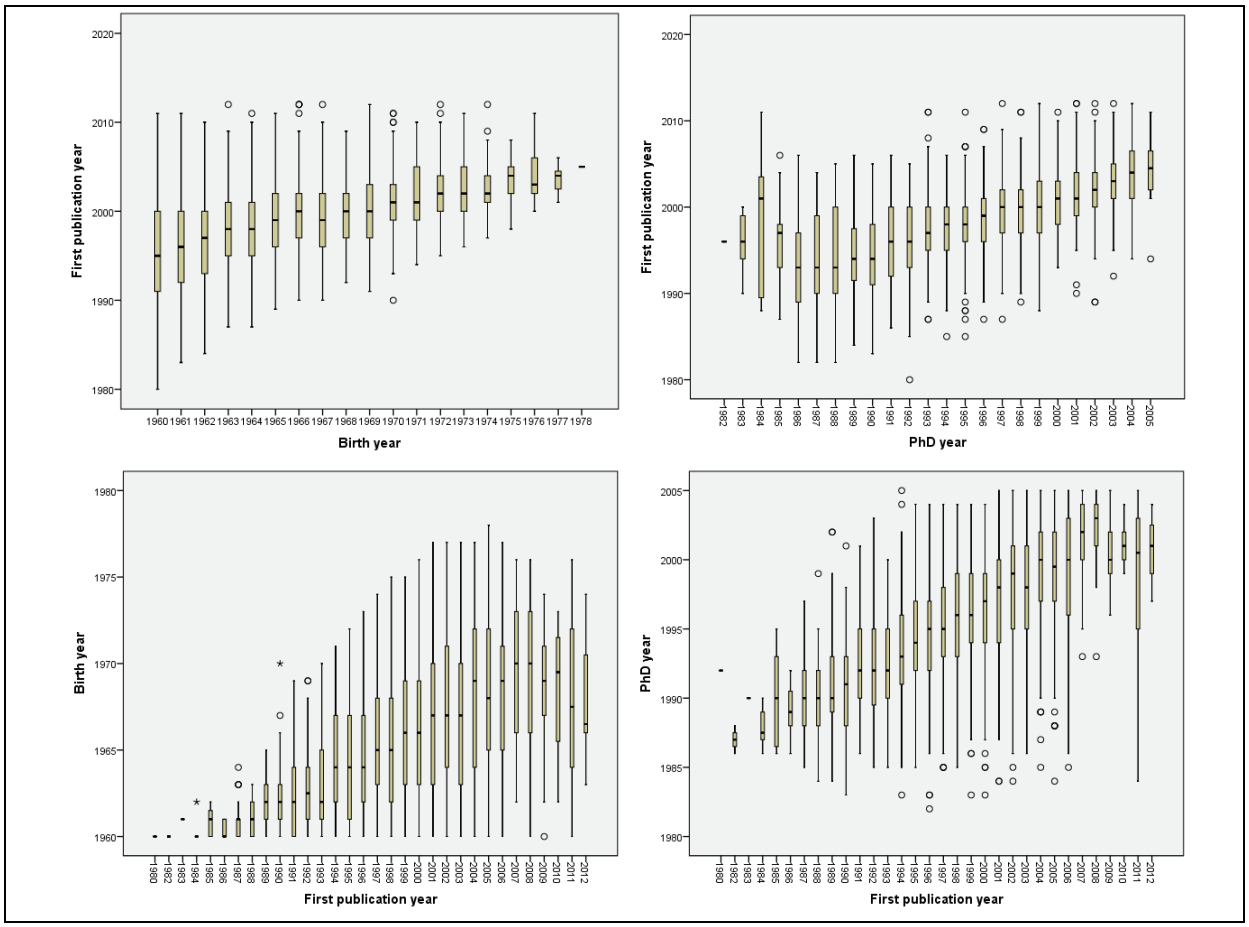


Figure 2. Box plot distribution of scholars across the different ages (all scholars together).

The two graphs on top of Figure 2 present boxplots of YFP observations grouped by each distinct birth year and PhD year. In the case of the birth year, it is possible to see how the earlier the year of birth the larger the variation of the YFP, thus indicating how researchers of all ages start their publication activities at different points in their lives, although the majority (i.e. the 'box' in the graph) tends to concentrate in a range of 5 to 10 years. The YFP median also tends to increase as the birth year increases. In the case of the PhD year we see also a quite disperse distribution of the first publication year of the scholars, although (with the exception of some irregularities among the scholars with the earliest PhD years) we notice a stepper increase in the median value of the YFP as the PhD year increases.

The graphs on the bottom of Figure 2 show the distribution of the two real ages (birth and PhD years) as a function of the YFP. Here we can also see an important dispersion of scholars across the two ages. However, in order to summarize the results of these two graphs, in Figure 3 we present the interquartile ranges (i.e. range of the number of years that include the 50% of all the observations), thus allowing to identify where most of the scholars are located in the distribution as a function of their first publication year.

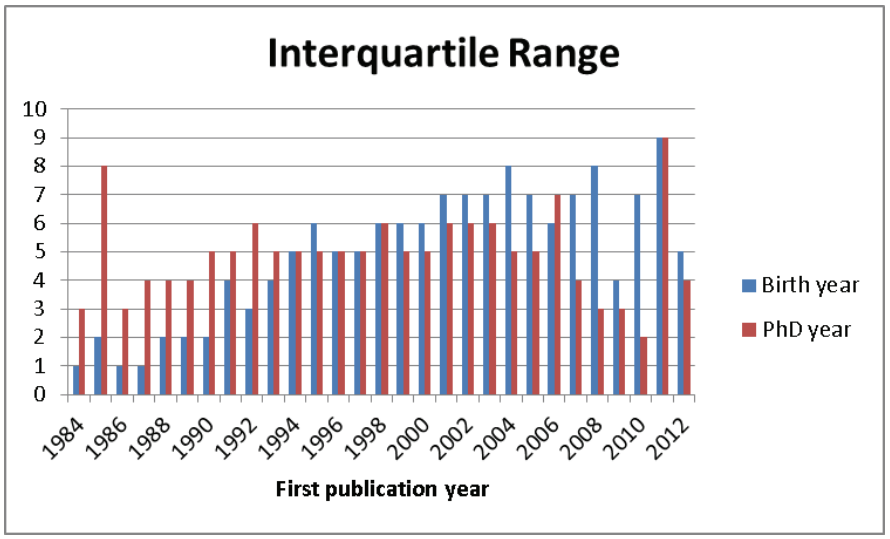


Figure 3. Interquartile range (in number of years) for Birth year and PhD year as a function of YFP.

Figure 3 shows that the interquartile range in all cases is smaller than 10 years for any of the two ages considered. Actually the average for all the YFP years considered is 4.9 years for both ages (with a median of 5). Thus, a possible interpretation of this result is that if we would only count with the YFP of the scholars, with a range of around 5-10 years we would be able to capture the real age of about 50% of all the scholars who started to publish that year.

Exploring a predictive model for the age of scholars based on bibliometric indicators

In this section a more predictive approach is presented. We are interested in estimating the birth and PhD years of a generic researcher by using the YFP indicator in our data sample. Numerous approaches can be taken, from the selection of different models and independent variables that could influence the two ages. In the present study we choose the simple linear regression model, with the average birth year and the average PhD year as dependent variables and the YFP as the independent variable. We will therefore infer on the average birth and PhD year of a scholar, and Figures 4 and 5 provide the linear regression fit of the two models, along with confidence and prediction intervals.

Using linear regression analysis the average ages (birth year and PhD year) of the whole list of scholars are fitted, including a 95% confidence interval as well as a 95% prediction interval. Although both intervals account for the uncertainty of the regression parameter estimates, there is an important distinction between the two intervals. The confidence interval is supposed to cover the true average birth year (of all the scholars in the statistical population) with high probability in 95% of the cases. The prediction interval provides limits on a future sampled observation that is an average of a given number of scholars from the set of all the scholars in the world. The prediction intervals refer then to actual observations in the data, and hence account also for the variation in the data, whereas the confidence intervals refer to the population's (of all scholars) average birth year. The prediction intervals are always larger than the confidence intervals.

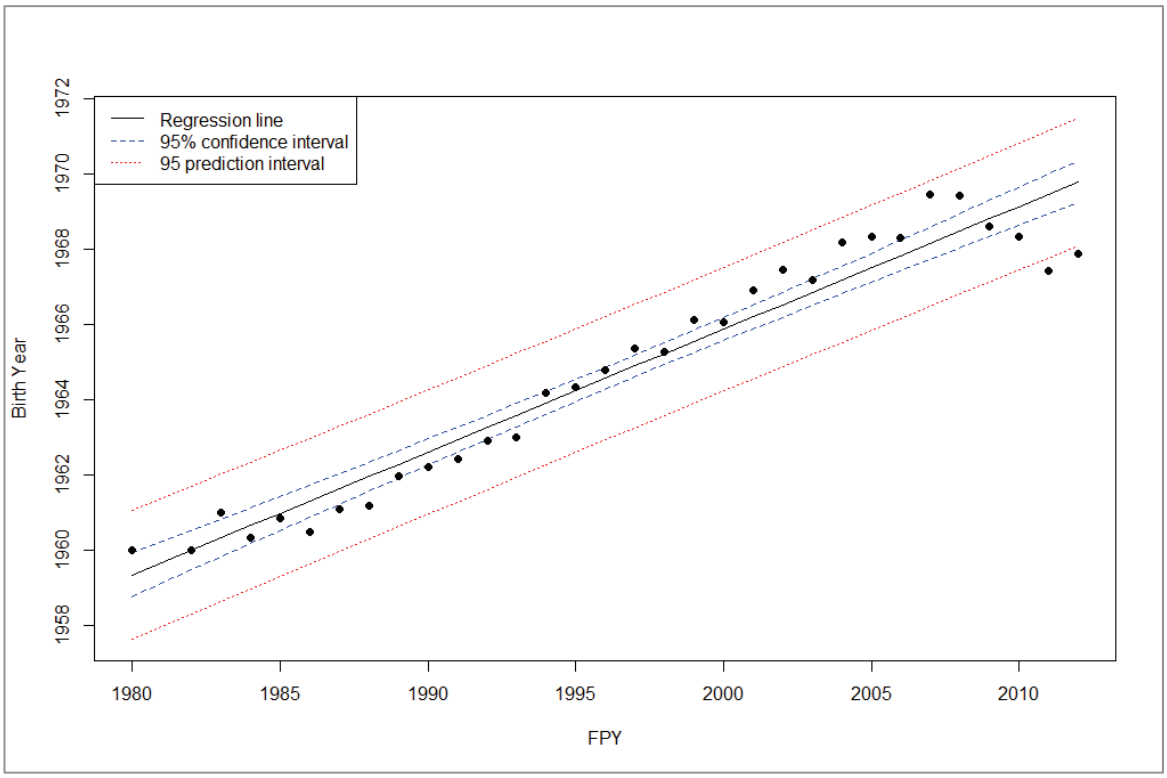


Figure 4. Average birth year by YFP, fitting a regression line and 95% confidence and prediction intervals.

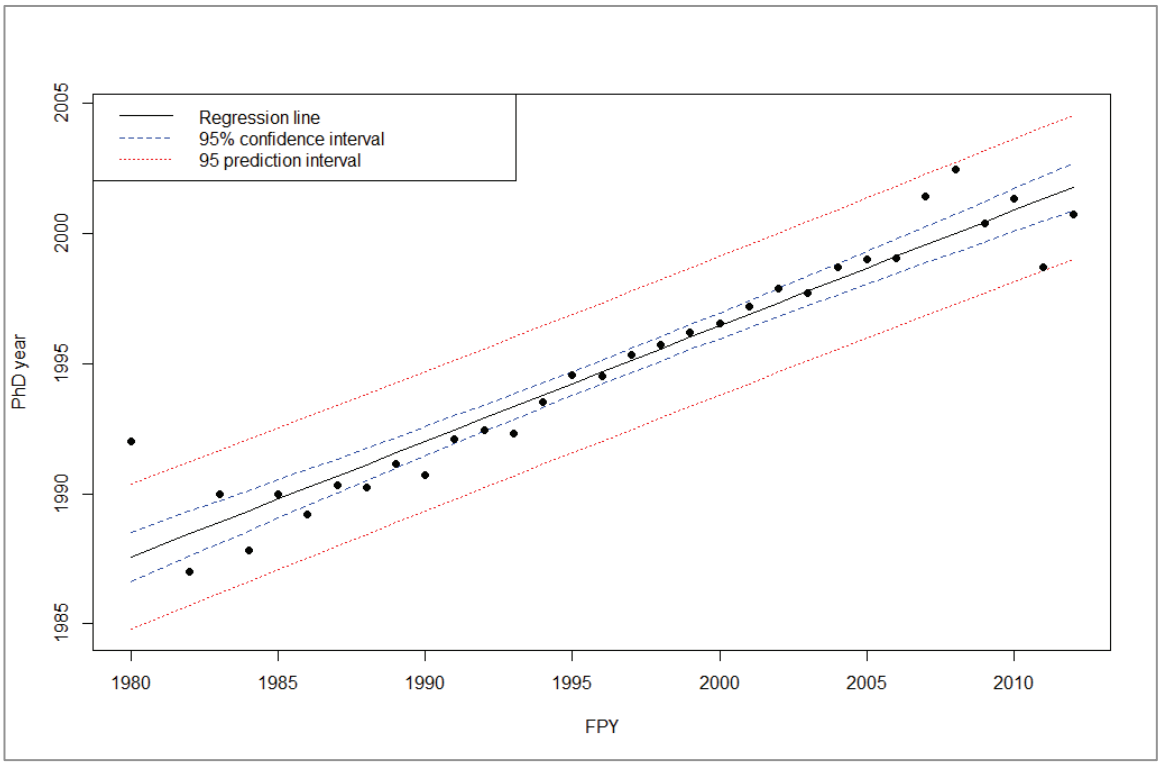


Figure 5. Average PhD year by YFP, fitting a regression line and 95% confidence and prediction intervals.

The main difference with bottom graphs in Figure 2 is that here the target is to estimate the average age of scholars from a given YFP. For example, in Figure 4 we can see how for scholars with a YFP=1995 their average birth year would be 1963, and the prediction interval ranges between 1961 and 1965. A similar pattern is observed in Figure 5 for PhD year (i.e.

with YFP=1995 the average PhD year would range around a period of five years). This suggests that we would be able to estimate the average ages of the scholars with a given YFP within an interval of 5 years. Of course, it is important to keep in mind that this analysis is based on the average values for all scholars, which is different from the individual prediction of individual scholars; however the relatively short prediction intervals (around 5 years) supports the importance of the YFP as relevant proxy for the ages of individual scholars.

Discussion and conclusions

Age is one of the most important socio-demographic determinants of researchers' activities, funding, output and impact. However, the lack of systematically recorded information on the age (real or academic) of researchers makes the need of developing reliable and valid proxies a priority. So far, the age of the first publication of individual scholars has been frequently considered as a proxy of the real age of scholars; however its validity has never been tested. Based on a sample of Quebec researchers for whom their actual birth year, PhD year as well as the year of their first publication are known, a study on the relationships among these ages has been performed.

The three ages correlate moderately well, birth year and PhD year have a good relationship, and YFP has moderate correlations with the other two ages, particularly with the PhD year. It is also possible to detect an important dispersion of scholars across the different ages, indicating that new authors (and new researchers) basically can come from a wide range of years. This means that, in spite of the moderate correlation between the YFP and the other ages, the YFP can only be considered as a proxy for researchers' age, as it does mix researchers with different birth and PhD years. The consideration of cohorts of years seems to be a more reasonable alternative. Thus, it is possible to argue that considering authors who started to publish in a given year, the majority of these scholars would have ages (birth and PhD) within a range of 5 to 10 years.

It is important also to highlight some of the limitations of this study. In the first place, we are working with a dataset of scholars from only one location (Quebec in Canada), so we need to keep in mind the limitations of the representativeness of our sample for the whole world. Thus, issues related with the changes and internal evolution of PhD programs could partly influence the results and hinder their generalization. Secondly, WoS is the only database considered for the determination of the YFP, however scholars can publish outputs not covered by this database, which is likely the case in Quebec, whose local literature in the social sciences and humanities is highly relevant (Larivière & Macaluso, 2011). Thirdly, in this study we haven't explored differences across fields, but arguably there are differences in the relationship between the ages and the first publication year of the scholars as disciplinary differences in individual productivity have been also discussed (Ruiz-Castillo & Costas, 2014).

All in all, considering the limitations previously exposed, our results are still policy-relevant and support the idea that the first publication year(s) of individual scholars can work as a reasonable proxy as their age, particularly when considering cohorts of researchers. For the final version of the paper other approaches will be also considered, including the analysis of the positions of the scholars in the papers (as these positions are related with the age of scholars (Costas & Bordons, 2011), other bibliometric indicators (e.g. the total number of publications of a scholar and total number of citations, which are age dependent) as well as the different disciplines of scholars. Finally, the consideration of other datasets from other countries and/or disciplines is an important development in order to globally validate the different tests and models obtained and to establish a more generalizable approach for the estimation of ages based on bibliometric data. A potential recommendation derived from this study is the relevance of incorporating information about the age, PhD year, gender and other

demographic characteristics in modern Research Information Systems (RIS). This would allow for more accurate studies of the demographics and changes in the trends of scientific productivity of individual scholars.

References

Bar-Ilan, J. (2014). Evaluating the individual researcher - adding an altmetric perspective. *Research Trends*, *37*, 31–34.

Bornmann, L. & Leydesdorff, L. (2014). On the meaningful and non-meaningful use of reference sets in bibliometrics. *Journal of Informetrics*, *8*(1), 273–275. doi:10.1016/j.joi.2013.12.006

Canibano, C., Otamendy, F. J. & Solis, F. (2011). International temporary mobility of researchers: a cross-discipline study. *Scientometrics*, *89*(2), 653–675.

Caron, E. & Van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In E. Noyons (Ed.), *19th International Conference on Science and Technology Indicators. "Context counts: pathways to master big data and little data."* Leiden: CWTS-Leiden University.

Costas, R. & Bordons, M. (2011). Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective. *Scientometrics*, *88*(1), 145–161. Retrieved from http://www.springerlink.com/index/10.1007/s11192-011-0368-z

Costas, R. & Noyons, E. (2013). Detection of different types of " talented " researchers in the Life Sciences through bibliometric indicators : methodological outline Sciences through bibliometric indicators : methodological outline 1. *CWTS Working Paper Series*, (CWTS-WP-2013-006). Retrieved from http://www.cwts.nl/pdf/CWTS-WP-2013-006.pdf

Falagas, M. E., Ierodiakonou, V. & Alexiou, V. G. (2008). At what age do biomedical scientists do their best work? *The FASEB Journal*, *22*(12), 4067–4070.

Franzoni, C., Scellato, G. & Stephan, P. (2012). Patterns of international mobility of researchers : evidence from the GlobSci survey. In *International Schumpeter Society Conference* (pp. 1–32). Retrieved from http://www.aomevents.com/media/files/ISS 2012/ISS SESSION 7/Scellato.pdf

Freeman, R. B. (2014). Strength in diversity. *Nature*, *513*, 305.

Fronczak, P., Fronczak, A. & Holyst, J. A. (2006). Publish or perish: analysis of scientific productivity using maximum entropy principle and fluctuation-dissipation theorem. *arXiv*.

Gingras, Y., Larivière, V., Macaluso, B. B., Robitaille, J.-P. & Lariviere, V. (2008). The Effects of aging on researchers' publication and citation patterns. *Plos ONE*, *3*(12), e4048. doi:10.1371/journal.pone.0004048

Lariviere, V., Archambault, E. & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: from exponential growth to steady-state science (1900-2004). *Journal of the American Society for Information Science and Technology*, *59*(2), 288–296. doi:10.1002/asi

Larivière, V., Gingras, Y., Cronin, B. & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*, 4–6.

Levin, S. G. & Stephan, P. E. (1989). Age and research productivity of academic scientists. *Research in Higher Education*, *30*(5), 531–549.

Mauleón, E. & Bordons, M. (2006). Productivity , impact and publication habits by gender. *Scientometrics*, *66*(1), 199–218.

Moed, H. F., Aisati, M. M. & Plume, A. (2013). Studying scientific migration in Scopus. *Scientometrics*, *94*, 929–942. doi:10.1007/s11192-012-0783-9

Moed, H. F. & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, 1987–2001. doi:10.1007/s11192-014-1307-6

Radicchi, F. & Castellano, C. (2013). Analysis of bibliometric indicators for individual scholars in a large data set. *Scientometrics*, *97*(3), 627–637. doi:10.1007/s11192-013-1027-3

Ruiz-Castillo, J. & Costas, R. (2014). The skewness of scientific productivity. *Journal of Informetrics*, *8*(4), 917–934. doi:10.1016/j.joi.2014.09.006

Appendix 1. Main descriptive values

Disciplinary division		Birth year	PhD year	YFP	P	Birth year to YFP	PhD year to YFP
Basic Medical Sciences	N	713	713	713	713	713	713
	Mean	1993.66	1964.72	1997.01	52.54	32.29	3.34
	Std. Deviation	4.503	3.427	4.835	67.27	4.58	5.54
	Median	1994.00	1964.00	1997.00	30.00	32	3
	Minimum	1983	1960	1980	1	20	-13
	Maximum	2005	1976	2008	788	46	21
Business & Management	N	243	243	243	243	243	243
	Mean	1997.50	1965.92	2000.56	10.92	34.64	3.05
	Std. Deviation	4.427	4.313	4.476	12.405	4.31	4.269
	Median	1998.00	1965.00	2001.00	7.00	35	3
	Minimum	1983	1960	1986	1	25	-10
	Maximum	2005	1976	2012	96	49	27
Education	N	47	47	47	47	47	47
	Mean	1997.38	1965.49	2001.04	8.43	35.55	3.66
	Std. Deviation	3.943	4.117	5.254	13.333	5.71	3.93
	Median	1998.00	1965.00	2001.00	4.00	35	3
	Minimum	1989	1960	1986	1	25	-5
	Maximum	2003	1974	2010	70	48	12
Engineering	N	514	514	514	514	514	514
	Mean	1996.38	1966.27	1998.67	38.08	32.40	2.30
	Std. Deviation	4.713	4.488	4.509	48.889	4.36	4.19
	Median	1996.00	1966.00	2000.00	24.50	32	2
	Minimum	1982	1960	1985	1	22	-11
	Maximum	2005	1977	2009	692	44	17
Health Sciences	N	292	292	292	292	292	292
	Mean	1996.89	1965.45	1998.10	49.80	32.65	1.20
	Std. Deviation	4.183	4.006	4.800	72.488	5.13	4.76
	Median	1997	1965	1998	30	32	1
	Minimum	1985	1960	1984	1	22	-13
	Maximum	2005	1976	2012	788	49	18
Humanities	N	347	347	347	347	347	347
	Mean	1996.78	1965.76	2001.11	3.91	35.35	4.32
	Std. Deviation	4.341	4.115	4.382	5.338	4.52	4.19
	Median	1997	1965	2001	2	35	4
	Minimum	1986	1960	1986	1	24	-6
	Maximum	2005	1978	2012	65	47	20
Non-Health Professional	N	112	112	112	112	112	112
	Mean	1997.84	1965.52	2001.21	10.30	35.70	3.36
	Std. Deviation	4.594	4.480	5.070	14.222	5.84	4.89
	Median	1998	1965	2001.5	4	35	3

Disciplinary division		Birth year	PhD year	YFP	P	Birth year to YFP	PhD year to YFP
	Minimum	1985	1960	1990	1	24	-6
	Maximum	2005	1977	2012	70	51	21
Sciences	N	826	826	826	826	826	826
	Mean	1995.35	1965.88	1997.92	36.45	32.04	2.57
	Std. Deviation	4.441	4.287	4.860	48.406	4.67	4.37
	Median	1996	1965	1999	25.00	32	3
	Minimum	1985	1960	1982	1	22	-11
	Maximum	2005	1977	2012	775	46	17
Social Sciences	N	502	502	502	502	502	502
	Mean	1997.36	1966.75	1999.66	15.87	32.9084	2.3008
	Std. Deviation	4.25	4.33	4.53	19.11	4.36	3.7
	Median	1998.00	1967.00	2000.00	10.00	33	2
	Minimum	1987	1960	1986	1	23	-11
	Maximum	2005	1977	2012	204	48	15
Total	N	3596	3596	3596	3596	3596	3596
	Mean	1995.95	1965.77	1998.73	32.04	32.97	2.78
	Std. Deviation	4.64	4.18	4.89	50.56	4.77	4.60
	Median	1996	1965	1999	17	33	3
	Minimum	1982	1960	1980	1	20.00	-13.00
	Maximum	2005	1978	2012	788	51.00	27.00

Appendix 2. Pearson correlations by ages

Division	Ages	Birth year	YFP	PhD year
Basic Medical Sciences	Birth year	1.000	0.426	0.627
	First publication year	0.426	1.000	0.297
	PhD year	0.627	0.297	1.000
Business & Management	Birth year	1.000	0.521	0.656
	First publication year	0.521	1.000	0.540
	PhD year	0.656	0.540	1.000
Education	Birth year	1.000	0.277	0.686
	First publication year	0.277	1.000	0.670
	PhD year	0.686	0.670	1.000
Engineering	Birth year	1.000	0.531	0.800
	First publication year	0.531	1.000	0.588
	PhD year	0.800	0.588	1.000
Health Sciences	Birth year	1.000	0.333	0.530
	First publication year	0.333	1.000	0.444
	PhD year	0.530	0.444	1.000
Humanities	Birth year	1.000	0.435	0.733
	First publication year	0.435	1.000	0.538
	PhD year	0.733	0.538	1.000
Non-Health Professional	Birth year	1.000	0.258	0.605
	First publication year	0.258	1.000	0.492
	PhD year	0.605	0.492	1.000
Sciences	Birth year	1.000	0.484	0.793
	First publication year	0.484	1.000	0.561
	PhD year	0.793	0.561	1.000
Social Sciences	Birth year	1.000	0.517	0.768
	First publication year	0.517	1.000	0.646
	PhD year	0.768	0.646	1.000
Total	Birth year	1.000	0.457	0.711
	First publication year	0.457	1.000	0.535
	PhD year	0.711	0.535	1.000

Corpus Specific Stop Words to Improve the Textual Analysis in Scientometrics

Vicenç Parisi Baradad[1] and Alexis-Michel Mugabushaka

[1]*Vicenc.PARISI-BARADAD@ec.europa.eu*
European Research Council Executive Agency, COV 24/161, B-1049 Brussels (Belgium)

Abstract
With the availability of vast collection of research articles on internet, textual analysis is an increasingly important technique in scientometric analysis. While the context in which it is used and the specific algorithms implemented may vary, typically any textual analysis exercise involves intensive pre-processing of input text which includes removing topically uninteresting terms (stop words). In this paper we argue that corpus specific stop words, which take into account the specificities of a collection of texts, improve textual analysis in scientometrics. We describe two relatively simple techniques to generate corpus-specific stop words; stop words lists following a Poisson distribution and keyword adjacency stop words lists. In a case study to extract keywords from scientific abstracts of research project funded by the European Research Council in the domain of Life sciences, we show that a combination of those techniques gives better recall values than standard stop words or any of the two techniques alone. The method we propose can be implemented to obtain stop words lists in an automatic way by using author provided keywords for a set of abstracts. The stop words lists generated can be updated easily by adding new texts to the training corpus.

Conference Topic
Methods and techniques

Introduction
Textual analysis -also referred to as "lexical analysis"," text mining", "co-word analysis" or "linguistic network"- has a long tradition in scientometric analysis. Earlier references can be found in the pioneering work of Eugene Garfield and others (see Garfield, 1967) studying the potential of citation analysis in information retrieval as compared to methods based on terms frequencies. Callon et al. (1983, 1986) introduced the concept of co-word analysis in science and technology studies. This technique was further developed and popularized in scientometrics by the work of Leydesdorff (1989) and researchers at the Center for Science and Technology Studies (CWTS) at the Leiden University (Noyons & van Raan, 1998).

With the availability of vast collections of research articles and better and faster computer tools, which help text analysis, the technique has firmly established itself in scientometric analysis. Nowadays it is used in various contexts: to study the thematic proximity in a collection of documents; to map scientific papers based on concept maps; to detect dynamics and trends of research based, for example, on centrality of concepts or to characterise a particular research community, by identifying relationships between the terms it uses.

While textual analytical techniques differ in degree of complexities and approaches they take, virtually all of them require relatively intensive pre-processing of the input texts. Typically, the following steps are involved in the pre-processing: (1) tokenization, (2) converting to lower case, (3) stemming and (4) removing stop words. For this last step, researchers typically use standard stop words lists obtained from texts in many different domains.

In this paper we argue that using corpus specific stop words might help the textual analysis. The paper is divided in four parts. The next section reviews briefly existing work on stop words and describes in detail two, relatively simple methods, to extract corpus specific stop words. In the subsequent, third, section we present a case study to illustrate the benefits of corpus specific stop words over more general stop words. The concluding remarks discuss limitations and point to future directions.

Related Work

When researchers in scientometrics started using textual analysis, they were standing in long tradition of information retrieval research. Early studies of word frequencies in a text or collection of documents appeared in the last century, when George K. Zip formulated an empirical law that relates terms frequencies (tf) to rank in a frequency ordered word list (Zip, 1932). This frequency characterisation was used later by Hans Luhn to obtain statistical information of words in texts and to compute a relative measure of the significance of individual words and phrases (Luhn, 1958). Using this measure Luhn hypothesized that the most discriminant words are those appearing in the middle of the frequency rank. Salton went a step further by incorporating the document frequency (df) as a measure of the discriminatory capacity of the words (Salton & Young, 1973). They suggested that words can appear in a document collection either in a random manner or concentrated in a few exemplars and they proposed the product of the term frequency times the inverse document frequency (tf • idf) as a measure of the degree of significance: the words appearing in many documents (df high) or with a low presence (tf low) are considered stop words. Based on these frequency descriptions Christopher Fox elaborated in the 90's a list containing stop words (Fox, 1990) extracted from the Brown Corpus of English literature. Although these stop words can be considered the standard or classical list and they have been frequently used, we note two limitations: first they are quite outdated and second they may be too general to take into account the specificities of a collection of texts. They may not be suitable to filter out words belonging to specific research fields or words of recent apparition. As Makrehchi & Kamel (2008) suggest, specific stop words differ from one domain to another.

Several methodologies have been proposed recently to create new stop words lists, customized to particular corpus. Among them, two proposals attracted our attention due to their relative simplicity.

On one hand, an unsupervised method to compute stop words lists arises from the study of the statistical distribution of words, by Church, K. and Gale, W. (1995) and their hypothesis that common stop words follow a Poisson distribution. This has been used to create a stop word list for particular Polish texts (Jungiewicz & Lopuszyński, 2014). We call this approach the *Poisson stoplist.*

Under this hypothesis one assumes that the document frequency of words (df) in a corpus can be estimated (dfe) from their term frequency (tf) and the total number of documents (N) by using the probability theory:

$$\frac{dfe}{N} = 1 - P(0),$$

where $P(0)$ is the probability of not appearing the word. Assuming a Poisson distribution for stop words, the probability of k instances of a word is given by:

$$P(k, \mu) = \frac{e^{-\mu} * \mu^k}{k!},$$

where μ is the average number of instances per document:

$$\mu = \frac{tf}{N}$$

The relation dfe/df is supposed to be close to 1 for randomly distributed terms (stop words) and shows an increase for highly cluttered terms (keywords); although this depends on the corpus, as Jungiewicz and Lopuszyński found when computing their stop word lists for legal texts from the public procurement domain. They realised that their most common stop words had a high variability in their distribution and replaced the Poisson assumption with a negative binomial distribution, which allows a larger variance.

On the other hand, S. Rose et al. (2010) proposed an unsupervised, domain and language independent method to extract keywords from individual texts called RAKE (Rapid Automatic Keyword Extraction) and a supervised method to elaborate stop word lists based on the intuition that words adjacent to keywords tend to be stop words.
RAKE uses stop words to parse the text and extract candidate key phrases (consisting in one or more words). The key phrases are then scored by computing word co-ocurrences and using a metric that favours words belonging to long key phrases. The top T candidates are chosen as keywords (key phrases).
The method proposed by S. Rose to extract stop words from a corpus resorts on accumulating for each word its 'adjacency frequency' (af) and 'keyword frequency' (kf), together with the term frequency (tf) and document frequency (df). Then, given a selection threshold n, the most frequent words with af > kf are chosen as stop words. This method is called by the author *keyword adjacency stoplist* (because it includes primarily words that are adjacent to and not within keywords: Rose et al. 2010, p. 14). We refer to this method as *RAKE stoplist* in this paper.

Case Study: stop list for a collection of abstracts of funded projects

To study the suitability of the above described methodologies and create our own stop words list we applied them to a corpus from abstracts of projects, funded by the European Research Council, in the Life Sciences domain. This corpus consists of 1579 projects covering diverse research areas. The table 1, shows the number of project abstracts by each research area (which corresponds to the scientific panel in which the project was evaluated).

Table 1. Overview of the corpus of abstracts used in the case study

Scientific areas	abstracts	%
Molecular and structural biology and biochemistry	176	11.1
Genetics, genomics, bioinformatics and systems biology	178	11.1
Cellular and developmental biology	164	10.4
Physiology, pathophysiology and endocrinology	176	11.15
Neurosciences and neural disorders	217	13.7
Immunity and infection	168	10.6
Diagnostic tools, therapies and public health	209	13.2
Evolutionary, population and environmental biology	168	10.6
Applied life sciences and biotechnology	115	7.3

Creating stop words

We randomly chose 80% of the abstracts as a training set and the other 20% as a test set.

Following the algorithms outlined in Rose et al. 2010, we wrote a program in Python to create a table (which we call Frequency table) with all the words (12621 in total) of the training set that contains the words, term frequencies (tf), document frequencies (df), keyword frequencies (kf) and adjacent frequencies (af).
This table was used to create both the *Poisson stoplist* and the *RAKE stoplist*. For the later, we set various thresholds to obtain the top n words with the highest term frequency.

Evaluating the stopwords

To evaluate if the corpus-specific stop words improve textual analysis, we use them in extracting keywords. We compare the keywords extracted using those stop words with

author-provided keywords. The idea is that, depending on the stop words used, the keywords extracted will match more or less the ones provided by the authors and the higher the share of matched keywords the better the stop words list.

It should be noted that author-provided keywords do not necessary contain words which also appears in the abstracts. In our corpus, out of 7845 keywords given by the authors only 3494 (44.5 %) where encountered in the abstracts. This means that the precision and F-measure need to be taken into account with care and thus we have not used them for the evaluation of the quality of the stop words list, resorting only to the recall measure, computed as the relation between the total number of correct extracted keywords and the total number of keywords given by the authors, that appear in the abstracts.

We compared the keywords provided by authors with the keywords extracted using the following lists of stop words

1. Standard *Fox stop words list*
2. Stop words list created using the Poisson distribution hypothesis (*Poisson stoplist*)
3. Stop words list computed using keyword adjacency (*RAKE stoplist*)
4. Stop words lists computed using combinations of *Fox,Poisson* and *RAKE*

For keywords extraction we used a Python implementation of the RAKE algorithm (https://github.com/aneesha/RAKE)

1. *Fox stoplist*

This list serves as a baseline for our work and the computation of the recall of the keywords extracted using RAKE algorithm does not need to tune any parameter. The recall obtained is 56.42%.

2. *Poisson stoplist*

To extract the stop words using this approach we need first to set the threshold for the relation dfe/df. To do that we computed the mean and standard deviation of the dfe/df for all the *Fox* stop words that appear in the training set. Figure 1 shows the plot of these values, where the mean (dfe/df) + std(dfe/df) is 1.55. There are only 14 Fox stop words excluded from the list and apart from the words (*ordering, right and small*) their term frequency is very low. We have used this threshold to obtain the stop words from our training data appearing in at least 10 documents (df>10) and we have obtained a list of 2008 words that gives a recall of 58.25% in our test set, which is better than the *Fox stoplist*.

3. *RAKE stoplist*

To use the RAKE approach we extracted all the words from the training set with af>kf and created an ordered table, sorted in descending order of word occurrence (tf). This table consisted in a list of 2045 candidate stop words. To choose the top best frequency rank we tested subsets of these lists and computed their recall values. The result obtained using all the words in the list was 45.42 % of recall and the results improved by removing words from the list, having a peak at a 53.31% of recall, when using the first 185 words of the rank.

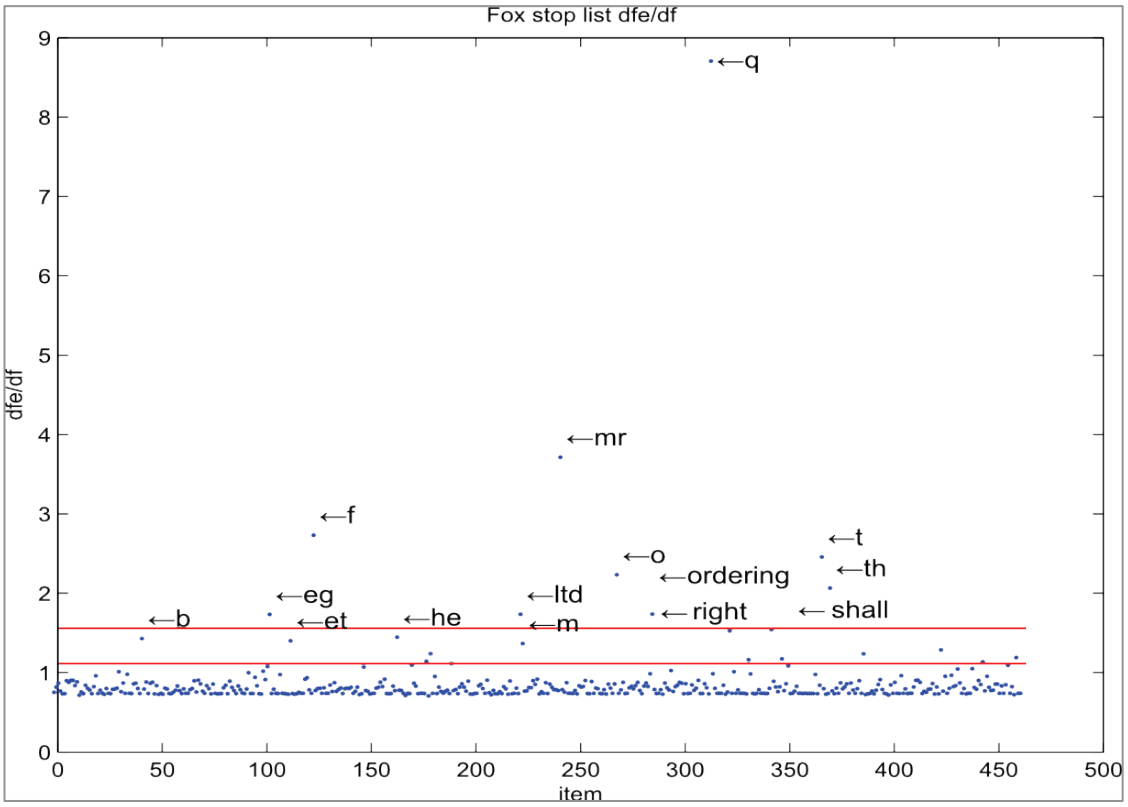


Figure 1. Fox stoplist dfe/df values found in the training corpus. Just a few words are above the standard deviation limit, and they are rarely found (tf very low).

4. *Combinations Poisson and RAKE*

Since the *RAKE stoplist* gave us worse results than the *Fox stoplist*, we tried to combine them with the Poisson approach (*RAKE-Poisson*) and we extracted the words with df>10,af>kf and dfe/df<1.55. This improved the previous results, giving a recall of 62.34%. Note that the condition dfe/df> r can also be seen as an adaptive threshold on tf, since, under the Poisson distribution, it can also be expressed as:

$$tf > N * \ln\left(\left(\frac{df}{N}\right)r - 1\right),$$

and instead of choosing a minimum common tf for all the words, we adapt the tf to each word's df. In Figure 2 we have plotted the df of the RAKE stop words (tf>0) , together with the Fox stop words found in the *RAKE stoplist*. Also we plotted the dfe/1.55 curve which shows the limit above which the words belong to *RAKE-Poisson stoplist*.

After inspecting the frequencies of the RAKE-Poisson stop words we found words expected to appear in Life Sciences texts and we questioned ourselves if their removal from the stoplist would improve the recall results. To check it we removed them by hand and the recall increased to 64.56 %. A more detailed inspection of the stoplist frequencies allowed us to see that just a few words (6 in total) belong to the life sciences domain (genetic, disease, protein, molecular, gene, cell), all them with kf>60 had a 1.1<dfe/df <1.55. In all them the af/kf relation was less than 5 (af/kf<5). This data gave us the intuition that we needed to decrease the dfe/df threshold and also to be more strict on the af/kf condition, so we tested a stoplist consisting in the RAKE intersection with Poisson stoplist (df>10 and af>5*kf and dfe/df<1.2) which gave a recall of 68.69%, being this the best result. We call it the *RAKEm-Poisson stoplist*.

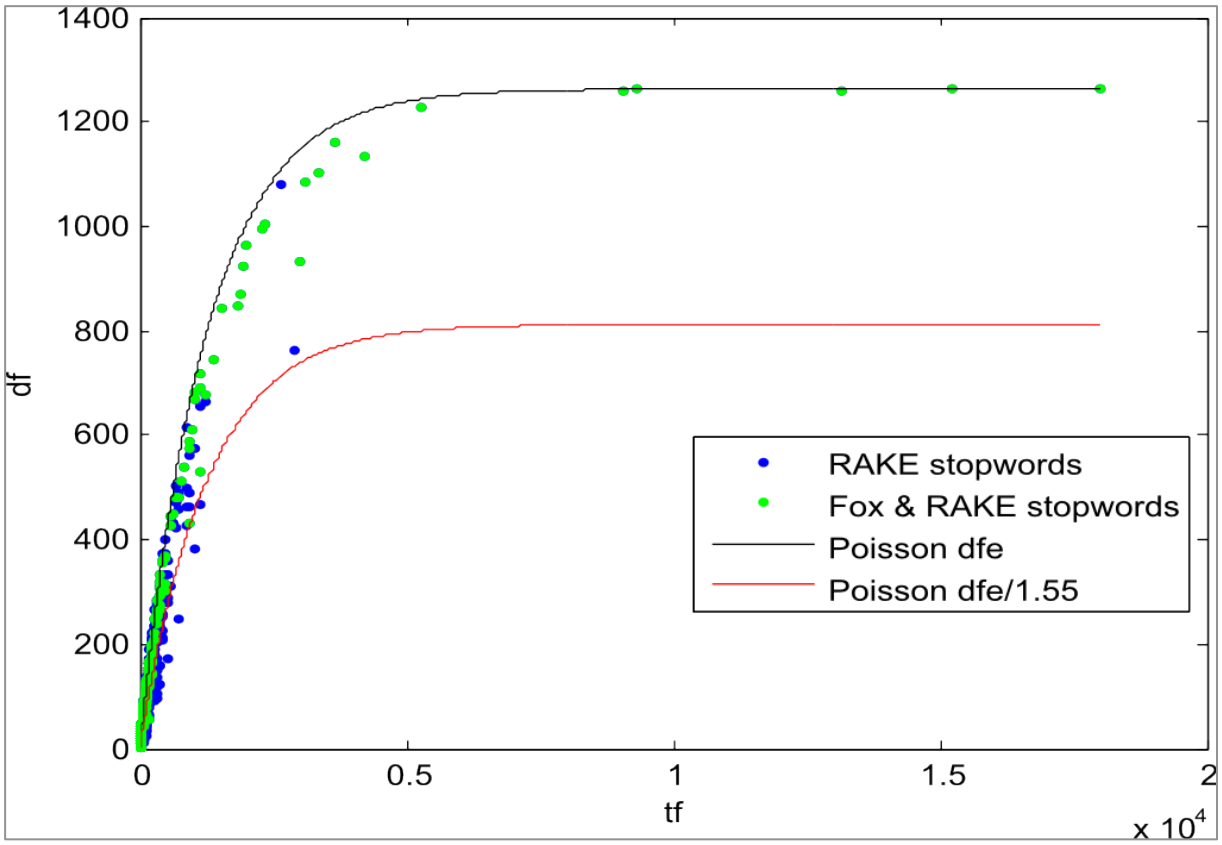


Figure 2. RAKE and Fox stop words. We can see that the Fox stop words follow the Poisson distribution better than the RAKE stop words, which appear more concentrated at low df values.

Conclusion

Our aim is to obtain stop words that help to provide meaningful and significant keywords that summarize the texts; the validation of the stoplists we did was based using the author given key phrases which most of the times had fewer words than the ones obtained using RAKE. We think that this circumstance is favouring standard stoplists since they will still produce single word keywords given by authors and end up yielding overall recall values similar to specific domain stoplists. Therefore we plan as a future work to use measures that evaluate semantic value of the key phrases.

We would like to remark that the RAKE-Poisson stoplist can be obtained from the word frequencies and the author keywords, without further human intervention. Our future work involves also the automatization of the computation of the best af/kf and dfe/df thresholds to generate the *RAKEm-Poisson stoplists*.

References

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping, *World Patent Information, 29*(4), 308-316.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: an introduction to co-word analysis. *Social Science Information, 22*, 191-235.

Callon, M., Law, J., & Rip, A. (Eds.). (1986). *Mapping the Dynamics of Science and Technology*. London: Macmillan.

Church, K. and Gale, W. (1995). Poisson mixtures. *Journal of Natural Language Engineering*, *2*, 163-190.

Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, *24*, 19-35.

Garfield, E. (1967). Primordial concepts, citation indexing and eIistorio-bibliography. *The Journal of Library History, 2*(3), 235-249. http://www.garfield.library.upenn.edu/essays/v6p518y1983.pdf

Jungiewicz, M. & Lopuszyński, M. (2014). Unsupervised keyword extraction from Polish legal texts. *Advances in Natural Language Processing*, 65–70. Springer LNCS.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy*, 18(4), 209-223.

Luhn, P. (1958). The automatic creation of literature abstracts, *IBM Journal of Research and Development, 2*(2) 159–165

Makrehchi, M. & Kamel, M. (2008) *Automatic Extraction of Domain-specific Stopwords from Labeled Documents*. Berlin / Heidelberg: Springer.

Noyons E. C. M. & Raan A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science*, *49*(1), 68–81.

Rose, S., Engel, D., Cramer, N. & W. Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*. John Wiley & Sons, Ltd.

Salton, G. & Yang, S. (1973). On the specification of term values in automatic indexing, *Journal of Documentation,* 29(4), 351–372.

Sinka, M.P. & Corne, D.W. (2003). Towards modernised and web-specific stoplists for web document analysis, *IEEE/WIC International Conference on Web Intelligence,* 396-402.

Zipf, K. (1932). S*elective Studies and the Principle of Relative Frequency in Language*. Cambridge: MIT Press.

Epistemic Diversity as Distribution of Paper Dissimilarities

Jochen Gläser[1], Michael Heinz[2] and Frank Havemann[2]

[1] *Jochen.Glaser@ztg.tu-berlin.de*
Center for Technology and Society, TU Berlin, Hardenbergstr. 16-18, Berlin, 10623 (Germany)

[2] *{Michael.Heinz, Frank.Havemann}@ibi.hu-berlin.de*
Berlin School of Library and Information Science, Humboldt-University of Berlin, Dorotheenstraße 26, 10099 Berlin (Germany)

Abstract
We continue our quest for measures of epistemic diversity that fit the inherent properties of thematic structures in science. Starting from theoretical considerations, we argue that currently available measures of diversity are not applicable to the epistemic diversity of published scientific knowledge because topics are fluid and overlap. Consequently, we abandon attempts to assign papers to topics and instead explore opportunities to measure diversity based on paper dissimilarities. Considerations of the exploitation of information and signal-to-noise ratios in networks of papers let us dismiss an earlier attempt to base a dissimilarity measure on the resistance distance between papers in the network of papers and their cited sources. In this paper, we explore a dissimilarity measure based on papers' 'views' on the whole network, with the 'view' of a paper consisting of all other papers in the network ranked according to the length of their shortest paths to the paper. We present test results on the diversity of topics, journals and country outputs for information science (2008) as well as on the diversity of country outputs in astronomy and astrophysics (2010).

Conference Topics
Methods and techniques; Indicators

Introduction
The epistemic diversity of research – the diversity of empirical objects, methods, problems, or approaches to solving them – has become a matter of concern for science policy. Attempts by science policy to increase the selectivity of research funding and the growth in strength and homogeneity of incentives for universities have led to concerns about an undue reduction of the diversity of research. Several specific warnings refer to the UK's research assessment exercise (Gläser et al., 2002, Molas-Gallart & Salter, 2002, Rafols et al., 2012). A similar concern has been raised in Germany, where profile-building activities at all universities may make the small subjects disappear (HRK, 2007). Laudel & Weyer (2014) observed in the Netherlands that universities' uniform responses to political signals contributed to the disappearance of one field and the stagnation of another.

Discussions about dangers to the epistemic diversity of research have in common that they lack both theoretical backing and empirical evidence. Epistemic diversity is an ill-understood topic in science studies. It is rarely clear what the concept is intended to refer to, how epistemic diversity might affect research, and how it can be operationalized. Theoretical reasoning drawing on analogies to biodiversity assumes diversity is good for science (e.g. Rafols et al., 2012). However, arguments lack empirical grounding, and no specific arguments about necessary and sufficient levels of diversity or about dangers of too much diversity can be made. The empirical studies of interdisciplinarity (e.g. Bordons et al., 2004; Rafols & Meyer, 2007; Rafols et al., 2012) were forced to use rather coarse indicators such as the journal classification of the web of science, and could not theoretically justify the measures they applied.

The aim of our paper is to present a systematic approach to the measurement of diversity that derives possible bibliometric measures of diversity from properties of the system whose diversity is to be measured, namely scientific knowledge.

We start from a theoretical definition of 'topics' in science and demonstrate that the properties of topics do not match the built-in assumptions of current indicators. While this does not necessarily invalidate the indicators, the assumptions underlying the measurement of diversity in science must be made explicit, and their applicability be argued. We suggest two additional strategies that may alleviate the problems resulting from the mismatch between properties of topics and prerequisites of indicators. The first strategy abandons the explicit identification of topics and measures the diversity of paper networks rather than scientific knowledge. We propose a measure of paper similarity that takes some of the properties of scientific knowledge into account, and demonstrate our approach by applying the measure to two data sets. The second strategy, which is outlined in this paper but not applied, uses the same similarity measure for determining the disparity of topics, thereby enabling the application of existing diversity measures.

Theoretical background

In the most general sense, 'diversity' is the property of a system, namely its heterogeneity, which is caused by the disparity of its elements. Among the many aspects of a science system to which the concept diversity can be applied, we are interested in the diversity of published scientific knowledge. Other aspects of a field's diversity such as the diversity of informal knowledge, instrumentation, empirical objects, or scientific training of researchers, will not be considered here. The *epistemic diversity of a research field* is thus defined here as the diversity of published knowledge claims about scientific problems, solutions, empirical objects, approaches and methods, which are communicated by the field's researchers in publications.

The definition of epistemic diversity as a property of published knowledge suggests using bibliometric methods for its measurement. These methods must support the reconstruction of knowledge structures from publications in a way that is both valid (i.e. returns knowledge structures researchers work with) and supports the measurement of diversity. Fulfilling both requirements is made difficult by inherent properties of knowledge structures in science. In the following, we first discuss the built-in assumptions of current measures of diversity. We then argue that properties of scientific knowledge and of its representation in publications do not meet these assumptions, and discuss opportunities to reconstruct knowledge structures from publications and to measure the epistemic diversity of research.

Built-in assumptions of current approaches to the measurement of diversity

Diversity has been an important topic of biological and environmental research for some time. These fields are mainly concerned with the impact of diversity on the stability and development of biotopes and species. Two approaches to the measurement of biodiversity can be distinguished:

a) The diversity of biotopes[1] composed of several species is measured with a three-level hierarchical approach. Biotopes are considered as consisting of species, which in turn consist of individuals. Three factors contribute to the diversity of such a system, namely

- variety (the number of species in the biotope),
- disparity (the extent to which the species differ from each other), and
- evenness (the distribution of individuals across the different species).

Depending on the research question, these factors can be assessed separately (e.g. if only the number of species is measured) or be combined in synthetic measures such as Shannon's Entropy (combining variety and evenness) or the Rao-Index (combining all three measures).

[1] A biotope is a physical environment (habitat) with a distinctive assemblage of conspicuous species (Olenin & Ducrotoy, 2006: 22).

This approach to diversity is applied in fields outside the biosciences as well (see Rafols et al., 2012, Stirling, 2007). It requires that

- the system whose diversity is to be measured can be analytically decomposed in three levels (system, categories, and elements),
- the contribution of differences between individuals of the same species to the biotope's diversity can be neglected,
- the categories can be constructed as disjunct by assigning each element to exactly one category or by fractional assignments of elements to categories, and that
- all categories share a property that can be used to calculate disparity.

b) The diversity of species composed of individuals is measured on the basis of a two-level approach. In this approach, variety and evenness become meaningless because there is no intermediate level of categories to which elements can belong. The only remaining basis for measuring the diversity of the system is the disparity of individuals. While this approach is used less frequently, it can be considered to be more fundamental because it conceptualizes diversity as the degree to which the elements of a system (here: a species) differ from each other. This approach is applicable as long as a system can be delineated and elements share a property that can be used to calculate disparity.

Both approaches share a premise concerning the disparity of categories and elements. Categories and elements are conceptualized as stable, and their pairwise disparities as independent, i.e. not affected by other categories respectively elements. New elements entering the system (i.e. individuals of a species being born or migrating to a biotope) do not affect the disparity between existing elements or between the categories, and new categories (i.e. species migrating to a biotope) do not affect the disparity between the categories or between the elements that are already present. The same applies to the disappearance of elements or categories.

Properties of topics in scientific knowledge

If the approaches to the measurement of diversity are to be applied to scientific knowledge, the system, categories and elements must be determined. For the three-level approach, the system would be the knowledge of a field, topics in this field would serve as categories, and knowledge claims (the claim for some empirical, theoretical or methodological statement to be true) would constitute the elements of the system. For the diversity measures discussed above to be applicable, these knowledge structures would need to fulfil the built-in assumptions of the measures. We therefore begin by briefly discussing the properties of scientific knowledge in its structures.

Scientific knowledge is produced by scientific communities whose members

- observe the community's shared body of knowledge,
- interpret this knowledge in the light of their own research experience,
- identify gaps in that knowledge and design research processes for producing the knowledge that closes the observed gap, and
- offer their interpretation and the new knowledge to their community.

The interpretation of the community's knowledge and claims about new knowledge are fully or partially shared by some members of the community. We define a topic as *a focus on theoretical, methodological or empirical knowledge that is shared by a number of researchers and thereby provides these researchers with a joint frame of reference for the formulation of problems, the selection of methods or objects, the organization of empirical data, or the interpretation of data* (on the social ordering of research by knowledge see Gläser, 2006). This definition resonates with Whitley's (1974) description of research areas but abandons the assumption that topics form a hierarchy. The only demand the definition makes is that some scientific knowledge is perceived similarly by researchers and influences their decisions.

Due to this nature as shared and collective perspectives, topics have structural and dynamic properties that affect the opportunities for measurement. *Structural properties* include the following:

1) All topics are *emergent meso- or macro-structures*, i.e. they are collective-level products of autonomous interpretations and uses of knowledge by individual researchers.

2) From this follows that topics are *local* in the sense that they are primarily topics to the researchers whose decisions are influenced and who contribute to them, and only secondarily topics to those colleagues who are outside observers.

3) Given the multiple objects of knowledge that can serve as common reference for researchers, it is inevitable that topics *overlap*. Overlaps are ubiquitous because any research is likely to address several topics at once, e.g. by including theories about an object, methodologies for investigating it, and empirical information about an object. They also occur when a knowledge claim belongs to several topics at once (e.g. formulae used in bibliometrics belonging to mathematics but also expressing bibliometric relationships).

4) Knowledge has a *fractal structure* (e.g. van Raan, 2000), and topics can have any size between small (emerging topics that in the beginning may concern just two or three researchers) and very large thematic structures such as bibliometrics. The 'size' of a topic can be defined in various ways – as scope (range of phenomena covered), level of abstraction (which is again linked to the range of phenomena covered), or number of research processes or researchers influenced by it. In all these dimensions there is a continuum from very small to very large topics.

5) The degree to which knowledge influences researchers' actions, and the strength of links between new findings and existing knowledge that are constructed by researchers, also vary between 'very weak' and 'very strong'. As a result, the '*distinctiveness'* of topics varies. Some topics are unambiguously seen as being different from other knowledge by most researchers of a field and are thus well separated from surrounding knowledge, while others are much less pronounced.

These structural properties of topics let them form an inconsistent poly-hierarchy for which not even meaningful levels can be determined. This also implies that no field or collection of papers has exactly one definite thematic structure. Different perspectives can be applied to fields and collections of papers and will return different topical structures. Topics may overlap in their boundaries or pervasively. They vary considerably in their size and 'distinctness', i.e. the extent to which they actually constitute a shared concern of researchers.

Dynamic properties of topics are shaped by their role in the knowledge production process. As coinciding perspectives of researchers, topics are perpetually changing. Researchers constantly revise their perspectives on the existing knowledge and thus the relationships of their perspectives to those of their colleagues. They also utilize and contribute to more than one topic (e.g. theoretical, methodological and empirical ones). Hence, their production of new knowledge may instigate at least one and in many cases all of the following changes:

* Enrichment: Since new knowledge is added to the system, the community's knowledge on a topic is likely to grow.

* Restructuring: The new knowledge is linked to existing knowledge and thereby links existing knowledge, i.e. the density of connections in the system of knowledge increases.

* Reduction: The new knowledge may devalue existing knowledge by proving it to be wrong or may reduce it by subordinating it to a generalisation.

Through these processes, the size of topics, their distinctness and relations between them are constantly changed. New topics may emerge at any time, and existing topics may disappear or radically change.

Representation of knowledge in publications and reconstructions of topics

Since bibliometric methods reconstruct knowledge structures from publications, the representation of knowledge in publications provides the opportunities and constraints for a bibliometric measurement of diversity, which we now discuss in more detail. In the sociology of science, knowledge claims are treated as the basic unit through which new knowledge is communicated (e.g. Cozzens, 1985, Pinch, 1985). Knowledge claims are claims that some new knowledge produced by the author is true; a publication usually contains several such claims.

For the new knowledge claims to be added to the community's body of knowledge, they must be used by other community members in their subsequent knowledge production. This requires the new knowledge to be available to all potential users, which is achieved by publication. With each publication, researchers construct

- an account of the state of the current knowledge on a topic,
- the claim that there is a specific gap in that knowledge,
- the claim to have developed an approach whose application can close that gap,
- the new knowledge produced with this approach, which is claimed to close the gap, and
- in many cases conclusions concerning implications of the new knowledge including the necessity of further specific research (Gläser, 2006: 125-126, Swales, 1986: 45).

These claims embed the new knowledge that is offered to the community in the existing knowledge. However, they do so selectively and *ad hoc*. The claims in a publication are organised in a way that maximises the chances of the new knowledge's further use by emphasizing originality, relevance, validity and reliability of the new knowledge. Links to the existing knowledge are crafted to further this impression.

The new knowledge claims shape subsequent knowledge production processes if they inform the formulation of problems, choice of methods or interpretation of results by readers of the publication. If they do so, the researchers using them are likely to indicate the link of the new knowledge they offer to these knowledge claims, thereby treating them as part of the community's knowledge. This 'elementary process' of adding knowledge causes the dynamic properties described in the previous section. If a new knowledge claim is added, the community's knowledge becomes enriched, and its structure changes because the claim creates new links between, reinforces or remove existing links. New knowledge claims may also invalidate existing claims or subsume them to more general statements if they are used by other community members in this way.

Consequences for the measurement of diversity

The properties of knowledge claims and topics affect the opportunities to reconstruct topics from publications with bibliometric methods, i.e. by using properties of publications such as authors, journals, references, or terms. To begin with, no method for the bibliometric reconstruction of individual knowledge claims has been proposed so far. Knowledge claims are represented in series of sentences and clauses that are distributed across a publication. Reconstructing them would be a task for linguistics but is still impossible for that field, too.

Bibliometric methods are better suited for the reconstruction of topics because the latter are larger and span many publications. However, from the properties of topics described earlier follows that none of the bibliometrically usable properties of a paper can be assumed to be thematically homogeneous in the sense of representing only one topic. Since research processes are influenced by and address more than one topic, topics overlap in research processes, publications (and thus references), terms, journals, and authors. Furthermore, researchers apply their individual perspectives on the scientific knowledge when constructing

and linking topics, which is why links to topics may occur unpredictably in a variety of scientific fields. Consequently, any finite sub-set of papers is unlikely to include all publications addressing a specific topic, which means that any hierarchy of topics is also only partially covered by the paper set.

Owing to the mismatch between properties of publications that can be used for the reconstruction of topics and the representation of topics in publications, bibliometric methods inevitably reduce the complexity of the underlying knowledge structures. This is not a problem in itself because all models reduce complexity. The question is not how the reduction of complexity can be avoided but whether a specific reduction of complexity is appropriate to the purpose. Answers to this question should be part of a bibliometric methodology that links specific purposes of topic reconstruction to specific strategies that are applied. The absence of such a methodology is one of the major obstacles for bibliometrics.

When we apply these methodological considerations to the measurement of epistemic diversity, we can distinguish three strategies for solving the problems posed by properties of scientific topics. The first strategy, which has been applied in all attempts to measure epistemic diversity so far, constructs distinct topics to which papers are assigned. The three-level approach is then used for the measurement of diversity.

A second possible strategy would be to construct overlapping topics to which papers belong partially. In order to apply three level-diversity measures, the topics would have to be made disjunct by fractionalising the papers. The disparity of topics would need to be measured based on the difference in paper membership. While this strategy still has some problems in the case of pervasive overlaps of topics, it would create a more precise representation of topics and still enable the application of three-level diversity measures.

The third strategy, which we apply in the remainder of the paper, circumvents the problem of topic reconstruction by applying the two-level approach. Since knowledge claims cannot be reconstructed from publications, the strategy measures paper diversity as a proxy for knowledge diversity. This strategy requires a similarity measure for published papers, which should reflect the properties of thematic structures in science discussed above.

Methods and Data

Network-based measures of paper similarity

Diversity measures for the two-level approach aggregate the pairwise similarities of all elements. Among the many ways in which the similarity of two papers in a network can be determined, we need to find those that strike a balance between utilizing as much information as possible and avoiding the inclusion of irrelevant information that contaminates the measure.

Bibliographic coupling is well-established, and is commonly considered as one of the best bibliometric measures of paper similarity (Ahlgren & Jarneving, 2008: 274-275). The strength of bibliographic coupling between two papers can be used directly as a measure of their similarity. However, bibliographic coupling is not a useful measure for the similarity of papers that are not coupled. All these papers must be considered equally dissimilar, which they are certainly not. Thus, bibliographic coupling is unsatisfactory as a measure of paper similarity in networks.

An alternative to using bibliographic coupling is the utilization of all connections in a network, e.g. by measuring similarity as resistance distance in networks of papers and their cited sources or in bibliographic coupling networks. In this approach, indirect links between papers are taken into account, i.e. information about the whole network is utilized for the calculation of all pairwise paper similarities (see Gläser et al., 2013 for an example). However, this approach inevitably uses information about detours through a network – i.e.

about connections that exist and can technically be made but are not meaningful in terms of paper similarities. In other words, the measure is distorted by paths that do not reflect thematic similarity. Furthermore, our own experiments showed the measure to favour papers with a high degree. Finally, using all paths in a paper network for the measurement of its diversity makes the measure particularly sensitive to changes in the network structure. If measures of paper similarity are based on the resistance distance, each paper that is added to the network changes the resistance distance and thus the similarities of all papers in the network. This is an extremely unrealistic assumption about the impact of new publications on the epistemic diversity of a field.

Between the use of only information about direct coupling and the use of information about all possible connections between papers lie measures such as length of the shortest path between two nodes. This measure makes little sense in networks of papers and their cited sources because each reference two papers have in common creates a path of the length two between them. For networks in which links reflect the relative strength of bibliographic coupling, the length of shortest paths captures more information.

By determining the length of the shortest path between two papers in a network, other connections are taken into account indirectly by dismissing them as longer paths. Still, the environment of a paper is largely neglected by such a measure. However, the length of shortest path can be used to construct an indirect measure of paper similarity that takes the environment of papers into account. We can construct the 'view' of a paper on its environment by ranking all other papers in the network according to their distance to that paper. The 'view' describes how dissimilar other papers in the network are in terms of their shortest paths. The similarity between two papers can be defined as the similarity of the two papers' 'views' on the network, which is measured by calculating the rank correlation of the two lists.

Thus, we measure the similarity of two papers by:

- determining the shortest paths between all pairs of papers in a bibliographically coupled network (weighted with the arccosine of Salton's Cosine),
- creating a 'view' of each paper by ranking all other papers according to increasing lengths of their shortest paths,
- calculating the similarity of two papers as the rank correlation (Spearman) between the two lists, and
- transforming the rank correlation in a similarity measure.

This measure, which can be interpreted as the similarity of the 'views' of the two papers on their scientific environment, avoids the influence of degrees. It is similar to the use of "preferences" in an "affinity" system by Balcan et al. (2012) in their construction of overlapping endogenous communities.

Data

To test our measure, we used two data sets. The first data set is the main component of publications (articles, letters and proceedings papers) in six information science journals, which consists of 492 papers (see Havemann et al., 2012 for a description of this data set). The second data set is the main component of 14,770 publications (articles, letters, and proceedings papers) published 2010 in 53 astronomy and astrophysics journals (see Havemann et al., 2015 for a description of this data set). For each data set, we constructed and analysed the bibliographic coupling network.

Methods

For each data set, we calculated pairwise paper similarities as transformed Spearman's rank correlation of the papers' 'views' on the network. The 'view' of a paper p_i on the network is the vector of shortest paths between p_i and the papers p_1 to p_n of the network. Thus, the dissimilarity of two papers – their distance – is calculated as

$$dist\ \left(view(p_i i), view\left(p_j j\right)\right) = 1 - \frac{r_{sp}\left(view(p_i), view\left(p_j\right)\right) + 1}{2}$$

Where r_{sp} is the Spearman's rank correlation coefficient of the two views.

We tested this similarity measure on our information science data set by using it for a Ward clustering and comparing the best matching Ward clusters to three topics we had previously identified by inspecting titles and keywords of the articles.

We then calculated the distributions of paper similarities for country subsets and journal subsets of papers in both data sets, and used the median of the distributions as single-number value of the subset's diversity.

Our diversity measure also enables the construction of 'collective views', i.e. of 'views' of paper sets on each other. We exploited this opportunity in a third step and constructed similarities between countries and journals in information science.

Results

Information science

Our Ward clustering with the similarity measure led to results that compare well to previous experiments with other algorithms (Table 1).

Table 1. Salton's Cosine of precision and recall of pre-defined information science topics by five algorithms.[2]

Table	*MONC*	*HLC*	*FHC*	*RDDC*	*SPBC*
h-index	0.71	0.93	0.59	0.92	0.95
Bibliometrics	0.79	0.82	0.83	0.87	0.86
Webometrics	0.58	0.60	0.46	0.65	0.53

The three best performing algorithms – HLC, RDDC and SPBC – perform best for the h-index, good for bibliometrics including the h-index, and worst for Webometrics. These differences may be linked to the topics' internal diversity (Figure 1). Internal diversity is lowest for the h-index (all papers are very similar) and highest for webometrics (a high proportion of webometrics papers is not very similar). The differences in internal diversity may explain the differential success of algorithms in recapturing the topics.

[2] MONC= Merging overlapping natural communities, HLC=Hierarchical link clustering, FHC=Fuzzification of hard clusters (see Havemann et al., 2012). RDDC= Ward clustering with a similarity measure using the rank correlation of 'views' based on the resistance distance in direct citation networks (Gläser et al., 2013). SPPC= Ward clustering with a similarity measure using the rank correlation of 'views' based on the length of shortest paths in bibliographic coupling networks (algorithm presented in this paper). Among the three topics, bibliometrics also includes the h-index papers.

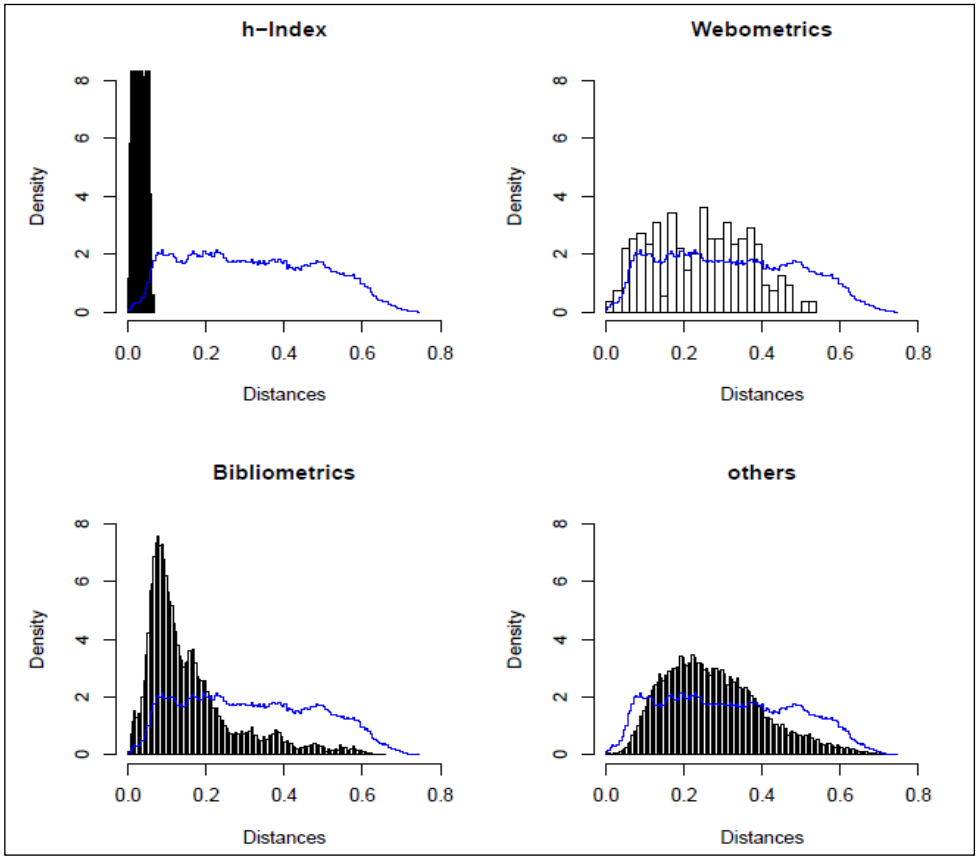


Figure 1. Internal diversity of three topics in the information science network (the blue lines represent the distribution for the whole network, the areas always equal one).

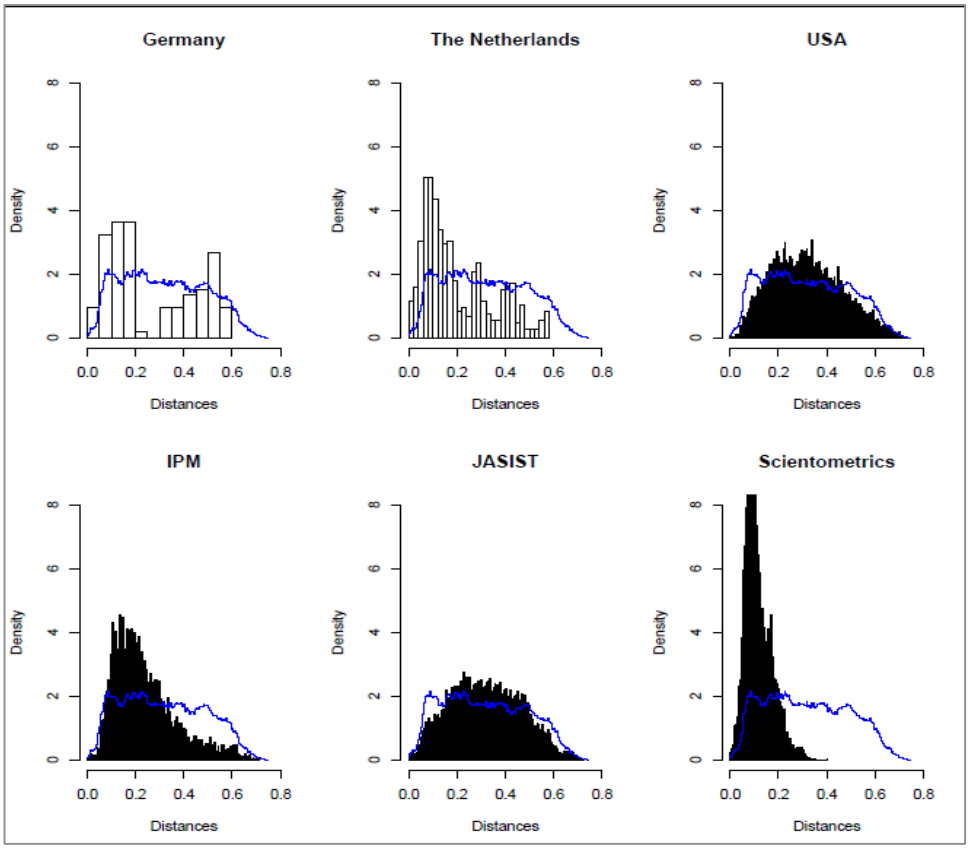


Figure 2. Diversity of information science publications from three countries and three journals.

Figure 2 shows the diversity of information science publications in three journals and of three countries. According to these distributions of distances,

a) Dutch information science publications are less diverse than the few German publications and the publications from the USA; and

b) *Scientometrics* was the least diverse (most focused) journal, followed by *JASIST* and *Information Processing and Management*.

Astronomy and astrophysics

The astronomy and astrophysics publication network is less diverse than the information science network. Taking the median as a single-number measure of diversity, the information science network (median = 0.32) is much more diverse than the astronomy and astrophysics network (median = 0.27). Owing to space limitations, we can provide only one comparison. Figure 3 compares the distribution of paper similarities for Chilean and US-American publications. Astronomy and astrophysics publications from Chile appear to be much less diverse (much more concentrated on one or few topics) than those from the USA.

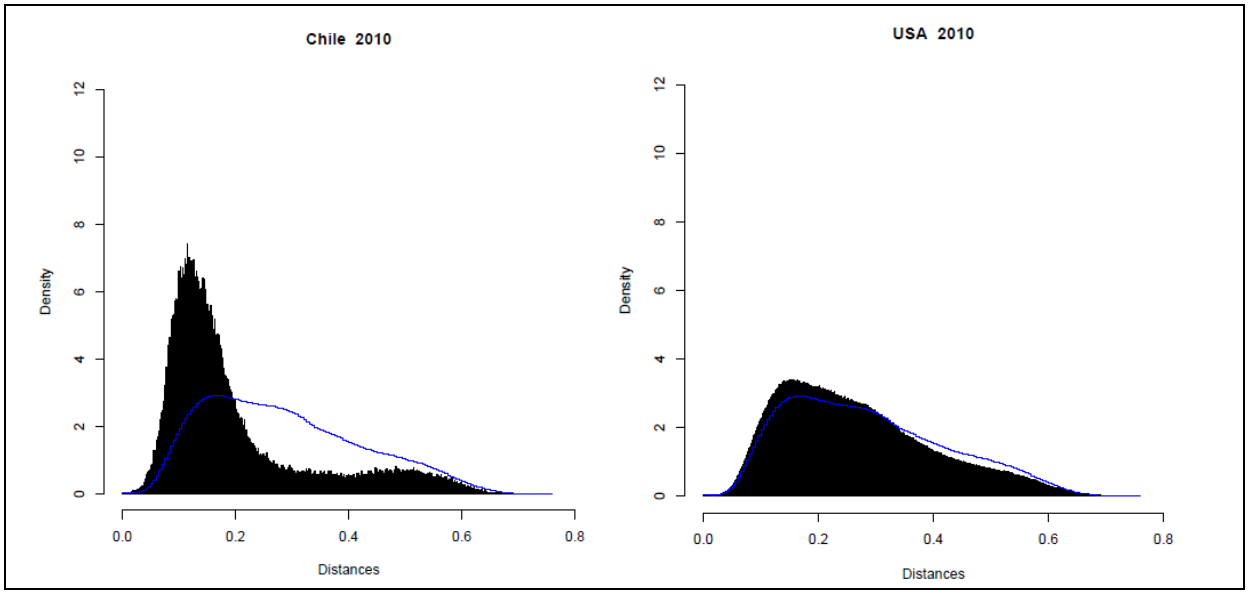


Figure 3. Diversity of astronomy and astrophysics publications from Chile and the USA (the blue lines represent the distribution for the whole network).

Discussion

A small but noxious problem for the application of our diversity measure is the occurrence of direct citations between publications from the same year. Direct citations can be considered a strong indicator of thematic similarity. However, it is not known how strong an indicator a direct citation is, and how it should be treated in comparison with bibliographic coupling of two publications. Our current solution is to add the citing and cited publication to each other's reference lists, i.e. integrating direct citation into bibliographic coupling. This solution is, however, as arbitrary as any other solution would be.

A more consequential limitation stems from our use of networks of papers as models of published knowledge. Adding a node with at least two links to a network indirectly changes connections between all nodes. This is not true for added knowledge, which can induce changes in similarities that remain local in that they affect only the knowledge to which it links directly. Although the length of the shortest path between two papers is not as sensitive to changes in networks as the measure we tried before (resistance distance), it remains to be seen whether time series of diversity constructed with our distance measure can be

interpreted. Since the literature in most fields keeps growing, time series of diversity have to cope with ever-growing paper networks.

Finally, a third limitation is inherent to our measure. Measuring the diversity of any set of papers with the approach suggested in this paper requires the set of papers to be embedded in a connected subgraph. If a research organisation has publications in many unrelated fields (as most universities do, providing an aggregate measure of the diversity of this organisations published output would be impossible. However, such an aggregate measure is likely to be meaningless in any case.

Conclusion

While further tests are of course necessary, the diversity measure proposed in this article appears to enable comparisons of paper sets from topics, journals, specialised organisations, or countries. The measure appears to use enough information to provide meaningful results without being sensitive to the noise created by network connections that have no bearing on the similarity of two papers. It is also compatible with sociological findings that ground the publication process in an author's personal experience and perspective. The 'view' of a paper on the network can easily be interpreted as the scientific perspective of its author.

Our discussion of diversity measures and their applicability to the epistemic diversity of published knowledge suggests two lines of further work. First, the problem of time series must be solved, i.e. the diversity of a field must be measured for networks of different sizes. This requires assessing the sensitivity of our diversity measure for changes in networks that are unrelated to epistemic diversity.

Second, a solution must be found for the measurement of diversity with a three-level approach. This is both theoretically and practically important because changes in the diversity of research are caused by the selective growth and shrinking of topics. Understanding the role of epistemic diversity for research requires causally attributing changes in the epistemic diversity to such processes of growth and decline, which in turn requires linking publications to topics. The obvious solution is making topics disjoint by fractionally assigning papers to overlapping topics. However, this does not solve all problems posed by thematic structures in science. Consider the following simple example: A paper on the h-index is simultaneously a paper in bibliometrics because the topic h-index is fully included in bibliometrics. How would one assign such a paper to the two topics?

Developing three-level measures for the diversity of overlapping topics might mean abandoning all established measures, and might prove a very challenging task.

References

Ahlgren, P. & Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document–document similarity approaches in the context of science mapping, *Scientometrics, 76*, 273-290.

Balcan, M.-F., C. Borgs, M. Braverman, J. Chayes & S.-H. Teng. (2012). Finding endogenously formed communities. arXiv:1201.4899v2.

Bordons, M., F. Morillo & I. Gómez (2004). Analysis of cross-disciplinary research through bibliometric tools. In: H. F. Moed , W. Glänzel & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research*, Dordrecht: Kluwer, 437-456.

Cozzens, S. E. (1985). Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science. *Social Studies of Science 15*(1), 127-153.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften: Die soziale Ordnung der Forschung*. Frankfurt am Main: Campus.

Gläser, J. G. Laudel, S. Hinze & L. Butler (2002). Impact of Evaluation-based Funding on the Production of Scientific Knowledge: What to Worry About, and how to Find Out. Report to the BMBF. Retrieved January 20, 2015, from http://www.laudel.info/wp-content/uploads/2013/pdf/research%20papers/02expertiseglaelauhinbut.pdf.

Gläser, J., F. Havemann, M. Heinz & A. Struck. (2013). Measuring the diversity of research using paper similarities. In: S. Hinze & A.Lottmann (Eds.), *Translational twists and turns: Science as a socio-economic endeavor. Proceedings of the 18th International Conference on Science and Technology Indicators, Berlin, Germany, September 4 — 6, 2013*, Berlin, 130-139.

Havemann, F., J. Gläser, M. Heinz & A. Struck (2012). Identifying overlapping and hierarchical thematic structures in networks of scholarly papers: A comparison of three approaches. *PLoS ONE, 7*(3), e33255.

Havemann, F., Gläser, J. & Heinz, M. (2015). A link-based memetic algorithm for reconstructing overlapping topics from networks of papers and their cited sources. *15th International Conference on Scientometrics and Informetrics*, Istanbul, 29 June -3 July 2015.

Hochschulrektorenkonferenz (HRK) (2007). Die Zukunft der kleinen Fächer: Potenziale – Herausforderungen – Perspektiven, Bonn: Hochschulrektorenkonferenz. http://ipts.jrc.ec.europa.eu/home/report/english/articles/vol66/ITP1E666.html.

Laudel, G. & E. Weyer. (2014). Where have all the Scientists Gone? Building Research Profiles at Dutch Universities and its Consequences for Research. In: Richard Whitley and Jochen Gläser (Eds.), *Organizational Transformation And Scientific Change: The Impact Of Institutional Restructuring On Universities And Intellectual Innovation*, Bingley, UK: Emerald Group Publishing Limited, 111-140.

Molas-Gallart, J. & A. Salter (2002). Diversity and Excellence: Considerations on Research Policy. IPTS Report.

Olenin, S. & J.-P. Ducrotoy (2006). The concept of biotope in marine ecology and coastal management. *Marine Pollution Bulletin 53*, 20–29.

Pinch, T. (1985). Towards an Analysis of Scientific Observation: The Externality and Evidential Significance of Observational Reports in Physics. *Social Studies of Science 15,* 3-36.

Rafols, I., L. Leydesdorff, A. O'Hare, P. Nightingale & A. Stirling (2012). How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management. *Research Policy, 41*, 1262-1282.

Rafols, I. & M. Meyer (2007). How cross-disciplinary is bionanotechnology? Explorations in the specialty of molecular motors. *Scientometrics, 70*(3), 633-650.

Stirling, A. (2007). A general framework for analyzing diversity in science, technology and society. *Journal of The Royal Society Interface, 4*, 707–719.

Swales, J. (1986). Citation analysis and discourse analysis. *Applied Linguistics, 7*, 39-56.

Van Raan, A. F. J. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics,* 47, 347-362.

Whitley, R. (1974). Cognitive and social institutionalization of scientific specialties and research areas. In: R. Whitley (ed), *Social Processes of Scientific Development* (pp.69–95). London.

Using Bibliometrics-aided Retrieval to Delineate the Field of Cardiovascular Research

Diane Gal[1], Karin Sipido[1] and Wolfgang Glänzel[2]

{diane.gal, karin.sipido}@med.kuleuven.be , wolfgang.glanzel@kuleuven.be
[1]Department of Cardiovascular Sciences, KULeuven, 3000 Leuven (Belgium)
[2]ECOOM and Dept. MSI, KU Leuven, 3000 Leuven (Belgium) & Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
A hybrid search strategy, using lexical and citation based methods, is presented in this paper as a robust method to delineate the broad field of cardiovascular research. Overall, this study aims to provide scientifically reliable and accurate data driven evidence about cardiovascular research by establishing a dataset of published research in this field. A workflow is presented that outlines the methods carried out to establish a core dataset based on a core set of journals, to identify and use search terms to detect a broader dataset, and then to apply measures of similarities between the citations of these two datasets to ensure relevance of the final dataset. The final core set of journals established comprises of 120 unique journals covered in Thomson Reuters *Web of Science Core Collection* (WoS) database including a total of 320,647 documents from 1991 to 2013. The search terms utilised include 107 cardio-specific terms that initially identify 1.8 million unique documents when searching the title, abstract and keywords. Upon application of the citation-based similarity measures the final combined dataset consists of 845,071 publications. Overall, establishing a relevant dataset of cardiovascular research means placing a greater emphasis on having a precise dataset, reducing recall in the process.

Conference Topic
Methods and techniques

Introduction
Experts in the cardiovascular field are concerned that there is a decline in quality and innovation in cardiovascular research and that fragmentation of this broad field is leading to loss of cross-pollination and missed opportunities for translation of research from bench to bedside. In this context we have launched a project to examine cardiovascular research output over a 23 year period to provide rigorous and reliable scientific information about cardiovascular research activities. The findings of this project are expected to serve as a complement to expert opinion and previously published studies (Huffman et al., 2013; Jones, Cambrosio, & Mogoutov, 2011; Sipido et al., 2009; van Eck, Waltman, van Raan, Klautz, & Peul, 2013; Yu, Shao, He, & Duan, 2013), to provide scientifically reliable and accurate data driven evidence about cardiovascular research.
The objectives of the project are to:

- Characterise the size, growth, topics and visibility of research outputs over 23 years;
- Analyse the geographical distribution of research outputs and its evolution;
- Visualise and analyse research collaboration; and
- Identify emerging topics in cardiovascular research.

To gain a comprehensive view of research in this field a broad scope and definition has been applied to include papers published in scientific journals from basic, clinical and epidemiological studies related to the cardiovascular system, including the heart, the blood vessels and/or the pericardium. The main source of data is the *Web of Science Core Collection*. The purpose of this paper is to describe the methods utilised, and the roadmap set, to establish a dataset of published research undertaken in the cardiovascular field.

Methods

Hybrid search strategies for subject delineation, previously described and published (Bolaños-Pizarro, Thijs, & Glänzel, 2010; Glänzel, Janssens, & Thijs, 2009; Zitt & Bassecoulard, 2006), have been adapted to establish a dataset of cardiovascular research. This includes (1) establishing a core dataset based on a core set of journals and core search terms, (2) identifying a broader dataset of publications through the use of search terms, and then, (3) applying measures of similarities by citations between the documents in these datasets to select a final dataset with acceptable precision and recall. A workflow/roadmap was developed to outline the main steps taken to establish the dataset, as can be seen in Figure 1.

Core Journal Dataset

All data have been retrieved from Thomson Reuters Web of Science Core Collection. The core set of journals was selected through expert review of the scope/aim of all 183 journals included in the 'Cardiac & Cardiovascular Systems' and the 'Peripheral Vascular Disease' Web of Science Categories. The scope/aim for each journal was obtained through online web-based searches. Using an online survey tool, two experts reviewed the title and scope/aim of each journal to assess the relevance of the journal and indicate whether they had experience with each journal (e.g. reading, editing, reviewing, submitting a document for publication). Journals that were assessed by at least 1 expert as being a core cardiovascular journal – defined as a journal publishing greater than 90% of its articles, reviews, letters and notes on the cardiovascular domain – were included in the core journal dataset. Disagreements between the experts were reviewed by the project team. Journals were excluded from the core dataset only when the expert excluding the journal was the only one that had previous experience with the journal. The final dataset was obtained by identifying all articles, letters, notes and reviews published journals that are covered in the 1991–2013 volumes of the WoS database.

Search Terms Datasets

A number of sources were reviewed to identify relevant cardiovascular-specific search terms, including:

- Medical Subject Headings (MeSH)
- International Classification of Diseases (ICD)-10
- Cochrane Hypertension/Heart/Peripheral Vascular Disease Groups/Systematic Reviews
- Cardioscape project taxonomy (European Society of Cardiology, 2014)
- Recent published research (Bolaños-Pizarro et al., 2010; Huffman et al., 2013; Jones et al., 2011; van Eck et al., 2013)

Subsequently, a group of eight topic experts representing a mix of clinical scientists, basic scientists and epidemiologists were invited to review the combined list of 105 search terms to assess their relevance in identifying as broad a range of cardiovascular research publications as possible. All search terms were included where at least half of the reviewers agreed that they were relevant search terms to include in the search strategy.

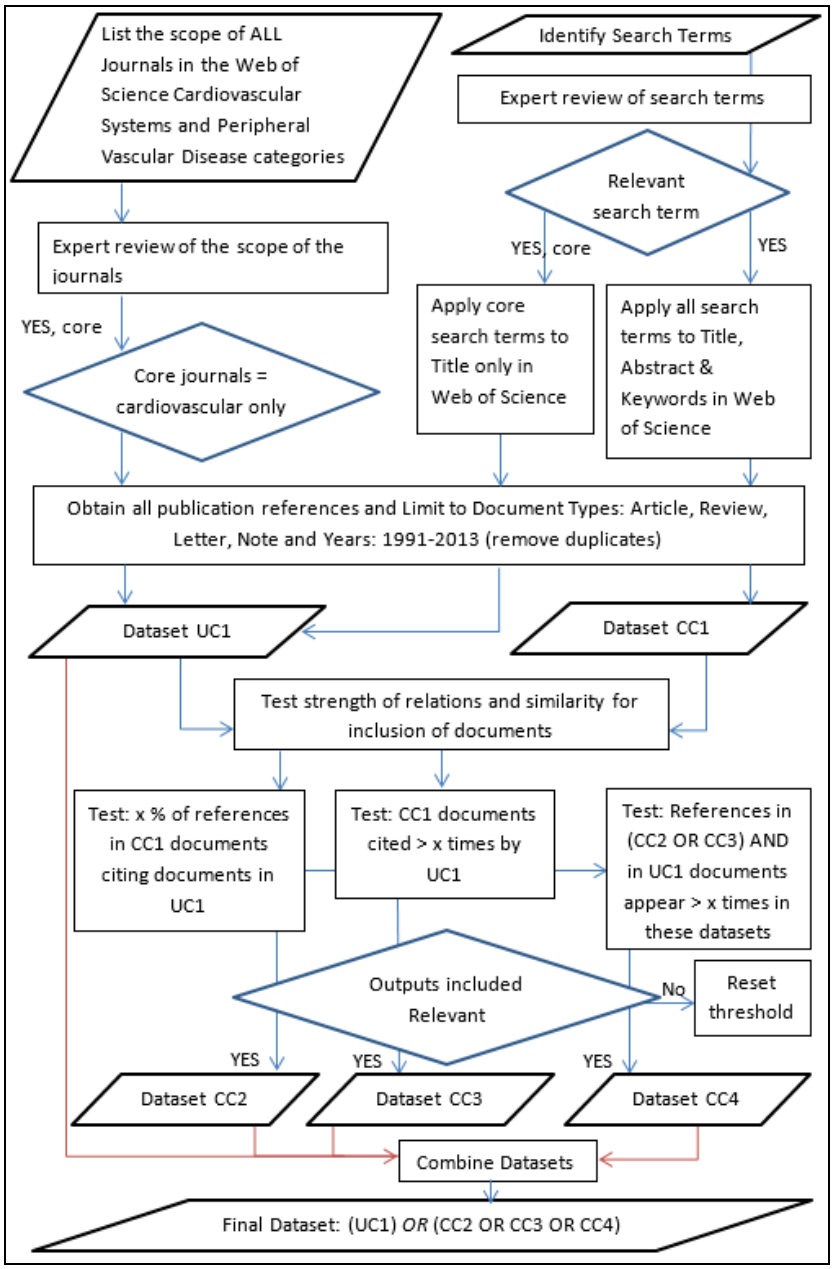


Figure 1. Workflow of field delineation of Cardiovascular Research

In addition, experts were asked to suggest any potentially missing search terms. New search terms suggested and disagreements were reviewed by the project team. The broad search terms dataset was obtained by applying the full search strategy to the complete Web of Science database, to identify all articles, letters, notes and reviews published between 1991 and 2013. To add to the core journal dataset, highly cardiovascular specific or core search terms were selected that when searched in the title would identify core cardiovascular publications.

Similarity Measures and Thresholds

For the extension of the core dataset, i.e., the seed of relevant literature, we followed an algorithm using a logical combination of *unconditional* and *conditional* criteria (Glänzel, 2014). In the present project we have linked literature retrieved based on conditional criteria (the broad search terms set) to the set of surely relevant documents (the core journals and core search terms set), using citation-based similarities. In particular, three measures of similarity

between the core dataset and the broad search terms dataset were utilised: a) the share of references of broad search terms documents that cite the core documents, b) the number of references of the core documents that cite the broad search terms documents and c) the number of shared references between the core dataset and the restricted search terms dataset. The thresholds for each measure were set following iterative testing, whereby a low threshold was first applied and a random sample of the titles and abstracts of 500 documents was reviewed for relevance to the cardiovascular field. The threshold was altered until the sample contained a high precision and the level of noise (peripheral and irrelevant documents) was reduced to an acceptable level, defined as a 5% level of noise. To confirm the relevance of the documents identified, the random samples considered to have acceptable thresholds were reviewed by one topic expert.

Findings

Core Dataset

After expert review, 120 journals were included as core journals. The two expert reviewers agreed on the exclusion of 61 journals and disagreed on the inclusion of 39 journals (21% of all 183 journals), of these only two journals were excluded as the expert who had experience with the journal was the one that excluded it. For the remaining 37 journals, they were included since both experts had previous experience for three journals and neither expert had experience for 34 journals. The final core journal documents therefore consist of 320,647 articles, letters, notes and reviews from 1991 to 2013. Thirteen of the search terms, identified below, were considered to be highly cardiovascular specific. The core search terms when searched only in the title, added 141,676 documents to the core journal documents, resulting in a core dataset of 462,323 documents. Review of this dataset confirmed that it provides a precise sample of cardiovascular-specific documents for this study.

Broad Search Terms Dataset

After expert review by 6 topic experts and the project team, 107 search terms were included in the final search strategy. Of the original 105 terms reviewed, three search terms were removed since more than half of the experts suggesting to remove them. A total of 22 unique terms were also suggested by three of the topic experts. The project team assessed and included four of these new terms. Then one additional term was added to the search strategy to include this term with and without its common prefix. The final broad search terms dataset consists of 1,656,278 unique articles, letters, notes and reviews from 1991 to 2013 where the search terms could be identified in the abstract, keywords or title. All documents in the core dataset were removed from this broad search term dataset.

A comparison of all documents obtained by searching the abstract, keywords and title is presented in Figure 2.

As a validation of the search strategy and selection of core journals, when the search strategy was applied to the 120 core journals, 95% of all core journal dataset documents were identified by the search terms.

Similarity Measures and Thresholds

An initial test was undertaken to limit the search terms dataset by removing all documents that had no links with the core journal documents. A total of 228,000 documents had no links meaning they did not cite the core journal set, they were not cited by the core journal set *and* they did not have any common references with the core journal set. This reduced the search terms set to less than 1.6 million documents, however upon review of random samples it was

clear that stronger measures of similarity would be needed to further restrict the search terms dataset to include the most relevant documents in the final dataset.

Iterative testing and review of random samples led to the selection of a combined dataset where at least 12% of the references in the broad search documents cited documents in the core dataset or where the broad search documents where cited greater than 4 times by the core documents. For this chosen dataset, no more than 10% of the random samples were considered not relevant or peripheral to the cardiovascular field. Documents from the third measure of similarity using bibliographic coupling was not included in the final dataset since it was not possible to achieve less than a 10% noise level through iterative testing and review of random samples. The final restricted broad search terms dataset consists of 382,748 unique articles, letters, notes and reviews from 1991 to 2013.

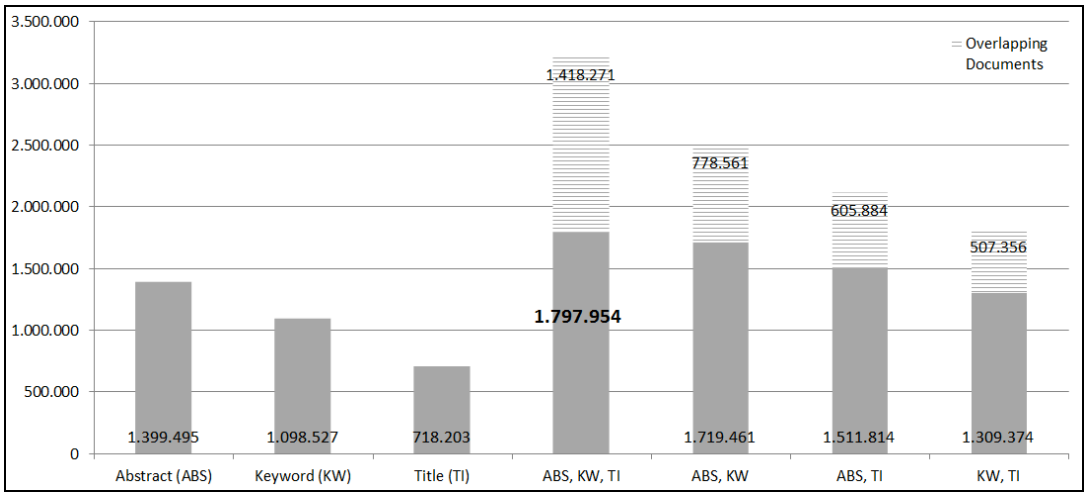


Figure 2. Number of documents identified when searching 107 search terms in Abstracts, Keywords and Titles [Data sourced from Thomson Reuters Web of Science Core Collection].

Final Combined Dataset

Combined, the core and restricted datasets create a final dataset of 845,071 unique documents from the cardiovascular field. Overall, the combined dataset has a 4.5% noise level (estimated).

Discussion

Only one previously published bibliometric study of cardiovascular research used a hybrid search strategy to establish its dataset (Bolaños-Pizarro et al., 2010). However, due to the broad scope of this study, which aims to include all types of research – from basic to clinical research, a broader list of cardio-specific search terms was created. Attention was also placed on ensuring that the search terms selected could identify cardiovascular research over the long time period of the study, as well as, enable the identification of new and emerging fields in cardiovascular research. The 107 search terms greatly increases the recall of documents, though this also means that a greater amount of noise was present in the broad search terms dataset. Hence, the importance of utilising measures of similarity between the two datasets to restrict the broad search terms dataset to include only the most relevant documents. This was done through testing various thresholds of citation-based similarities, as the final step of this robust method to delineate complex fields of research. Including both directions of citation-based similarities (ie. documents from core journals dataset citing documents in search terms dataset and vice versa) also ensures that the distribution of documents sampled is representative over time. The initial threshold of 5% noise was re-evaluated through testing

and due to the broad nature of the cardiovascular field a higher level of noise (10%) was considered acceptable as this includes peripheral research that has a component linked to cardiovascular research. The broad search terms dataset has been reduced to less than a quarter of initial documents identified to ensure the final dataset is as precise as possible and can be considered a representative sample of cardiovascular research over the 23 year period.

Conclusions

Bibliometrics-aided retrieval is a robust method to delineate the field of cardiovascular research. Through using this method, a representative dataset of cardiovascular research was established irrespective of changes in the field, such as vocabulary used, over the time-frame of this study. Overall, establishing a relevant dataset of cardiovascular research means placing a greater emphasis on having a precise dataset, reducing recall in the process.

Acknowledgments

Thank you to Bart Thijs for his input into the study methods.

References

Bolaños-Pizarro, M., Thijs, B., & Glänzel, W. (2010). Cardiovascular research in Spain. A comparative scientometric study. *Scientometrics*, *85*(2), 509–526. doi:10.1007/s11192-009-0155-2

European Society of Cardiology. (2014). *CardioScape: A survey of the European cardiovascular research landscape* (p. 52). Retrieved June 2, 2015 from: http://www.cardioscape.eu/static_file/CardioScape/PNO%20report/CardioScape_Summary%20Report_30092014.pdf

Glänzel, W. (2014). Bibliometrics-aided retrieval – where information retrieval meets scientometrics. *Scientometrics*. doi:10.1007/s11192-014-1480-7

Glänzel, W., Janssens, F., & Thijs, B. (2009). A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. *Scientometrics*, *79*(1), 109–129. doi:10.1007/s11192-009-0407-1

Huffman, M. D., Baldridge, A., Bloomfield, G. S., Colantonio, L. D., Prabhakaran, P., Ajay, V.S., Lewison, G., & Prabhakaran, D. (2013). Global cardiovascular research output, citations, and collaborations: A time-trend, bibliometric analysis (1999–2008). *PLoS ONE*, *8*(12), e83440. doi:10.1371/journal.pone.0083440

Jones, D. S., Cambrosio, A., & Mogoutov, A. (2011). Detection and characterization of translational research in cancer and cardiovascular medicine. *Journal of Translational Medicine*, *9*(1), 57. doi:10.1186/1479-5876-9-57

Sipido, K. R., Tedgui, A., Kristensen, S. D., Pasterkamp, G., Schunkert, H., Wehling, M., Dambrauskaite, V. (2009). Identifying needs and opportunities for advancing translational research in cardiovascular disease. *Cardiovascular Research*, *83*(3), 425–435. doi:10.1093/cvr/cvp165

Van Eck, N. J., Waltman, L., van Raan, A. F. J., Klautz, R. J. M., & Peul, W. C. (2013). Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. *PLOS ONE*, *8*(4). doi:10.1371/journal.pone.0062395

Yu, Q., Shao, H., He, P., & Duan, Z. (2013). World scientific collaboration in coronary heart disease research. *International Journal of Cardiology*, *167*(3), 631–639. doi:10.1016/j.ijcard.2012.09.134

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing & Management*, *42*(6), 1513–1531. doi:10.1016/j.ipm.2006.03.016

Locating an Astronomy and Astrophysics Publication Set in a Map of the Full Scopus Database

Kevin W. Boyack[1]

[1] *kboyack@mapofscience.com*
SciTech Strategies, Inc., 8421 Manuel Cia Pl NE, Albuquerque, NM 87122 (USA)

Abstract
A dataset containing 111,616 documents in astronomy and astrophysics has been created and is being partitioned by several research groups using different algorithms. In this paper, rather than partitioning the dataset directly, we locate the data in a previously created model in which the full Scopus database was partitioned. Given that the other research groups are partitioning the data directly, use of this method will allow comparisons between using local and global data for community detection. In other words, use of this method will allow us to start to answer the question of how much the rest of a large database affects the partitioning of a journal-based set of documents. We find that the astronomy document set, while spread across hundreds of partitions in the Scopus map, is located in only a few regions of the map. Thus, the use of a global map to partition astronomy documents is likely to give very similar results to partitioning using local approaches because of the insularity of the field of astronomy. However, we do not expect local and global data to give as similar results for other fields, because most other fields are less insular than astronomy.

Conference Topic
Methods and techniques

Introduction
Partitioning of a dataset into groups of similar objects – alternatively known as clustering, community detection or topic detection – is a current research topic in a number of fields, including scientometrics and network science. A number of different algorithms are used to partition scientific literature into topics or clusters. While many of these are based on the property of modularity (cf., Blondel, Guillaume, Lambiotte, & Lefebvre, 2008; Newman & Girvan, 2004; Waltman & van Eck, 2013), others are based on graph layout and pruning (Martin, Brown, Klavans, & Boyack, 2011) or on complex network flows (Rosvall & Bergstrom, 2008). Despite the obvious differences between these algorithms, they are all based on a common principle – that of dividing a literature set into partitions where the within-partition signals are much stronger or denser than the between-partition signals.

It is well known that different topic detection algorithms give somewhat different results for the same data set. What is not known is the specifics of why particular algorithms give particular results, or exactly what operations of a particular algorithm lead it to give different results than those obtained by another algorithm. In general, we know very little about what types of features result from different algorithms, and how these affect the output structures. This can make it difficult to interpret the partitions and maps that are produced by an algorithm. Are the partitions produced by an algorithm representative of actual structures in science, are they merely artifacts resulting from the algorithm and its parameters, or are they something in between? This is a difficult question to which, we suspect, the answer is far beyond the scope of even a large study. Nevertheless, we are hopeful that a comparison of partitioning methods and their results using a single dataset might lead to some general understanding of the types of features that result from different methods and algorithms. This type of understanding has the potential to enable both researchers and decision makers to more clearly understand and interpret the results of a particular partitioning.

To this end, a number of researchers (see papers from this special session) have come together to explore this question. As a first step, each research group has created a partitioning of a

single dataset using their own algorithms. The work-in-progress papers in this session describe the partitioning algorithms and results from each group. The multiple results will be combined and compared in a next phase of the project to determine similarities and differences in the features resulting from the different methods and algorithms. Beyond that, we collectively hope to learn more about both common and unique structural features that result from the different algorithms.

This paper details the method used by SciTech Strategies to partition an "astronomy and astrophysics" literature dataset. It differs from the other methods in one significant aspect – the other groups have all created local solutions (partitioning the dataset directly), while we have created a global model (partitioning the entire Scopus database) and have located the astronomy dataset within those partitions (Klavans & Boyack, 2011). Use of this method enables us to start to answer the question of how much the rest of the database affects the partitioning process.

Global Model

Our global model of science consists of 48,533,301 documents from Scopus. Of these, 24,615,844 documents are indexed source documents from Scopus 1996-2012, while the remaining 23,917,457 are non-source documents that were each cited at least twice by the set of source documents. The method used to generate the document set and citing-cited pairs list is very similar to that used for the recent "non-source" map of Boyack and Klavans (2014).

The model was created by taking the over 582 million citing-cited pairs within this set of 48.5 million documents, calculating similarity values between pairs of documents based on direct citation, and then partitioning the documents using the new CWTS smart local moving algorithm (Waltman & van Eck, 2013). The citing-cited pairs were provided by SciTech Strategies (STS) to Ludo Waltman at CWTS, who ran the similarity calculation and partitioning steps. The CWTS smart local moving algorithm was used to create a four-level hierarchical solution, with resolution values chosen to result in a solution with roughly 100k, 10k, 1000, and 100 clusters. Details of the partitioning are given in Table 1.

Table 1. Multi-level partitioning of the Scopus database using the CWTS smart local moving algorithm.

Level	*Partitions Desired*	*Resolution*	*Partition Min Size*	*# Partitions*	*Partitions > Min Size*	*# Pubs*	*% Pubs Lost*
1	100000	3e-5	50	114679	91726	48399235	0.28%
2	10000	3e-6	500	13157	10059	47323189	2.49%
3	1000	3e-7	5000	1048	849	46929303	3.30%
4	100	5e-8	50000	122	114	46705047	3.77%

Visual maps of the partition solutions at level 1 and level 2 were created using the following process. At each level, 1) pairwise similarity between partitions was calculated from the titles and abstracts of the documents in each partition using the BM25 textual similarity measure, 2) each resulting similarity list was filtered to retain the top-n (5-15) similarities per partition, and 3) layout of the partitions on the x,y plane was done using the DrL algorithm. These steps are ones we commonly use to create science maps, and are described in more detail in Boyack & Klavans (2014). In each case, only those partitions that were of the minimum size desired (91,726 for level 1, and 10,059 for level 2) were included in the map. Field counts for each cluster in each map were calculated using UCSD map of science journal-to-field assignments (Börner et al., 2012), and each cluster was assigned to its dominant field and correspondingly colored in the map. The two maps are similar in that they show that the 12 large fields (see

legend at the bottom of Figure 1) occupy similar positions in both maps. The change in granularity of the partitions does not change the overall look and feel of the map.

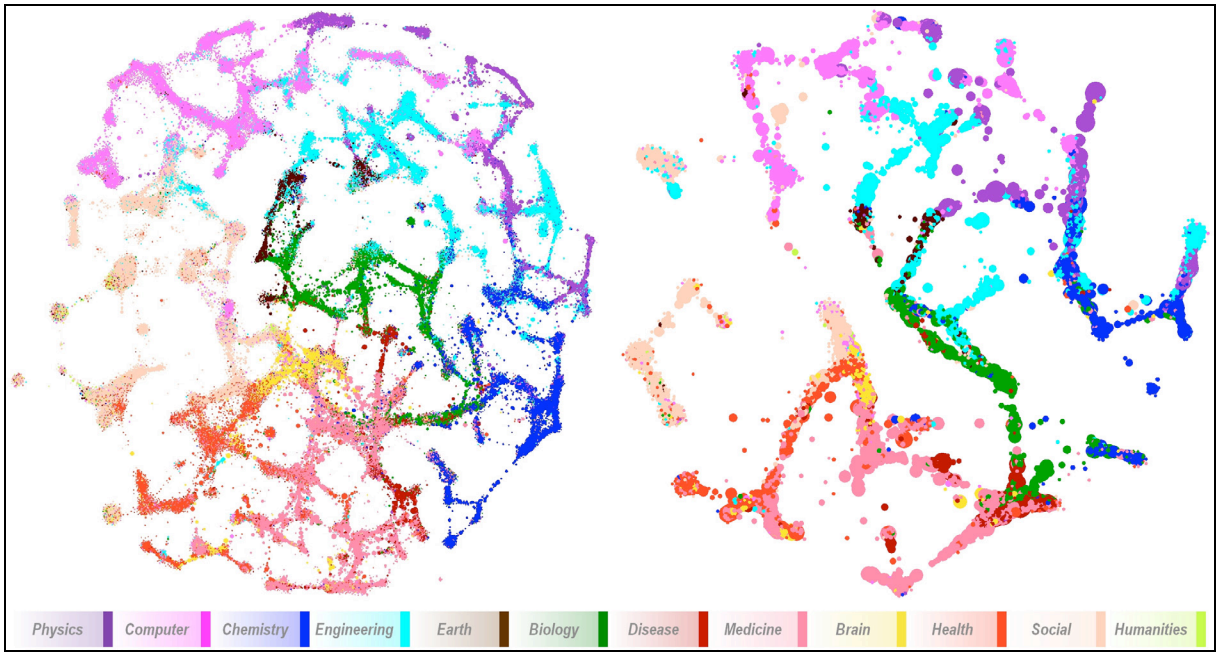


Figure 1. Visual maps of the Scopus database using level 1 (left) and level 2 (right) partitions.

Astronomy Dataset

The astronomy dataset used by each research group consists of 111,616 document records with accompanying data from the Web of Science. This dataset was created by researchers at Humboldt University for use by project participants, and is comprised of documents published from 2003-2010 in a set of 59 astronomy and astrophysical journals, limited to articles, letters, and proceedings papers. Over half of the documents were from four journals, as shown in Table 2.

Table 2. Dominant journals in the astronomy and astrophysics dataset.

Journal	Count
Astrophysical Journal	19582
Physical Review D	19061
Astronomy & Astrophysics	14668
Monthly Notices of the Royal Astronomical Society	11599

In order to use the Scopus-based global model and map, Scopus identifiers for the WoS records were identified to the extent possible by matching source data (journal, title, volume, page, year). Definitive matches were obtained for 107,888 (96.66%) of the documents. Of the 3,728 documents that were not matched, roughly half were in source titles that are not covered by Scopus (such as the IAU Symposium), and thus could only be matched if they were cited non-source materials. The remaining unmatched records were in source titles that are covered by Scopus, but that we could not match. This lack of uniformity between databases is primarily due to differences in the way titles are listed (particularly for non-ASCII characters) and missing records. Despite the unmatched records, we consider a match rate of nearly 96.7% to be very good, and certainly sufficient for reasonable comparison with the partitions from other groups. Once the matching was done, documents from the astronomy dataset were located in global map at three levels (1, 2, and 3 from Table 1).

Astronomy is known to be a relatively insular discipline, with fewer links (percentage basis) to and from other disciplines than for most other disciplines. Thus, we expected the effect of including an additional 48 million documents in a cluster solution to have only a modest effect on the partitioning of the astronomy document set. We did not expect the astronomy documents to be scattered throughout the map. As expected, the astronomy documents are heavily concentrated in the global model. At level 1, 50% of the astronomy documents are in partitions where the astronomy set documents comprise at least 66.5% of the partition contents (limited to the years of study, 2003-2010). In other words, when sorting partitions by concentration of the astronomy document set within the partition, 50% of the total papers are accounted for in partitions with a concentration of over 66.5%. Using an alternative measure, when partitions are sorted by the number of papers from the astronomy document set, the number of non-set papers equals the number of set papers only when 90,000 of the 111,616 papers are accounted for, as shown in Figure 2.

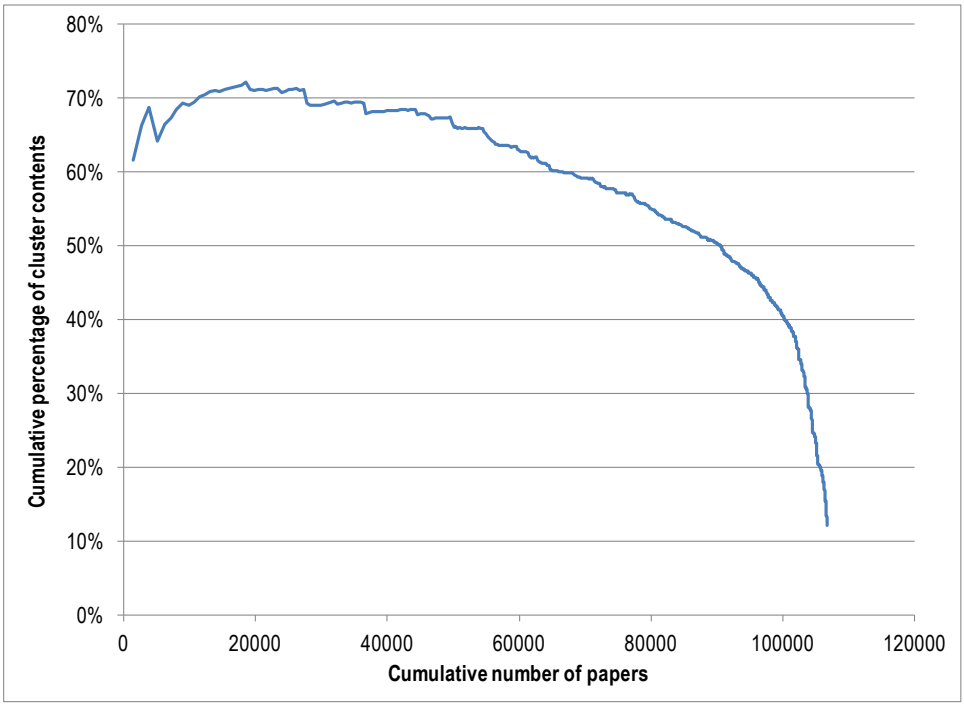


Figure 2. Distribution of the astronomy dataset across partitions in the level 1 solution.

Overlays showing the positions of the partitions with at least 50 documents from the astronomy set are shown for both the level 1 and level 2 maps in Figure 3. For level 1, this comprises 408 partitions and 90,763 documents (84.1% of the matched documents), while for level 2 it comprises 119 partitions and 101,895 documents (94.4% of the matched documents). Both maps make it clear that while the documents are parsed out into hundreds of partitions, each representing distinct topics, these topics are concentrated in only a few areas in the map.

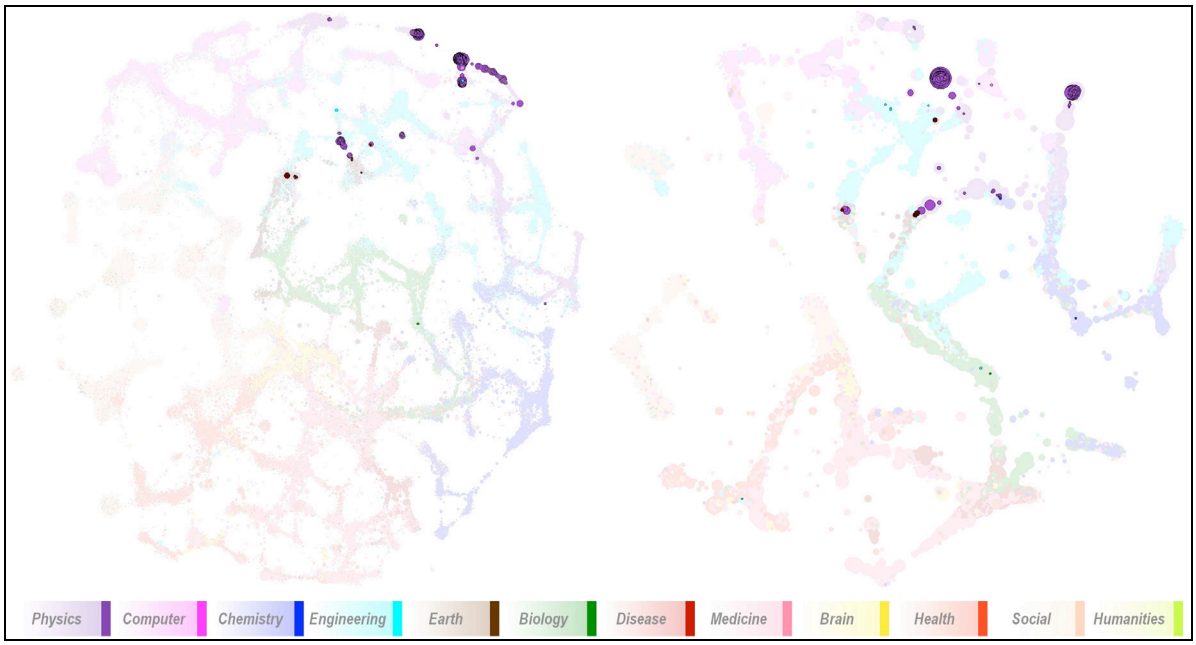


Figure 3. Overlays of the positions of the astronomy set documents on the Scopus level 1 (left) and level 2 (right) maps of Figure 1.

Discussion

Recalling that the astronomy document set was based on a set of journals, the high level of concentration of the overlays shown in Figure 3 suggests that use of journals is a very reasonable strategy for building a dataset in the field of astronomy. Astronomy journals have a very tight profile on a document-based map. By contrast, high profile journals in other fields, such as JACS, Physical Review Letters, and New England Journal of Medicine, have very broad profiles, and overlays for these journals (not shown here) spread across large regions of the map. Thus, while a dataset based on journals is useful to characterize astronomy, journals may be far less useful for characterizing other fields. Correspondingly, the use of a global map to partition astronomy documents is likely to give very similar results to partitioning using local approaches because of the insularity of the field of astronomy. We would not expect the use of a global map to partition a local document set from another field to work as well. Or, rather, we would expect the journal-based approach to fall short in other fields because it would leave out so much of the relevant contextual literature.

References

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 10*, P10008.

Boyack, K.W., & Klavans, R. (2014). Including non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics, 8*, 569-580.

Börner, K., Klavans, R., Patek, M., Zoss, A.M., Biberstine, J.R., Light, R.P., Larivière, V., & Boyack, K.W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464.

Klavans, R., & Boyack, K.W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 1-18.

Martin, S., Brown, W.M., Klavans, R., & Boyack, K.W. (2011). OpenOrd: An open-source toolbox for large graph layout. *Proceedings of SPIE - The International Society for Optical Engineering, 7868*, 786806.

Newman, M.E.J. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 026113.

Rosvall, M. & Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the USA, 105*(4), 1118-1123.

Waltman, L. & van Eck, N.J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B, 86*, 471.

Scientific Workflows for Bibliometrics

Arzu Tugce Guler[1], Cathelijn J. F. Waaijer[2] and Magnus Palmblad[1]

[1]*a.t.guler@lumc.nl, n.m.palmblad@lumc.nl*
Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden (The Netherlands)

[2]*c.j.f.waaijer@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Faculty of Social and Behavioural Sciences, Leiden University, Leiden (The Netherlands)

Abstract
Scientific workflows organize the assembly of specialized software into an overall data flow and are particularly well suited for multi-step analyses using different types of software tools. They are also favourable in terms of reusability, as previously designed workflows could be made publicly available through the myExperiment community and then used in other workflows. We here illustrate how scientific workflows and the Taverna workbench in particular can be used in bibliometrics. We discuss the specific capabilities of Taverna that makes this software a powerful tool in this field, such as automated data import via communication with Web services, smooth data extraction from XML by XPath and various data analyses and visualizations with the statistical language R. The support of the latter allows integration of a number of recently developed R packages for bibliometric analysis. A number of simple examples illustrate the possibilities of Taverna in the field of bibliometrics and scientometrics.

Conference Topic
Methods and techniques

Introduction
Information processing permeates the scientific enterprise, generating and organizing knowledge about nature and the universe. In the modern era, computational technology enables us to automate data handling, reducing the need for human labor in information processing. Often information is processed in several discrete steps, each building on previous ones and utilizing different tools. Manual orchestration is then frequently required to connect the processing steps and enable a continuous data flow. An alternative solution would be to define interfaces for the transition between processing layers. However, these interfaces then need to be designed specifically for each pair of steps, depending on the software tools they use; which compromises reusability. Whether the data flow is automated or done by the researcher manually, the latter still has to deal with many low-level aspects of the execution process (Gil, 2008).

Scientific workflow managers connect processing units through data and control connections and simplify the assembly of specialized software tools into an overall data flow. They smoothly render stepwise analysis protocols in a computational environment designed for the purpose. Moreover, the implemented protocols are reusable. Existing workflows can be shared and used by other workflows, or they can be modified to solve different problems. Several general purpose scientific workflow managers are freely available, and a few more optimized for specific scientific fields (De Bruin, Deelder, & Palmblad, 2012). Most of these managers provide visualization tools and have a graphical user interface, e.g. KNIME (Berthold et al., 2007), Galaxy (Goecks, Nekrutenko, & Taylor, 2010) and Taverna (Oinn et al., 2004). Not surprisingly, scientific workflows are now becoming increasingly popular in data intensive fields such as astronomy and biology.

In this paper, we describe the use of scientific workflows in bibliometrics using the *Taverna Workbench*. Taverna Workbench is an open source scientific workflow manager, created by

the myGrid (Stevens, Robinson, & Goble, 2003) project, and now being used in different fields of science. Taverna provides integration of many types of components such as communication with Web Services (WSDL, SOAP, etc.), data import and extraction (XPath for XML, spreadsheet import from tabular data), and data processing with Java-like Beanshell scripts or the statistical language R (Wolstencroft et al. 2013). Beanshell services allow the user to either program a small utility from scratch and towards a specific goal, or to integrate already existing software in the workflow. The R support is a particularly powerful feature of Taverna. Although R was initially developed as a language for statistical analysis, its widespread use has seen it adopted for many tasks not originally envisioned—a fate not unlike its commercial cousin, MATLAB. One such task is text mining. The R package *tm* (Feinerer, Hornik, & Meyer, 2008) provides basic text mining functionality and is used by a rapidly growing number of higher-level packages, such as *RTextTools* (Jurka, Collingwood, Boydstun, Grossman & van Atteveldt, 2014), *topicmodels* (Grün & Hornik, 2011) and *wordcloud* (Fellows, 2013). Similarly, there are many toolkits and frameworks for text mining in Java that could also be called from within a Taverna workflow.

An Example Workflow

We designed a simple workflow, *compare_two_authors* (see below), to generate a histogram for the number of publications over time and a co-word map for the titles of the two authors' publications. The workflow takes as inputs PubMed results in XML, the names of two authors, a list of excluded words and a minimum number of occurrences.

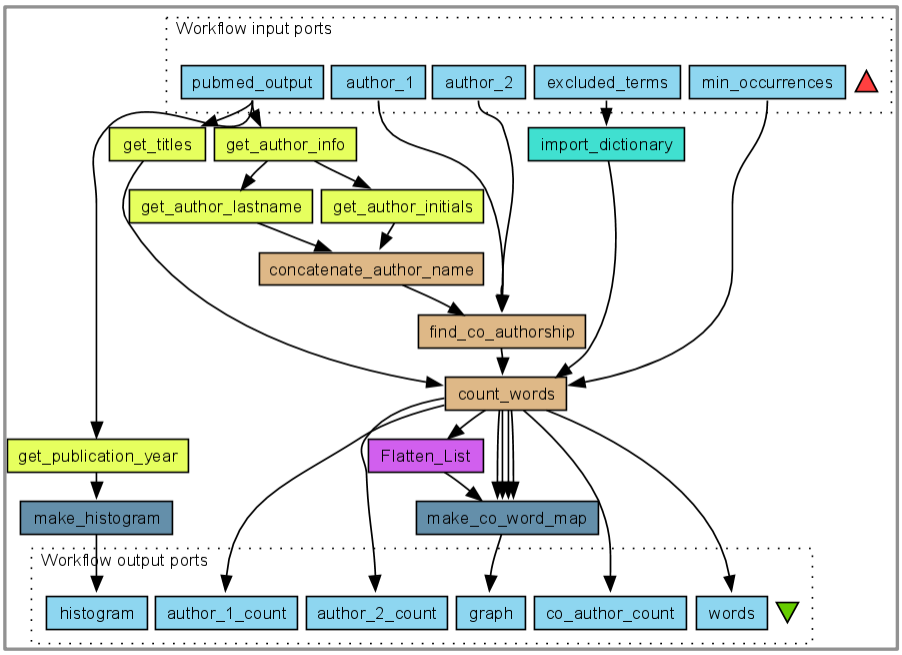


Figure 1. A workflow designed in Taverna for analyzing scientific output over time and comparing word usages of two authors.

The excluded terms are contained in a text file, so the *spreadsheet import* service in Taverna is used to extract each word in the file, line by line. The PubMed results are in XML format, and the extraction of publication years, titles and author names are done by *XPath* services. XPath is a query language for selecting elements and attributes in an XML document. The XPath service in Taverna eases this process by providing a configuration pane to render an XML file of interest as a tree and automatically generate an XPath expression as the user

selects a specific fragment from the XML (Fig. 2). The results of the query can either be passed as text or as XML to other workflow components.

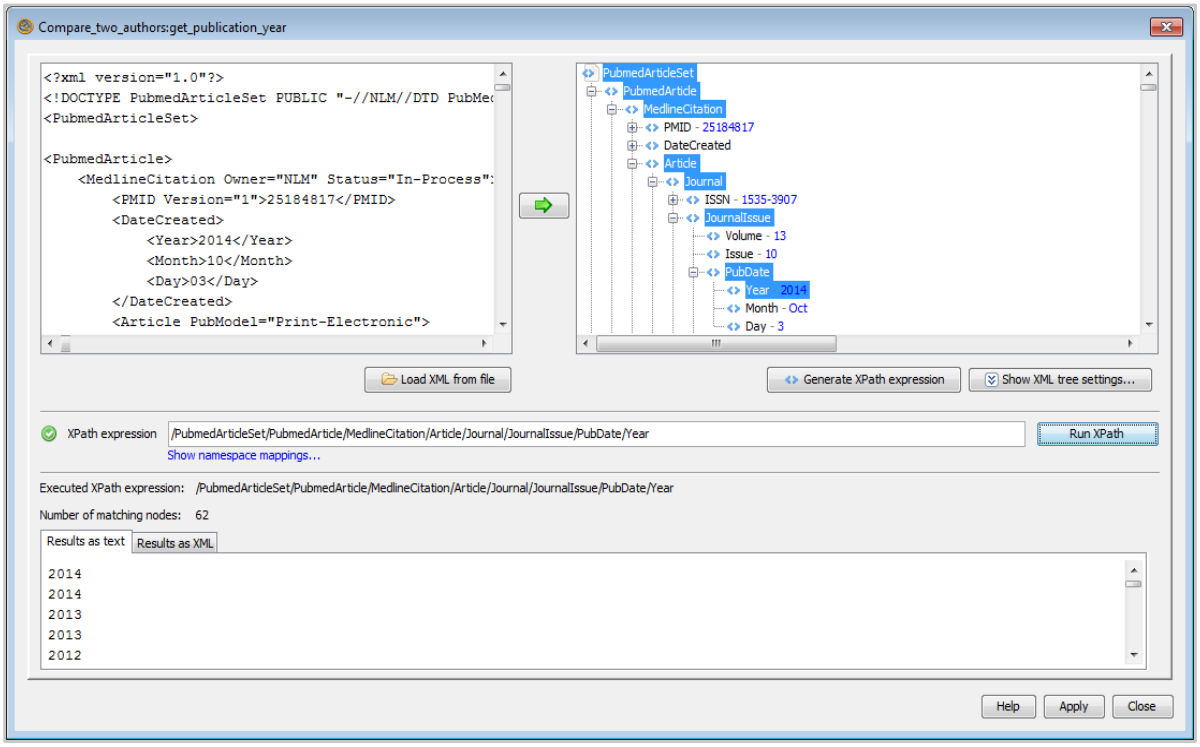


Figure 2. XPath configuration pane for extracting publication year from PubMed XML.

The data extracted by the spreadsheet import and XPath services is fed to a series of Beanshell components that find co-authorships and count co-occurrence of words in the extracted titles. Beanshell is a light-weight scripting language that interprets Java. In our workflow, the Beanshell services do simple operations on strings, such as concatenation of surnames and initials that are extracted separately using XPath (*concatenate_author_names*), matching strings to find co-authorships (*find_co_authorship*) and counting the number of words occurring in each title authored by one or both authors (*count_words*). The two authors' usage of the words, excluding *excluded_terms*, that appear at least *min_occurrences* times in total, are then used to draw a co-word map using the *igraph* (Csárdi & Nepusz, 2006) R package. It is generally up to the workflow designer what part of the workflow to code in Java (Beanshell), in R, or in third language called via the *Tool* command-line interface. More types are available for data connectors between R components (logical, numeric, integer, string, R-expression, text file and vectors of the first four types) than between Beanshell components, where everything is passed as strings. When dealing with purely numerical data, we recommend R over Beanshells within Taverna.

After all the necessary inputs are provided, the workflow is ready to be executed. In the Taverna Workbench *Results* perspective (Fig. 3), each completed process is grayed out to show the progress of the workflow run. The execution times, errors and results are also visible in this perspective.

We ran the workflow for two scientists active in our own field, mass spectrometry, Gary L. Glish and Scott A. McLuckey, whom we knew to have worked on similar topics and also co-authored a number of papers. However, the workflow will work on any two authors with publications indexed by PubMed. The co-word map in Figure 4 visualizes the co-occurrence of words in titles by the location and thickness of the connecting edge, while the relative frequency of usage by the two authors is indicated by the color (from white to gray).

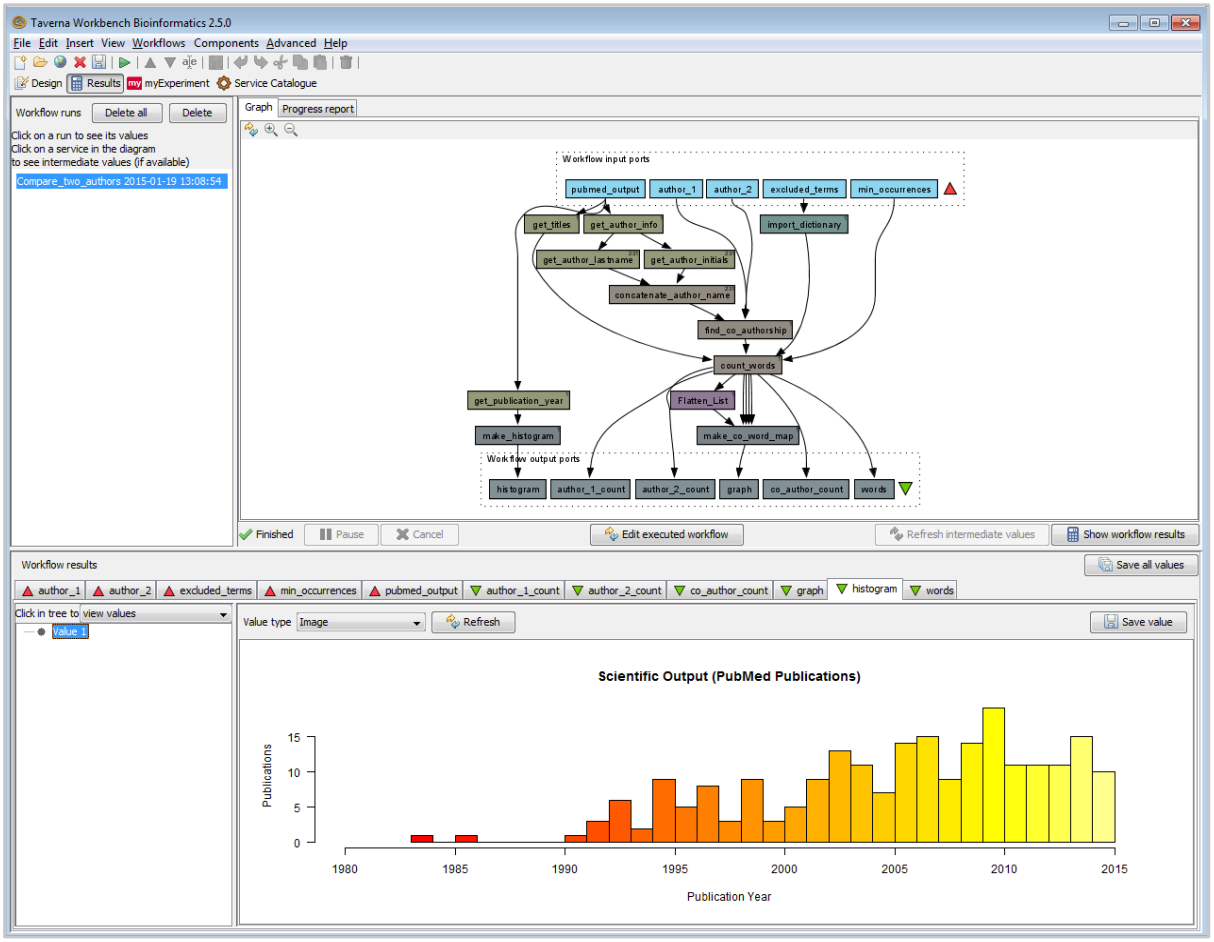


Figure 3. Workflow progress and output in the Taverna workbench Results perspective.

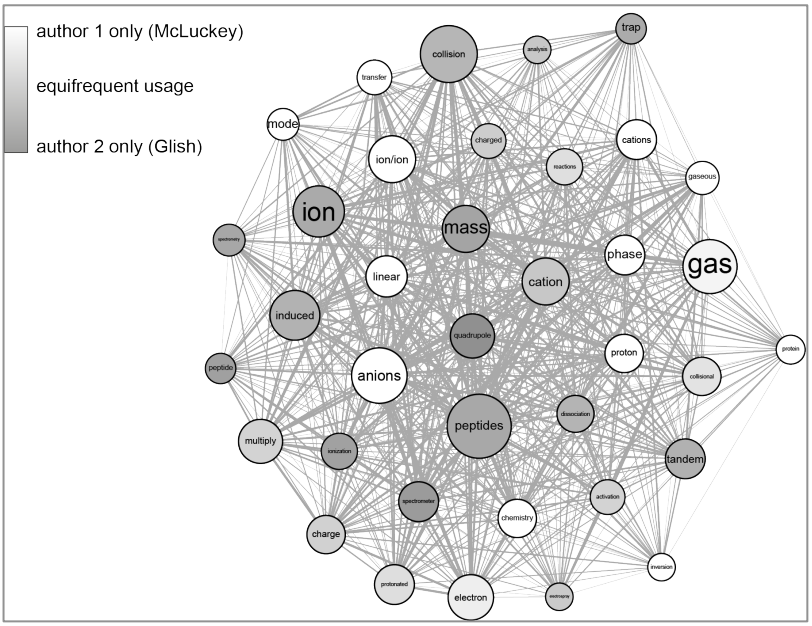


Figure 4. Co-word map output from the *compare_two_authors* workflow.

Connecting to Web Services and External Databases

Automatically generating networks directly from online data is also possible in Taverna workbench. Taverna can invoke WSDL (Web Services Description Language) style Web services given the URL of the service's WSDL document. The WSDL is an XML-based interface description language often used together with a SOAP (Simple Object Access protocol) to access the functions and parameters of a service. Many bibliographic resources are available through Web services, such as Web of Science (WoS). Some services, including the WoS, require authentication. An entire bibliometric study can be contained inside a single Taverna workflow that takes the user queries, or questions of the study, generate the Web service requests, execute these, retrieve the data and proceed with further (local) bibliometric and statistical analysis, and visualization.

A Taverna workflow that invokes WSDL services from WoS to automatically execute a query may look like in the figure below. This Taverna workflow takes as input common search parameters and a generic WoS query string, and passes these to the Web service via the WoS WSDL interface. Values that have only one possible value, such as the language (English, "en") are here hard-coded in the workflow as *Text constants*.

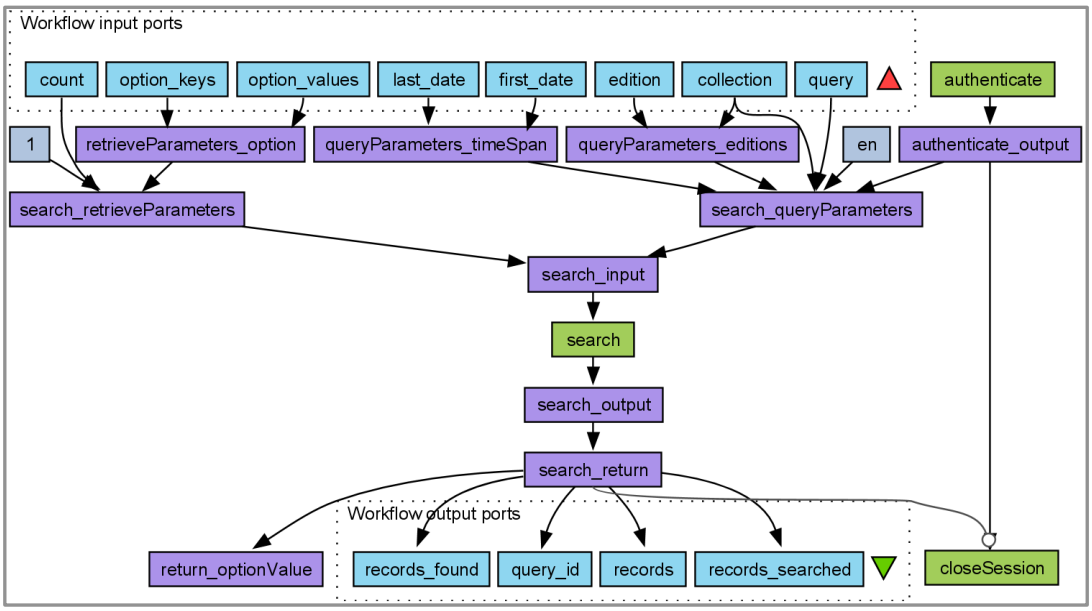


Figure 5. A simple workflow for retrieving bibliometric data using Web services.

Future Work

The use of scientific workflows in bibliometrics is still in its infancy. Modules that accomplish basic bibliometric tasks could be designed and combined in various ways for different studies, thus benefiting from modularity and reusability of scientific workflows. As mentioned above, the direct support of R inside Taverna workflows is particularly useful for bibliometrics. A number of R packages for bibliometric analysis have recently been released, ranging from simple data parsers such as the *bibtex* package (Francois, 2014) for reading BibTeX files to libraries or collections of functions for scientometrics, such as the *CITAN* package (Gagolewski, 2011). The latter package contains tools to pre-process data from several sources, including Elsevier's Scopus, and a range of methods for advanced statistical analysis. The *igraph* package itself comes with some functions specifically for bibliometric analysis, e.g. *cocitation* and *bibcoupling*. Clustering or rearranging the graph spatially so that strongly connected words appear closer together is possible with *igraph*, but may also be assisted by other packages. More crucially, the example workflow here does not yet

implement any advanced text mining functionality for homonym disambiguation or natural language processing. The *openNLP* R package provides an interface to openNLP (Hornik, 2014) and may be used to extract noun phrases and clean up the co-word maps.

Several of our Taverna workflows for bibliometrics and scientometrics, including the two workflows in Figure 1 and Figure 5, can be found in the myExperiment (Goble et al., 2010) group for Bibliometrics and Scientometrics (http://www.myexperiment.org/groups/1278.html). As always, we are grateful for any feedback on these workflows.

Acknowledgements

The authors would like to thank Dr. Yassene Mohammed for technical assistance and Thomson Reuters for granting access to the Web of Science Web services lite.

References

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* (pp. 319-326). Heidelberg: Springer.

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, *1695*, 1695.

De Bruin, J. S., Deelder, A. M., & Palmblad, M. (2012). Scientific Workflow Management in Proteomics. *Molecular & Cellular Proteomics*, *11*, M111.010595–M111.010595.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal Of Statistical Software*, *25*(5), 1–54.

Francois, R. (2014). bibtex: bibtex parser. R package version 0.4.0. Retrieved from http://CRAN.R-project.org/package=bibtex.

Fellows, I. (2013). wordcloud: Word Clouds. R package version 2.4. Retrieved from http://CRAN.R-project.org/package=wordcloud.

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics*, *5*(4), 678–692.

Gil, Y. (2008). From Data to Knowledge to Discoveries: Scientific Workflows and Artificial Intelligence. *To Appear in Scientific Programming*, *16*(4), 1–25.

Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., & De Roure, D. (2010). myExperiment: A repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research*, *38*(May), 677–682.

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, *11*, R86.

Grün, B., & Hornik, K. (2011). topicmodels : An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*, 1–30.

Hornik, K. (2014). openNLP: Apache OpenNLP Tools Interface. R package version 0.2-3. Retrieved from http://CRAN.R-project.org/package=openNLP.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E,. & van Atteveldt, W. (2014). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.4.2. Retrieved from http://CRAN.R-project.org/package=RTextTools.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., & Li, P. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045–3054.

Stevens, R. D., Robinson, A. J., & Goble, C. a. (2003). myGrid: personalised bioinformatics on the information grid. *Bioinformatics (Oxford, England)*, *19 Suppl 1*(1), i302–i304.

Wolstencroft, K. et al. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research* 41(W1), W557-W561.

Expertise Overlap between an Expert Panel and Research Groups in Global Journal Maps

A.I.M. Jakaria Rahman[1], Raf Guns[2], Ronald Rousseau[3] and Tim C.E. Engels[4]

[1] *jakaria.rahman@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp (Belgium)

[2] *raf.guns@uantwerpen.be*
Institute for Education and Information Sciences, University of Antwerp, Venusstraat 35, B-2000 Antwerp (Belgium)

[3] *ronald.rousseau@uantwerpen.be*
ronald.rousseau@kuleuven.be
Institute for Education and Information Sciences, University of Antwerp, Venusstraat 35, B-2000 Antwerp (Belgium); and KU Leuven, Dept. of Mathematics, B-3000 Leuven (Belgium)

[4] *tim.engels@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Political and Social Sciences, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, and
Antwerp Maritime Academy, Noordkasteel Oost 6, B-2030 Antwerp (Belgium)

Abstract
There are no available methods to measure overlap in expertise between a panel of experts and evaluated research groups in discipline-specific research evaluation. This paper explores a bibliometric approach to determining the overlap of expertise, using the 2009 and 2011 research evaluations of ten Pharmaceutical Sciences and nine Biology research groups of the University of Antwerp. We study this overlap at the journal level. Specifically, journal overlay maps are applied to visualize to what extent the research groups and panel members publish in the same journals. Pharmaceutical Sciences panel members published more diversely than the corresponding research groups, whereas the Biology research groups published more diversely than the panel. Numbers of publications in the same journals vary over a large scale. A different range of coverage was found for different research groups; there is also a significant difference between maximum and minimum coverage based on discipline. Future research will focus on similarity testing, and a comparison with other disciplines.

Conference Topic
Methods and techniques

Introduction
Expert panel review is considered the standard for determining research quality of individuals and groups (Nedeva et al., 1996; Rons, et al., 2008; Butler & McAllister, 2011; Lawrenz et al., 2012), but also, for instance, for research proposals submitted to research funding organizations. The principal objective of such evaluations is to improve the quality of scientific research. Currently, there are no available methods that can measure overlap in expertise between a panel and the units of assessment in discipline-specific research evaluation (Engels et al., 2013). Rahman et al. (2014) explored expertise overlap between panel and research groups through publishing in the same Web of Science subject categories. Since one category may comprise a wide array of different subfields and topics (Bornmann, et al., 2011), it is up for discussion how relevant it is to have panel members and research group members publishing in the same subject categories. This paper presents a journal level analysis to explore this issue. Journals cover more closely related subfields and topics (Tseng & Tsay, 2013). This paper uses overlay maps at the journal level (Leydesdorff & Rafols,

2012), with special attention to the quantification of similarity between groups and panel for two disciplines.
In 2007, the University of Antwerp (Belgium) introduced site visits by expert panels that promise communication and participation between expert and research groups. It is expected that each research group's expertise is well covered by the expertise of the panel members.
We have used the data collected in the frame of research evaluation by the University of Antwerp. This research in progress paper explores the expertise overlap between expert panel and research groups of the department of Biology and Pharmaceutical Sciences. Hence, the research questions are:

1) To what extent is there overlap between the panel's expertise and the expertise of the groups as a whole?
2) To what extent is each individual research group's expertise covered by the panel's expertise?

Data and Method

In this paper, we present an analysis of the 2009 assessment of ten research groups (2001-2008) of the Department of Pharmaceutical Sciences, and the 2011 assessment of the nine research groups (2004-2010) belonging to the Department of Biology, University of Antwerp. The citable items from the Science Citation Index Expanded of the Web of Science (WoS) published by the research groups in the reference period were considered.
Both panels were composed of five members (including the chair). All the publications of the individual panel members up to the year of assessment were taken into account. The combined publication output of the Pharmaceutical Sciences panel members is 1,029 publications. In total, these publications appeared in 300 different journals. The number of publications per panel member ranges from 124 to 353, in 39 to 93 different journals. The Biology panel members' publication output amounts to 786 publications in 217 different journals. The number of publications per panel member ranges from 76 to 262, in 36 to 76 journals. There are no co-authored publications between panel members in both cases.

Table 1: Publication profile of the Pharmaceutical Sciences and Biology research groups

Pharmaceutical Sciences research groups (2001-2008)			**Biology research groups (2004-2010)**		
Group code	**Number of Publications**	**Number of Journals**	**Group code**	**Number of Publications**	**Number of Journals**
PSRG - A	40	22	BRG - A	168	53
PSRG - B	62	32	BRG - B	58	33
PSRG - C	61	35	BRG - C	212	212
PSRG - D	32	17	BRG - D	175	68
PSRG - E	64	42	BRG - E	168	69
PSRG - F	34	21	BRG - F	58	35
PSRG - G	67	31	BRG - G	280	139
PSRG - H	39	27	BRG - H	67	42
PSRG - I	29	10	BRG - I	86	52
PSRG - J	11	09	**----**	**----**	**----**
All groups together	**372**	**180**	All groups together	**1,153**	**372**

PSRG = Pharmaceutical Sciences Research Group; BRG = Biology Research Group.

Table 1 lists the number of publications of the research groups. The Pharmaceutical Sciences research groups published 372 publications in 180 journals, including 67 joint publications

between the groups, while the Biology research groups generated 1,153 publications in 372 journals, and there are 119 joint publications between the groups.

For this paper, we adopted the overlay mapping methods based on a global journal map from Web of Science data (Leydesdorff & Rafols, 2012). Journals overlay maps were created for the panels, all individual research groups, and the combined research groups of each department. To this end, all Source titles (Journal titles hereafter) pertaining to the entire citable journal output of the panel members and the groups were retrieved and entered into network software, and overlay information was added to the global journal map. The overlap of research group and panel publications was visualized on a global journal map based on the retrieved journal titles, using the visualization program VOSviewer (van Eck & Waltman, 2010).

Analysis and Results

Panel profiles versus Group profiles

Pharmaceutical sciences panel publications are found in 300 different journals, whereas those of the combined Pharmaceutical Sciences groups cover 180 journals. The journal overlay maps for the Pharmaceutical Sciences combined groups (Fig. 1) and the panel (Fig. 2) clearly show that the publication scope of the panel is wider than that of the combined groups. The panel publications are strong (11.86%) in 'Pharmaceutical Research', 'British Journal of Clinical Pharmacology', and 'Archiv der Pharmazie' journals, whereas the research group publications are clustered (8.6%) in 'Kidney International', 'Planta Medica', 'Environmental Science & Technology' journals.

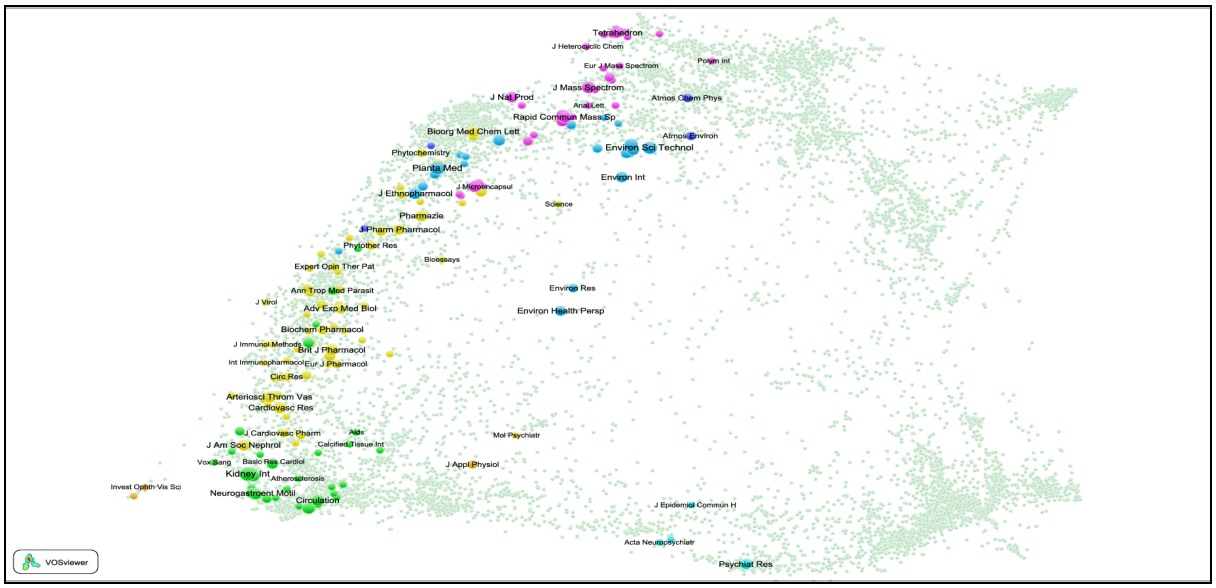


Figure 1. Pharmaceutical Sciences groups' publications overlay to the global journal maps.

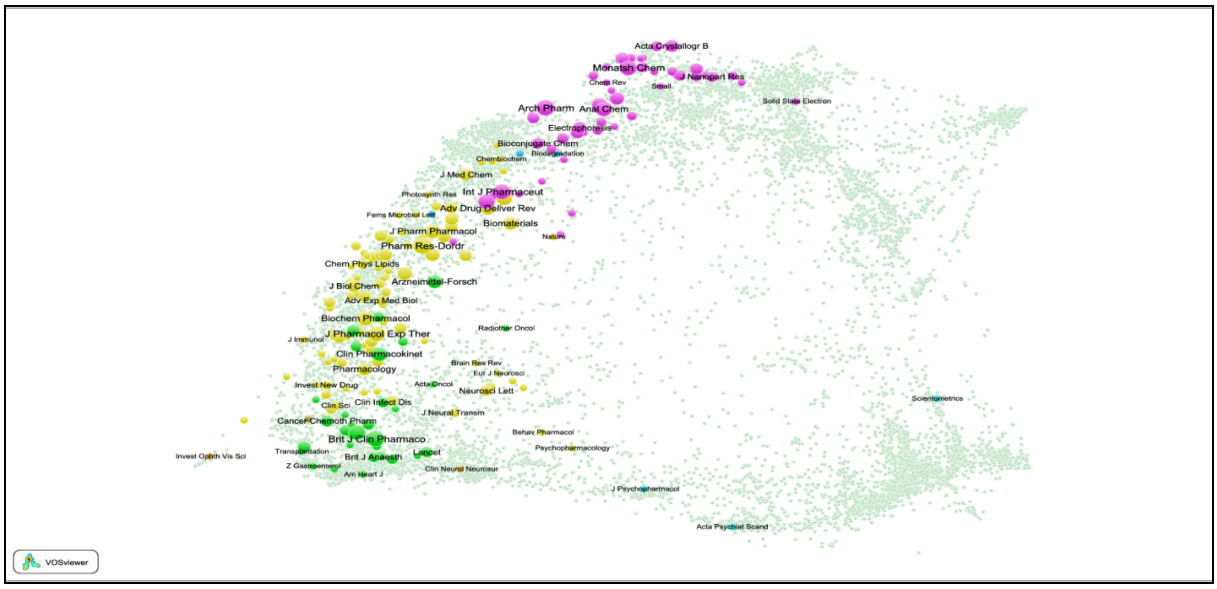


Figure 2. Pharmaceutical Sciences Panel publications overlay to the global journal maps.

Contrariwise, Biology panel publications appeared in 218 journals, while those of the combined Biology groups cover 372 journals. The overlay maps for the Biology department (Figs. 3 and 4) revealed a wider publication scope for the combined research groups compared to the Biology panel. The panel's publications are strong (8.58%) in 'Environmental Pollution', 'Biological Journal of the Linnean Society', and 'Journal of Experimental Biology', whereas the groups' publications tend to be mainly clustered (12.47%) in 'Experimental and Applied Acarology', 'General and Comparative Endocrinology', 'Journal of Experimental Biology'.

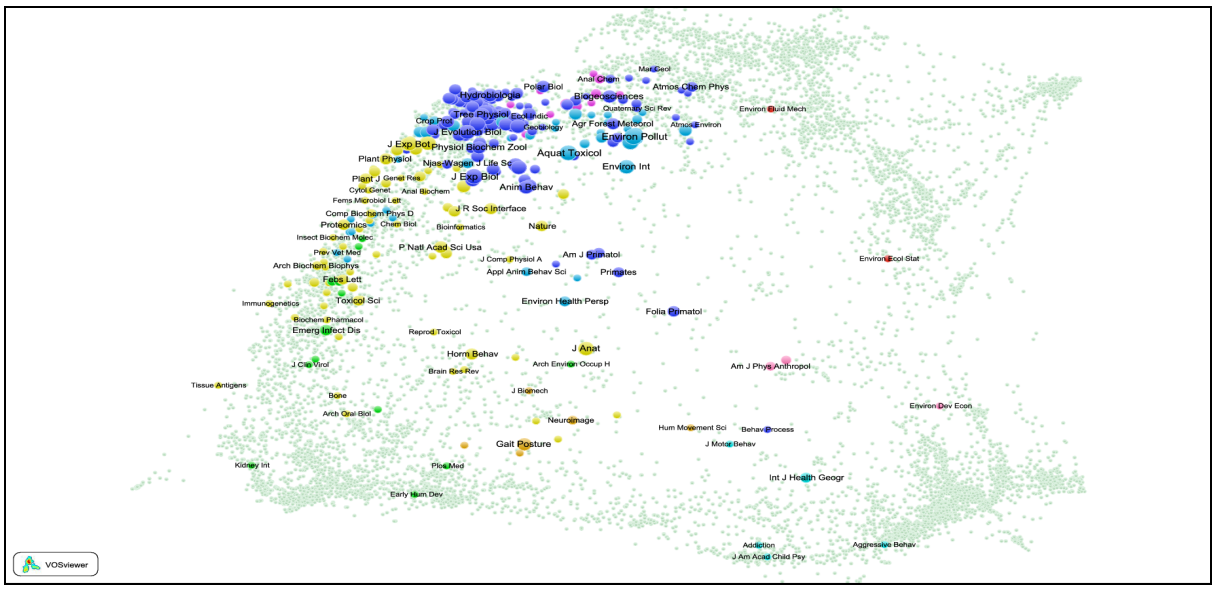


Figure 3. Biology groups' publications overlay to the global journal maps.

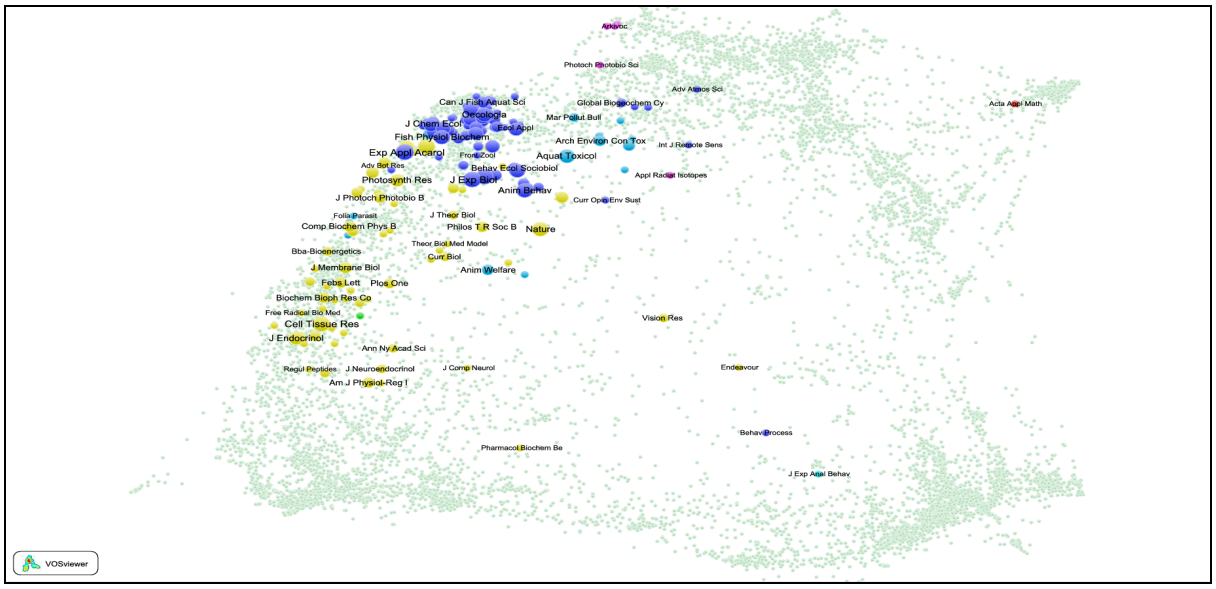


Figure 4. Biology Panel members' publications overlay to the global journal maps.

Table 2 shows that there is no common journal in the top five journals between the Pharmaceutical Sciences panel and groups. Table 2 further shows that there is only one common journal, *Journal of Experimental Biology*, (panel 3.82%, groups 2.26%) in the top five journals between Biology panel and groups.

Table 2: Top five Journals title for the panels and the groups

Panel publications			Group publications		
Pharmaceutical Sciences Department					
Journals Title	**Records**	**% of 1029**	**Journals Title**	**Records**	**% of 372**
Pharmaceutical Research	52	5.05	Kidney International	13	3.5
British Journal of Clinical Pharmacology	35	3.4	Planta Medica	11	2.96
Archiv der Pharmazie	35	3.4	Environmental Science Technology	8	2.15
Clinical Pharmacology Therapeutics	27	2.62	Journal of Mass Spectrometry	7	1.88
Monatshefte Fur Chemie	23	2.23	Chemosphere	7	1.88
Biology Department					
Journals Title	**Records**	**% of 786**	**Journals Title**	**Records**	**% of 1153**
Experimental and Applied Acarology	35	4.45	Environmental Pollution	40	3.47
General and Comparative Endocrinology	33	4.2	Biological Journal of the Linnean Society	33	2.86
Journal of Experimental Biology	30	3.82	Journal of Experimental Biology	26	2.26
Proceedings of the Royal Society B:Biological Sciences	22	2.8	Aquatic Toxicology	23	1.2
New Phytologist	22	2.8	Environmental Science Technology	22	1.91

Together, the Pharmaceutical Sciences panel and groups have 60 journals in common. In addition, 240 journals have panel publications but no group publications, while 120 journals contain group publications but no panel publications. Further, Biology panel and group publications were common in 93 journals. Moreover, 125 journals contained panel publications but no group publications and 279 journals have group publications but no panel publications.

These findings demonstrate that Pharmaceutical Sciences panel published more diversely than the groups, whereas the opposite is true for the Biology department. However, the Pharmaceutical Sciences panel overlaps in one third of the journals of groups' publications, whereas the Biology panel overlaps almost half the journals where biology groups have publications too.

Panel profile versus Individual group profile

Overlay maps of the publications of the individual groups were created, and subsequently compared with the two panel overlay maps. Most Pharmaceutical Sciences research groups have at least one journal in common with the panel; this is the case for PSRG-A (50%), PSRG-B (40.63%), PSRG-C (31.42%), PSRG-D (58.82%), PSRG-E (40.78%), PSRG-F (61.9%), PSRG-G (16.13%), PSRG- H (37.03%), and PSRG-J (20%). Only PSRG-I has none. All Biology research groups have one or more journals in common with the panel: BRG-A (41.51%), BRG-B (18.75%), BRG-C (33.33%), BRG-D (35.29%), BRG-E (42.65%), BRG-F (48.57%), BRG-G (35.97%), BRG-H (19.05%), BRG-I (25%).

These data show that the research outputs of three of the ten Pharmaceutical Sciences research groups (A, D, F) are 50–62 percent, four groups (B, C, E, H) are 30–40 percent, two groups (G, J) are 20 to 15 percent covered by the panels' expertise thematically, whereas one group (group I) is not covered at all. At the same time, three out of nine Biology research groups (A, E, F) are 40-50 percent, three research groups (C, D, G) are 30-40 percent, and another three research groups (B, H, I) are below 25 percent covered by the panel's expertise.

Conclusion

The results indicate that the Biology research groups published more diversely than the panel, which is similar to the findings in Rahman et al. (2014). However, the Pharmaceutical Sciences panel published more diversely than research groups, which is opposite to what was found in Rahman et al. (2014) where the research groups published more diversely in Web of Science subject categories than the panel did. The most likely reason is that all panel members' publications are taken into account (published over the course of over 20 years, often working in different countries and on different topics), whereas the research groups have a specific focus and choose the journals accordingly.

Pharmaceutical Sciences panel overlaps in one third of the journals of the corresponding group's publications, whereas the Biology panel overlaps in close to half the journals where Biology groups have publications. In addition, the number of publications in the same journals by the expert panel and research group varied, and a different range of coverage was found for different research groups. There is also a significant difference between maximum and minimum coverage based on discipline. To quantify which overlap leads to the best standard for evaluation, a considerable range of percentage of common journals between the panel and research group needs to be identified. The considerable range of percentage will express a well-covered, partially covered, and hardly covered expertise based on journal level matching. In subsequent analysis, we will compare results with corresponding results for other disciplines and explore other criteria for adequate relations between evaluation panels and groups.

Acknowledgments

This investigation has been made possible by the financial support of the Flemish Government to ECOOM, among others. The opinions in the paper are the authors' and not necessarily those of the government. The authors thank Nele Dexters for assistance.

References

Bornmann, L., Mutz, R., Marx, W., Schier, H., & Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication? Journal of the Royal Statistical Society: Series A (Statistics in Society), *174*(4), 857–879.

Butler, L., & McAllister, I. (2011). Evaluating University research performance using metrics. *European Political Science*, *10*(1), 44–58.

Engels, T. C. E., Goos, P., Dexters, N., & Spruyt, E. H. J. (2013). Group size, h-index, and efficiency in publishing in top journals explain expert panel assessments of research group quality and productivity. *Research Evaluation*, *22*(4), 224–236.

Lawrenz, F., Thao, M., & Johnson, K. (2012). Expert panel reviews of research centers: The site visit process. E*valuation and Program Planning*, *35*(3), 390–397.

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from web-of-science data. *Journal of Informetrics*, 6(2), 318–332.

Nedeva, M., Georghiou, L., Loveridge, D., & Cameron, H. (1996). The use of co-nomination to identify expert participants for Technology Foresight. *R&D Management*, *26*(2), 155–168.

Rahman, A. I. M. J., Guns, R., Rousseau, R., & Engels, T. C. E. (2014). Assessment of expertise overlap between an expert panel and research groups. In Ed Noyons (Ed.), Context Counts: Pathways to Master Big and Little Data. Proceedings of the Science and Technology Indicators Conference 2014 Leiden (pp. 295–301). Leiden: Universiteit Leiden.

Rons, N., De Bruyn, A., & Cornelis, J. (2008). Research evaluation per discipline: a peer-review method and its outcomes. *Research Evaluation*, *17*(1), 45–57.

Tseng, Y.-H., & Tsay, M.-Y. (2013). Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR. *Scientometrics*, *95*(2), 503–528.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.

Contextualization of Topics - Browsing through Terms, Authors, Journals and Cluster Allocations[1]

Rob Koopman[1], Shenghui Wang[1] and Andrea Scharnhorst[2]

[1] *rob.koopman@oclc.org*, [1] *shenghui.wang@oclc.org*
OCLC Research, Schipholweg 99, Leiden (The Netherlands)

[2] *andrea.scharnhorst@dans.knaw.nl*
DANS-KNAW, Anna van Saksenlaan 51, The Hague (The Netherlands)

Abstract
This paper builds on an innovative Information Retrieval tool, *Ariadne*. The tool has been developed as an interactive network visualization and browsing tool for large-scale bibliographic databases. It basically allows to gain insights into a topic by contextualizing a search query (Koopman et al., 2015). In this paper, we apply the Ariadne tool to a far smaller dataset of 111,616 documents in astronomy and astrophysics. Labeled as the *Berlin dataset*, this data have been used by several research teams to apply and later compare different clustering algorithms. The quest for this team effort is how to delineate topics. This paper contributes to this challenge in two different ways. First, we produce one of the different cluster solutions and second, we use *Ariadne* (the method behind it, and the interface - called *LittleAriadne*) to display cluster solutions of the different group members. By providing a tool that allows the visual inspection of the similarity of article clusters produced by different algorithms, we present a complementary approach to other possible means of comparison. More particularly, we discuss how we can - with *LittleAriadne* - browse through the network of topical terms, authors, journals and cluster solutions in the *Berlin dataset* and compare cluster solutions as well as see their context.

Conference Topic
Methods and techniques; Mapping and Visualization

Introduction
What are essence and boundary of a scientific field? How can a topic be defined? Those are questions that are core to bibliometrics. Rigour and stability in defining boundaries of a field are important for research evaluation and funding distribution. However, if you as a researcher would seek for information about a certain topic of which you are not an expert yet, your information needs are quite different. Among the many possible hits for a search query you might want to know which are core works (articles, books) and which are rather peripheral. You might want to use different rankings (Mutschke & Mayr, 2014) or get some context. On the whole you would have less need to define a topic and a field in a bijective, univocal way. The same holds if you want to compare different clustering algorithms. Here again, you are in need to illustrate similarities and differences between different allocations of documents to clusters. Ways to contextualize them and browse through these contexts would be desirable. This is our starting point.
Decades of bibliometrics research have produced many different algorithms to cluster bibliographic records. They often focus on one entity of the bibliographic record. For example, articles and terms those articles contain (in title, abstract and/or full text) form a

[1] This paper is submitted as part of the Special Session at the ISSI conference 2015 "Same data – different results? The performative nature of algorithms for topic detection in science".

bipartite network from which we can either build a network of related terms (co-word analysis) or a network of related articles (based on shared words). The first method, sometimes also called lexical, has been often applied in scientometrics to produce so-called topical or semantic maps. The same exercise can be applied to authors and articles, articles and journals, in effect each element of the bibliographic record for an article (Havemann & Scharnhorst, 2012). If we extend the bibliographic record with the list of references, we enter the area of citation analysis. Here two methods are widely used: direct citations (known as delivering often sparse matrices) and co-citation maps (known as a good method to identify research fronts). Hybrid methods combine citation and lexical analysis (e.g., Zitt & Bassecoulard, 2006; Janssens et al., 2009). The majority of studies applies one technique. But, sometimes analysis and visualization of multi-partite networks can be found (cf. Van Heur, Leydesdorff, & Wyatt 2013).

Each of the possible different network representations of articles stands for another aspect of connectivity between published scientific works. Co-authorship networks shed light on the social dimension - the invisible colleges - of knowledge production (Mali et al., 2012; Glänzel & Schubert, 2004). Citation relations are interpreted as traces of flows of knowledge (Price, 1965; Radicchi, Fortunato, & Vespignani, 2012). Depending on which element of the bibliographic record is used, we obtain different perspectives on how a field or a topic is to be conceived - as conceptional, cognitive unit; as a community of practice; or as institutionalized in journals. We can call this a *measurement effect*. Another source of variety next to differences resulting from what to analyze is how to analyze it. Finding clusters is part of network analysis. But, clusters can be defined in different ways, and aside of different possible definitions of cluster to determine them for a large-scale network can be algorithmically challenging. Consequently, we find different solutions for one algorithm (if parameters in the algorithm are changed) and different solutions for different algorithms. One could call this an effect of the *choice of instrument for the measurement*. Last but not least, we can ask ourselves, if topics clearly delineated from each other really exist. Often in science very different topics still are related to each other. There exist unsharp boundaries and almost invisible long threads in the *fabric of science* (Boyack & Klavans, 2010), which might inhibit to find a contradiction-free solution. There is a seeming paradox between the fact that experts often can rather clearly identify what belongs to their field or a certain topic, and that it is so hard to quantitatively represent this with bibliometrics methods. However, a closer look into science history and science and technology studies reveals that what belongs to a field or a topic can still differ substantially also in the opinions of different experts; it changes over time; and even a defined canon or body of knowledge determining the essence of a field or a topic might be still subject to controversies and changes.

In the quest to define a topic two things collide. The principal, methodological and data-based ambiguity of what a topic is and the necessity to define a topic for purposes of education, knowledge acquisition and evaluation. This makes it such an intriguing problem to be solved. Because different perspectives can be valid, there is also a need to preserve the above sketched diversity or ambiguity. Having said this, for the sake of scientific reasoning it is also necessary to be able to further specify the validity and appropriateness of different methods to define topics and fields. This paper contributes to the development of methods to compare algorithms and to visualize their different results.

We contribute to this sorting out process in two different ways. First, we apply standard clustering techniques to a specific article matrix built in a specific way from what we call a semantic matrix, in which rows are formed by entities from the bibliographic records of the articles (author names, journal ISSNs, topical terms, subjects, and other characteristics), columns by reduced dimensions from co-occurrence of entities and topical terms (one subset of the entities) over the whole set of articles. While we explain this in detail later, let us note

here that the approach is conceptually more similar to classical information retrieval techniques based on Salton's vector space model than to usual bibliometrical mapping techniques (Salton & McGill, 1983).
In a second step, we present an interactive visual interface called *LittleAriadne* that allows to display the context around those extracted and networked entities. The interface responds to a search query with a network visualization of most related terms, authors, journals and (other) cluster numbers. The query entry can be words, authors, but also cluster solutions. The displayed nodes or entities around a query term represent to a certain extent the context of this query. Depending on the query entry, we will see more or less other terms, journals, or authors. The interface allows to foreground one of entity types by selecting them. The interface has been originally developed for a much larger bibliographic database. In this paper our research questions are:

- Q1: How does the *Ariadne* algorithm work on a much smaller, field specific dataset? What possibility do we have to relate the produced contexts to domain knowledge?
- Q2: Can we use *Ariadne* to label the clusters produced by the different methods?
- Q3: Can we use *Ariadne* to compare different cluster assignments of papers, by treating those cluster assignments as additional entities? What can we visually learn about the topical nature of these clusters?

Data

The dataset used in this paper – called *Berlin dataset* - entails papers published in the period 2003-2010 in 59 astrophysical journals. Those papers have been downloaded from the Web of Science in the context of a German-funded research project called "Measuring Diversity of Research," conducted at the Humboldt-University Berlin - hence the coined name *Berlin dataset*. It contains 120,007 records in total. Eventually, 111,616 records of the document types Article, Letter and Proceedings Paper have been treated with different clustering methods (see the other contributions for this special session).
Some of those cluster outcomes have been shared and are later displayed in the visual interactive interface. Table 1 shows the label of the different sets of clusters *x* we have included in *LittleAriadne*, whereby *x={a, b, ..., f}*. We have noted by which group cluster solutions were produced in the *Source* column. Each clustering method produced a set of clusters, whereby *y* stands for the number of clusters in a set. In our paper we used cluster solutions from CWTS (label: cwts 1.8), Cornell, Humboldt-University Berlin (hu), SciTech (sts-rg), KU Leuven (bc15) and one of our own (oclc_20). Except of cluster set *e*, they are all of the same order of magnitude. Because *Ariadne* relies on statistics across a corpus of articles as large as possible to produce semantic relatedness, we decided to discard clusters with less than 4 articles. But, from the solutions with many clusters (*d, e*) we decided not to display all. The last column in Table 1 gives the final numbers of the clusters from different clustering solutions.

Method

Ariadne - an interactive visualization to navigate entities from large bibliographic databases

The *Ariadne* algorithm has been developed on top of the article database, *ArticleFirst* of OCLC. The interface, accessible at http://thoth.pica.nl/relate, allows users to visually and interactively browse 35 thousand journals, 3 million authors, 1 million topical terms associated with 65 million articles (Koopman et al., 2015). For the purpose of this paper, we applied the same method on the Berlin dataset and built an instantiation, *LittleAriadne*, accessible at http://thoth.pica.nl/astro/relate.

Table 1. Statistics of clusters generated from different methods.

x	Source	*y=#Cluster*	*#Cluster in Ariadne*
a	cwts 1.8	23	23
b	cornell	23	23
c	oclc_20	20	20
d	hu	139	48
e	sts-rg	5664	229
f	bc15	15	15

Table 2. An article from the Berlin dataset.

Article ID	ISI:000276828000006
Title	On the Mass Transfer Rate in SS Cyg
Abstract	The mass transfer rate in SS Cyg at quiescence, estimated from the observed luminosity of the hot spot, is log M-tr = 16.8 +/- 0.3. This is safely below the critical mass transfer rates of log M-crit = 18.1 (corresponding to log T-crit(0) = 3.88) or log M-crit = 17.2 (corresponding to the ""revised"" value of log T-crit(0) = 3.65). The mass transfer rate during outbursts is strongly enhanced
Author	[author:smak j]
ISSN	[issn:0001-5237]
Subject	[subject:accretion, accretion disks] [subject:cataclysmic variables] [subject:disc instability model] [subject:dwarf novae] [subject:novae, cataclysmic variables] [subject:outbursts] [subject:parameters] [subject:stars] [subject:stars dwarf novae] [subject:stars individual ss cyg] [subject:state] [subject:superoutbursts]
Cluster label	[cluster:a 19] [cluster:b 16] [cluster:c 15] [cluster:d 51] [cluster:e 17] [cluster:f 1]

Table 2 shows for one example article from the *Berlin dataset* those fields of the bibliographic record that we used for *LittleAriadne*. It also shows which categories of entities we have. The ISI record ID has been used among the teams to compare solutions. For *Ariadne* as an interface, it does not matter. *Ariadne* is different from a usual Information Retrieval search engine because it does not primarily deliver lists of documents matching a query, but a network of those entities which profile in the whole corpus 'resonate' most with the query entry. We come back to this aspect later. We further define so-called topical terms. Topical terms are frequent single or two-word phrases extracted from all titles and abstracts, for example, "mass transfer" and "quiescence" in our example. Next to the topical term, each author name is treated as an entity. In Table 2 we display the author name (and other entities below) in a syntax that can be used in the search field of the interface to search for a specific author. The next type of entities is the ISSN number of a journal. One can search for a single journal using the ISSN number, in the visual interface the journal title is used as label for a node representing a journal. Further, we have so-called subjects as separate entity type. Those subjects origin from the fields "Author Keywords" and "Keywords Plus" of the original Web

of Science records. As last type of entities we add - and this is specific for *LittleAriadne* - to each of the articles cluster labels from their assignments to clusters produced by different teams. For example, the article in Table 2 has been assigned to cluster number 19 by source a (cwts 1.8) number 16 by source b (cornell), and so on. In other words, we treat the cluster assignments of articles as they would be classification numbers or additional subject headings.

With the above detailed parsing of the bibliographic records we then build the matrix C (see Figure 1). In C, frequent topical terms, subjects, author names, cluster labels and journals appearing in the *Berlin dataset* form the rows, and topical terms as well as subjects are listed in columns. The relatedness between all entities is computed based on the *context* they share, instead of direct co-occurrences in the data. The context of these entities is captured by their co-occurrences with topical terms and subjects, that is, we count how often an author, or a cluster label co-occurs with a certain topical term or subject in an article, summing up over all articles in the corpus. In the Berlin dataset, we have in total 90,343 entities, including 59 journals, 27,027 author names (single instances, no author disambiguation applied), 358 cluster IDs, 39,577 topical terms and 23,322 subjects. This would produce a sparse matrix of roughly 90K x 63K that is expensive for computation.

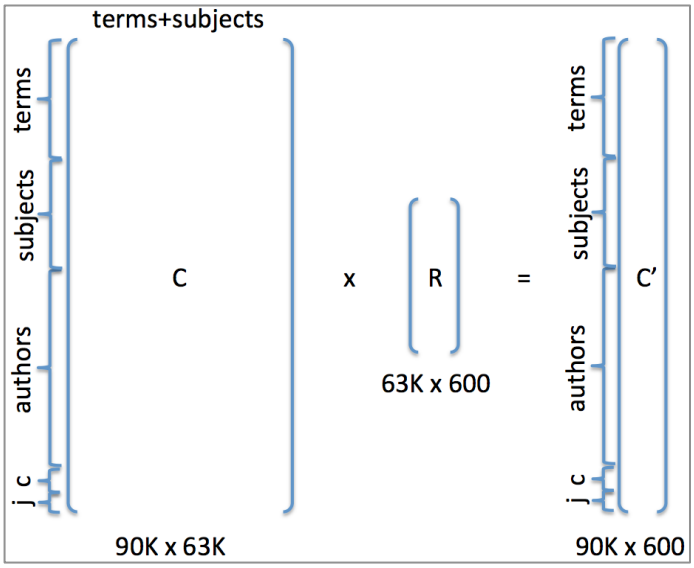


Figure 1. Dimension reduction using Random Projection.

To make the algorithm scale and produce a responsive visual interface, we applied *Random Projection* (Johnson & Lindenstrauss, 1984; Achlioptas, 2003) to reduce the dimensionality of the matrix. As shown in Figure 1, by multiplying C with a 63K x 600 matrix of randomly distributed -1 and 1, the original 90K x 63K matrix C is reduced to a *Semantic Matrix C'* of the size of 90K x 600, with each row vector representing the semantics of an entity. With this Semantic Matrix, the interactive visual interface dynamically computes the most related entities (e.g. ranked by cosine similarity) to a search query and presents a networked visualization of the context of a query term whereby entities are positioned closer to each other if they are more related to each other.

OCLC clusters production - Clustering the Berlin dataset using the Semantic Matrix

The Ariadne interface provides a networked view about entities associated with articles, but it does not produce article clusters straightaway. In order to cluster articles, we need to build a semantic representation of each article. We receive the semantic representation for an article by the following steps. For each article, we look up all entities related to this article in the

Semantic Matrix C'. For our example in Table 2 we have one vector representing the single author of that article in the whole Semantic Matrix, 12 vectors representing the subjects, one vector for the journal, 6 vectors representing the cluster labels and *n* vectors for all extracted topical words. In other words, each article is represented by a subset of vectors and the vector components correspond to the dimensions of the Semantic Matrix. We then take the average of those single entity vectors as the semantic representation of a specific article. All articles together form a matrix M with 111,616 rows and 600 columns. We applied a standard clustering technique - the MiniBatchKmeans method (Sculley 2010) - to M. We used the scikit-learn python library (http://scikit-learn.org/) for this. Applied to the *Berlin dataset* we receive a cluster solution with a comparable size of k=20 clusters, labeled as oclc_20, and a unique assignment of articles to this cluster.

Results - The Berlin dataset in *LittleAriadne*

We used the visual, interactive interface built for the *Berlin dataset* to the context around a specific cluster solution and the similarity between different ones. For this we performed different experiments, which correspond to the research questions Q1-Q3 of the introduction

- Experiment 1: We used *LittleAriadne* as information retrieval tool. We searched with query terms, inspected and navigated through the resulting network visualization. (Q1)
- Experiment 2: We used the semantic matrix to provide the most related topical terms for each cluster as an approximation of cluster labels. (Q2)
- Experiment 3: We used the query syntax to display two or more cluster solutions together in one overview. (Q3)

Experiment 1 - Information retrieval

In *LittleAriadne* we can now study the *Berlin dataset* as any other dataset. Figure 2 gives a snapshot of the context about "magnetic flux" used as query term.[2] The most related topical terms and subjects are shown, together with 3 most related clusters provided by CWTS, Cornell and SciTech (coded in different colors). Each node is clickable which leads to another visualization of the context of the selected node. When mousing over a node, one sees how often this entity occurs in the whole corpus. Given that different statistical methods are at the core of the Ariadne algorithm, this gives an indication of the reliability of the suggested position and links. In the interface one can further refine the display. For instance, one can choose the number of nodes to be shown or decide to limit the display to only authors, journals, topical terms or clusters. Within the interface, one can navigate the context of entities in the *Berlin dataset* by seamlessly travelling between authors, journals, topical terms and clusters in a visual and interactive way.

Experiment 2 -Labeling clusters

Please note, that in *LittleAriadne* we cannot see the position of articles in relations to the different entities. One could say that the articles produce the elements of the networked context, but they themselves are distributed over it. What we can do is to switch to a view that shows most related topical terms, subjects, journals, authors, and other clusters. The outcome of such a *click-through* action is shown in Figure 3.[3] In this example, the most related topical terms, subjects, one journal, and four other clusters are presented as the contextual information about the cluster "a 2".

[2] Figure 2 is accessible at http://thoth.pica.nl/astro/relate?input=magnetic+flux.

[3] Figure 3 is accessible at http://thoth.pica.nl/astro/relate?input=%5Bcluster%3Aa+2%5D.

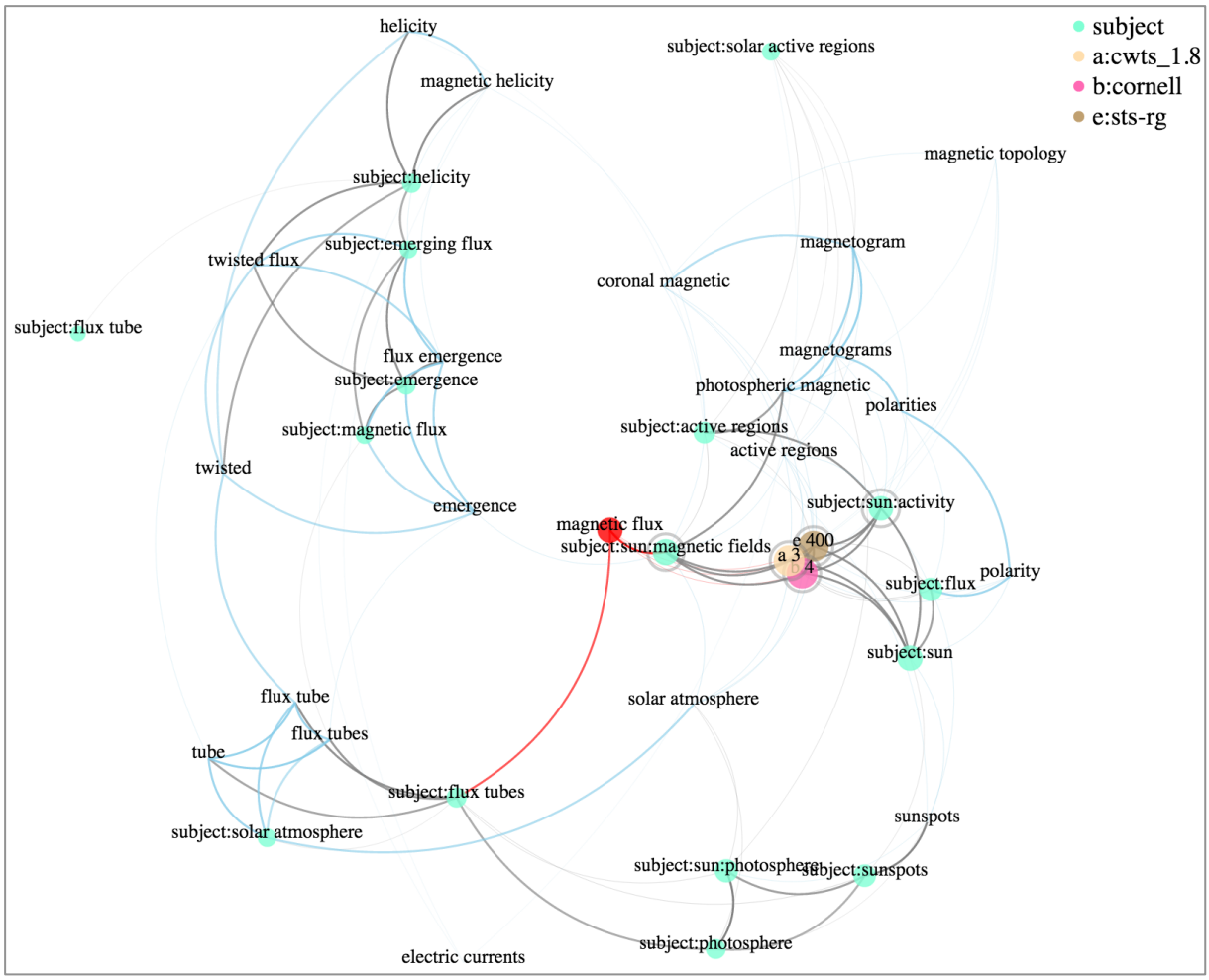


Figure 2. Context around "magnetic flux".

It is now possible to label each cluster using the most related topical terms. As shown in Table 3, the 9 topical terms most related to cluster "a 2" are "cosmology," "dark energy," "density perturbations," "cosmologies," "planck," "cosmological," "spatial curvature," "inflationary," and "inflation." Together they give a rough idea about what this cluster with 8,954 articles is about, but it requires domain expertise to evaluate and transform them into real cluster labels, meaning representing names of specialties, topics or fields used by the scientific community, a well-known problem of bibliometric mapping (Noyons, 2005).

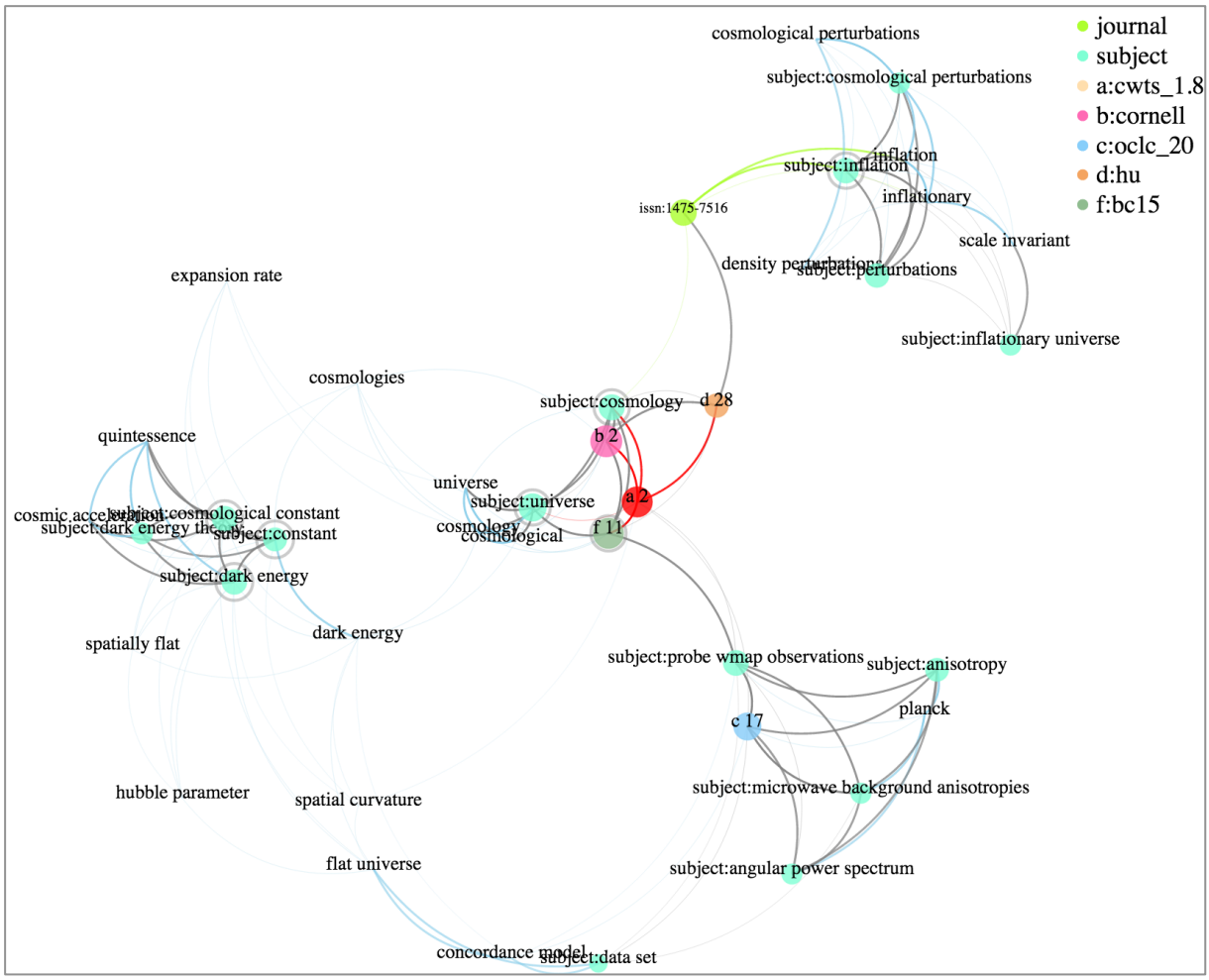


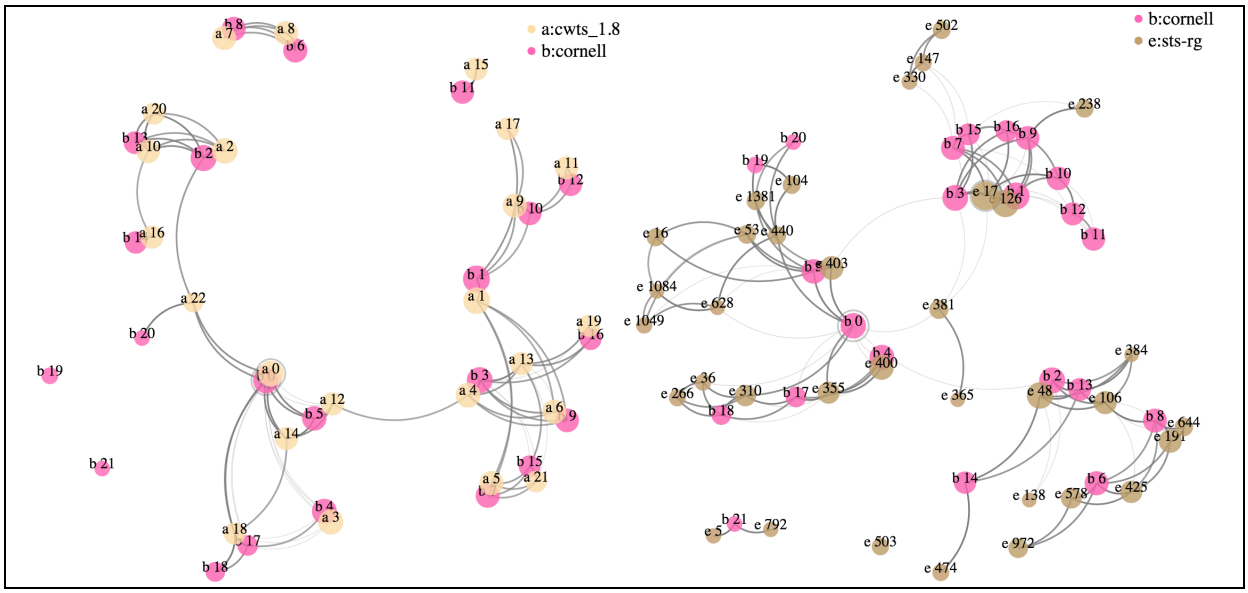
Figure 3. The contextual view of cluster “a 2”.

Table 3. Top related topical terms.

Cluster ID	Top 9 most related topical terms
a 2	"cosmology" "dark energy" "density perturbations" "cosmologies" "planck" "cosmological" "spatial curvature" "inflationary" "inflation"
b 2	"cosmology" "cosmological constant" "cosmologies" "cosmological" "universes" "dark energy" "quadratic" "tensor" "planck"
c 17	"power spectrum" "cosmological parameters" "cmb" "last scattering" "anisotropies" "microwave background" "power spectra" "planck" "cosmic microwave"
d 28	"density perturbations" "inflationary" "inflation" "dark energy" "scale invariant" "spatial curvature" "cosmological perturbations" "inflationary models" "cosmologies"
f 11	"cosmology" "cosmological" "dark energy" "universe" "planck" "density perturbations" "cosmologies" "spatial curvature" "flat universe"

Experiment 3 - Comparing cluster solutions

In *LittleAriadne* we extended the interface with a possibility to compare sets of clusters. In Figure 4 (a) we can visually see the high similarity between clusters from CWTS and those from Cornell.[4] Nearly each CWTS cluster is accompanied by a Cornell cluster. Figure 4 (b) shows two other sets of clusters which partially agree with each other but also clearly have different capacity in distinguishing different clusters.[5] Figure 5 shows all the cluster entities from all six clustering solutions. Given the amount of the clusters, it is difficult to grasp the detailed difference between solutions. However, this visualization does provide a general overview of all the clustering solutions, based on their similarities to each other.



(a) Highly similar (between CWTS 1.8 and Cornell)

(b) Partially agreeing (between Cornell and SciTech)

Figure 4. Comparison between sets of clusters.

Discussion and Conclusion

We present a method and an interface that allows browsing through the contexts of entities, such as topical terms, authors, journals and subjects associated with a set of documents. We have applied the method to the problem of topic delineation addressed in this special session. Because the tool shows (local) context and not the position of single documents in relation to clusters we think it has a potential to be complementary to any other method of cluster comparison. In particular, we have asked how the *Ariadne* algorithm works on a much smaller, field specific dataset. Not surprisingly, compared with our exploration in the ArticleFirst interface, we find more consistent representations. That means that specific vocabulary is displayed, which can be cross-checked in Wikipedia or Google Scholar, for which the interface offers a direct click through.

[4] Figure 4(a) is accessible at http://thoth.pica.nl/astro/relate?input=%5Bcluster%3Aa%5D%5Bcluster%3Ab%5D&type=S&show=50.

[5] Figure 4(b) is accessible at http://thoth.pica.nl/astro/relate?input=%5Bcluster%3Ae%5D%5Bcluster%3Ab%5D&type=S&show=300.

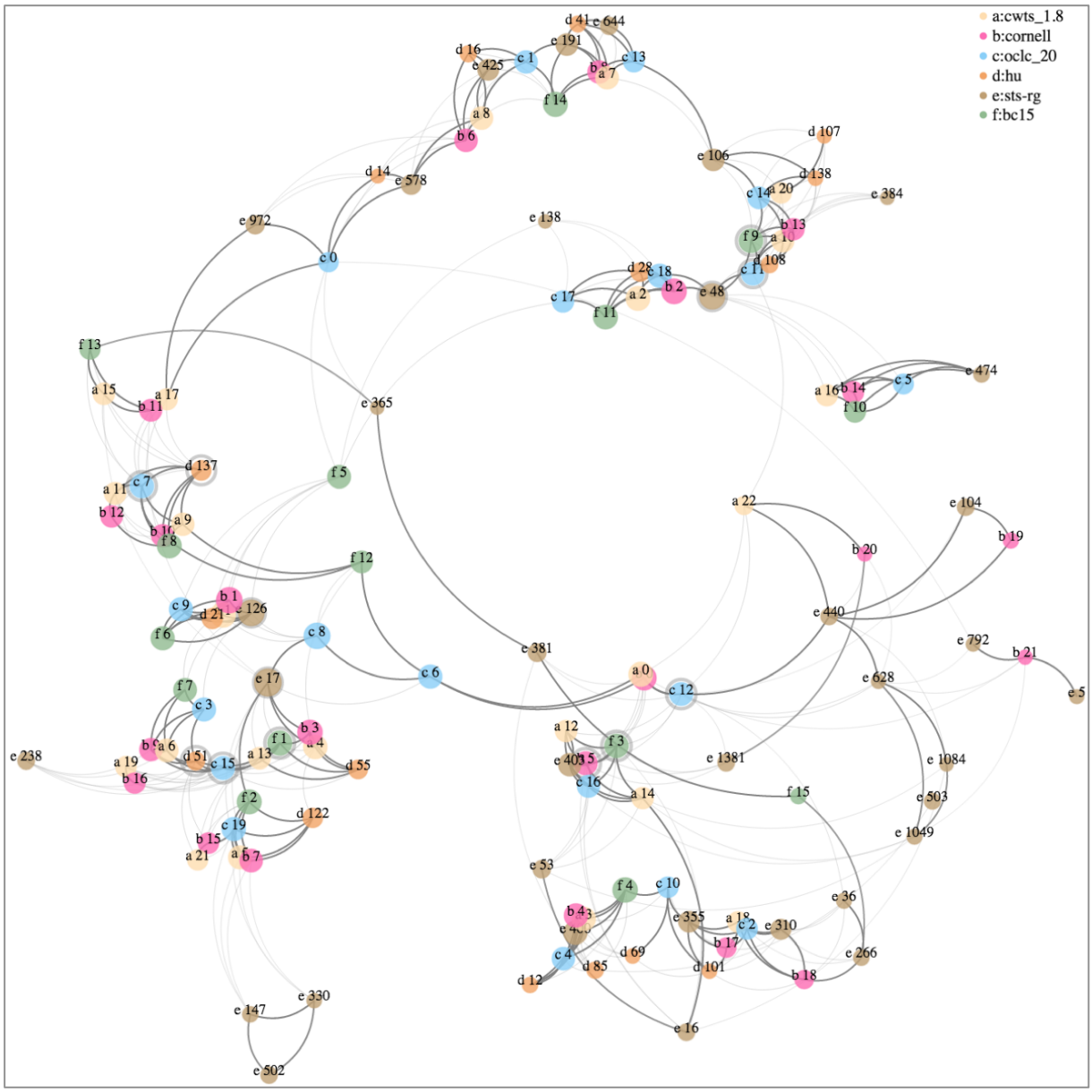


Figure 5. Comparing clusters from 6 clustering solutions.

On the other hand, the bigger number of topical terms in the larger database leads to a situation where almost every query term produces a response. In *LittleAriadne* searches for e.g., literary persons such as Jane Austen retrieve nothing - a blank screen. In preparation of this paper we surfed through the interface, and compared the most relevant topical terms around a cluster to other classifications used in Astrophysics, such as Physics and Astronomy Classification Scheme (PACS®[6]). In this punctual exploration we did find correlations between the names of PACS classes (subclasses, and related controlled vocabulary) and the selected topical terms in *LittleAriadne*. We will further compare the context around clusters and the suggested related topical terms with labels produced by other teams in this special session. Ultimately, the discussion with domain experts belongs to a proper evaluation of the interface. We demonstrated that we can use *LittleAriadne* to compare different cluster solutions mutually and even generate a wider overview. We will discuss in the special session how Ariadne can further be of use in the comparison of clustering and delineation of topics.

[6] http://www.aip.org/publishing/pacs/pacs-2010-regular-edition

At least, we hope that this interactive tool supports discussion about different clustering algorithms and helps to find the right meaning of clusters, and appropriate labels for them.
We also have plans to further develop the Ariadne algorithm. The Ariadne algorithm is general enough to accommodate additional types of entities to the semantic matrix. In the future, we plan to add citations, publishers, conferences, etc. with the aim to provide a richer contextualization of entities. We also plan to add links to articles that contribute to the contextual visualization, this way strengthening the usefulness of *Ariadne* not only for the associative exploration of contexts similar to scrolling through a systematic catalogue, but also as a direct tool for document retrieval. In this context we plan to further compare *LittleAriadne* and *Ariadne*. In a first attempt, we 'projected' the astrophysical documents into *ArticleFirst* by looking them up in the large semantic matrix built for Ariadne. We found the resulting representations less consistent when browsing through. That is not a surprise, because when merging them you see how field-specific content fits and miss-fits into many other contextualizations. The advantage of *LittleAriadne* is the confinement of the dataset to one scientific field and topics within. We hope by continuing such experiments also to learn more about the relationship between genericity and specificity of contexts, and how that can be best addressed in information retrieval.

Acknowledgments

Part of this work has been funded by the COST Action TD1210 KnoweScape.

References

Achlioptas, D. (2003) Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, *66*, 4, 671-687.

Boyack, K.W. & Klavans, R. (2010). Weaving the Fabric of Science. Courtesy of Kevin W. Boyack and Richard Klavans, SciTech Strategies, Inc. In K. Börner & E.F. Hardy (Eds), *6th Iteration (2009): Science Maps for Scholars*, Places & Spaces: Mapping Science. http://scimaps.org.

Glänzel, W., & Schubert, A. (2004). Analysing Scientific Networks through Co-authorship. In *Handbook of Quantitative Science and Technology Research* (pp. 257–276). doi:10.1007/1-4020-2755-9_12

Havemann, F., & Scharnhorst, A. (2012). Bibliometric Networks. *Arxiv Preprint: arXiv:1212.5211 [cs.DL]*, 20. Digital Libraries; Physics and Society. Retrieved from http://arxiv.org/abs/1212.5211

Janssens, F., Zhang, L., Moor, B. De, & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing and Management*, *45*(6), 683–702. doi:10.1016/j.ipm.2009.06.003

Johnson, W., & Lindenstrauss, J. (1984) Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics, 26*, 189–206.

Koopman, R., Wang, S., Scharnhorst, A., & Englebienne, G. (2015). Ariadne's Thread — Interactive Navigation in a World of Networked Information. In *CHI '15 Extended Abstracts on Human Factors in Computing Systems*. Seoul, South Korea, April 18-23, 2015 ACM 978-1-4503-3146-3/15/04, Preprint available at http://arxiv.org/abs/1503.04358

Mali, F., Kronegger, L., Doreian, P., & Ferligoj, A. (2012). Dynamic scientific co-authorship networks. In A. Scharnhorst, K. Börner, & P. van den Besselaar (Eds.), *Models of Science Dynamics* (pp. 195–232). Berlin, Heidelberg: Springer International Publishing. doi:10.1007/978-3-642-23068-4_6

Mutschke, P., & Mayr, P. (2014). Science models for search: a study on combining scholarly information retrieval and scientometrics. *Scientometrics*, 102(3): 2323-2345. doi:10.1007/s11192-014-1485-2

Noyons, C. (2005). Science Maps within a Science Policy Context. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 237–255). Springer International Publishing. doi:10.1007/1-4020-2755-9_11

Price, D. J. de Solla. (1965). Networks of Scientific Papers. *Science (New York, N.Y.)*, *149*, 510–515. doi:10.1126/science.149.3683.510

Radicchi, F., Fortunato, S., & Vespignani, A. (2012). Citation networks. In A. Scharnhorst, K. Börner, & P. van den Besselaar (Eds.), *Models of Science Dynamics* (pp. 233–257). Berlin, Heidelberg: Springer International Publishing. doi:10.1007/978-3-642-23068-4_7

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 1177-1178. DOI=10.1145/1772690.1772862 http://doi.acm.org/10.1145/1772690.1772862

Van Heur, B., Leydesdorff, L., & Wyatt, S. (2013). Turning to ontology in STS? Turning to STS through "ontology." *Social Studies of Science*, *43*(3), 341–362. doi:10.1177/030631271245814

Zitt, M., & Bassecoulard, E. (2006). Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences. *Information Processing and Management*, *42*(6), 1513–1531. doi:10.1016/j.ipm.2006.03.016

A Link-based Memetic Algorithm for Reconstructing Overlapping Topics from Networks of Papers and their Cited Sources

Frank Havemann[1], Jochen Gläser[2] and Michael Heinz[1]

[1] *frank.havemann@ibi.hu-berlin.de* [1] *michael.heinz@rz.hu-berlin.de*
Humboldt-Universität zu Berlin, Berlin School of Library and Information Science, Dorotheenstraße 26, 10099 Berlin (Germany)

[2] *jochen.glaser@ztg.tu-berlin.de*
TU Berlin, Center for Technology and Society, Hardenbergstr. 16-18, D-10623 Berlin (Germany)

Abstract
In spite of recent advances in field delineation methods, enduring problems such as the impossibility to justify necessary thresholds and the difficulties in comparing thematic structures obtained by different algorithms leave bibliometricians with a sense of uneasiness about their methods. In this paper, we propose and demonstrate a new approach to the delineation of thematic structures that attempts to fit the methods for topic delineation to the properties of topics. We derive principles of topic delineation from a theoretical discussion of thematic structures in science. Applying these principles, we cluster citation links rather than publication nodes, use predominantly local information and grow communities of links from seeds in order to allow for pervasive overlaps of topics. The complexity of the clustering task requires the application of a memetic algorithm that combines probabilistic evolutionary strategies with deterministic local searches. We demonstrate our approach by applying it to a network of 14,954 Astronomy & Astrophysics papers and their cited sources.

Conference Topic
Methods and techniques (special session on algorithms for topic detection)

Introduction
The identification of thematic structures (topics or fields) in sets of papers is one of the recurrent problems of bibliometrics. It was deemed one of the challenges of bibliometrics by van Raan (1996) and is still considered as such despite the significant progress and a plethora of methods available. Major developments since van Raan's paper include approaches that cluster the whole Web of Science based on journal-to-journal citations, co-citations, or direct citations, the advance of hybrid approaches that combine citation-based and term-based techniques, and term-based probabilistic methods (topic modelling). However, methodological problems endure and leave bibliometricians with a sense of uneasiness about their methods. Advanced methods still apply thresholds that must be arbitrarily set and adapted to the specific structures that shall be obtained. The relevance of the structures identified by bibliometric methods are difficult to verify independently, and the relationships between thematic structures are difficult to assess. A recent analysis by Hric et al. (2014) found that current algorithms for the detection of communities in network of papers respond to topological properties of networks but not necessarily to the underlying real-world properties of nodes clustered. This observation casts further doubts on the fundamental assumption underlying bibliometric methods for topic delineation, namely that the topics reconstructed using *structural properties of networks of papers* reflect *thematic properties of the research published in those papers*.

In this paper, we propose and demonstrate a new approach to the delineation of thematic structures. We derive principles of topic delineation and criteria for the assessment of algorithms from a theoretical discussion of properties of thematic structures in science. Applying these principles, we cluster citation links rather than publication nodes, use predominantly local information, and grow communities from seeds in order to allow for

pervasive overlaps of topics. The complexity of the clustering task requires the application of a memetic algorithm that combines nondeterministic evolutionary strategies with deterministic local searches. We demonstrate our approach by applying it to a network of 14,954 Astronomy & Astrophysics papers and their cited sources.

Strategy, Methods and Data

Theoretical considerations and strategy

We define topics as theoretical or empirical knowledge about objects or methods of research that is a common focus for a set of research processes because it provides a reference for the decisions of researchers – the formulation of problems, the selection of methods or objects, the organisation of empirical data, or the interpretation of data (on the social ordering of research by knowledge see Gläser 2006). This definition resonates with Whitley's (1974) description of research areas but abandons the assumption that topics form a hierarchy. It only demands that some scientific knowledge is perceived similarly by researchers and influences their decisions.

This weak definition is linked to three properties of topics that create the problems for bibliometrics:

1) The fractal nature of knowledge has been described by van Raan (1991) and Katz (1999). Topics can have any 'size' (however measured) between the smallest (emerging topics that just concern one researcher) and very large thematic structures (fields or even themes cutting across several fields). Methods for topic identification should thus not be biased against any particular topic size.

2) Given the multiple objects of knowledge that can serve as common reference for researchers, topics inevitably overlap. Publications commonly contain several knowledge claims, which are likely to address different topics (Cozzens, 1985; Amsterdamska & Leydesdorff, 1989). Methods for topic identification should thus take into account that bibliometric objects (publications, authors, journals, and cited sources) are likely to belong to several topics simultaneously. Methods also should enable the reconstruction of topics that overlap pervasively (i.e. not only in their boundaries).

3) All topics emerge from coinciding autonomous interpretations and uses of knowledge by researchers (see e.g. the case studies discussed by Edge and Mulkay, 1976, pp. 350-402). While individual researchers may launch topics and advocate them, the latter's content and fate depends on the ways in which they are used by others. From this follows that topics are local in the sense that they are primarily topics to the researchers whose decisions are influenced by and who contribute to them. Methods for topic identification can reconstruct this insider perspective by using local information. Global approaches create different representations of topics by finding a compromise between insider perspectives and all outsider perspectives on topics.

Methods

For a detailed description of the method see Havemann, Gläser, & Heinz (2015). We operationalise 'topic' as a set of thematically related papers but cluster citation links instead of papers because the former can be assumed the thematically most homogenous bibliometric objects (see Evans & Lambiotte, 2009; and Ahn, Bagrow & Lehmann, 2010 on link clustering).

Cost Function: We followed the suggestion by Evans and Lambiotte (2009) to obtain link clusters by clustering vertices in a network's line graph and defined a local cost function $\Psi^*(L)$ of link set L in the line-graph approach. The internal degree $k_i^{\text{in}}(L)$ of node i is defined as the number of links in L attached to i. The external degree of a node is obtained by

subtracting the internal from the total degree: $k_i^{\text{out}}(L) = k_i - k_i^{\text{in}}(L)$. External degrees k_i^{out} are weighted with subgraph membership-grade k_i^{in}/k_i of boundary node i to obtain a measure of external connectivity of link set L:

$$\sigma(L) = \sum_{i=1}^{n} \frac{k_i^{out}(L) k_i^{in}(L)}{k_i} \quad (1)$$

where n is the number of all nodes. The sum can be restricted to boundary nodes because only for boundary nodes of L is $k_i^{\text{out}} k_i^{\text{in}} > 0$. A simple size normalization that accounts for the finite size of the network is achieved by adapting the ratio cut suggested by Wei and Cheng (1989) for link communities, which leads us to the cost function *ratio node-cut* $\Psi^*(L)$:

$$\Psi^*(L) = \frac{\sigma(L)}{k_{in}(L)(1 - \frac{k_{in}(L)}{2m})} \quad (2)$$

where m is the number of all links and $k_{in}(L)$ is the sum of all internal degrees $k_i^{\text{in}}(L)$. $\Psi^*(L)$ essentially relates external to total connectivity of link set L. It can be used to identify link communities (sets of links that are well connected internally and well separated from the rest of the graph) by finding local minima in the cost landscape.

Since the cost landscape is often very rough—has many local minima that sometimes correspond to very similar subgraphs—the resolution of the algorithm must be defined by setting a minimum distance (number of links that differ) between subgraphs corresponding to different local minima. We define the range of a community as the environment in which no subgraph exists that has a lower Ψ^* value. For our experiments with the citation network of astrophysical papers we set a community's minimum range at one third of its size.

Algorithm: The cost function Ψ^* is used in a clustering algorithm that grows communities from seeds. This approach fulfils two more principles derived from our definition of a topic. The independent construction of each community prevents a size bias of the algorithm and enables pervasive overlaps.

```
choose a connected subgraph as a seed
initialize population P by mutating the seed with high variance several times and adapt mutants
while the best community is not too old do
    mutate the best community with low variance and adapt the mutants
    if a mutant is new and its cost is lower than highest cost then
        add it to population P
    end if
    cross the best community with other communities and adapt the offspring
    if offspring is new and its cost is lower than highest cost then
        add it to population P
    end if
    select the best individuals so that the population size remains constant
    if there is no better best community for some generations and innovation rate is low then
        renew the population (mutate the best community with high variance and adapt it)
        select the best individuals so that the population size remains constant
    end if
end while
```

Figure 1. Pseudocode of memetic evolution.

The task of finding communities in large networks is always very complex and requires the use of heuristics. We chose a memetic algorithm that accelerates the search by combining non-deterministic evolution with a deterministic local search in the cost landscape (Neri,

Cotta, & Moscato, 2012). In our algorithm, populations of subgraphs evolve because after a random initialization of a population of some definite size, the genetic operators of crossover, mutation, and selection are repeatedly applied (Fig. 1). Each crossover and mutation is followed by a local search.

Data

The algorithm is applied to the citation network of 14,954 papers published 2010 in 53 journals listed in the category Astronomy & Astrophysics of the Journal Citation Reports 2010 (the journal *Space Weather* with 45 articles was accidentally left out). We downloaded all articles, letters and proceedings papers from the Web of Science. Reference data had to be standardised with rule-based scripts. To reduce the complexity of the network, we omitted all sources that are cited only once because they do not link papers and their removal should not unduly influence clustering. We excluded 184 papers that are not linked to the giant component of the citation network and proceeded with a network of 119,954 nodes that are connected by 536,020 citation links. We neglected the direction of citation links and analysed an undirected unweighted connected graph.

Experiments and Preliminary Results

Constructing the seed population

Since topics can assume all possible sizes, the algorithm should start from differently sized seed graphs. In our experiments, we combined two strategies for obtaining seeds. First, we used Ward clustering with a similarity measure derived from theoretical considerations (Gläser, Heinz & Havemann, 2015). We ordered all hard clusters by their stability (the length of their branch in the dendrogram) and selected the most stable but not too large clusters (a total of 63) as seeds. In addition, we used the citation links of 969 randomly selected papers as seed graphs.

Each seed was first adapted by a local search and then used to initialise the population of 16 different communities by mutating the seed with a variance of 15%.

Owing to the randomness of the evolutionary mechanisms the choice of seed graphs is unlikely to affect the clustering results. However, it is likely to effect the efficiency of the algorithm.

Running the memetic algorithm

Up to ten experiments were run with each seed. The standard mutation variance in each experiment was 5%, i.e. up to 5% of the nodes were randomly exchanged. The variance was increased to 15% for one mutation if Ψ^* values did not improve for 10 generations. Again, we assume these parameters to effect the algorithm's efficiency rather than its outcomes.

Table 1. Examples of experiments with the memetic algorithm.

Community	*Seed sub-graph*		*Number of generations*	*Community*		*Remaining nodes from seed*
	Size	*Ψ* value*		*Size*	*Ψ* value*	
1	13,469	.0692	339	10,586	.0339	10,380
2	19,697	.1174	233	35,159	.0397	18,860
3	35	.4075	232	33	.0047	0
4	76	.5498	203	28	.0975	0

Experiments with the seeds described above resulted in a total of 3,944 distinct communities, 1,375 of which were disregarded because there were better communities within a distance of

less than one third of their size. The remaining 2,569 communities were ordered by increasing Ψ^* values. Table 1 provides exemplary descriptions of some of the experiments. We then calculated the relative coverage of the network as a function of Ψ^* by successively uniting the *L*-sets of the ranked communities. Relative coverage is the ratio of the union's size to the number of all links *m* (Fig. 2). This function has a sharp bend at Ψ^*=0.10458, shortly below maximum coverage. We used this Ψ^* value as cutoff point, which gives us a preliminary result of 154 communities that cover 98.9 % of all links.

Currently, each of these 154 best communities is used as a seed for a refined local search that adds or removes single links instead of nodes with all their links. For some of the 154 communities this additional local search has already led to better communities.

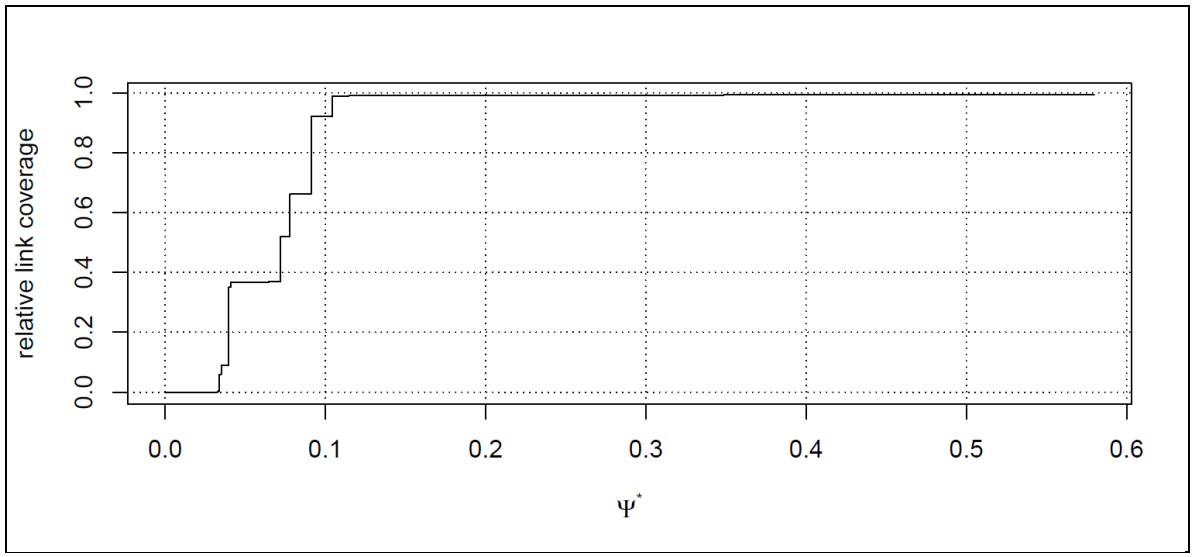


Figure 2. Relative coverage of the network by communities as a function of a Ψ* threshold.

Preliminary results

The 154 communities vary in their size between 9 and 49,324 nodes. Some of the communities overlap pervasively. Seventy communities were not a subset of any other community. The other 84 communities were subsets of one (12 communities) to 28 other communities (1). In Figure 3 we plot sizes and cost of the 154 best communities. Blue circles represent communities that are subsets of others. Green circles represent communities that overlap with another community in 95% of their nodes. All other communities are represented by red circles. The numbers in four circles refer to the communities described in Table 1.

The communities form a poly-hierarchy because some smaller communities are subsets of two larger communities that have no hierarchical subset relation. A community can also have a rest of nodes which are not members of any of its sub-communities.

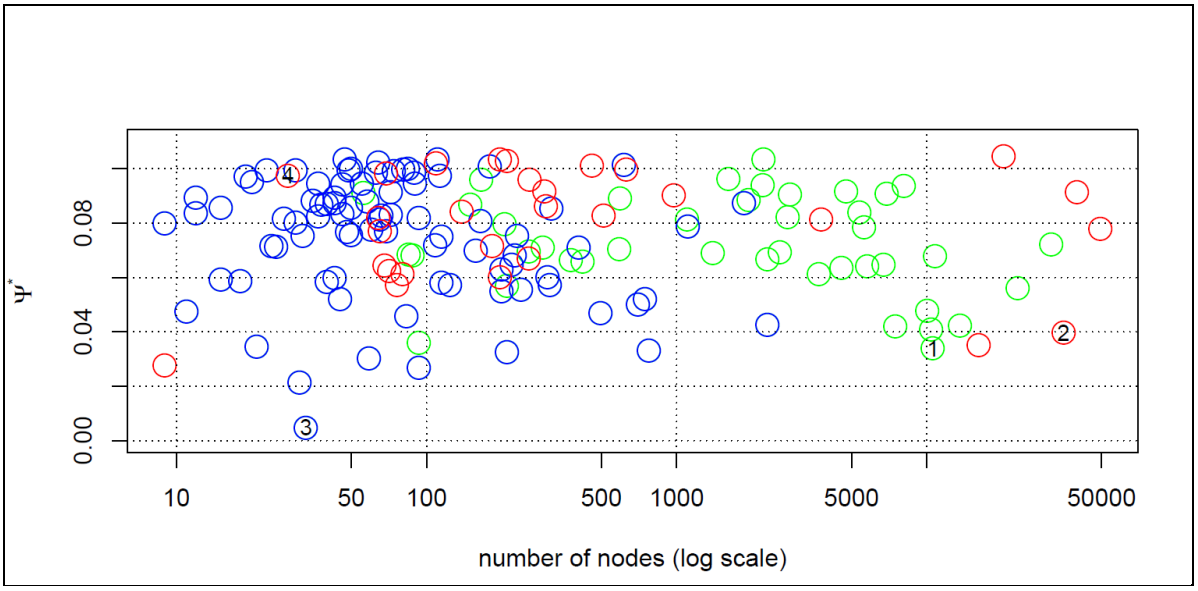


Figure 3. Sizes and Ψ* values of a set of communities covering 98.9% of the graph.

Conclusions

The communities have the structural properties of topics that were derived from the definition. Comparisons with other cluster solutions and tagging of communities will show whether the communities are consistent. We will test the dependence of results on parameter and seed choice with a smaller network. Ultimately, only a discussion with experts can show whether the communities obtained provide one of the possible scientifically meaningful cluster solutions of the astronomy and astrophysics dataset.

Acknowledgments

The work published here was funded by the German Research Ministry (01UZ0905). We thank Andreas Prescher for programming a fast C++-based R-package for parallel node-wise memetic search in the Ψ^* - landscape.

References

Ahn, Y., Bagrow, J. & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761-764.

Amsterdamska, O. & Leydesdorff, L. (1989). Citations: Indicators of Significance? *Scientometrics*, 15, 449-471.

Cozzens, S. E. (1985). Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science. *Social Studies of Science, 15*(1), 127-153.

Edge, D. & Mulkay, M. J. (1976). *Astronomy Transformed: The Emergence of Radio Astronomy in Britain*. New York: John Wiley & Sons, Inc.

Evans, T. & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E*, *80*(1), 16105.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften. Die soziale Ordnung der Forschung*. Frankfurt a. M.: Campus.

Gläser, J., Heinz, M. & Havemann, F. (2015). Measuring the diversity of research. Paper submitted to the 15th International Conference on Scientometrics and Informetrics, Istanbul, 29 June -4 July 2015.

Havemann, F., Gläser, J. & Heinz, M. (2015). Detecting Overlapping Link Communities by Finding Local Minima of a Cost Function with a Memetic Algorithm. Part 1: Problem and Method. arXiv:1501.05139.

Hric, D., Darst, R. K. & Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E*, *90*, 062805.

Katz, J. S. (1999). The self-similar science system. *Research Policy*, *28*, 501-517.

Neri, F., Cotta, C. & Moscato, P. (Eds.) (2012). *Handbook of Memetic Algorithms*, Volume 379 of Studies in Computational Intelligence. Berlin: Springer.

Van Raan, A. F. J. (1991). Fractal Geometry of Information Space as Represented by Co-Citation-Clustering. *Scientometrics*, *20*, 439-449.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*, 397-420.

Wei, Y.-C. & Cheng, C.-K. (1989). Towards efficient hierarchical designs by ratio cut partitioning. In *IEEE International Conference on Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers*, pp. 298–301.

Whitley, R. D. (1974). Cognitive and social institutionalization of scientific specialties and research areas. In: R. Whitley (Ed.), *Social Processes of Scientific Development* (pp. 69-95), London: Routledge & Kegan Paul.

Re-citation Analysis: A Promising Method for Improving Citation Analysis for Research Evaluation, Knowledge Network Analysis, Knowledge Representation and Information Retrieval

Dangzhi Zhao[1] and Andreas Strotmann[2]

[1] *dzhao@ualberta.ca*
School of Library and Information Studies, University of Alberta, Edmonton (Canada)

[2] *andreas.strotmann@gmail.com*
ScienceXplore, F.-G.-Keller-Str. 10, D-01814 Bad Schandau (Germany)

Abstract
Citation analysis is used in research evaluation exercises around the globe, directly affecting the lives of millions of researchers and the expenditure of billions of dollars. It is therefore crucial to seriously address the problems and limitations that plague it. Central amongst critiques of the common practice of citation analysis has long been that it treats all citations equally, be they crucial to the citing paper or perfunctory. Weighting citations by their value to the citing paper has long been proposed as a theoretically promising solution to this problem. *Re-citation analysis* proposes to tune out the large percentage of perfunctory citations in a paper and tune in on crucial ones when performing citation analysis, by ignoring uni-citations (mentioned just once in a paper) and counting and analyzing only re-citations (used again and again in a citing paper). By focusing on core connections in knowledge networks, re-citation analysis can help research evaluation become more sensitive to the distinction between essential and perfunctory impact of research. It may benefit citation-link based knowledge representation and retrieval systems with improved precision by better capturing "aboutness" of articles, the essence of subject indexing in knowledge representation and retrieval, rather than merely providing "relatedness" information.

Conference Topic
Theory; Methods and techniques

Introduction
Citation analysis is used in research evaluation exercises around the globe, directly affecting the work and lives of millions of researchers and the expenditure of billions of dollars. It is therefore crucial to seriously address the problems and limitations that plague it. Central amongst critiques of the current practices of citation analysis has long been that it treats all citations equally, be they crucial to the citing paper or perfunctory. This problem is especially serious when tracing or assessing research impact.

Weighting citations by how they are used in the citing paper has therefore long been proposed as a theoretically promising solution to this problem, but in practice it has not been studied closely at a large scale until recently. Increasingly available digital full-text documents and advances in text processing technologies are now making it feasible to conduct large-scale studies on citation counting weighted by in-text citation frequency, location or context. As a result, interest in this type of studies is growing.

Re-citation analysis as defined here may be viewed as a large sub-class of the class of in-text frequency weighted citation analysis schemes, a class which has recently been found to be the most effective one among many features of in-text citations at characterizing essential citations (Zhu, Turney, Lemire, & Vellino, 2014). We discuss in this paper why we consider re-citation analysis a promising method for improving citation analysis for research evaluation, knowledge network analysis, knowledge representation and information retrieval.

Weighted Citation Counting

Citation analysis examines citation patterns and networks in the scholarly literature through statistical analysis and network visualization. It is applied widely in the social sciences to trace knowledge flows, to evaluate research impact, to study the characteristics of scholarly communities and knowledge networks, and to create citation link based knowledge representation and retrieval systems (Borgman & Furner, 2002; Hall, Jaffe, & Trajtenberg, 2005).

The basic assumption underlying citation analysis is that a citation represents the citing author's use of the cited work, and that it therefore indicates that the citing and cited works are related in subject matter or methodological approach (Garfield, 1979; White, 1990). The total number of citations that a document or any aggregate of documents (e.g., author oeuvre, journal) receives (or a score derived from it, e.g., h-index) is therefore used to assess its impact on research in research evaluation. Citation links are used to signify knowledge flow from the cited to the citing group and, along with scores derived from these links, to measure the relatedness between documents or their aggregates in the study of knowledge networks and in the representation and retrieval of related documents.

The assumptions of citation analysis are believed to be in line with Merton's normative view of science (Garfield, 1979; Merton, 1942; White, 1990). Like other activities of science, citation behaviour is assumed to be governed by a set of norms which require authors to cite documents that have influenced them in developing their current works in order to give credit where credit is due (Edge, 1979; Griffith, 1990; Peritz, 1992; Tranöy, 1980). Although citations for reasons other than giving due credit do exist (Cronin, 1984; Edge, 1979), citation analysis has generally been found to produce valid results because it is based on a statistical analysis of the collective perceptions of large numbers of citing authors, most of whom do adhere to the norms most of the time (Small, 1977; White, 1990). This is especially true with citation network analysis and citation link based knowledge representation and retrieval, as even non-normative citations will not refer to unrelated works.

Researchers do cite for various reasons and citations do serve many different functions in citing papers, however (Brooks, 1985, 1986; Case & Higgins, 2000; Chubin & Moitra, 1975; Liu, 1993; Moravcsik & Murugesan, 1975; Shadish, Tolliver, Gray & Sengupta, 1995; Vinkler, 1987). Small (1982), for example, identified five typical distinctions in citation classification schemes: (1) negative or refuted, (2) perfunctory or noted only, (3) compared or reviewed, (4) used or applied, and (5) substantiated or supported by the citing work.

The importance of weighing citations by their role in the text has therefore long been recognized (Herlach, 1978; Narin, 1976). In recent years, with increasingly available digital full-text documents and advances in technologies for text processing, interest in studying weighted citations has finally picked up. Studies have experimented with weighing citations by the frequency with which they are referred to in the text (e.g., Ding, Liu, Guo, & Cronin, 2013; Hou, Li, & Niu, 2011; Zhu, Turney, Lemire, & Vellino, 2014), by the citation impact of citing papers (Ding & Cronin, 2011), or by the location and context in which they are cited (Boyack, Small, & Klavans, 2013; Jeong, Song, & Ding, 2014). It has been found that frequency-weighted citation ranking can outperform traditional citation ranking of top authors, and that in-text citation frequency was the best of many other full-text features to help spot citations that were considered crucial to the citing papers by their authors, at least in a hard science field studied (Zhu, Turney, Lemire, & Vellino, 2014).

Depending on what functions they serve in a given citing paper, citations likely appear more or less frequently there: perfunctory ones once only, negative or contrastive ones a couple of times, and used or substantiated ones many times. By weighing citations by their frequency of appearance in a scholarly paper, it is hoped that essential citations could be assigned greater weight than perfunctory ones so that citation analysis can focus on the more profound

influences and on organic relationships. If so, this could improve traditional citation analysis significantly as a high incidence of perfunctory citations has been observed (Small, 1982). For example, Teufel, Siddharthan, & Tidhar (2006) found that only a fifth of the references are essential for the citing papers, and Moravcsik & Murugesan (1975) noted that 40% references were perfunctory, frequently simply copied from other papers without ever having been read (Dubin, 2004).

Re-citation analysis: motivation and innovation

Perfunctory citations can thus be considered a serious source of noise if the signal that one wants to detect is the direct and substantial flow of knowledge in the literature. There are two obvious types of approaches to dealing with this problem: (1) to amplify the signal or (2) to filter out the noise. The ultimately best approach is likely some combination of the two. All frequency-based weighing schemes studied so far used the former approach by assigning a weight based on the in-text citation frequency such as assigning a weight of N or N^2 to a citation that appears N times in a citing paper.

By contrast, re-citation analysis, a concept we introduced recently (Zhao & Strotmann, 2015), uses the latter approach: it attempts to filter out perfunctory citations from the analysis by removing uni-citations (i.e., documents referenced only once in the text of a work) in order to analyze only re-citations (i.e., references that appear more than once in the text of a citing paper). The degree to which a cited work is used or has impacted research can be further differentiated by assigning weights to different re-citation frequencies. Re-citation analysis can thus combine the noise filtering and signal amplification approaches, offering the potential to find an optimal weighing scheme for in-text citation frequency.

Thus, the fundamental difference between re-citation analysis and all other frequency-based weighing schemes and hence the innovation of re-citation analysis is that the former attempts to make the fundamental qualitative distinction between those citations that represent real use by, or core impact on, the citing paper (which it tends to retain for analysis) and those that are merely mentioned in passing as related work that the author is aware of but did not directly rely on (which it tends to remove). The basic assumption of re-citation analysis is that papers are very likely to be cited again and again in a publication that relies heavily on them, while perfunctory citations should appear once only in a citing paper almost by definition.

Re-citation analysis can also avoid potential technical problems associated with simply amplifying multi-citations. Since the noise created by perfunctory citations is very strong (40% or more), the signal amplification required to counter it tends to be so strong that it can cause serious distortions. For example, Zhao & Strotmann (2015) found that a simple weight of N does not suffice to make non-perfunctory citations stand out. N^2 is the minimal power of N that fulfills this requirement, but tends to be seriously affected by ultra-meticulous in-text citing styles of a few authors as it overweighs high in-text frequencies. Weighing re-citations avoids this problem.

Promises of Re-citation Analysis

Re-citation analysis can be expected to contribute significantly to the theory and methods of citation analysis. It addresses head-on an old and fundamental concern with citation analysis, especially with evaluative citation analysis. By proposing to filter out the strong noise caused by a high incidence of perfunctory citations rather than simply amplifying multi-citations, it also opens up a new way of thinking about weighing citations at a time when the study of weighted citation counting based on full-text analysis is still in its infancy.

Re-citation analysis is promising in improving citation analysis for research evaluation, knowledge network analysis, knowledge representation and information retrieval.

- Evaluative citation analysis ranks authors, journals, institutions or other components of the scholarly communication system by their citation counts or by derivative scores such as the h-index. Scores based on re-citation counting can be expected to boost those researchers or groupings whose publications receive close scrutiny and to introduce a bias against those whose work mainly provides convenient background information. Such re-citation metrics should thus be better at measuring research impact than traditional citation metrics.
- In citation-based knowledge network analysis and visualization, results based on re-citations can be expected to be significantly more detailed and "crisp" than those based on citations since re-citation based relations (e.g., direct re-citation, co-recitation, or re-citation coupling) should represent core relationships where citation-based relations include many peripheral ones. The price might be an underestimation of interrelatedness between distant parts of a science map.
- For information retrieval (IR), re-citation based similarity metrics can likely provide a considerably enhanced precision of the "Similar documents" or "More like this" feature that many IR systems provide nowadays, compared to citation-based ones. The latter can be expected to show better recall, however, so that a (weighted) combination of the two may work better than either one alone.
- For knowledge representation, it is well understood that citations in scholarly publications serve as concept symbols (Small, 1978). One would expect the presence of a certain set of citations in a paper to translate fairly straightforwardly to the assignment of that paper to a specific subject category. However, subject categories are meant to capture the paper's "aboutness", but a large percentage of citations merely provide "relatedness" information. We suspect that re-citations, on the other hand, do correspond to a considerable degree to concept symbols with an "aboutness" semantics. A re-citation based form of computer-aided subject indexing might therefore be feasible.

Re-citation analysis may thus have a profound impact on the future of the scholarly communication system and of Scientometrics as re-citation analysis values and thus encourages research that is worth following in depth, whereas traditional citation analysis has encouraged review publications that tend to be cited widely.

Finally, as they rely on access to the full text of scholarly publications rather than on citation databases such as Web of Science and Scopus, re-citation analysis methods and metrics are as easily available to the study and evaluation of the social sciences and humanities as to that of the natural and life sciences. Unlike the latter, the former have never been treated fairly by traditional citation analysis due to the insufficient coverage of their literature by these databases.

References

Borgman, C.L. & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*, 3-72.

Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co‐citation clustering using full text. *Journal of the American Society for Information Science and Technology, 64*(9), 1759-1767.

Brooks, T. A. (1985). Private acts and public objects: an investigation of citer motivations. *Journal of the American Society for Information Science, 36*(4), 223-229.

Brooks, T. A. (1986). Evidence of complex citer motivations. *Journal of the American Society for Information Science, 37*(1), 34-36.

Case, D. O. & Higgins, G. M. (2000). How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science, 51*(7), 635-645.

Cronin, B. (1984). *The Citation Process. The Role and Significance of Citations in Scientific Communication*. London: Taylor Graham.

Chubin, D. E. & Moitra, S. D. (1975). Content analysis of references: adjunct or alternative to citation counting? *Social Studies of Science, 5*(4), 423-441.

Ding, Y. & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management, 47*, 80–96.

Ding, Y., Liu, X., Guo, C., & Cronin, B. (2013). The distribution of references across texts: Some implications for citation analysis. *Journal of Informetrics, 7*(3), 583-592.

Dubin, D. (2004). The Most Influential Paper Gerard Salton Never Wrote. *Library trends, 52*(4), 748-764.

Edge, D. (1979). Quantitative measures of communication in science: A critical review. *History of Science Cambridge, 17*(36), 102-134.

Garfield, E. (1979). *Citation indexing – Its Theory and Application in Science, Technology, and Humanities.* New York: John Wiley & Sons.

Griffith, B. C. (1990). Understanding science: Studies of communication and information. In C. L. Borgman (ed.). *Scholarly Communication and Bibliometrics*, 33-45. Newbury Park, CA: Sage Publications, Inc.

Hall, B.H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of Economics, 36* (1), 16–38.

Herlach, G. (1978). Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science, 29*(6), 308-310.

Hou, W., Li, M., & Niu, D. (2011). Counting citations in texts rather than reference lists to improve the accuracy of assessing scientific contribution. *BioEssays, 33*, 724-727.

Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. *Journal of Informetrics, 8*(1), 197-211.

Liu, M. (1993). The complexities of citation practice: A review of citation studies. *Journal of Documentation, 49*, 370–408.

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology, 1*, 115-126.

Moravcsik, M. J., & Murugesan, P. (1975). Some results on the function and quality of citations. *Social Studies of Science, 5*(1), 86-92.

Narin, F. (1976). *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Washington, D. C.: Computer Horizons.

Peritz, B. C. (1992). On the objectives of citation analysis: Problems of theory and method. *Journal of the American Society for Information Science, 43*(6), 448-451.

Shadish, W. R., Tolliver, D., Gray, M., & Gupta, S. K. S. (1995). Author judgements about works they cite: three studies from psychology journals. *Social Studies of Science, 25*(3), 477-498.

Small, H. (1977). A co-citation model of a scientific specialty: A longitudinal study of collagen research. *Social Studies of Science, 7*(2), 139-166.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*(3), 327-340.

Small, H. (1982). Citation context analysis. *In* B. J. Dervin & M. J. Voigt (eds.)*, Progress in Communication Sciences, 3* (pp. 287-310). Norwood, NJ: Ablex.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function*. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (pp. 103-110)*. Stroudsburg, PA, USA.

Tranöy, K. E. (1980). Norms of inquiry: Rationality, consistency requirements and normative conflict. In *Rationality in Science* (pp. 191-202). Springer Netherlands.

Vinkler, P. (1987). A quasi-quantitative citation model. *Scientometrics, 12*(1), 47-72.

White, H. D. (1990). Author co-citation analysis: Overview and defense. *In* C. L. Borgman (ed.), *Scholarly Communication and Bibliometrics* (pp. 84-106). Newbury Park, CA: Sage.

Zhao, D. & Strotmann, A. (2015). Dimensions and uncertainties of author citation rankings: Lessons learned from frequency-weighted in-text citation counting. *Journal of the Association for Information Science and Technology*, doi: 10.1002/asi.23418.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2014). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*. Early view (DOI: 10.1002/asi.23179).

Topic Affinity Analysis for an Astronomy and Astrophysics Data Set

[1]Theresa Velden, Shiyan Yan, and Carl Lagoze

tvelden@umich.edu, shiyansi@umich.edu, clagoze@umich.edu
[1]School of Information, University of Michigan, 105 S. State Street, Ann Arbor, MI 48109 (USA)

Abstract
In this paper we map the affinity between topics extracted from a body of literature published in Astronomy and Astrophysics journals between 2003-2010. The topics are extracted using the popular information theoretic Infomap clustering algorithm (Rosvall & Bergstrom, 2008) iteratively on the giant component of the direct citation network constructed from the data. The affinity network shows what topics are disproportionally well connected (by citations) to other topics. The topology of the network highlights a large division into astrophysics versus astronomically oriented publications. Bridging between those two domains is a population of smaller topics. Going forward, we plan to create and analyze topic affinity network maps for alternative solutions to the topic extraction challenge on that same data set that are produced by our colleagues and that will be discussed and compared at the proposed special session on 'Same data? Different results? The performative nature of algorithms for topic detection in science' at ISSI 2015. We expect that topic affinity mappings will help to examine the nature of differences between different topic extraction solutions.

Conference Topic
Methods and techniques (special session on algorithms for topic detection)

Introduction
The mapping of research topics and collaborative ties in scientific research fields (Morris 2008) is flourishing for a number of reasons. Increasingly, scholarly publications and their metadata are available from a variety of sources (digital libraries, institutional and disciplinary repositories, along with bibliographic abstracting services such as the long established Web of Knowledge and more recently, Scopus). Complementing this is the emergence of sophisticated algorithms for the analysis of complex networks (Newman 2003b) and the wide availability of advanced user-friendly network analysis and visualization tools like pajek, gephi, or VOS Viewer.

However, many different algorithms for community extraction and topic detection exist and offer different suggestions what the most prominent groupings of publications or authors may be. The special session at ISSI 2015 sets out to systematically compare and evaluate the origin, extent, and implication of differences between topic extraction methods. In this paper we describe the results of our approach to topic detection and topic affinity analysis to the shared 'astronomy and astrophysics' data set. This approach has emerged from research program on studying behavioral patterns in scientific communities and comparing them across fields, and may help to shed light on the nature of differences between topic extraction solutions.

Background
As described in (Velden 2009), we take a mixed method approach to studying field-specific practices and cultures of scientific communities, integrating ethnographic field studies with network analytic methods. The network analytic method we apply here to the 'astronomy and astrophysics' data set is part of an ongoing effort to combine network analytic with ethnographic methods (Velden, Haque & Lagoze, 2010; Velden, 2013). This evolves a tradition of close-up analysis of scientific networks and communication practices started by Crane's work (1972) on invisible colleges and taken up more recently by Zuccala (2006).

Scientific research specialties are a complex social and cognitive phenomenon. Sociologically, they can be characterized as collective production communities that emerge from the indirectly coordinated activity of autonomous actors (research groups) who aim to contribute to a shared knowledge base (Gläser, 2006; Velden, 2013). Therefore, the combined analysis of social and cognitive structures is of particular interest (Ding, 2011). In our work we achieve this in two steps: first by algorithmically extracting major research topics in a research specialty from the direct citation network and generating an affinity network that shows what topics are disproportionally well connected through citations to other topics. In a second step, we overlay the topic information on the group collaboration network (Velden, Haque & Lagoze, 2010) extracted from the co-author network of the research specialty. The resulting maps show how collaborative ties connect groups active in a particular topic area. This paper reports work in progress. At this point, we have produced and analyzed the topic affinity network. Producing the overlay with the group collaboration network will be one of the next steps.

Method

Our approach to topic extraction and topic affinity analysis is discussed in detail in Velden (2013). Below we briefly review the relevant details for the analysis reported in this paper.

Data

The data set used in this study includes papers published 2003-2010 in 59 astrophysical journals indexed by Web of Science. By accepting only documents of type 'Article', 'Letter', and 'Proceedings Paper', the data set comprised the bibliographic data of 111,616 publications.

Network construction

Various citation-based approaches have been used in the past to detect topics in research fields. These include bibliographic coupling, co-citation and direct citation, including or excluding citation environments. The advantages and disadvantages of these approaches have been discussed in Boyack (2010). We base our topic extraction on the direct citation network.

Clustering

We use the Infomap clustering algorithm (Rosvall & Bergstrom, 2008) twice to iteratively extract clusters of clusters of documents. The repeated clustering is necessary to obtain sufficiently large entities (topics) for further visual inspection and analysis. In the resulting topic network, nodes represent clusters of publications based on the direct citation links between them.

Topic affinity network

We evaluate the strength of citation links between topic areas relative to a null model that assumes a random distribution of citation links proportional to topic area sizes. Hence, the existence of a link between topics in the affinity indicates a surplus of connectivity between the two topic areas in question, whereas the absence of a link may either mean 'normal' (random) background connectivity or a negative affinity value ('antagonism').

The affinity between a source topic area and a target topic area is calculated as shown in Figure 1 below.

Assume:
A_{11-i}: Top 11 Areas expect area i
$N_{p(j)}$: Number of papers in topic area j
C_{ij}: Number of Citation from topic area i to topic area j

We define the citation based affinity A between two topic areas i and j as the residual:

$$A_{ij} = \frac{\text{Actual Count}_{ij} - \text{Expected Count}_{ij}}{\sqrt{\text{Expected Count}_{ij}}}$$

where:

$$\text{Actual Count}_{ij} = C_{ij}$$

$$\text{Expected Count}_{ij} = \frac{N_{p(j)}}{\sum_{k \in A_{11-i}} N_{p(k)}} \times \left(\sum_{k \in A_{11-i}} C_{ik}\right)$$

Figure 1. Affinity between a source topic area and a target topic

Topic affinity as defined here is a relative property. It expresses the relative preference for documents in one topic area to cite documents in another area given the choice of topic areas included in the data set and in the affinity calculation. Theoretically, the relative affinity to document clusters outside the set of topic areas selected for this analysis or even outside of the data set (external citations) could be greater than to the ones in the set.

Topic Labeling

To support the interpretation of the resulting topic affinity network, we use a semi-automatic approach to labeling topic areas. To this end, we analyze the frequency of journals that the documents in each topic area are published in. Using a measure based on the concept of *term frequency - inverse document frequency (tf-idf)* to combine popularity with distinctiveness of a journal title within the data set, we produce a ranked list of the 15 most popular journals in each topic area. From those journal titles we then derive labels that typically reflect sub disciplinary orientation of topic areas. A more detailed and specific identification of topic area content either algorithmically or through expert evaluation or would be desirable.

Results

The topic extraction from the giant component of the direct citation network results in 22 document clusters ('topics'). For pragmatic reasons, to support interpretation of the visualized network, we include only the largest eleven topic areas in the affinity network. Given the uneven size distribution of clusters (Fig. 1), these largest clusters account for the large majority of publications in the giant component of the direct citation network, namely 84% (see Table 1 for details on the sizes of various network components).

Table 1. Properties of direct citation network.

	# of nodes (documents)	**% of network**	**% of giant component**
entire network	111,616	100	N.A.
giant component	101,831	91.2	100
11 largest topic areas	85,562	84.0	76.7

The topic affinity network for the largest 11 document clusters is shown in Figure 2. The most striking topological feature regards the relationship between the three largest topics. Notably,

topic 3 (Astronomy/Solar System) is not directly connected with the other two topics, topic 1 (Astronomy/Astrophysics) and topic 2 (Gravitational Physics, Cosmology). Topic 2 has a strong directed link to topic 1, indicating that it borrows disproportionally from the literature in topic 2. Topics 1 and 3 are indirectly linked, via small, astronomically oriented 'proxy topics', essentially topics 7 and 9, and to l lesser degree topics 10 and 11. However, there exists only a very faint indirect affinity link between topic 2 and topic 3, via topic 11.

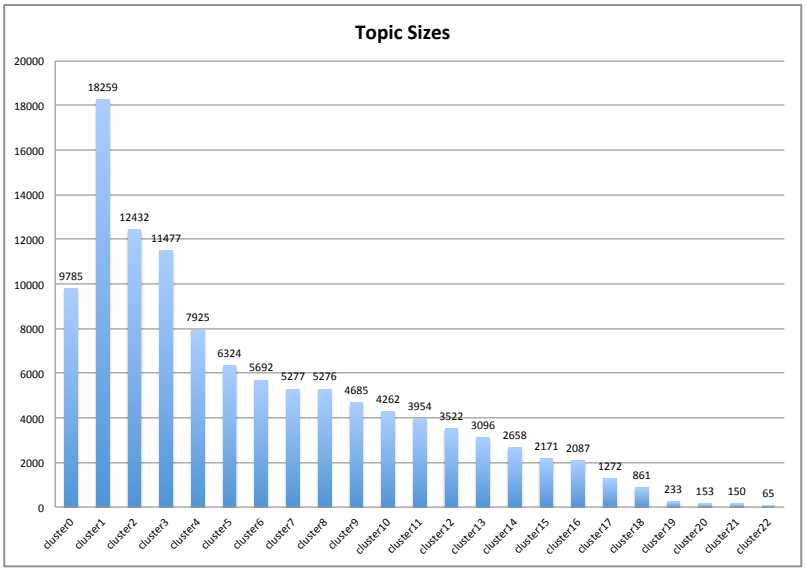


Figure 1. Sizes of the 22 document clusters ('topics') that constitute the giant component of the direct citation network. Cluster '0' shows the number of documents not included in the giant.

Discussion

Based on our own, if limited, expertise in this larger domain of research, we would offer the following speculations about the interpretation of the tripartite structure of the current 2003-2010 literature in the astronomy and astrophysics data set that is suggested by the topology of the affinity network in figure 2. The literature is subdivided into three large domains, with distinct research focus, namely astrophysics - the quest for developing a theoretical understanding of physical and chemical properties of celestial bodies (topic 1), gravitational physics - the quest for understanding the workings of gravitational forces in the universe (topic 2), and planetary science - the quest for understanding the composition, dynamics and history of planets and solar systems (topic 3). As reflected by the affinity network, in the 2003-2010 period, the three domains rely to varying degrees on astronomical observation; this is least the case for gravitational physics. An interesting open question is to what degree the observational astronomy literature has been integrated through citations into these larger topics rather than being identifiable as separate topics. The topic affinity network further underlines that whereas there are strong connections between astrophysics and gravitational physics (such as the role of gravitational forces in the formation of black holes and the puzzle of the nature of black matter), the cognitive links between gravitational physics and planetary science are weak.

Table 2. Ranking of the 15 most popular journals in each topic. This list of journal titles is used to help identify the subject matter of a topic in terms of its subdisciplinary orientation.

Journal titles	# of publications	tf*idf score
Area1		
ASTRONOMICAL JOURNAL	1098	0.104672985
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	4415	0.091614001
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	401	0.06435346
ASTRONOMISCHE NACHRICHTEN	314	0.062939775
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF AUSTRALIA	116	0.056289489
NEW ASTRONOMY REVIEWS	347	0.043675217
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	152	0.037652069
ASTRONOMY REPORTS	164	0.032873003
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	171	0.027442498
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	284	0.027073887
ASTROPHYSICAL JOURNAL LETTERS	510	0.022086996
PHYSICAL REVIEW D	164	0.020641889
ASTROPHYSICS AND SPACE SCIENCE	290	0.006017681
ASTROPHYSICAL JOURNAL	5565	0
ASTRONOMY & ASTROPHYSICS	3148	0
Area2		
PHYSICAL REVIEW D	5616	0.700439718
JOURNAL OF COSMOLOGY AND ASTROPARTICLE PHYSICS	1416	0.533389555
CLASSICAL AND QUANTUM GRAVITY	1533	0.376292436
GENERAL RELATIVITY AND GRAVITATION	543	0.204541334
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	655	0.081693023
GRAVITATION & COSMOLOGY	75	0.036063565
ASTROPARTICLE PHYSICS	78	0.023617218
NEW ASTRONOMY	46	0.017327627
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	783	0.016100189
NEW ASTRONOMY REVIEWS	122	0.015216105
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	49	0.007792228
ASTROPHYSICS AND SPACE SCIENCE	286	0.005880784
ASTROPHYSICAL JOURNAL LETTERS	40	0.001716582
ASTROPHYSICAL JOURNAL	506	0
ASTRONOMY & ASTROPHYSICS	325	0
Area3		
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	364	0.160723328
ICARUS	150	0.129745662
ASTRONOMISCHE NACHRICHTEN	361	0.128983753
ASTRONOMICAL JOURNAL	732	0.124387179
NEW ASTRONOMY	107	0.072503543
ASTROPHYSICS	89	0.060306686
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	1461	0.054039787
ASTRONOMY REPORTS	108	0.038587937
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	111	0.031752855
ASTROPHYSICAL JOURNAL LETTERS	318	0.024548551
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	127	0.021580836
NEW ASTRONOMY REVIEWS	85	0.019070268
ASTROPHYSICS AND SPACE SCIENCE	385	0.014240464
ASTROPHYSICAL JOURNAL	2773	0
ASTRONOMY & ASTROPHYSICS	3122	0
Area4		
SOLAR PHYSICS	1248	2.133094119
ANNALES GEOPHYSICAE	228	0.222784668
ADVANCES IN SPACE RESEARCH	372	0.153453831
GEOPHYSICAL AND ASTROPHYSICAL FLUID DYNAMICS	77	0.131609172
ASTRONOMISCHE NACHRICHTEN	187	0.096348379
SPACE SCIENCE REVIEWS	96	0.093804071
ASTRONOMY REPORTS	119	0.061312605
ASTROPHYSICAL JOURNAL LETTERS	333	0.037069606
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	77	0.031763293
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	95	0.03073523
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	102	0.024994227
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	189	0.010080921
ASTROPHYSICS AND SPACE SCIENCE	75	0.004000365
ASTROPHYSICAL JOURNAL	2165	0
ASTRONOMY & ASTROPHYSICS	1609	0
Area5		
ICARUS	2102	2.700439718
PLANETARY AND SPACE SCIENCE	850	1.091995129
ASTROBIOLOGY	258	0.454192886
EARTH MOON AND PLANETS	257	0.330167939
CELESTIAL MECHANICS & DYNAMICAL ASTRONOMY	170	0.299274383
SOLAR SYSTEM RESEARCH	167	0.29399307
SPACE SCIENCE REVIEWS	115	0.115737336
ADVANCES IN SPACE RESEARCH	263	0.111741818
ANNALES GEOPHYSICAE	104	0.104666808
ASTRONOMICAL JOURNAL	231	0.058301094
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	219	0.012031166
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	38	0.009590656
ASTROPHYSICAL JOURNAL LETTERS	72	0.008255273
ASTROPHYSICAL JOURNAL	286	0
ASTRONOMY & ASTROPHYSICS	598	0
Area6		
PHYSICAL REVIEW D	4101	0.700439718
JOURNAL OF COSMOLOGY AND ASTROPARTICLE PHYSICS	353	0.182093016
ASTROPARTICLE PHYSICS	430	0.178295313
CLASSICAL AND QUANTUM GRAVITY	33	0.011092632
ADVANCES IN SPACE RESEARCH	45	0.00979976
ASTROPHYSICS	16	0.008253508
NEW ASTRONOMY REVIEWS	45	0.007685878
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	45	0.007685878
COMPTES RENDUS PHYSIQUE	16	0.006634244

Journal titles	# of publications	tf*idf score
Area 6 (contd)		
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	15	0.002561959
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	69	0.001942923
ASTROPHYSICAL JOURNAL LETTERS	29	0.001704276
ASTROPHYSICS AND SPACE SCIENCE	25	0.000703958
ASTROPHYSICAL JOURNAL	241	0
ASTRONOMY & ASTROPHYSICS	107	0
Area7		
BALTIC ASTRONOMY	64	0.093118611
REVISTA MEXICANA DE ASTRONOMIA Y ASTROFISICA	39	0.077757085
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	131	0.063035619
ASTRONOMICAL JOURNAL	218	0.062312612
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	686	0.042681773
ASTRONOMY REPORTS	65	0.039065743
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	45	0.033422965
SPACE SCIENCE REVIEWS	26	0.029634918
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	90	0.02572539
ASTROPHYSICAL JOURNAL LETTERS	160	0.02077656
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	36	0.013586115
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	23	0.011067322
ASTROPHYSICS AND SPACE SCIENCE	176	0.010950426
ASTROPHYSICAL JOURNAL	1856	0
ASTRONOMY & ASTROPHYSICS	1359	0
Area8		
PHYSICAL REVIEW D	5208	0.700439718
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	31	0.004169284
CLASSICAL AND QUANTUM GRAVITY	8	0.002117529
JOURNAL OF COSMOLOGY AND ASTROPARTICLE PHYSICS	5	0.002030988
GENERAL RELATIVITY AND GRAVITATION	3	0.001218593
ASTROPHYSICS	3	0.001218593
NUOVO CIMENTO DELLA SOCIETA ITALIANA DI FISICA C-GEOPHYSICS AND SPACE PHYSICS	3	0.001218593
COMPTES RENDUS PHYSIQUE	3	0.000979516
ASTROPARTICLE PHYSICS	2	0.000653011
GRAVITATION & COSMOLOGY	1	0.000518518
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	3	0.00051445
NEW ASTRONOMY REVIEWS	1	0.000134493
ASTROPHYSICS AND SPACE SCIENCE	2	4.43E-05
ASTRONOMY & ASTROPHYSICS	2	0
ASTROPHYSICAL JOURNAL	1	0
Area9		
ASTRONOMICAL JOURNAL	571	0.282314914
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF AUSTRALIA	86	0.216437853
ACTA ASTRONOMICA	50	0.172434283
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	104	0.133611565
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	909	0.097827391
NEW ASTRONOMY	48	0.094634582
ASTRONOMISCHE NACHRICHTEN	79	0.0821274
ASTRONOMY REPORTS	43	0.044702256
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	58	0.037861606
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	45	0.037454628
ASTROPHYSICAL JOURNAL LETTERS	159	0.035713223
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	42	0.020765721
ASTROPHYSICS AND SPACE SCIENCE	81	0.008717292
ASTROPHYSICAL JOURNAL	1073	0
ASTRONOMY & ASTROPHYSICS	1051	0
Area10		
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	217	0.086427698
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	783	0.067881878
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	82	0.054979695
ASTRONOMISCHE NACHRICHTEN	65	0.054433948
ADVANCES IN SPACE RESEARCH	72	0.048274854
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	64	0.033654761
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	49	0.03285372
NEW ASTRONOMY REVIEWS	58	0.030499627
PHYSICAL REVIEW D	49	0.025766927
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	49	0.025766927
ASTROPHYSICAL JOURNAL LETTERS	135	0.024426497
ASTRONOMICAL JOURNAL	50	0.019914216
ASTROPHYSICS AND SPACE SCIENCE	106	0.009189628
ASTROPHYSICAL JOURNAL	1332	0
ASTRONOMY & ASTROPHYSICS	897	0
Area11		
NUOVO CIMENTO DELLA SOCIETA ITALIANA DI FISICA C-GEOPHYSICS AND SPACE PHYSICS	105	0.152244762
PHYSICAL REVIEW D	117	0.0561696
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF THE PACIFIC	59	0.055745158
MONTHLY NOTICES OF THE ROYAL ASTRONOMICAL SOCIETY	596	0.047172325
CHINESE JOURNAL OF ASTRONOMY AND ASTROPHYSICS	70	0.042848482
ASTRONOMICAL JOURNAL	88	0.031998146
ASTRONOMY LETTERS-A JOURNAL OF ASTRONOMY AND SPACE ASTROPHYSICS	58	0.027844759
INTERNATIONAL JOURNAL OF MODERN PHYSICS D	56	0.026884595
ASTROPHYSICAL JOURNAL LETTERS	162	0.026760324
ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES	43	0.026321211
NEW ASTRONOMY REVIEWS	51	0.024484185
PUBLICATIONS OF THE ASTRONOMICAL SOCIETY OF JAPAN	49	0.017817149
ASTROPHYSICS AND SPACE SCIENCE	115	0.009102042
ASTROPHYSICAL JOURNAL	1459	0
ASTRONOMY & ASTROPHYSICS	589	0

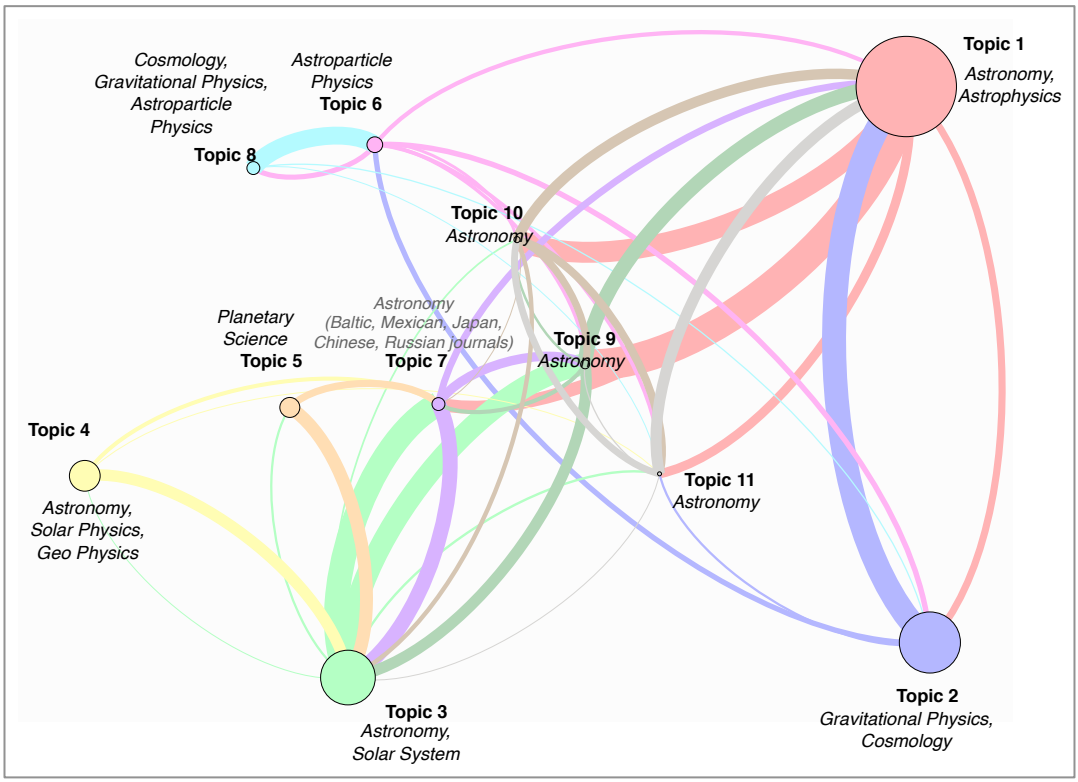


Figure 2: Topic affinity network. Node size indicates number of documents. Link strength indicates relative preference given by publications in one topic to cite publications in another. Links are directed: they are colored by their source node and curve clockwise away from it.

To further validate these hypotheses, a review of the topic contents and interpretation of the topic affinity links by experts could be insightful. Further, an extension of the data set backward in time to show the temporal evolution of affinity links could be informative. This would allow matching the evolution of affinity links over time to reports by experts about major research developments in this domain that may affect the interlinking between topics. One challenge in such an undertaking is that not just the linkages between topics evolve over time, but so does the identity of topics itself.

Conclusions

The topology of the affinity network highlights cognitive links between the topics extracted by our method from the astronomy and astrophysics data set. The interesting question in the context of the special session on the comparison of topic extraction algorithms will be what other cognitive features of this literature will be highlighted, if the affinity network is constructed for alternative groupings of documents into topics produced by other topic extraction algorithms. We suggest that this method of investigating the nature of differences between alternative topic extraction results is useful, in particular for cases where the topic size distribution is such that the large majority of documents, 80-90% is concentrated in 10-30 topics. For more granular topic extraction results the affinity network visualization is likely to become too unwieldy to interpret.

Acknowledgements

We gratefully acknowledge funding from SMA 1258891 EAGER: Collaborative Research: Scientific Collaboration in Time, as well as a travel grant by the intergovernmental framework for European Cooperation in Science and Technology (COST, Action: TD1210).

References

Boyack, K. W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404.

Crane, D. (1972). *Invisible Colleges - Diffusion of Knowledge in Scientific Communities*. The University of Chicago Press.

Ding, Y. (2011). Community detection: Topological vs. topical. *Journal of Informetrics*, *5*(4), 498–514.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften - die soziale Ordnung der Forschung*, Volume 906 of Campus Forschung. Frankfurt: Campus Verlag.

Morris, S. & Van der Veer Martens, B. (2008). Mapping research specialties. *Annual Review of Information Science and Technology*, *42*(1), 213–295.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.

Rosvall, M. & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123.

Velden, T. (2013). Explaining field differences in openness and sharing in scientific communities. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 445–458. ACM.

Velden, T. & Lagoze, C. (2013). The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology*, *64*(12), 2405–2427.

Velden, T. & Lagoze, C. (2009). Patterns of collaboration in co-authorship networks in chemistry - mesoscopic analysis and interpretation. In Larsen, B. & Leta, J. (eds.), *Proceedings of ISSI 2009 - the 12th International Conference of the International Society for Scientometrics and Informetrics, Rio de Janeiro, Brazil, July 14-17, 2009* (2 volumes).

Velden, T., Haque, A. & Lagoze, C. (2010). A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. *Scientometrics*, *85*(1), 219–242.

Zuccala, A. (2006) Modeling the invisible college. *Journal of the American Society for Information Science and Technology*, *57*(2), 152 – 168.

Time & Citation Networks

James R. Clough and Tim S. Evans

{james.clough09, t.evans} @ imperial.ac.uk
Imperial College London, Centre for Complexity Science, South Kensington Campus, London SW7 2AZ (U.K.)

Abstract
Citation networks emerge from a number of different social systems, such as academia (from published papers), business (through patents) and law (through legal judgements). A citation represents a transfer of information, and so studying the structure of the citation network will help us understand how knowledge is passed on. What distinguishes citation networks from other networks is time; documents can only cite older documents. We propose that existing network measures do not take account of the strong constraint imposed by time. We will illustrate our approach with two types of causally aware analysis. We apply our methods to the citation networks formed by academic papers on the arXiv, to US patents and to US Supreme Court judgements. We show that our tools can reveal that citation networks which appear to have very similar structure by standard network measures, turn out to have significantly different properties. We interpret our results as indicating that many papers in a bibliography were not directly relevant to the work and that we can provide a simple indicator of the important citations. We suggest our methods may highlight papers which are of more interest for interdisciplinary research. We also quantify differences in the diversity of research directions of different fields.

Background
Bibliometrics has a long tradition of dealing with citation networks from a network point of view as Price's model (Price, 1965) shows. The recent explosion of interest in network analysis in other fields has led to development of existing methods and introduced many new techniques. However most network methods assume static graphs where time plays no explicit role even if the underlying data is almost always evolving. Time can be incorporated into a network representation in two main ways. If we assign a single time to each edge we have a *Temporal Edge Network*. Such networks have received considerable attention (Holme & Saramäki, 2012). For instance they form a useful representation for the pattern of communications between individuals. Alternatively in *Temporal Vertex Networks* each node carries a single time. The citation network provides a natural example of the latter as each paper has its publication date. Here then we will focus on the analysis of this second type of temporal network, using the bibliometric context of citation networks to motivate our work.

The causal structure of citations plays a central role in bibliometric analyses. At the simplest level understanding the different time scales for citation patterns seen in different research fields is known to be essential. In Price's model (Price, 1965) vertices appear in a fixed order, reflecting the order of publication of real citation networks. Price's model captures the essential nature of a citation; they are always from newer to older papers. Applying Price's growing network model to other contexts where time plays a different role makes no sense e.g. links between web pages are not constrained by the age of a web site.

The constraints imposed by time are very different from the spatial constraints. Network science has few tools specifically developed to work with temporal vertex networks. However as part of our work we adapt results found in other areas: discrete mathematics, quantum gravity, and in computer science. Bibliometrics asks very different questions about such networks so applying these ideas is not always straightforward.

Our hypothesis is that existing network measures do not account for the constraint of time. So we have embarked on a programme to develop new temporally aware network measures and to prove their utility in the context of citation networks.

Methods and Data

Our networks are defined such that each node has a unique time. Edges can only exist from a younger to an older node, see Figure 1. Citations between academic papers are a good example, patents and court rulings have similar citation structures. All edges are directed, but the arrow of time also ensures that such networks will have no loops (acyclic) provided you follow the direction of the edges. The formal name for such a network is a *Directed Acyclic Graph* or *DAG* for short.

In practice, citation data is not exactly a DAG but we found that citations in the 'wrong' direction form less than 1% of our data so they should have a limited effect on any conclusions. We construct a true DAG by dropping any such acausal citations.

We have used a variety of data sets in our work (Clough et al., 2015, Clough & Evans, 2014). We have used citation information on the arXiv repository taken from two independent different sources. This allows us to check that our results are robust against any differences in citation extraction. First we use the KDD cup data (2003) which covers the first ten years of the hep-ph and hep-th sections (theoretical and phenomenological particle physics respectively). We have also looked at a separate version which covers all sections of arXiv up to 2013 which was derived from paperscape.org they also form a citation network.

We have also studied the citation network of around 4,000,000 US patents between 1975 and 1999 (Hall, Jaffe, & Trajtenberg, 2001). Finally we worked with the network defined by about 25,000 judgements of the US Supreme court 1754 to 2002 (Fowler & Jeon, 2008).

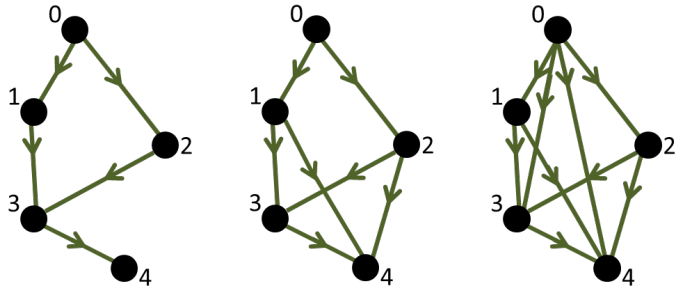


Figure 1 The unique transitively reduction (left) and transitive completion (right) of the citation network (a Directed Acylic Graph or DAG) shown in the centre. All casual relationships implied by an edge in the central network appear as an explicit edge in the right hand network. The edges in the left hand network are the least required to capture all these causal relationships.

Transitive Reduction (TR)

Our first example of a network operation, which takes account of the constraint of time, is Transitive Reduction (*TR*). In TR, links are removed provided that they leave the connectivity of every pair of nodes unchanged. That is if there was a path between a given pair of nodes (respecting the direction of the links) before TR, there will still be at least one such path after TR. This process can be defined on any network but for DAGs it is guaranteed to produce a unique result, see Figure 1. Algorithms for this procedure are well known in computer science but we found basic implementations in python were sufficient even for our largest networks (Clough et al., 2015)

Once we have this essential causal core of our citation network we illustrate our approach with two simple measures: the fraction of edges lost in the TR process and a comparison of the citation count of papers before and after TR.

Dimension

In bibliometrics, we often place papers in different fields as there is great interest in understanding the relationships between topics, as illustrated by maps-of-science (such as

Börner et al., 2012). It is natural to ask if we can assign a sense of dimension to such 'topic' spaces. A high dimension would indicate that researchers can develop work in several independent directions, a low dimension indicates that all the work in that field is tightly linked with little independence. There are some standard ways to assign an effective dimension to a network but these all assume that all directions are similar, just as moving left/right or forwards/backwards is the same for a ball on a flat table. Unfortunately, none of the measures used in the network science literature take account of time, which is a very different sort of dimension. Given that temporal information is an essential part of the definition of a citation network, we must work with a different type of measure. Our work (Clough & Evans, 2014) draws on inspiration from work in discrete mathematics on *posets* (partially ordered sets, e.g. Bollobás & Brightwell, 1991) and from the Causal Set programme of quantum gravity (e.g. Reid, 2003).

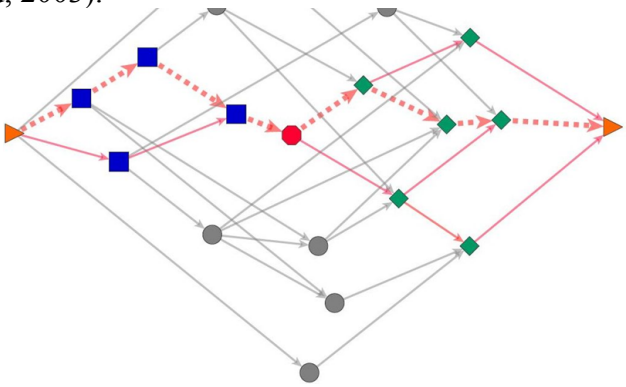


Figure 2 An illustration of the box counting method to find dimension. Here the source and the target papers (triangles at left and right respectively) define an interval of N=19 papers - the other vertices shown here. The edges represent the transitively reduced citation network of all twenty paper. The midpoint is shown as the red circle in the centre. It defines two sub-intervals N_1=4 (blue squares) on the left and N_1=6 on the right (green diamonds). This gives D=2.16 and D=1.61 as our dimension estimates. The example was generated by throwing points down with one space and one time coordinate chosen at random, i.e. D=2.

Our first approach is a simple box counting method (Reid, 2003). We first choose a pair of papers, the source and target nodes, at random. We then find the *interval* defined by the source and target nodes, which is the set of all N papers which lie on a path between source and target. As always our paths must respect the direction of time. Next we find the midpoint, a node chosen such that two sub-intervals defined by source and midpoint, and by midpoint and target nodes, are roughly equal size $N_1 \approx N_2$. It then follows that we should expect the 'length' scale of our two smaller intervals interval to be roughly half that of the large interval. Assuming papers are scattered at equal density in our data, we can use the number of points in an interval as a measure of the volume in the space-time. It then follows that the ratio of the number of points from small to large interval should scale as $N_1/N \approx N_2/N \approx 2^{-D}$. By analysing many intervals within one academic field the space-time dimension D (one time and $(D-1)$ topic space dimensions) of that field may be found.

The second method we use here is the Myrheim-Meyer dimension estimator (see Reid, 2003 for references). To do this we again pick a source and target paper. We then count the number causally connected pairs P in the interval defined by our source and sink which contains N nodes and these are related by $(P/N^2)=\Gamma(D+1)\,\Gamma(D/2)\,/\,(4\,\Gamma(3D/2))$ where $\Gamma(x)$ is the standard Gamma function. This formula is derived for a large N by assuming points are sprinkled at uniform density in Minkowskii space-time. We have also used the same approach

to show that in a different type of space, the cube box space of Bollobás & Brightwell (1991) the formula is simply $P=N(N-1)/2^D$.

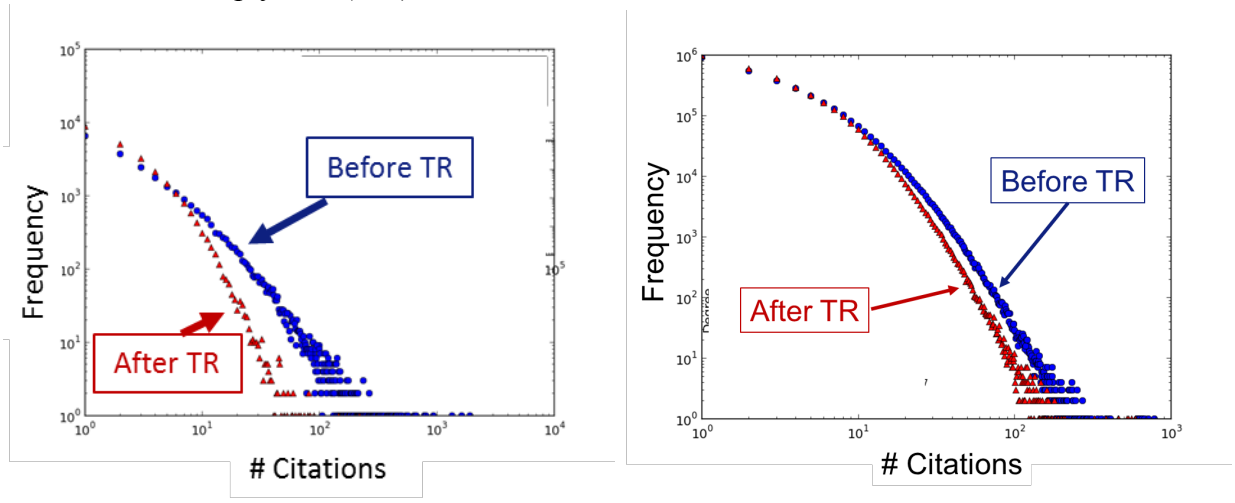


Figure 3 The citation count distribution before and after TR. On the left the results for the quant-ph section of arXiv (paperscape dataset) shows a significant change and an overall loss of around 80% of the edges. On the other hand, US patents shown on the right lose around 15% of edge and the citation distribution remains similar.

Findings

One of the most striking findings is that different types of citation network show very different behaviour under TR. All the citations networks of academic papers we have studied have shown a dramatic loss in the number of edges, typically around 70% to 80%. Further, it is the high cited papers which suffer the most as can be seen in Figure 3 for the hep-th arXiv where the citation distribution becomes noticeably steeper. On investigation it is clear that the edges which remain are those with the age difference between cited and citing papers. Interestingly citations in US supreme court judgements show a similar pattern (not shown) but US patents show only a moderate loss as shown in Figure 3.

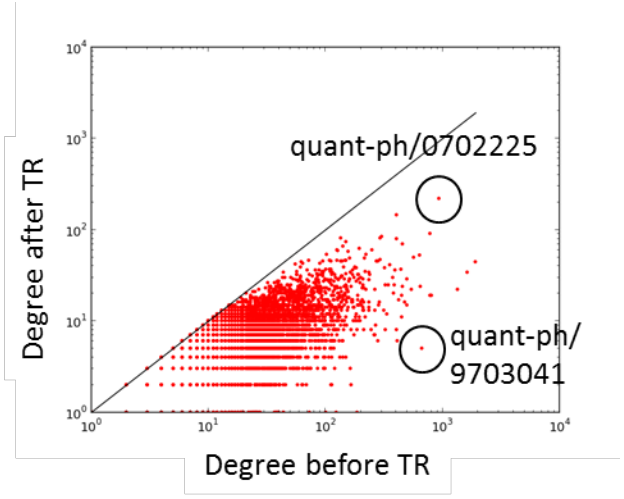


Figure 4 The citation count before and after TR for each paper in the quant-ph paperscape data.

Rather than looking at these bulk statistics we can look at the effect of TR on individual papers. Of course there are winners and losers. The example of the astro-ph arXiv section from paperscape.org highlights the different fates of two papers, see Figure 4. Paper quant-ph/9703041 (an older research paper on quantum entanglement) is one of the most highly cited papers with 664 citations yet TR shows that anyone using quant-ph/9703041 also took

information (directly or indirectly) from five other papers. On the other hand, paper quant-ph/0702225 (a more recent review of quantum entanglement) begins with a similar number of citations, 937, yet after TR it retains 219 of these.

We have also run our dimension measures on a variety of data sets. Our results are consistent whichever of the measures we use. What emerges is that we can generally give each field a well-defined dimension and that these are significantly different. For instance Figure 5 shows how papers in two parts of the arXiv repository have distinctive dimensions. For the arXiv we have found dimensions of about for hep-th (string theory), 3 for both hep-ph (particle physics) and quant-ph (quantum physics), and around 3.5 for while astro-ph (astrophysics).

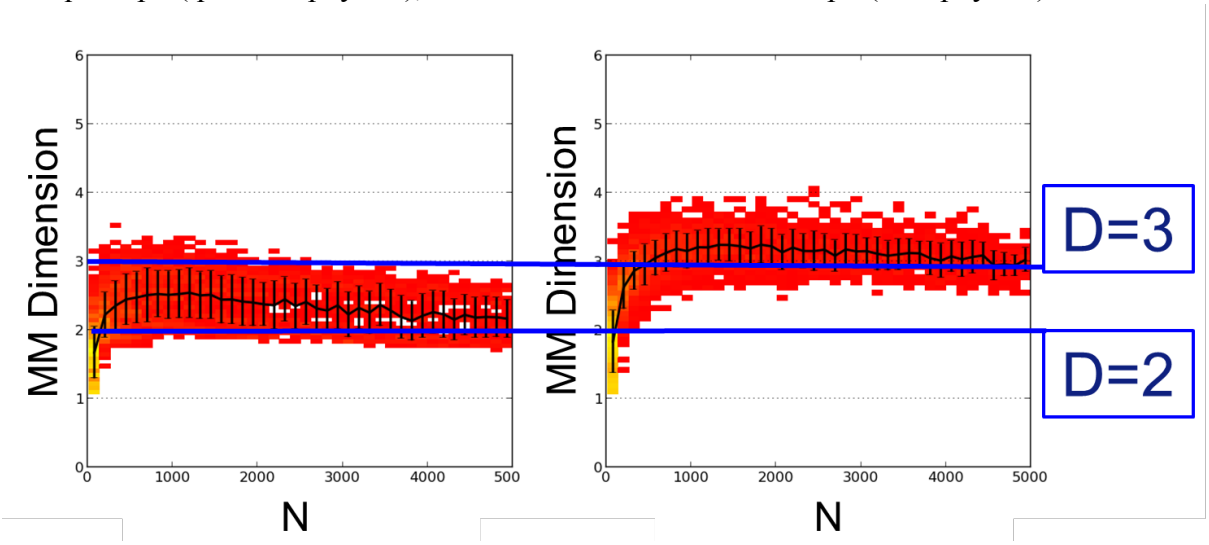


Figure 5 Dimension of two parts of the arXiv repository (KDD cup dataset) using the MM (Myrheim-Meyer) dimension estimator. Each point represents the dimension estimated from an number of intervals defined by two randomly chosen papers. On the left the hep-th section is seen to be of lower dimension than the hep-ph section shown on the right.

Discussion

For us TR captures the essential causal skeleton underlying the citation network. If information is flowing from older papers to newer papers and this is reflected in the bibliographies, then all the links in the transitively reduced network are the minimum needed for such a process. Of course in practice authors may use 'short cuts' and derive information directly from older papers, but equally such short cuts were not essential and therefore there is no reason to suppose they were important. We see TR as providing a lower bound on the actual route used by the flow of important information. To go beyond this, some sort of expensive semantic analysis is needed, be it via automatic methods or by hand.

In fact we believe the transitively reduced network may be much closer to the actual set of citations of direct relevance to a publication. We have found that around 80% of links between academic papers are removed by TR. Interestingly this matches the figure given by Simkin & Roychowdhury (2003, 2005) who suggest around 80% of citations are copied from intermediate works. Any citation which was copied will always be removed by TR.

Our suggestion is that TR could be an important way to reveal which papers were essential for the developments described in a new paper. Not surprisingly, these tend to be recent papers but it is still a surprise to find such a large fraction are removed. We have shown that there are big differences in the post-TR citation count of papers in similar fields with similar high citation counts. This could be a way to discriminate between papers and could provide an alternative basis for a recommendation system. For instance searches could be ordered by post-TR citation count. One hypothesis is that papers which retain a high citation count after

TR have been used across a wider range of topics. These are works which might be of more interest to researchers looking for papers outside their normal field of interest.

The behaviour of our patents and court citations also shows how TR can be a useful way to highlight different citation practices. The court data behaves in a way which is similar to that of academic papers with a large number of edges lost under TR. On the other hand, patents lose only a small fraction of their edges. The difference reflects the fact that for a patent, citations are a recognition of prior art, a legal necessity when writing a patent. However, as a patent is meant to be a novel development, they presumably try not to refer to earlier work so as to appear to be as different as possible from the literature. On the other hand, US Supreme Court judges seem to act like academic authors, citing older documents, which may have no direct relevance, along with the more recent documents, which have the latest distillation of this knowledge and are the real source of any innovation.

Our dimension measures again highlight difference between fields. We interpret the low dimension of the hep-th arXiv to suggest that string theory is a rather narrow field feeding off a few strands of research, at least when compared to hep-ph, quant-ph and astro-ph where research appears to be moving in a wider range of directions.

Conclusions

We have argued that citation networks require a new type of measure which takes account of the constraint imposed by time. We have given some examples of how this can be done and shown that they reveal some interesting features in real citation networks. We hope to add other measures and to improve the interpretation of our results by comparing them with non-network derived measures.

Acknowledgments

We would like to thank Damien George and Robert.Knegjens who provided us with access to their paperscape.org arXiv citation data. We also would like to acknowledge useful conversations with K.Christensen, J.Gollings, A.Hughes and T.Loach.

References

Bollobás B, & Brightwell G. (1991). Box-Spaces and Random Partial Orders. *Transactions of the American Mathematical Society*, 324, 59-72.

Börner, K.; Klavans, R.; Patek, M.; Zoss, A.M.; Biberstine, J.R.; Light, R.P.; Larivière, V. & Boyack, K.W. (2012). Design and update of a classification system: The UCSD map of science *PloS one*, 7, e39464

Clough, J.R.; Gollings, J.; Loach, T.V. & Evans, T.S. (2015). Transitive reduction of citation networks *Journal of Complex Networks,* 3, 189-203 http://dx.doi.org/10.1093/comnet/cnu039.

Clough, J.R. & Evans, T.S. (2014). What is the dimension of citation space? arXiv:1408.1274.

Fowler, J.H. &Jeon, S. (2008). The authority of Supreme Court precedent, *Social Networks*, 30, 16-30.

Hall, B., Jaffe, A. & Trajtenberg, M. (2001). The NBER Patent Citations Data File, NBER Working Paper Series.

Holme, P. & Saramäki, J. (2012). Temporal Networks. *Physics Reports*, 519, 97-125.

KDD Cup, 2003: *Network mining and usage log analysis*. Retrieved 1st October 2012 from http://www.cs.cornell.edu/projects/kddcup/datasets.html.

Price, D.S. (1965). Networks of Scientific Papers. *Science*, 149, 510-515.

Reid, D.D. (2003). Manifold dimension of a causal set: Tests in conformally flat spacetimes. *Phys. Rev. D*, 67, 024034

Simkin, M.V. & Roychowdhury, V.P. (2003). Read before you cite! *Complex Systems*, 14, 269-274.

Simkin, M.V. & Roychowdhury, V.P. (2005). Stochastic modeling of citation slips. *Scientometrics*, 62, 367-384.

Coming to Terms: A Discourse Epistemetrics Study of Article Abstracts from the Web of Science

Bradford Demarest[1], Vincent Larivière[2], Cassidy R. Sugimoto[3]

[1] *bdemares@indiana.edu*
Indiana University – Bloomington, School of Informatics and Computing, Bloomington, IN (United States)

[2] *vincent.lariviere@umontreal.ca*
Université de Montréal, École de bibliothéconomie et des sciences de l'information, Montreal, QC (Canada)

[3] *sugimoto@indiana.edu*
Indiana University – Bloomington, School of Informatics and Computing, Bloomington, IN (United States)

Abstract
This study investigates the relative power and characteristics of a set of social and epistemic terms to distinguish among disciplines of research article abstracts, using a corpus of 928,572 abstracts from 13 disciplines indexed by Web of Science in 2011. Applying the machine-learning approach to discourse epistemetrics using a sequential minimal optimization (SMO) algorithm, and a feature set of terms derived from Hyland's (2005) metadiscourse studies per Demarest and Sugimoto (2014), the current paper reports subsets of terms that best (and least) distinguish among disciplines, finding that the terms least able to distinguish among disciplines are rarely used and overwhelmingly adjectival or adverbial markers of authorial attitude, reflecting personal positioning, while terms best able to distinguish disciplines are mostly verbs frequently used as engagement markers, framing the generation of knowledge for the readership in ways that are standardized within disciplines (while varying among them). We plan to analyze the findings of the current research-in-progress from discipline-based as well as term-based perspectives, incorporating both into a two-mode network, as well as incorporating finer grained data for specific specializations to compare with the current higher-level disciplinary findings.

Conference Topic
Methods and techniques, altmetrics

Introduction
Understanding and depicting the relationships among different academic realms (whether disciplines, fields, specialisms, or a host of other divisions using some combination of social, epistemological, and institutional aspects) is a well-studied subarea of scientometric (Leydesdorff & Rafols, 2009). Initial forays into modeling disciplinary differences based on a core set of social and epistemic terms have yielded potentially promising results (Demarest & Sugimoto, 2013; Demarest & Sugimoto, 2014). However, no studies to date have used computational approaches to compare the abilities of specific social and epistemic terms to distinguish among disciplines. The current work-in-progress seeks to enact such a comparison, using a machine-learning approach to derive term differences between pairs of disciplines and by extension between a given discipline and all other disciplines under study. In finding the social and epistemic terms that best distinguish among academic disciplines, we hope to open new dimensions of analysis of the sciences through their texts.

Literature Review
There have been very few previous attempts to map the relatedness of academic disciplines based upon common social and epistemic terms. However, previous research of social and epistemic discourse usage in different academic disciplines as well as previous studies of document, journal, author, and discipline similarity or relatedness based on a variety of other measures guide the current study.

Differences in how academic disciplines employ language that positions the author in relation to the reader, the text itself, and previous scholars and works have been studied under various monikers, including stance (Biber & Finegan, 1989), metadiscourse (Hyland & Tse, 2004), appraisal (Martin & White, 2008), and attitude (Halliday, 1985). For the most part these differences have not been studied using automated quantitative methods (although cf. Argamon and Dodick, 2004), and in no cases have the resulting metrics been used as a basis for mapping the relatedness of disciplines. The current study draws upon Hyland's (2005) study of metadiscourse in a number of different disciplines, leveraging a set of words and phrases that Hyland (2005) found to be widely occurring in academic writing as our feature set for machine learning-based modeling of term differences among disciplines.

Previously, scholars have sought to map science based upon patterns of co-citation (Boyack, Klavans, & Börner, 2005) as well as topic, via ISI subject headings (e.g., Leydesdorff & Rafols, 2009). Other studies of similarity or relatedness have sought to compare multiple kinds of networks, including "bibliographic coupling, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks" (Yan & Ding, 2012, p. 1313). While the current work-in-progress focuses on a single type of similarity, it is with the intention of eventually adding to and comparing with these previously established measures of comparison. Furthermore, in order to create results that are comparable to previous work, we will also draw our data from the Web of Science, focusing specifically on the genre of scholarly articles, and use the high-level subject categories (although in future iterations of this study we hope to look at both higher and lower-level subject categories).

Methods

The current study analyzes all journal article abstracts from 13 disciplines contained in the Web of Science from 2011, totaling 928,572. Table 1 provides an overview of disciplines and counts of abstracts in the data corpus.

Table 1. Counts of abstracts by discipline.

Discipline	Abstracts
Engineering and Tech	172949
Biomedical Research	153166
Chemistry	129685
Physics	121702
Biology	93765
Earth and Space	70018
Mathematics	42685
Social Sciences	40463
Professional Fields	34590
Health	28343
Psychology	25802
Humanities	13673
Arts	1731
TOTAL	**928572**

For each abstract, relative frequencies were computed for 307 words or phrases taken from Hyland (2005). These terms fall into one or another of the following categories: hedges, boosters, attitude markers, engagement markers, and self-mentions. Hedges (e.g., "perhaps", "possible", "approximately") mitigate the certainty of an assertion, while boosters (e.g., "clearly", "obvious") amplify it. Attitude markers, such as "unexpectedly" or "unfortunately", frame assertions affectively, expressing the author's emotion regarding the

asserted facts, as distinct from their assurance of the facts' certainty. Engagement markers (such as "the reader" and "you", but also imperative verbs such as "consider" or "observe") address the reader explicitly or implicitly, and guide the reader to specific social and epistemic framing of an assertion (e.g., as an externally observable fact or as an idea intended for mental simulation). Finally, self-mentions, such as "I", "we", or "the author", serve as means for authors to insert themselves into the text, either as subjective actors or as social players (whether alone or as part of an authorial cohort).

After preparing the data, the Sequential Minimal Optimization algorithm (SMO) (Platt, 1998), a support-vector model classifier implemented in the WEKA v3.6.6 tool (Hall et al., 2009), was employed to create models distinguishing between each pair of disciplines based on the socio-epistemic features' relative frequencies. The resulting term weights for each model of discipline pairs were then normalized across the model, such that the absolute values of weights for a given discipline pair model would sum to 1. Model-normalized weights for each term were then averaged for each discipline across all discipline pairs for which the given discipline was a pair member. For the sake of standardization, negative term weights indicate a positive correlation with a given discipline (i.e., the more frequently the term appears in a text, the more likely this text belongs to the given discipline), while positive term weights indicate a negative correlation (i.e., the more frequently the term appears in the text, the less likely this text belongs to the given discipline).

Results

Due to space limitations, we eschew reporting the full 307 term set of results, focusing instead on the terms that most and least distinguish among disciplines. We discern these terms based upon the standard deviation of model-normalized average weights, as terms that discern well among disciplines will result in strong positive as well as negative weights, depending on which discipline is being modeled, while terms whose weights have small absolute values will in turn have smaller standard deviations, as all weights approach the 0 point.

Table 2 reports the 20 terms with the highest standard deviations of model-normalized average weights, as well as the 20 terms with the lowest standard deviations. While the results might at first blush suggest that the terms with the lowest standard deviations are part of a universal academic discourse, it is worth noting that many of the terms in the Bottom 20 list are exceedingly rare in the sample – out of 928,572 abstracts, "unbelievable" appears in 3 of them (although "shockingly" also appears in 3 abstracts; however, "unbelievable" is found in 2 engineering abstracts and one humanities abstract, suggesting that the scant data that exists shows no distinction between two otherwise fairly different disciplines). Also worth noting is that any terms that appeared in no abstracts at all are eschewed from the reported results.

However, the bottom 20 terms do provide some information about scholarly writing across the disciplines – the vast majority of these terms (19 out of 20) act as attitude markers; given the wide range of adjectives and adverbs available to describe the affective state of the author (and given that adjectives and adverbs are linguistic "open classes", i.e., new words can and are generated for these classes regularly), it is not surprising that such terms would be diffuse, rare, and not strongly indicative as individual terms.

Pivoting to consider the top 20 terms, the first notable characteristic is that where the bottom 20 terms tend toward adjectives and adverbs (as well as attitude markers), 19 of the top 20 terms are either self-mentions or engagement markers (and the latter for the most part are verbs). While nouns and verbs are also linguistic open classes, the use of verbs to describe the epistemic frame of scientific work here as well as the terms with which scientific authors refer to themselves can be seen to be more standardized within disciplinary communities, whereas the attitude markers of the bottom 20 terms are more personalized. The indicative

strength of self-mentions such as “we”, “my”, and “author”, as well as verbs like “argue” and “measure” also resonates with previous findings of Demarest and Sugimoto (2014), with “argue” and “my” serving as a strong indicator of philosophy and “measure” and “we” a better indicator of psychology and physics in dissertation abstracts as well.

Table 2. The top and bottom 20 social and epistemic terms for distinguishing among disciplines (ranked by standard deviation).

Top 20		Bottom 20	
Term	Standard Deviation	Term	Standard Deviation
we	0.009848	shockingly	0.0009166
argues	0.009686	view	0.0008793
prove	0.009614	disappointed	0.0008707
argue	0.009098	astonishingly	0.0008043
author	0.009063	!	0.0007801
showed	0.008494	incontestable	0.0007541
about	0.008138	knowledge	0.0007406
let	0.008044	incontrovertible	0.0007283
proved	0.008019	presumable	0.0007005
my	0.007908	unclearly	0.0006577
recall	0.007684	desirably	0.0006524
estimate	0.007646	amazed	0.0006068
review	0.007592	disappointingly	0.0006046
measure	0.007268	uncertainly	0.0004573
pay	0.007173	undisputedly	0.0003956
thought	0.007102	unbelievably	0.0003247
claims	0.006978	incontrovertibly	0.0002968
consider	0.006879	incontestably	0.0002821
shown	0.006687	astonished	0.0002649
set	0.006672	unbelievable	0.0001121

Another aspect of the findings to consider is that while the standard deviation values derive from the full set of model-normalized average weights, in some circumstances high standard deviation values can derive from a single outlier, while in others it derives from a more uniform spread of weights. Figure 1 depicts the model-normalized average weights for the top 20 terms ranked by standard deviation. Visual inspection reveals terms whose weights are more uniformly distributed (e.g., “author”), which suggest that they may serve as robust terms to distinguish among a variety of disciplines, while other terms (e.g. “let”, “prove”, and “proved”) serve as strong indicators of a single outlier discipline, with all other disciplines much more tightly clustered. As it happens, the terms “let”, “prove”, and “proved” provide a strong indication of mathematics as they occur more frequently in a text, in contrast to all other disciplines.

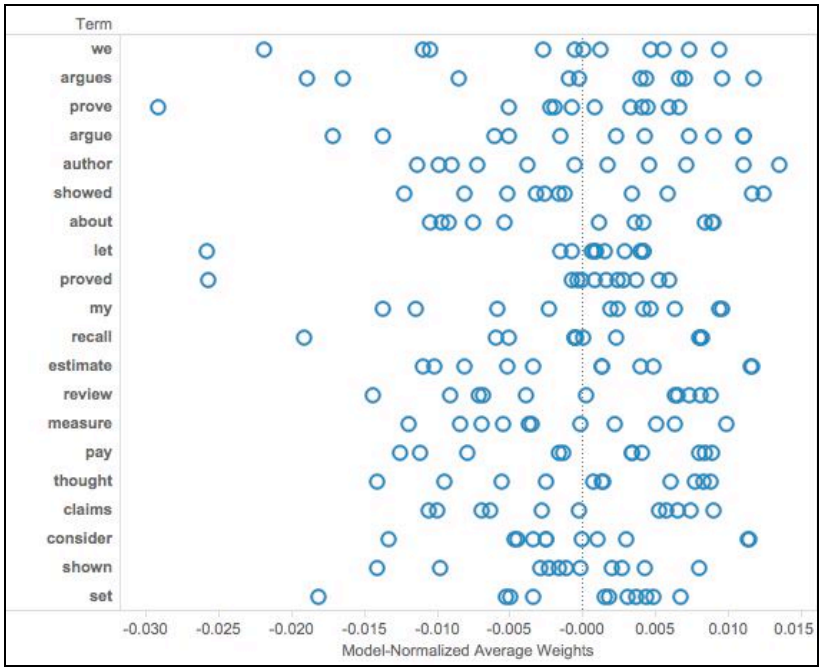


Figure 1. Model-normalized average weights (Top 20, ranked by standard deviation).

Future Directions

While the results of the current study-in-progress have focused on summary ranking and overall patterns of distribution of weights per term, our next goals in the near term are to more deeply tease apart trends as they appear for single disciplines as well as groups of disciplines, including the traditional groupings of soft vs. hard and pure vs. applied (Biglan, 1973). Further, we can derive overall measures of similarity among disciplines from the overall accuracy measures of the machine-learning models from which these terms are taken (per Demarest & Sugimoto, 2014), or more ambitiously we could seek to cast disciplines and terms in a bipartite network, to more fully grasp the interplay between different disciplinary communities and the words they use.

More distantly, we intend to use this same approach, in light of patterns and trends perceived at the current level of aggregations, to consider specializations, so that we may ask questions such as how broad the social and epistemic spread of specialized areas of study are within disciplines – are some disciplines more socially or epistemically diverse, and others more centralized? Do these degrees of variety reflect patterns of fragmentation and specialization in subject area? It is questions such as these that compels the current research-in-progress.

References

Argamon, S. & Dodick, J. (2004). Conjunction and modal assessment in genre classification: A corpus-based study of historical and experimental science writing. In *AAAI Spring Symposium on Attitude and Affect in Text*. Retrieved from http://www.aaai.org.proxyiub.uits.iu.edu/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-001.pdf?origin=publication_detail

Biber, D. & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, *9*(1), 93–124.

Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, *57*(3), 195.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*(3), 351–374. doi:10.1007/s11192-005-0255-6

Demarest, B. & Sugimoto, C. R. (2013). Interpreting epistemic and social cultural identities of disciplines with machine learning models of metadiscourse. *In Proceedings of ISSI 2013 (Vol. 2, pp. 2027–2030)*. Vienna. Demarest, B., & Sugimoto, C. R. (2014). Argue, observe, assess: Measuring disciplinary identities and

differences through socio-epistemic discourse. *Journal of the Association for Information Science and Technology*. doi: 10.1002/asi.23271

Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2011). Language and Ideology in Congress. *British Journal of Political Science*, *42*(01), 31–55. doi:10.1017/S0007123411000160

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold Press.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum International Publishing Group.

Hyland, K. & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics*, *25*(2), 156–177.

Klavans, R. & Boyack, K. W. (2009). Toward a consensus map of science. *Journal of the American Society for Information Science and Technology*, *60*(3), 455–476.

Leydesdorff, L. & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362.

Martin, J. R. & White, P. R. R. (2008). *Language of Evaluation: Appraisal in English* (First Edition.). Palgrave Macmillan.

Platt, J. C. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines. In *Advances in Kernel Methods - Support Vector Learning*. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4376

Yan, E. & Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology*, *63*(7), 1313–1326. doi:10.1002/asi.22680

Using Hybrid Methods and 'Core Documents' for the Representation of Clusters and Topics: The Astronomy Dataset

Wolfgang Glänzel[1,2] and Bart Thijs[1]

[1] *wolfgang.glanzel@kuleuven.be, bart.thijs@kuleuven.be*
KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
[2]Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
Based on a dataset on Astronomy & Astrophysics a hybrid cluster analysis has been conducted. Hybrid clustering was based on a combination of bibliographic coupling and textual similarities using Louvain method at two resolution levels. The procedure resulted in seven and thirteen clusters, respectively. The statistics reflect a high quality of classification. For labelling and interpreting clusters, *core documents* are used. The results of these two scenarios are presented, discussed and compared with each other. The two scenarios clearly result in hierarchical structures that are analysed with the help of a concordance table. Furthermore, the core documents help depict the internal structure of the complete network and the clusters.
This work has been done as part of the international project 'Measuring the Diversity of Research' and in the framework a special workshop on the comparative analysis of algorithms for the identification of topics in science organised in Berlin in August 2014.

Conference Topic
Methods and techniques (special session on algorithms for topic detection)

Introduction
Within the framework of the event series on 'Measuring the Diversity of Research' a special workshop on the comparative analysis of algorithms for the identification of topics in science was organised in Berlin in August 2014. A dataset downloaded from Thomson Reuters Web of Science covering the annual volumes 2003–2010 was shared with all contributors in order to test the various algorithms and techniques and to compare the results of the different approaches. On the basis of the shared Astronomy & Astrophysics dataset the following analysis has been conducted at our institute. In particular, the topic structure of the subject defined by the set was analysed using two different but related techniques. A cluster analysis was based on bibliographic coupling and textual similarity. And *core documents* (Glänzel & Czerwon, 1996) defined on the same links were used to represent topics within the subject and to depict the internal structures of both subject and clusters (cf. Glänzel & Thijs, 2011). Main results are presented in the following, but changing parameters of the algorithm and of the combination of the components leads to further results.
Currently a new and more robust method for the measurement of textual similarities and thus for the revision of the lexical component is in development. A comparison of the results of the present study with those of the new algorithm is part of the ongoing project and will be presented on a later occasion, when available.

Methodological aspects
The advantage of using hybrid lexical–citation based methods, notably of combinations of term-frequency and bibliographic coupling, has already been discussed in previous studies (e.g., Glenisson et al., 2005; Boyack & Klavans, 2010). However, at this level of aggregation (topics within the same field or discipline) we have encountered several specific problems that have already been reported in earlier studies in the context of the detection of emerging topics (e.g., Glänzel & Thijs, 2012). Terms and phrases might become less specific since they express common knowledge base and vocabulary while others might gain more 'information

value'. The most important TF-IDF keywords and terms alone are often not specific enough for topic description and labelling. Thus a larger set of terms is needed to describe topics at this level. A possible solution has already be discussed already in earlier studies (e.g., Glänzel & This, 2011): On one hand, depending on the level of aggregation *and* the discipline under study, the weight of the two components can be adjusted and, on the other hand, instead of the best TF-IDF terms *core documents* can be used to describe and label clusters. In order to apply the hybrid clustering we have only vertices with *positive* degree (i.e., documents with at least one link) taken into account. Furthermore, we have removed all papers with publication years outside the period 2003–2010. Table 1 shows the description of the dataset.

Table 1. The input dataset.
[Data sourced from Thomson Reuters Web of Science Core Collection]

Data	Documents	Percentage
Original dataset	111514	100.00
Not present in ECOOM Database	103	0.09
Publications in 2003-2010	110412	99.01
Excluded from all analysis	1205	1.08

We applied Louvain method (Blondel et al., 2008) using Pajek (Batagelj & Mrvar, 2003) to this dataset. The reason for this choice was that hierarchical clustering with Ward used in previous projects (e.g., Thijs et al., 2013) often results in a heterogeneous "hotchpotch" cluster of objects that can otherwise not be assigned. Therefore we decided to apply Louvain method. We conducted a hybrid clustering with two components: *bibliographic coupling* (BC) and *textual similarity* (TS), where we used a weight of 0.75 for BC and 0.25 for TS according to the algorithm described in Glänzel & Thijs (2011). In particular, the underlying similarity measure r is defined as the cosine of the linear combination of the underlying angles between the vectors representing the corresponding documents in the vector space model, i.e.,

$$r = \cos\left(\lambda \cdot \arccos(\eta) + (1-\lambda) \cdot \arccos(\xi)\right), \quad \lambda \in [0,1],$$

where η is the similarity defined on bibliographic coupling and ξ the textual similarity. The λ parameter defines the convex combination, $\arccos(\eta)$ and $\arccos(\xi)$, respectively, denote the two underlying angles. Furthermore, we have conducted the clustering at two resolution levels, namely 0.7 and 1.4. The results of these two scenarios will be presented and briefly discussed in the following section.

Results

The results using both resolution levels are briefly summarised in Table 2. The number of documents, that could not been clustered, is marginal. The number of clusters has almost doubled (from 7 to 13) with growing resolution. The solutions for the two resolution levels are presented in Tables 3 and 4. Except for the tiny cluster (#13) on atmospheric turbulence in the second solution, all clusters are of reasonable size. This is expressed by the frequency, i.e., the number of documents per cluster (columns 2–4). The description of the clusters, shown in the last column of the tables, have been derived from the most important TF-IDF terms and the titles of the *core documents*, where the core documents have been determined according to see Glänzel (2012) on the basis of the *degree h-index* of the hybrid document network. In particular, core documents are represented by core nodes, which, in turn, are defined as nodes with at least h degrees each, where h is the h-index of the underlying graph. Or, to express this simpler, degrees of documents are ranked in descending order and the h-core is formed by the documents the degrees of which do not undercut their rank value. This method has proved

efficient in *local* clustering, that is, in clustering of fields or disciplines, where the network h-core usually represents the order of magnitude of 1% of the total document set (see Glänzel, 2012).

Table 2. Description of parameters and results. [Data sourced from Thomson Reuters Web of Science Core Collection].

Number of vertices		108937
Number of edges		87602281
Density		1.5%
All Degree Centralization		0.13
Method		Louvain (Pajek)
Hybridity parameter		$\lambda = 0.75$
Resolution	0.7	1.4
Number of Clusters	7	15
Documents not Clustered	360	360
Modularity	0.61	0.49

Table 3. Scenario 1 (description of structures in the seven-cluster structure). [Data sourced from Thomson Reuters Web of Science Core Collection].

Cluster	Freq	Freq%	CumFreq	CumFreq%	Label
1	20634	18.7%	20634	18.7%	Star Clusters
2	12149	11.0%	32783	29.7%	Terrestrial planets/Extra Solar Planets
3	14365	13.0%	47148	42.7%	Solar Flares
4	17036	15.4%	64184	58.2%	Star Formation
5	20173	18.3%	84357	76.5%	Dark Energy
6	15023	13.6%	99380	90.1%	Gamma Ray Burst
7	10820	9.8%	110200	99.9%	Neutrino

Table 4. Scenario 2 (description of structures in the 13-cluster structure). [Data sourced from Thomson Reuters Web of Science Core Collection].

Cluster	Freq	Freq%	CumFreq	CumFreq%	Label
1	11569	10.5%	11569	10.5%	Star Clusters / Globular Clusters
2	9470	8.6%	21039	19.1%	Disk around a brown dwarf or young star
3	12163	11.0%	33202	30.1%	Extrasolar planetary systems
4	15060	13.7%	48262	43.8%	Solar Flares
5	6481	5.9%	54743	49.6%	Dark Matter Halo: Formation of galaxies
6	10075	9.1%	64818	58.8%	Star formation
7	7523	6.8%	72341	65.6%	Dark Energy
8	9005	8.2%	81346	73.8%	Astrophysical jets and accretion discs
9	10298	9.3%	91644	83.1%	Brane-world black hole
10	5503	5.0%	97147	88.1%	Radio Pulsars
11	2336	2.1%	99483	90.2%	Gamma Ray Burst
12	10224	9.3%	109707	99.5%	Neutrino
13	477	0.4%	110184	99.9%	Atmospheric turbulence

Table 5. Core-document representation of Cluster #5 based on h-core. [Data sourced from Thomson Reuters Web of Science Core Collection].

UT	Degree	Rank	Title
000261696000006	111	1	Non-linear isocurvature perturbations and non-Gaussianities
000278201600003	99	2	Non-Gaussianity of quantum fields during inflation
000260529800008	96	3	Conditions for large non-Gaussianity in two-field slow-roll inflation
000261260200020	88	4	A curvaton with a polynomial potential
000278201600004	86	5	Local non-Gaussianity from inflation
000238060100019	84	6	Non-Gaussianities in two-field inflation
000186983100013	83	7	Generalized chaplygin gas with alpha=0 and the Lambda CDM cosmological model
000246571300004	82	8	Cleaned 3 year Wilkinson Microwave Anistropy Probe cosmic microwave background map: Magnitude of the quadrupole and alignment of large-scale modes
000253980700030	82	9	Non-Gaussianity analysis on local morphological measures of WMAP data
000276102300001	81	10	Scale dependence of local f(NL)
000270036800016	79	11	Non-Gaussianity beyond slow roll in multi-field inflation
000235669800017	78	12	Testing primordial non-Gaussianity in CMB anisotropies
000259692800055	77	13	Anomalous CMB North-South asymmetry
000185760100005	76	14	WMAP and the generalized Chaplygin gas
000250363000004	75	15	Alignment and signed-intensity anomalies in Wilkinson Microwave Anisotropy Probe data
000221258900057	74	16	Numerical analysis of quasinormal modes in nearly extremal Schwarzschild-de Sitter spacetimes
000264762500065	74	17	Modeling gravitational recoil from precessing highly spinning unequal-mass black-hole binaries
000220092300012	73	18	Non-Gaussianity in the curvaton scenario
000242409800004	72	19	Non-Gaussianity of the primordial perturbation in the curvaton model
000242449600008	72	20	A numerical study of non-Gaussianity in the curvaton scenario
000245928000021	70	21	Exploring the properties of dark energy using type-Ia supernovae and other datasets
000248953800006	70	22	Primordial non-Gaussianity in multi-scalar slow-roll inflation
000253764800075	70	23	Further insight into gravitational recoil
000243171800001	68	24	Inflationary trispectrum for models with large non-Gaussianities
000252864000020	68	25	Non-Gaussianity in the modulated reheating scenario
000243171800040	67	26	Primordial trispectrum from inflation
000275514800001	67	27	Disks in the sky: A reassessment of the WMAP ""cold spot""
000278201600005	67	28	Use of delta N formalism-difficulties in generating large local-type non-Gaussianity during inflation
000221258900023	66	29	Curvature and isocurvature perturbations in a three-fluid model of curvaton decay
000221277400044	66	30	Dirac quasinormal modes of the Reissner-Nordstrom de Sitter black hole
000266501900050	66	31	Trispectrum versus bispectrum in single-field inflation
000272271900003	66	32	The subdominant curvaton
000243725400002	65	33	The non-Gaussian cold spot in the 3 year Wilkinson Microwave Anisotropy Probe data
000244679500013	65	34	Mapping the large-scale anisotropy in the WMAP data
000255424300029	65	35	Generation and characterization of large non-Gaussianities in single field inflation
000235939700023	64	36	On the large-angle anomalies of the microwave sky
000250954900032	64	37	A note on the large-angle anisotropies in the WMAP cut-sky maps
000257290600085	64	38	Anti-de Sitter universe dynamics in loop quantum cosmology
000245405900001	63	39	Constraints on the generalized Chaplygin gas model from recent supernova data and baryonic acoustic oscillations
000183377200050	62	40	Generation of dark radiation in the bulk inflaton model
000188864800011	62	41	Large scale structure and the generalized Chaplygin gas as dark energy
...	...	...	...
000256378700020	60	59	A low cosmic microwave background variance in the Wilkinson Microwave Anisotropy Probe data
000259700200011	60	60	Consistency relations for non-Gaussianity

Table 5 lists the core documents of Cluster #5 of the first scenario with seven clusters as an example. The degrees given in the table also illustrates the role of core documents in the cluster: Core documents are by definition strongly interlinked with many other documents and therefore play a representative and central part in a network. And they are suited to depict the internal structure of the complete network, of a cluster or of parts of it. In this context Cluster #5 has not been chosen by chance. The core documents of this cluster form the centre of the structure. Links connecting core documents reveal the internal structure of both the field under study and the clusters as the links with other core documents of the same cluster as well as with those of other clusters are distinctly apparent. Beside this cluster, also cores documents of cluster 7 play a central part. This is shown in Figure 1. Core documents of cluster 5 are marked in pink, those of Cluster 7 in auburn.

By contrast, Figure 2 presents the concordance between the two scenarios. Indeed the two resolutions results in a different number of clusters as already have been shown in Tables 3 and 4. Now the question arises of whether the two approaches yield completely different structures or almost concordant hierarchic structures, where the choice of the resolution would go with merging and splitting clusters, respectively. The first case would, of course, be problematic and point to the possible inappropriateness of methodology, while latter case testifies consistency of the chosen method. Cluster concordance of the results of the two scenarios are visualised in Figure 2.

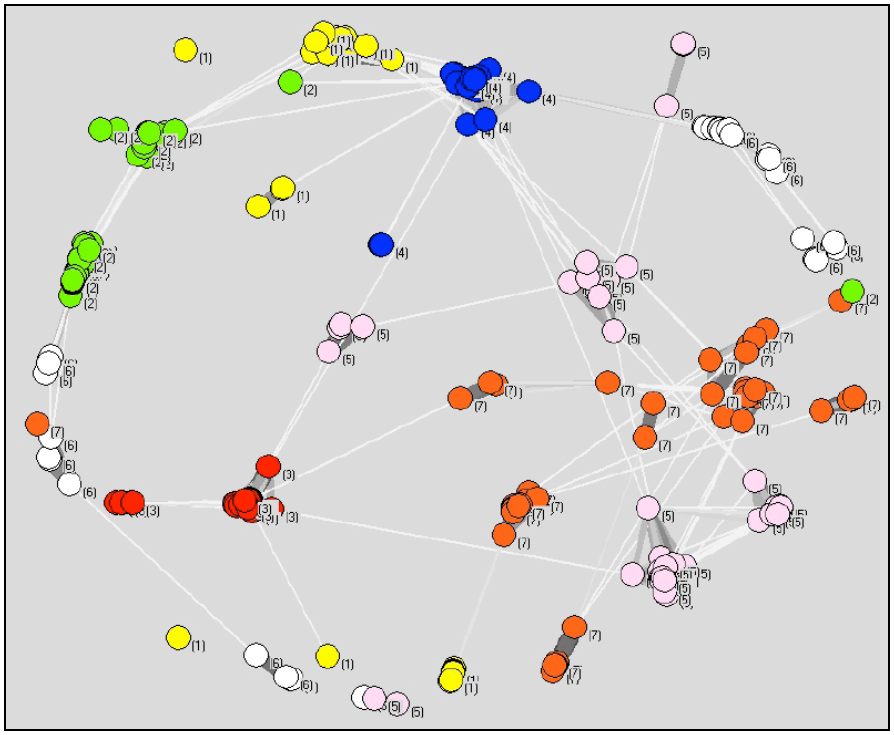


Figure 1. Structure of core documents in 7 clusters according to scenario 1 (Pajek with Fruchterman-Rheingold layout) [Data sourced from Thomson Reuters Web of Science Core Collection].

Hybrid Approach 13 Clusters \ Hybrid Approach 7 Clusters	1	2	3	4	5	6	7
1	94	0	0	5	0	1	0
2	96	1	0	2	0	1	0
3	1	95	3	1	0	0	0
4	1	0	91	0	0	7	0
5	1	0	0	86	2	1	10
6	0	0	0	100	0	0	0
7	0	0	0	2	98	0	0
8	2	0	0	4	0	94	0
9	0	0	0	0	97	0	2
10	1	3	1	1	42	51	0
11	1	2	0	1	0	95	0
12	0	0	0	1	2	1	95
13	5	61	24	0	2	0	7

Figure 2. Cluster concordance: scenario 1 – scenario 2 (overlap in %). [Data sourced from Thomson Reuters Web of Science Core Collection]

The document overlap in the corresponding clusters is expressed in per cents and, in order to facilitate interpretation, marked in different colours. Percentages sum up to 100% by rows. If one neglects the light-weight Cluster #13 in the second scenario, which actually represents just 0.4% of the total, one observes an almost perfect concordance of three clusters in scenarios 1 and 2 (#2 = #3, #3 = #4 and #7 = #12), one cluster splits up into two others (#4 = #5+#6) and finally two clusters split up into three clusters each, namely #5 = #7+#9+#10 and #6 = #8+#10+#11. Thus Cluster #10 in scenario 2 is the only one that breaches the strict hierarchy in the structures of the two scenarios. Its documents are almost equally distributed over Clusters #5 and #6 in scenario 1. The tiny one (#13) in the second

scenario can be considered a small sub-cluster of #2 in the first one, where it represents just slightly more than 2% of the documents of the total cluster.

Conclusions

Our main conclusions refer to two issues, firstly to the *clustering results* and secondly to the role of *core documents*. As to the clustering, both scenarios resulted in an almost perfect hierarchic structure. Cluster concordance and hierarchy was strong except for the cluster on 'Radio Pulsars' in the 13-cluster solution. This cluster was almost evenly spread over the clusters on 'Dark Energy' and 'Gamma Ray Burst' in the seven-cluster solution. Nevertheless, hierarchical assignment of 'Atmospheric Turbulence' in scenario 2 was also somewhat "fuzzy", but had a main concordance of more than 60% of documents with 'Coronal Loop' in the first scenario. In all other cases concordances were around or even above 90% document overlap.

The second group of remarkable observations refer to core documents. These documents represent the links across clusters as well as the internal topic structure of the clusters. In this context we have to repeat that core-document identification is in principle *independent* of clustering and thus does not require any cluster analysis or community detection, but it can be seamlessly integrated into clustering exercises, provided the same type of links, i.e., bibliographic coupling, co-citation, text similarity or hybrid, are used. Core documents reinforce the observation concerning centric results of the hybrid clustering. Core documents of the clusters on 'Dark Energy' and 'Neutrino' actually form the centre of the structure. The choice of the two resolution levels resulted in a hierarchic structure confirming the appropriateness of the applied method.

Acknowledgements

This work has been done as part of the international project 'Measuring the Diversity of Research' and in the framework a special workshop on the comparative analysis of algorithms for the identification of topics in science organised in Berlin in August 2014. The project and workshop series was jointly organised by the Humboldt Universität and Technische Universität Berlin. We would like to acknowledge their support of our study.

References

Batagelj, V. & Mrvar, A. (2003). *Pajek–Analysis and visualization of large networks*. In: M. Jünger & P. Mutzel (Eds.), Graph drawing software (pp. 77–103). Berlin: Springer.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.

Boyack, K. W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404.

Glänzel, W. & Czerwon, H.J. (1996), A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, *37*(2), 195–221.

Glänzel, W. & Thijs, B. (2011), Using `core documents' for the representation of clusters and topics. *Scientometrics*, *88*(1), 297–309.

Glänzel, W. & Thijs, B. (2012), Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics*, *91*(2), 399–416.

Glänzel, W. (2012), The role of core documents in bibliometric network analysis and their relation with h-type indices. *Scientometrics*, *93*(1), 113–123.

Thijs, B., Schiebel, E. & Glänzel, W. (2013), Do second-order similarities provide added-value in a hybrid approach? *Scientometrics*, *96*(3), 667–677.

Mining Scientific Papers for Bibliometrics: a (very) Brief Survey of Methods and Tools

Iana Atanassova[1], Marc Bertin[2] and Philipp Mayr[3]

[1] *iana.atanassova@univ-fcomte.fr*
Centre Tesniere, University of Franche-Comte, (France)

[2] *bertin.marc@gmail.com*
Centre Interuniversitaire de Rercherche sur la Science et la Technologie (CIRST),
Université du Québec à Montréal (UQAM), (Canada)

[3] *philipp.mayr@gesis.org*
GESIS, Leibniz Institute for the Social Sciences (Germany)

Introduction

The Open Access movement in scientific publishing and search engines like Google Scholar have made scientific articles more broadly accessible. During the last decade, the availability of scientific papers in full text has become more and more widespread thanks to the growing number of publications on online platforms such as ArXiv and CiteSeer (Wu, 2014). The efforts to provide articles in machine-readable formats and the rise of Open Access publishing have resulted in a number of standardized formats for scientific papers (such as NLM-JATS, TEI, DocBook).

Corpora

Different projects have been carried out to respond to the need of full-text datasets for research experiments (PubMed, JSTOR, etc.) and corpora.
E.g. the *iSearch* dataset was designed to facilitate research and experimentation in information retrieval, and specifically in aspects of task-based and integrated (a.k.a. aggregated) search. Its compressed size is about 46GB of documents in English from the physics domain that were collected from public libraries and open archive resources.

Semantic Web and Information Retrieval

Scientific papers are highly structured texts and display specific properties related to their references but also argumentative and rhetorical structure. Recent research in this field has concentrated on the construction of ontologies for citations and scientific articles.
CiTO (Shotton, 2010), the Citation Typing Ontology, is an ontology for the characterization of citations, both factually and rhetorically. It is part of SPAR, a suite of Semantic Publishing and Referencing Ontologies. Other SPAR ontologies are described at http://purl.org/spar/.

Statistical Analysis of Textual Data

Text Mining in R

Temis, an R Commander plugin (Bastin, 2013) provides integrated tools for text mining. Corpora can be imported in raw text. Another package is IRaMuTeQ (Ratinaud, 2009), a python application which uses the R libraries.

Correspondence Analysis

Correspondence analysis is a technical description of contingency tables and is mainly used in the field of text mining (Morin, 2006).
These tools could be very useful on the perspectives for the development of new text analytics approaches for bibliometrics.

Natural Language Processing Tools

Research in the field of Natural Language Processing (NLP) has provided a number of open source tools for versatile text processing.
The Apache *OpenNLP* library (Baldridge, 2005) is a machine learning based toolkit for the processing of natural language text. Written in Java, it is open source and platform-independent.
Stanford *CoreNLP* (Manning, 2014) integrates many NLP tools, including a part-of-speech (POS) tagger, a named entity recognizer (NER), a parser, a coreference resolution system, a sentiment analysis tool, and bootstrapped pattern learning tools. Stanford CoreNLP is written in Java and licensed under the GNU General Public License.
MALLET (McCallum, 2002) is a Java-based package for statistical NLP, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. It includes sophisticated tools for document classification: efficient routines for converting text to "features", a wide variety of algorithms (including Naïve Bayes, Maximum Entropy, and Decision Trees), and code for

evaluating classifier performance using several common metrics.

GATE (Cunningham, 2002) is open source free software for all types of computational tasks involving human language. It includes components for diverse NLP tasks, e.g. parsers, morphology, tagging, Information Retrieval tools, Information Extraction components for various languages.

CiteSpace (Chen, 2006) is a freely available Java application for visualizing and analyzing trends and patterns in scientific literature. It is designed to answer questions about a knowledge domain, which is a broadly defined concept that covers a scientific field, a research area, or a scientific discipline.

What is next?

Several studies examine the distribution of references in papers (Bertin, 2013). However, up to now full-text mining efforts are rarely used to provide data for bibliometric analyses. An example is the special issue on Combining Bibliometrics and Information Retrieval (Mayr, 2015). Novel approaches to full-text processing of scientific papers and linguistic analyses for Bibliometrics can provide insights into scientific writing and bring new perspectives to understand both the nature of citations and the nature of scientific articles. The possibility to enrich metadata by the full-text processing of papers offers new fields of application to bibliometrics studies like e.g. text reuse patterns in specific disciplines.

Working with full text allows us to go beyond metadata used in Bibliometrics. Full text offers a new field of investigation, where the major problems arise around the organization and structure of text, the extraction of information and its representation on the level of metadata. Unlike text-mining from titles and abstracts, full-text processing allows the extraction of rhetorical elements of scientific discourse, such as results, methodological descriptions, negative citations, discussions, etc. Scientific abstracts, by summarizing the text, provide only short, synthetic and thematic information.

Furthermore, the study of contexts around in-text citations offers new perspectives related to the semantic dimension of citations. The analyses of citation contexts and the semantic categorization of publications will allow us to rethink co-citation networks, bibliographic coupling and other bibliometric techniques.

Our aim is to stimulate research at the intersection of Bibliometrics and Computational Linguistics in order to study the ways Bibliometrics can benefit from large-scale text analytics and sense mining of scientific papers, thus exploring the interdisciplinarity of Bibliometrics and Natural Language Processing. Typical questions of this emerging field are: How can we enhance author network analysis and Bibliometrics using data obtained by text analytics? What insights can NLP provide on the structure of scientific writing, on citation networks, and on in-text citation analysis?

References

Baldridge, J. (2005). The Apache OpenNLP library. https://opennlp.apache.org/

Bastin, G., Bouchet-Valat, M. (2013). RcmdrPlugin. temis, a Graphical Integrated Text Mining Solution in R. *The R Journal 5*(1): 188–196

Bertin, M., Atanassova, I., Larivière, V., Gingras, Y. (2013). The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. *Proceedings of the 14th ISSI Conference (ISSI-2013)*, Vienna

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *JASIST, 57*(3): 359-377

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: an architecture for development of robust HLT applications. *Proceedings of the 40th Annual Meeting of the ACL*, pp. 168–175

Lykke, M., Larsen, B., Lund, H. & Ingwersen, P. (2010). Developing a Test Collection for the Evaluation of Integrated Search. Advances in Information Retrieval. *32nd European Conference on IR Research*, UK.

McCallum, A.-K. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J. & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations*, pp. 55-60

Mayr, P. & Scharnhorst, A. (2015). Scientometrics and Information Retrieval - weak-links revitalized. *Scientometrics, 102*(3): 2193-2199

Morin, A. (2006). Intensive use of factorial correspondence analysis for text mining: application with statistical education publications. *Statistics Educational Research Journal (SERJ)*

Ratinaud, P. (2009). IRaMuTeQ:Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires, http://www.iramuteq.org

Shotton, D. (2010). CiTO, the Citation Typing Ontology. *Journal of Biomedical Semantics, 1* (Suppl 1), S6.

Wu J., Williams K., Chen H.-H., Khabsa M., Caragea C., Ororbia A, Jordan D., Giles C. L. (2014). "CiteSeerX: AI in a Digital Library Search Engine," *Innovative Applications of AI, Proceedings of the 28th AAAI Conference*, pp. 2930-2937

A Multi-Agent Model of Individual Cognitive Structures and Collaboration in Sciences

Bulent Ozel

ozel@uji.es
Universitat Jaume I, Department of Economy, Castellon De La Plana (Spain)

Motivation

This research takes a multi agent perspective while simulating knowledge diffusion mechanism in science. Multi agent systems are systems that are composed of a large number of autonomous agents that are capable of interacting with each other. The autonomous agents are not controlled by a central mechanism, instead, their decision taking logics are part of their actions and they are decentralized, hence, they are able to make decisions in order to accomplish individual tasks (Wooldridge, 2009). In this research, a scientist who is situated within a co-authorship network is considered as an individual autonomous agent. Her decision process at picking another scientist to co-author a paper and outcome of such an interaction builds up our multi-agent system.

In a science network, if two scientists work on the same paper, then they are considered connected. The social interaction linkage between them is a possible channel for knowledge diffusion. In our model, each author is considered as an agent that is capable of working with other authors, choosing whom to work with and what subject to work on. In order to set-up initial environment of our multi-agent system we need to identify initial co-authorship network, as well as, we need to represent knowledge space of each individual author in the network. In order to capture a representation of an individual's expertise a set of keywords, which is driven from publications of the author is used to form the node set of the semantic network of that very individual. The semantic relations, namely the links, in between the keywords in the set are established by their co-occurrence on a published article.

There are a number of challenges at designing interaction and evolution of such multi agent system. The challenges are (i) being able to incorporate a dynamic social network perspective while modelling interactions in between agents, (ii) designing, simulating and examining various knowledge creation and diffusion mechanisms as the outcomes of agent-agent interactions.

The first challenge addresses a problem within multi-agent modelling research area. Computational simulation of social systems falls short at covering dense and multitude interactions in between actors. Majority of agent-agent interactions are implicitly and limitedly modelled via agent-agent interactions using environmental variables. This limitation is partly due to complexities at agent-agent interactions and mainly due to lack of empirically validated interaction mechanisms. In this work, we borrow and adopt models from social network literature. More specifically, we examine co-authorship networks and empirically validated interaction models within the field.

In the second challenge, we take a socio-cognitive approach. We model and exploit cognitive structure of each agent both at the incentives of individuals to select other agents to collaborate and at modelling the outcome of resulting interactions. Namely, agents purposefully interact to create and transfer new knowledge.

In addition to challenges mentioned above there are several implementation challenges to be addressed for the simulation model. First of all, not all agents in the population interact with each other at each run and preferences of interaction cannot be uniformly random. In the model, those ones who decide to collaborate compute the set of candidate collaborators autonomously. An agent's current knowledge space, and his/her ego network are taken into consideration at incentives to collaborate. For instance, literature suggests that repetition of joint collaborations follows a power law distribution (Morris & Goldstein, 2007) mimicking power law distribution of individual publication productivity. Likewise, propensity to collaborate with collaborator of an existing co-author is incorporated adopting transitivity property of social ties (Wellman, 1988). Another empirically validated model of social tie formation mechanism that is adopted is "preferential attachment". It is known that in a complex social network probability of a node to have a new connection is proportional to the connections it already has (Barabasi, 2002). At each round of the simulation each agent independently determines a candidate set of collaborators. This candidate set is formed employing above-mentioned mechanisms.

A second implementation challenge is how to incorporate knowledge of individual agents. Dynamic social network mechanism does not take actual knowledge space of individual into consideration. In other words, knowledge space of individuals does not play a direct role on the interactions. Besides, while social interaction mechanisms hint whom to pick to collaborate it

does not explain outcome of interactions. It is necessary to come up with empirically validated and sound models to represent what knowledge will be exchanged as the outcome of such social interactions.

Literature suggests that there are two competing social mechanisms, which may help to consider cognitive structure of individuals on the preferences of collaborators. They are 'cognitive distinctiveness' and 'cognitive similarity'. Cognitive distinctiveness or cognitive similarity of two agents is measured by comparing their knowledge bases. For a pair of agents when the distinctiveness is high then there are more possibilities for them to learn from each other. If their knowledge bases overlaps widely, the knowledge they can get from each other is limited (Carley, 1991). However, it is known that people, in some cases, tend to interact with people they are similar to; a tendency, which is known as homophily (McPherson et al., 2001). The experiments are devised to observe impact of these two competing models.

Implementation

As we have mentioned above, each author is represented as an agent. Each agent has its own individual memory, where its knowledge base and its co-authorship history is kept and updated throughout the simulation. Knowledge base of an agent is formed by set of keywords based on agent's publication records. This set of keywords is interrelated to each other. It is represented by a symmetric matrix. The matrix is a representation of cognitive structure of an agent. The entries of the matrix encode co-occurrence frequency of respective keywords. Co-authorship memory of an agent is a set of authors with whom the agent worked with on a publication.

Set of all the keywords that are gathered from all of the publications is represented as a weighted graph. If two keywords belong to the same publication, then they have a connection and weight of the connection is the number of the times they are used together. When entire set of publications for all agents is considered, then this graph is the cognitive structure of the entire network and it will be represented as an environmental component in the simulation.

It is certain that real agents learn from each other via collaboration, but this is not the only way of learning new things. They also learn from their readings, the workshops they attend and many other resources, etc. In order to represent all such various source of knowledge accumulation by agents, knowledge injection method is used. At each simulation time point, which is set as a year, a set of new keywords is added to the cognitive structure of entire population. A probabilistic model is adopted to update cognitive structures after injection of new keywords to the set. Betweenness centrality of existing keywords is used. The higher betweenness of a keyword, the higher chance it receives a new link.

Initial Findings and Future Work

Results from our initial experiments hint that in scenarios where agents are inclined to collaborate with cognitively dissimilar agents, then resulting collaboration structure rather mimics co-authorship relations seen within a research center. On the other hand, when cognitive similarity leads the incentives to pick a collaborator, then resulting co-authorship rather mimics network structures observed within domain of a journal in a field.

A large set of experiments is to be conducted to fully verify and validate our initial results, as well as, to discuss challenges addressed above.

There are a number of additional implementation challenges, which will be addressed and attempted as part of this ongoing research. They are (i) how to model when and in what circumstances multiple co-authorship occurs; (ii) at each run, not only new knowledge pieces but also new agents will be injected to the simulation. Knowledge base of those new agents will be composed of partially by a subset of keywords that is already in the current set and partially by new keywords that is not in the set. This approach will mimic arrival of new scientists in a field.

Bibliography

Barabasi, A. L., H. Jeong, Z. Neda, E. Ravasz, A. Schubert, & T. Vicsek. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications 311*(3-4), 590–614.

Carley, K. (1991). A theory of group stability. *American Sociological Review 56*, 331–354.

McPherson, M., Smith-Lovin, L. & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks, *Annual Review of Sociology 27*, 415-444.

Morris, S.A., & Goldstein, M.L. (2007). Manifestation of research teams in journal literature: a growth model of papers, authors, collaboration, coauthorship, weak ties, and Lotka's law. *JASIST*, *58*(12), 1764-1782.

Wellman, B. (1988). *Social Structures a Network Approach*, 19–61. Cambridge University Press.

Wooldridge, M. (2009). *Introduction to Multi-agent Systems*. John Wiley & Sons.

Hypothesis Generation for Joint Attention analysis on Autism

Jian Xu[1], Ying Ding[2], Chaomei Chen[3], Erjia Yan[3]

[1] *issxj@mail.sysu.edu.cn*
School of Information Management, Sun Yat-sen University, Guangzhou, Guangdong (China)

[2] *dingying@indiana.edu*
Department of Information and Library Science, Indiana University, Bloomington, Indiana (USA)

[3] *chaomei.chen@ drexel.edu and erjia.yan@drexel.edu*
College of Computing and Informatics, Drexel University, Philadelphia, Pennsylvania (USA)

Introduction

Every 20 minutes a new case of autism is diagnosed worldwide, which affects around 6% of the population of children. One of the major challenges in autism is how to reliably diagnose autism as early as possible so that early intervention can be imposed to dramatically change the whole situation, even lead to cure. Joint attention is among these early impairments that distinguish young kids with autism from normal kids. Joint attention is a transdisciplinary area which was studied in robotics, psychology, autism, and neuroscience. However, Due to the unaware of similar or related researches in different domains, researchers are unknowingly duplicating studies that have already been done elsewhere. On the other hand, due to the lack of domain knowledge in other domains, researchers can experience difficulties to understand the advances in other domains. To deal with this dilemma, generating hypotheses is considered a potentially effective way. It is a crucial initial step for scientific breakthroughs, and usually relies on prior knowledge, experience and deep thinking. Especially for transdisciplinary domains, generating hypothesis from literature in different but related disciplines can be exciting and highly demanded because it is no longer possible for domain experts in one domain to fully master the knowledge in another domain.

Although marked with several decades of research history, it is until recent years that hypotheses generating attracts more attention in transdisciplinary research domains. Swanson (1986) proposed ABC model to inference the literature-based hypotheses. Later on, Srinivasan (2004) presented open and closed text mining algorithms that are built within the discovery framework established by Swanson and Smallheiser. Their algorithms successfully generated ranked term lists where key terms representing novel relationships between topics are ranked high. Zhang et al. (2014) established the semantic Medline which biomedical entities and association are semantically annotated using concepts in UMLS. They assumed that the network motifs in the network can represent basic interrelationships among diseases, drugs and genes and reflect a framework in which novel associations can be derived as hypotheses to be further validated by domain experts. Spangler et al. (2014) presented a prototype system KnIT, which can mine the information contained in the scientific literature and represent it explicitly in a queriable network, and then further reason upon these data to generate novel and experimentally testable hypotheses. They applied their method to mine the publications related to p53 (a protein tumor suppressor) and are able to identify new protein kinases that phosphorylate p53. Malhotra et al. (2013) proposed a pattern matching approach for the detection of speculative statements in scientific text that uses a dictionary of speculative patterns to classify sentences as hypothetical. Their application on the domain of Alzheimer's disease showed that the automated approach captured a wide spectrum of scientific speculations and derived hypothetical knowledge leads to generation of a coherent overview on emerging knowledge niches. Song et al. (2007) constructed a Gene-Citation-Gene (GCG) network of gene pairs implicitly connected through citation and indicated that the GCG network can be useful for detecting gene interaction in an implicit manner. In this initiative, we use text mining approach to analyze related publications on joint attention from robotics, psychology, autism and neuroscience, to generate hypotheses which will be tested in the lab which collects eye contact and body movement sensor data. Here some preliminary results were reported and discussed.

Methodology

Due to the transdisciplinary character of "joint attention" research, we elaborately selected eight data sources (Wiley Online Library, ProQuest PsycINFO, Science Direct, Scopus, Web of Science, PubMed Central, Springer Link and Google Scholar) to maximize the coverage of the final dataset. The phrase "joint attention" is used to search separately on each data source.

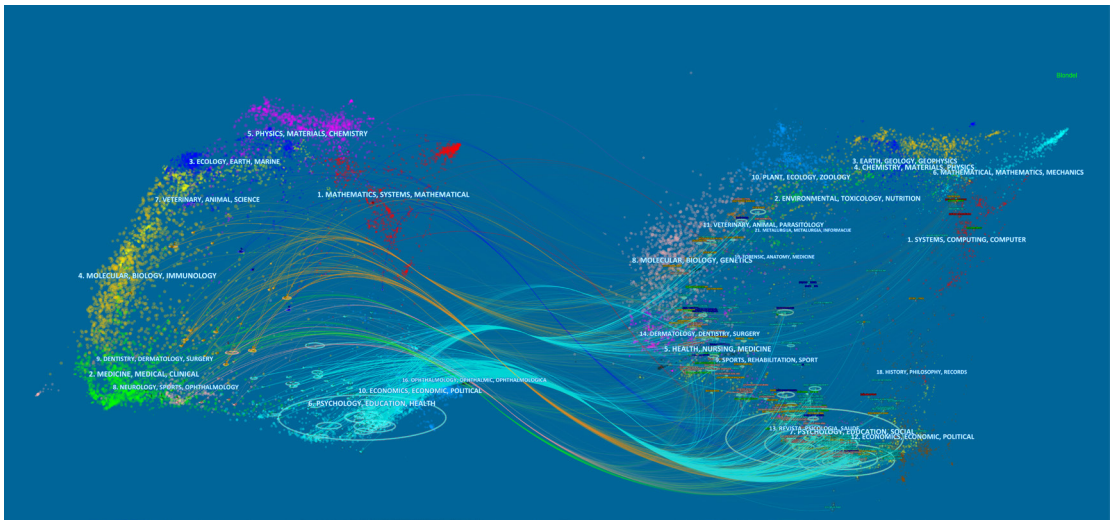


Figure 1. A dual-map overlay of "joint attention" search result from Web of Sciences.

Under the different download limitations, there are totally 39,845 records downloaded and 6,660 records left after remove duplicate records by the field "title". In the next step, keywords of each article in the dataset were extracted by using TF-IDF method. Then based on Keywords and other fields such as "journal name" and "citations", clustering were processed and relations among different clustering were analysed. By drawing the overall "research topic map", we can easily distinguish hot topics and their connections, and get to know their locations on the overall map. Then different dimensions (e.g., age, speech, language, and communication) were defined to analyse the distribution of current researches. Finally, from different dimension analysis aspects, research blind points were uncovered and new hypotheses were inferred, which will be tested in the lab.

Preliminary results

We tested a Web of Science query of "joint attention" (1,479 records) as a single dual-map overlay (Figure 1). Figure 1 shows the distribution of citing papers (left part) and cited papers (right part). Visualizations at this level are between journals, journal clusters, and overall maps. From the citation distribution and clustering results, we can identify the overall distribution of relevant sources and the most relevant targets (both ends with reference arcs). The label clustering result shows that the most popular domain discussing "joint attention" are Psychology, Education, Health, Medicine, Molecular, Economics, Mathematics, and Biology. It suggests that the Web of Science data is overwhelmingly dominated by a single journal *Journal of autism and developmental disorders*, with 169 papers. On the cited side, it is also the most cited journal in the dataset (6,640 citations). Other highly cited journals include *Child Development* (3,581 cites) and *Developmental Psychology* (2,328 cites).

Conclusions

This paper reports the ongoing effort on generating hypotheses in the transdisciplinary area of the joint attention research. We downloaded data from 8 separate data sources to maximize the coverage of "joint attention" related researches. Then text mining and visualization approaches were used to analyze related publications. Later stages of this research will generate hypotheses, which will be tested in the lab based on current research distributions on different predefined dimensions.

References

Swanson, DR. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, *30*(1), 7–18.

Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, *55*(5), 396-413.

Zhang, Y., Tao, C., Jiang, G., Nair, A.A., Su, J., et al. (2014). Network-based analysis reveals distinct association patterns in a semantic Medline-based drug-disease-gene network. *Journal of Biomedical Semantics*, 5:33.

Spangler, S., Wilkins, A.D., Bachman, B.J., Nagarajan, M., Dayaram, T., et al. (2014). Automated hypothesis generation based on mining scientific literature. *20th ACM SIGKDD* (pp. 1877-1886). New York:ACM

Malhotra, A., Younesi, E., Gurulingappa, H., & Hofmann-Apitius, M. (2013) 'HypothesisFinder:' A Strategy for the detection of speculative statements in scientific text. *PLoS Comput Biol*, *9*(7): e1003117.

Song, M., Han, N.G., Kim, Y.H., Ding, Y., & Chambers, T. (2014) Correction: Discovering Implicit Entity Relation with the Gene-Citation-Gene Network. *PLoS ONE*, *9*(1).

Chen, C., & Leydesdorff, L. (2014) Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the American Society for Information Science and Technology*, *65*(2), 334-351.

"What Came First – *Wellbeing* or *Sustainability*?" A Systematic Analysis of the Multi-dimensional Literature Using Advanced Topic Modelling Methods

Mubashir Qasim[1] and Les Oxley[1]

[1] *mq21@students.waikato.ac.nz,* [1]*loxley@waikato.ac.nz*
Waikato Management School, University of Waikato, Hamilton, New Zealand

Introduction

Both sustainability and well-being (SaW) are inter-dependent, inter-disciplinary, multi-dimensional, and international subject areas. However, people tend to interpret the subjects significantly differently based on their professional affiliation, academic background, geographical location etc., (Brunn, 2014; Roberts et al., 2013). A search of the SaW literature, using any scholarly search engine, generates results ranging from the thousands to millions creating a challenge for the researcher in picking the right papers; constructing a reasonable structure and synthesizing the vast material in order to conduct a comprehensive review of the literature. The work presented here relates to the use of a sophisticated method to exploit the explanatory power of metadata, attached to the results of a search query, to identify hidden patterns in the universe of given articles. The methods and metadata used to conduct the systematic analysis are briefly discussed under following headings.

Components of systematic literature analysis

Acquisition of data

Our quest begins with the analysis of key characteristics of metadata obtained from JSTOR *Data for Research* (DFR), which enables exploration of >9.2 million articles. We collected and analysed the metadata for a sample of 68,817 papers from DFR which related to SaW for this exercise. Metadata were generated against four queries with different sets of keywords as listed in Table 1. Analysis of the metadata was conducted in three steps: Step 1., analysis of keywords, subject and subject groups, disciplines and discipline groups, journals, authors and trends of publications (as presented in a recent study by (Brunn, 2014) but with slightly different approach). In Step 2., we applied the Latent Dirichlet Allocation (LDA) to study language differentiation between SaW themes. The main aim of this exercise was to identify complex hidden patterns in the data and present them in easily understandable ways. In Step 3., we used a reference manager software package called *Qiqqa* to identify key themes in the personal library and to identify seminal and frontier studies within each theme using cross references in the collection.

Table 1: Detail of search queries.

Query	Results	Search keywords	Search in
A	4,903	wellbeing OR well-being	Abstract
B	57,681	sustainability OR sustainable development	Title
C	5,472	sustainability; sustainable development; wellbeing; well-being	Any
D	761	sustainability OR sustainable development; well-being OR wellbeing	Abstract

Analysis of keyterms

We sampled 300 top keywords appearing in the corpus of each query to represent the frequently used language patterns in the subjects of SaW. The results are presented in the form of word-clouds in which the terms with high frequencies of occurrence are represented by the larger size of the word. Each word in the cloud indicates a dimension or issue in a subject (Jaewoo & Woonsun, 2014). Broadly discussed dimensions in the well-being literature include income, health, relationships, family, child, psychology etc., are correctly identified in our word-clouds.

Type of journals and subject group

Inter-relatedness of the SaW literature is established by confirming the large number of journals shared by SaW papers as suggested by (Mimno, 2012). Here, we extracted the names of the top 20 journals by number of articles in each query. Our analysis validates the assumption that many journals include papers on both aspects of the SaW literature. The interdisciplinary nature of the SaW literature is further established by similar categorization of SaW papers with respect to different subject groups.

Trends in publications

Many modern databases are devoted to tracking publications e.g., as Google Scholar, ISI Web of Science, JSTOR, SCOPUS, etc., and enable

scholars to perform quick and broad browsing of the literature (Hood & Wilson, 2003). Their expansions or contractions over time can indicate the interest of scholars in an area and the evolution of novel approaches (Adam, 2002; Casagrandi & Guariso, 2009).

In our analysis, we find the first article related to Query A, appears in 1919 and the number of publications remains trivial until the 1970's. Thereafter, a huge influx of papers begins in the late 1970's with 30 papers per year, peaking at 311 papers in 2012. In contrast, papers related to sustainability in Query B started much earlier with the first paper published in 1800. This number reaches to 50 papers per year in the next 100 years and steadily increase thereafter for another 50 years to around 250 papers per year in 1950. Post-1950, the number of scholarly articles grew five fold over the next five decades and peaked in 2005 at 1304 papers per year. Articles related to both SaW in Query C emerge in the late 1970's and grow exponentially over the next 40 years. As Query D is a subset of Query C they exhibit similar trends. A comparison of these trends with the papers in the entire DRF corpus of 9.3 million articles indicates the level of interest of the scholars over different years.

Authors of publications and places

Another way to consider the SaW literature is to analyse the country of the main author(s) of an article in order to answer the key question "what countries are leading the SaW agenda?" We select the top 20 authors in each set of documents based on their number of publications. Their country is established from the place of their affiliation at the time of publication. Our results show 74 unique authors from 12 different countries wrote 1,869 SaW paper. Not unexpectedly, 9 of these countries are developed OECD countries with the United States the home of 61% of SaW authors and 29% of this literature is produced by people from Europe, Canada and South Africa and rest of them are from Australia, India and Botswana.

Differentiating language using LDA

Finally, we conducted probabilistic analysis of the SaW literature using Latent Dirichlet Allocation (LDA) in order to establish underlying topics within the corpus of documents in each query (a topic is a set of co-occurring words). Our analysis helps understanding what sort of language is used within and across disciplines; what clusters of words happen to occur together; and how the use of language changes overtime. Results are shown by java based interactive visuals made in the programing language R. Each topic provides a clear structure to build a paragraph in a literature review and the cluster of topics gives a clear indication of the categories/themes within each set of documents.

Identification of seminal and frontier studies

Most dominant papers in our set of documents are identified using in-bound references assuming that heavily cited and highly ranked articles are the key papers in each collection. Identification of these articles provides the best starting point to begin the traditional literature review with. We used network diagrams using a reference manager called Qiqqa to conduct this exercise.

Validation of results

The results are validated using the metadata from another widely used scholarly source called Web of Science. Most of our results exhibit the same characteristics as the results of DFR data.

References

Adam, D. (2002). Citation analysis: The counting house. *Nature*, *415*, 726–729. doi:10.1038/415726a

Brunn, S. D. (2014). Cyberspace Knowledge Gaps and Boundaries in Sustainability Science: Topics, Regions, Editorial Teams and Journals. *Sustainability*, *6*(10), 6576–6603. doi:10.3390/su6106576

Casagrandi, R. & Guariso, G. (2009). Impact of ICT in Environmental Sciences: A citation analysis 1990–2007. *Environmental Modelling & Software*, *24*(7), 865 – 871. doi:http://dx.doi.org/10.1016/j.envsoft.2008.11.013

Hood, W. & Wilson, C. (2003). Informetric studies using databases: Opportunities and challenges. *Scientometrics*, *58*(3), 587–608. Kluwer Academic Publishers. doi:10.1023/B:SCIE.0000006882.47115.c6

Mimno, D. (2012). Computational Historiography: Data Mining in a Century of Classics Journals. *J. Comput. Cult. Herit.*, *5*(1), 3:1–3:19. New York, NY, USA: ACM. doi:10.1145/2160165.2160168

Roberts, L., Brower, A., Kerr, G., Lambert, S., McWilliam, W., Moore, K., Quinn, J., et al. (2013). A Good Life: How nature's ecosystem services contribute to the wellbeing of New Zealand and New Zealanders. Department of Conservation.

Jaewoo, C. & Woonsun, K. (2014). Themes and Trends in Korean Educational Technology Research: A Social Network Analysis of Keywords . *Procedia - Social and Behavioral Sciences, 131*(0), 171 – 176. doi:http://dx.doi.org/10.1016/j.sbspro.2014.04.099

Multi-Label Propagation for Overlapping Community Detection Based on Connecting Degree

Xiaolan Wu[1] and Chengzhi Zhang[2]

[1]*wuxiaolananhui@163.com*, [2]*zhangchz@istic.ac.cn*
Dept. of Information Management, Nanjing University of Science and Technology, Nanjing 210094 (China)

Introduction

With the growth of social media, social network analysis draws a great attention and becomes a hot research topic in the field of complex network, web mining, information retrieval, etc. An important aspect of social networks analysis is community structure (Newman, 2003).

In general, community detection methods are classified into two categories: overlapping methods (and non-overlapping methods (Hofman & Wiggins, 2008)). The former allows communities overlap, while the latter assumes that a network only contains disjoint communities. In this paper, we focus on the overlapping community detection. To find overlapping community, researchers use a wide variety of techniques, such as Clique Percolation Method, COPRA (Gregory, 2010), etc. COPRA is very fast, but the result of COPRA is nondeterministic, so we propose an improved COPRA with high determinacy in this paper.

An Improved COPRA Algorithm Based on Connecting Degree

To eliminate the nondeterministic of COPRA, we use Connecting Degree as definition 1.

Definition 1: Let v be a node on the undirected Graph $G(V;E)$, C is the set of overlapped communities on Graph, the connecting degree between node v and community $c(c \in C)$, denoted $C(v,c)$, be computed by the following formula (Duanbing, Mingsheng, Xia, 2013).:

$$C(v,c) = \frac{\sum_{u \in c} w_{vu}}{k_v} \qquad (1)$$

Where k_v is the degree of node v, w_{vu}=1 if there is an edge between node v and node u, zero otherwise. *

Connecting Degree can reflect the community tendency for a node to its neighbour communities, so we proposed a COPRA Based on Connecting Degree, named COPRA-CD. COPRA-CD works as follows: 1) To start, all nodes are initialized with a unique community identifier and a belonging coefficent setting to 1; 2) Each node updates its community identifier by the union of its neighbours labels, the corresponding belonging coefficient is obtained by normalizing the sum of the belonging coefficients of the communities over all neighbours. Then, comparing all the belonging coefficients and the parameter v, if all the belonging coefficients are less than v, calculating the connecting degree between node and its neighbour community, then only retain neighbour community with greatest connecting degree, else keeping these belonging coefficients that are more than v, then renormalize these belonging coefficients of remaining communities so that they sum to 1. After several iterations, if the stop criteria proposed by Gregory is satisfied, the propagation procedure stops; 3) Remove communities that are totally contained by others; 4) Split disconnected communities.

Experimental Results and Discussion

Test networks

At first, we do experiments on four real-world networks, whose information are shown in Table 1.

Table 1. General information of real networks

Networks	Description	Node&Edge
Karate	Zachary's karate club (Zachary, 1977)	34 &78
Dolphin	Lusseau's Dolphins (Lusseau, 2003)	62 & 159
Books	Books about US politics	105 & 441
Football	American College football union (Girvan, Newman, 2002)	115 & 616

Then we also test the performance of COPRA-CD on six LFR synthetic networks with various mixing parameter μ ranging from 0.1 to 0.6, the other standard configuration of LFR synthetic network used in this experiment is: n=1000, t_1=2, t_2=1, k=10, $\max k$=30, $\min c$=10, $\max c$=50, O_n=100, O_m=2.

Test metrics

To measure overlapping communities detection, Q_{ov} was be proposed by Nicosia et al (2009). The formulation of Q_{ov} as following:

$$Q_{ov} = \frac{1}{m} \sum_{c \in C} \sum_{i,j \in V} \left[\beta_{l(i,j),c} A_{ij} - \frac{\beta^{out}_{l(i,j),c} k_i^{out} \beta^{in}_{l(i,j),c} k_j^{in}}{m} \right] \qquad (2)$$

Where A_{ij} is the adjacency matrix of Direct Graph $G(E,V)$, C is the set of overlapped

* Corr. author: C. Zhang, Tel: +86-25-84315963.

communities, $l(i,j)$ is a link which starts at node i and ends at node j. $\beta_{l(i,j),c}$ is the belonging coefficient of $l(i,j)$ for community c, $\beta^{out}_{l(i,j),c}$ is the expected belonging coefficient of any possible link $l(i,j)$ starting from a node into community c, $\beta^{in}_{l(i,j),c}$ is the expected belonging coefficient of any link $l(i,j)$ pointing to a node going into community c. k_i^{out} is the out degree of node i, while k_j^{in} is the in degree of node j.

Test results and discussion

In order to show its performance, we compare three multi-label propagation algorithms, i.e., COPRA, COPRA-CD, and RC-COPRA. RC-COPRA stands for the version of COPRA with initialization using RC proposed by Wu et al. (2012). In our test, we run each algorithm 100 times on each network for the same value of parameter v. The average modularity result on real-world network was shown in Table 2, and the comparison performance on LFR synthetic networks was shown in Figure 1.

Table 2. Test Results on real-world Networks.

Networks	**COPRA (v=2)**	**COPRA-CD (v=2)**	**RC_COPRA (v=2)**
Karate	0.428	0.745	0.703
Dolphins	0.645	0.759	0.761
Books	0.826	0.815	0.830
Football	0.684	0.661	0.668
Networks	**COPRA (v=3)**	**COPRA-CD (v=3)**	**RC_COPRA (v=3)**
Karate	0.408	0.717	0.725
Dolphins	0.652	0.710	0.713
Books	0.830	0.822	0.827
Football	0.677	0.665	0.670

From Table 2, we find the modularity of CORPA is lower than that of other algorithms at the same v. At v =3, RC_COPRA algorithm gives better average modularity for every network, but at v=2, the modularity of RC_COPRA algorithm on Karate network is not better than that of COPRA-CD.

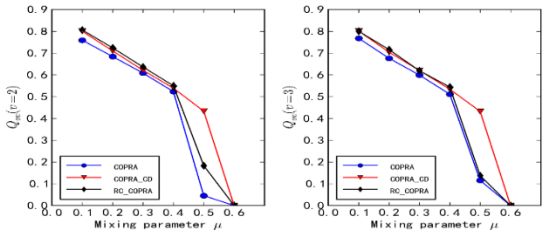


Figure 1. Experiment on synthetic networks.

As Figure 1 shows, when $\mu \leq 0.4$, all three algorithms show good performance. When $\mu =$ 0.5, LFR synthetic networks are very fuzzy, the overlapping community structure is not detected by COPRA and RC_COPRA, but detected by COPRA-CD, so we can conclude that for the given parameter, COPRA-CD is the most stable algorithm in these overlapping community detection algorithms.

Conclusions

In this paper, we propose COPRA-CD to uncover overlapping communities in social networks. Then we test it on four real-word networks and a group of synthetic networks. Experimental results show that both RC initialization and the connecting degree update strategy can bring improvements in quality, especially COPRA-CD has the best stability for fuzzy networks. In the future, COPRA-CD can be applied to analyze the community of co-author in paper.

Acknowledgments

This work is supported by the National Social Science Fund Project (grant number 14BTQ033).

References

Duanbing, C., Mingsheng, S., & Xia, L. (2013). Two-phase strategy on overlapping communities detection. *Computer Science, 40*(1), 225-228.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *PNAS, 99*(12), 7821-7826.

Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, *12*(10), 103018.

Hofman, J. M., & Wiggins, C. H. (2008). Bayesian approach to network modularity. *Physical review letters, 100*(25), 258701.

Krebs, V. (2008). A network of co-purchased books about US politics, www.orgnet.com.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270 (Suppl 2), S186-S188.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167-256.

Newman, M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E, 69*(6), 066133.

Nicosia, V., Mangioni, G., Carchiolo, V., & Malgeri, M. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics: Theory and Experiment,* (03), P03024.

Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, *435*(7043), 814-818.

Wu, Z.-H., Lin, Y.-F., Gregory, S., Wan, H.-Y., & Tian, S.-F. (2012). Balanced multi-label propagation for overlapping community detection in social networks. *Journal of Computer Science and Technology, 27*(3), 468-479.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research,* 452-473.

Reproducibility, consensus and reliability in bibliometrics

Raul I. Mendez-Vasquez[1] and Eduard Suñen-Pinyol[2]

[1]raul.mendez@fundaciorecerca.cat, [2]eduard.sunen@fundaciorecerca.cat
Fundació Catalana per a la Recerca i la Innovació (FCRI), Bibliometrics. Pg. Lluís Companys, 23 E-08010 Barcelona (Spain)

Introduction

Bibliometrics, and scientometrics in general, have been enjoying what seems to be an endless party. Far from stopping, the demand for bibliometric indicators from governmental bodies, administrators and researchers, is continuously growing. During this "give me the indicators" phase several solutions have been provided by the community, let say new and more sophisticated indicators, which in turn geared the transition to the present "give me the indicators, but really?" phase. The impressive penetration of bibliometric indicators in decision making processes, some of which are crucial in the development of researchers' careers, has also brought the necessity for credibility on bibliometrics, and more specifically, on how it is practiced. Examples of improper use of bibliometric indicators have raised skepticism among users of bibliometric reports[1].

As a scientific discipline, bibliometrics is subject to the principle of replication and corroboration of results, just like any other discipline. Precisely, the credibility of scientists goes hand in hand with the reproducibility of their results.

The objective of this contribution is to bring attention to the importance of the reproducibility of the number of publications as an indicator of the quality of bibliometric reports.

Methods

We compared the numbers of publications estimated by three units following this schema: CTWS vs. BAC (us) and SCIMAGO vs. BAC. Sixteen universities reported in the CTWS Leiden Ranking 2011/2012, and 20 universities reported in the Iberoamerican Ranking SIR 2012 produced by SCIMAGO were selected for the study. Source, type of document, language and period were matched in each comparison. The numbers of publications produced by the BAC were sourced with the National Citation Report for Spain (NCR), an *ad hoc* database built in July 2012 as a live extraction from the Web of Science that compiles all the publications between 1970 and 2011, with at least one address in Spain. The unification was performed by hand based solely on the information contained in the address field of the NCR. Hierarchy relationships such as university campuses and institutes, affiliated hospitals, etc, were reconstructed in the system. All the addresses were also located to a specific administrative unit (a city in the majority of cases). Both, the information on the organizational hierarchy and location of the addresses were used to unify the name variants of subunits whenever mother organizations were not present in the addresses. Changes in the structure of the organizations within the analyzed period were recorded in the system. The unification terminated when a precision higher than 97% was achieved.

Results

A simple examination of the number of publications of a small set of universities revealed important reproducibility issues, even when controlling for source dataset, period of time and the document type (Table 1. several rows and columns were removed). A positive and statistically significant correlation ($p<0.01$) was observed between the numbers of publications produced by the three units (CTWS & BAC, *rho* 0.785; SCIMAGO & BAC, *rho* 0.860). The dispersion around the regression line was smaller in the comparison between SCIMAGO & BAC, than between CTWS & BAC, suggesting the presence of an outlier observation, whose removal increased the correlation between CTWS and BAC (*rho* 0.975, $p<0,001$). The concordance between the rankings produced by the three units was also positive and high, (CTWS & BAC, *tau* 0.733, $p<0.001$; SCIMAGO & BAC, *tau* 0.705, $p<0.001$). Removing the mentioned outlier observation increased the concordance between the CTWS and BAC (*tau* 0.905, $p<0.001$)

Discussion

These technical issues may explain the observed variability in the number of publications.

1) Completeness of the unification. The CTWS unit selected the universities with at least 500 publications per year and extended the unification to the name variants occurring at least five times in the source dataset. The BAC unit aims at attributing all variants to corresponding universities. However, mistakenly attributed name variants and non-identified variants were allowed to a maximum of 3%. The CTWS unit attributed the publications

[1]The title of a number of articles published in Nature in 2010 reflect this position: "Assessing assessment", "Do metrics matter?", "How to improve the use of metrics", "Let's make science metrics more scientific". Available at: http://www.nature.com/news/specials/metrics/index.html.

based on author names, a procedure not performed by the BAC. SCIMAGO provides no information on the unification in the website of the report.

Table 1. Differences in the number of publications produced by three units.

	(A)	(B)	A-B	(A-B)/A	(C)	(D)	C-D	(C-D)/C
UB	7,672	11,804	-4,132	-53,86	15,290	16,222	-932	-6,10
UAB	5,992	9,319	-3,327	-55,52	13,262	13,200	62	0,47
UCM	6,616	8,863	-2,247	-33,96	13,240	12,160	1,080	8,16
UPM	2,323	8,813	-6,490	-189,2	7,458	11,096	-3,638	-48,78
UAM	5,236	8,034	-2,798	-53,44	10,591	10,873	-282	-2,66
UV	5,077	7,892	-2,815	-55,45	11,191	10,458	733	6,55
UGR	3,966	5,918	-1,952	-49,22	9,128	8,117	1,011	11,08
USC	3,589	5,181	-1,592	-44,36	7,132	6,854	278	3,90
US	3,848	4,909	-1,061	-27,57	7,933	6,366	1,567	19,75
UPC	3,067	4,900	-1,833	-59,77	11,068	6,502	4,566	41,25
UZAR	3,394	4,612	-1,218	-35,89	7,607	6,102	1,505	19,78
EHU	3,047	4,536	-1,489	-48,87	7,520	6,535	985	13,10
n			**16**	**16**			**20**	**20**
Avg[1]			**-2,165**	**-51,40**			**659**	**7,30**
SDev.[2]			**1,508**	**-39,37**			**1,722**	**19,56**
CI[3]			**-739**	**-19,29**			**755**	**8,57**

A, data reported in the Leiden Ranking 2011/2012; **B**, number of publications estimated by BAC; **A-B**, magnitude of the difference between CTWS and BAC; (**A-B)/A**, percentage of change between CTWS and BAC; **C**, data reported in the Iberoamerican Ranking SIR 2012; **D**, number of publications estimated by BAC applying SCIMAGO criteria, but sourcing the analysis with the WOS; **C-D**; magnitude of the difference between SCIMAGO and BAC; (**C-D)/C**, percentage of change between SCIMAGO and BAC. 1; average; 2, standard deviation; 3, 95% confidence interval of the average. Acronyms: UB, Univ. de Barcelona; UAB, Univ. Autònoma de Barcelona; UCM, Univ. Complutense de Madrid; UPM, Univ. Politécnica de Madrid; UAM, Univ. Autónoma de Madrid; UV), Univ. de València; UGR, Univ. de Granada; USC, Univ. de Santiago de Compostela; US, Univ. de Sevilla; UPC, Univ. Politècnica de Catalunya; UZAR, Univ. de Zaragoza; EHU, Univ. del País Vasco.

2) Exactness of the unification. The CTWS unit estimated a 5% of false negative cases, while the BAC ensures a maximum percentage of error of 3%. SCIMAGO provides no information on this regard.

3) Proximity to the units under analysis. Two observations support the notion that local knowledge may explain a substantial part of the observed discrepancies: 1) the difference between SCIMAGO & BAC was smaller than between CTWS & BAC, and 2), SCIMAGO attributed more publications to their neighboring universities (UGR & US) than BAC, and vice versa in the case of the UB & UAB). A comparison of the number of publications of the Dutch universities between CTWS and BAC may shed some light on the effect that local knowledge or "regional peculiarities" (Moed, 1996) have on this indicator.

4) Delineation of the universities. The CTWS unit took into account "important university institutes" and changes in the structure of universities, while BAC took into account institutes, but also faculties, technical schools, locations, and structural changes. Failing to aggregate the publications of subunits could also explain the observed differences (de Mesnard, 2012).

5) Completeness and accuracy of the database (location of addresses). There is a difference between the sources used by the CTWS unit and BAC. The NCR may compile fewer records than the WOS, as addresses have to be located to Spain and errors are likely to happen during this process. This inconsistency may also play a lesser role in the comparison between CTWS and BAC.

Final considerations

Discrepancies in the number of publications of universities in the order of 10^2 or 10^3 are irrelevant when comparing the figures produced by different units. However, the magnitude of the difference might represent half of the output in some cases. Fortunately, the numbers of publications produced by the three units correlated pretty well, and the rankings were concordant. Technical issues can no longer be used as arguments to explain divergences of this magnitude, as none of the factors presented here are completely dependent on the technical capacity of a unit, rather than on procedural decisions: 1) completeness and 2) exactness of the unification, 3) knowledge of the surrounding environment, 4) completeness and accuracy of the source or 5) the type of document and period of time. The findings suggest that a consensus addressing these factors would do more in reaching a methodological "greatest common denominator" between the different units enabling improving the reproducibility of the indicators.

References

de Bruin R.E. & Moed H.F. (1990). The unification of addresses in scientific publications. In: Egghe L., Rousseau R. (editors.), *Informetrics*, 65-78.

Butler L. (1999). Who 'Owns' this Publication?, *Proc. ISSI*, 87-96.

de Mesnard L. (2012). On some flaws of university rankings: The example of the SCImago report. *The Journal of Socio-Economics 41*(5), 495–9

Moed H.F. (1996). Differences in the construction of SCI based bibliometric indicators among various producers: A first overview. *Scientometrics, 35*(2):177-191 DOI: 10.1007/BF02018476

Van Raan A. F. J. (2005a). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133-143.

Semantometrics: Fulltext-based Measures for Analyzing Research Collaboration

Drahomira Herrmannova[1] and Petr Knoth[2]

[1] *d.herrmannova@open.ac.uk,* [2]*petr.knoth@open.ac.uk*
Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes (United Kingdom)

Introduction

The aim of this article is to demonstrate some of the possible uses of a novel set of metrics called *Semantometrics* in relation to the role of "bridges" in scholarly publication networks. In contrast to the existing metrics such as Bibliometrics, Altmetrics or Webometrics, which are based on measuring the number of interactions in the scholarly network, Semantometrics build on the premise that full-text is needed to understand scholarly publication networks and the value of publications.

Up to date many studies of scientific citation, collaboration and coauthorship networks have focused on the concept of cross-community ties (Shi et al., 2010; Guimerà et al., 2005; Silva et al., 2014). It has been observed that in citation networks, bridging or cross-community citation patterns are characteristic for high impact papers (Shi et al., 2010). This is likely due to the fact that such patterns have the potential of linking knowledge and people from different disciplines. Likewise, in collaboration and coauthorship networks, it has been shown that newcomers in a group of collaborators can increase the impact of the group (Guimerà et al., 2005).

The studies up to date have been focusing on analysing citation and collaboration networks without considering the content of the analysed publications. Our work has focused on analysing scholarly networks using semantic distance of the publications in order to gain insight into the characteristics of collaboration and communication within communities. Our hypothesis states that the information about the semantic distance of the communities will allow us to better understand the importance and the types of the cross-community ties (bridges).

More specifically, in order to gain insight into the type of collaboration between authors we are currently investigating the possibility of utilising semantic distance in a coauthorship network together with the concept of *research endogamy*. In social sciences, endogamy is the practice or tendency of marrying within a social group. This concept can be transferred to research as collaboration with the same authors or collaboration among a group of authors. The concept of research endogamy has been previously used to evaluate conferences (Montolio et al., 2013) as well as journals and patents (Silva et al., 2014).

Furthermore, in (Knoth & Herrmannova, 2014) we have introduced and tested the first Semantometric measure which we call *contribution(p)* and which can be used to estimate research publication contribution. Our results suggested that measuring semantic similarity of publications can be utilised to provide meaningful information about the value of a research publication, which is not captured by traditional bibliometric measures.

Types of research collaboration in a coauthorship network

We are currently investigating the possibility of combining semantic distance and research endogamy in the publication's collaboration network. The rationale behind this approach is based on how research collaboration happens. In case the authors of a publication come from different disciplines, their research is likely to link the two disciplines and to build a bridge between them. This bridge can help to provide vision and ideas otherwise unseen and help to transfer knowledge between the disciplines.

We propose to measure the semantic distance of coauthors of a publication based on semantic distance of all pairs of the coauthors, where the distance of a pair of authors can be expressed similarly as the *contribution(p)* measure (Knoth & Herrmannova, 2014). This situation is depicted in Figure 1, where the sets A and B correspond to the publication records of the two authors.

Table 1. Types of research collaboration based on semantic distance and research endogamy.

	High endogamy	**Low endogamy**
High distance	Established interdisciplinary collaboration	New interdisciplinary collaboration
Low distance	Expert group	New expert collaboration

In order to distinguish between emerging, short-term and established research collaboration, we propose to combine the semantic distance with research endogamy value of the publication as defined in (Silva et al., 2014). We assume that based on the combination of semantic distance and research endogamy the types of research

collaboration can be divided into four groups (Table 1).

We believe this classification is a useful tool in characterising the types of research collaboration that goes beyond the traditional understanding of the concept of bridges as used in scholarly communication networks. While semantic distance allows distinguishing between inter- and intra-disciplinary collaboration, research endogamy allows differentiating between emerging and established research collaborations.

Using semantic distance to measure research contribution in a citation network

A similar Semantometric approach based on the concept of semantic distance can be applied in citation networks. We have used this approach in (Knoth & Herrmannova, 2014) to develop a measure which we call *contribution(p)*. This measure is based on a hypothesis, which states that the added value of publication *p* can be estimated based on the semantic distance from the publications cited by *p* to the publications citing *p*. This situation is depicted in Figure 1.

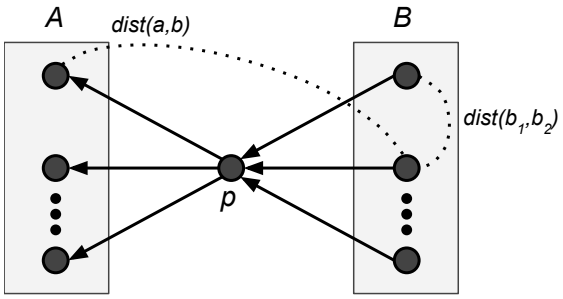


Figure 1. Explanation of contribution(p) calculation.

This hypothesis is based on the process of how research builds on the existing knowledge in order to create new knowledge on which others can build. A publication, which in this way creates a bridge between existing knowledge and something new, which will be developed based on this knowledge, brings a contribution to science. A publication has a high contribution if it connects more distant areas of science. Building on these ideas, we have developed a formula, which can be used for assessing research contribution of a publication. In order to adjust the contribution value to a particular domain and publication type, the metric uses a normalisation factor, which is based on the semantic distance of publications within the set of publications citing *p* and the publications cited by *p*. The measure and our experiments are in detail described in (Knoth & Herrmannova, 2014).

Conclusion

In this paper we proposed to apply the Semantometric idea of using full-texts to recognise types of scholarly collaboration in research coauthorship networks. We have applied semantic distance combined with research endogamy to classify research collaboration into four broad classes. This classification can be useful in research evaluation studies and analytics, e.g. to identify emerging research collaborations or established expert groups. Furthermore, we have presented another Semantometric measure, which we call *contribution(p)* and which is based on the idea of the importance of bridges in a citation network.

While bridges have been the concern of many research studies, their identification has been limited to the structure of the interaction networks. In contrast to these approaches, our approach takes into account both the interaction network (coauthorship, citations) as well as the semantic distance between research papers or communities. This provides additional qualitative information about the collaboration, which hasn't been previously considered.

References

Bornmann, L. & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1).

Guimerà, R., Uzzi, B., Spiro, J. & Nunes Amaral, L. A. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science*, *308*(April), 697–702.

Knoth, P. & Herrmannova, D. (2014). Towards Semantometrics: A new Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. *D-Lib Magazine*, *20*(11).

Montolio, S. L., Dominguez-Sal, D. & Larriba-Pey, J. L. (2013). Research Endogamy as an Indicator of Conference Quality. *SIGMOD Record*, *42*(2), 11–16.

Priem, J. & Hemminger, B. M. (2010). Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web. *First Monday*, *15*(7).

Seglen, P. O. (1992). The Skewness of Science. *Journal of the American Society for Information Science*, *43*(9), 628–638.

Shi, X., Leskovec, J. & Mcfarland, D. A. (2010). Citing for High Impact. Proceedings of the 10th Annual Joint Conference on Digital Libraries - *JCDL '10* (p. 49). New York, New York, USA.

Silva, T. H. P., Moro, M. M., Silva, A. P. C., Meira Jr., W. & Laender, A. H. F. (2014). Community-based Endogamy as an Influence Indicator. *Digital Libraries 2014 Proceedings.* London, United Kingdom.

Uncovering the Mechanisms of Co-authorship Network Evolution by Multirelations-based Link Prediction

Jinzhu Zhang, Chengzhi Zhang, Bikun Chen

{zhangjinzhu, zhangcz,chenbikun}@njust.edu.cn
Nanjing University of Science and Technology, Dept of Information Management, Xiaolinwei Str 200, Nanjing (China)

Introduction and literature review

Co-authorship network, a proxy of research collaboration, reveals the collaboration patterns and the determining factors through social network analysis perspective, with nodes representing authors and links representing co-authorships (Ortega, 2014; Yan & Ding, 2009). If we know what mechanisms push the evolution of co-authorship network, we could predict which authors may collaborate in future.

Most of the studies correlate co-authorship evolution mechanisms to similarity indicators which quantitatively compared by link prediction in homogeneous network (Lu & Zhou, 2010). In order to integrate multirelations between authors, path-based similarity indicators are proposed for co-authorship prediction in DBLP heterogeneous network (Sun et al., 2011; Sun & Han, 2013). However, what is the role of each mechanism plays and how to combine multiple mechanisms to suit the co-authorship network evolution need to be clarified, moreover, the method need to be verified in different domains.

Therefore, we integrate similarity indicators based on multirelations in heterogeneous network and quantitatively evaluate them by link prediction justly, to uncover and infer the mechanisms of co-authorship network evolution. Firstly, similarities between authors are represented by a matrix where the rows are multirelations and the columns are multirelations' measures. Secondly, the evaluation of similarities is processed based on link prediction, to reveal the importance of each mechanism which is the weight for combining multiple mechanisms. Finally, experiments are presented in the domain of Library and Information Science (LIS), which reveals the best appropriate mechanism, the significance of each mechanism and the combination strategy of different mechanisms.

Data and method

Data

We collect the data from the SCIE (Science Citation Index Expanded) databases in Thomson Reuters' Web of Science, using journal publications on subject category of LIS across 2000 to 2009.

We choose the authors that the frequency greater than or equal to five as the experiment data, which includes 669 authors, 3,948 articles, 6,476 keywords, 14 subject categories, 29 journals and 79,717 references.

We eliminate the subject categories because of too small numbers and references because of computing complexity. The co-author network has 1052 edges that indicate co-authorship, where we randomly choose 946 (90%) edges as training set and the remaining 106 edges as the testing set.

Multirelations-based link prediction

(1) Representation of co-authorships via multi-relations: Co-authorships via multirelations are systematically represented and extracted in a heterogeneous bibliographic network shown in Figure 1. Part of multirelations between authors could be represented in Table 1.

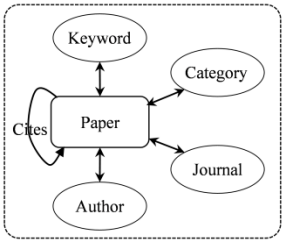


Figure 1. The nodes and relations in heterogeneous bibliographic network.

Table 1. Multirelations between authors.

Relations	*Description*
A-P-A-P-A	Common neighbours
A-P-A-P-A-P-A	Common neighbours' neighbours
A-P-J-P-A	Publish paper at the same journal
A-P-K-P-A	Authors have the same keyword
A-P-K-P-K-P-A	Authors' keywords co-word in same paper
A-P→P-A	Author x cite author y
A-P←P-A	Author x is cited by author y
A-P→P←P-A	Authors x and y cite the same paper
A-P←P→P-A	Authors x and y co-cited by same paper
A-P→P→P-A	Author x cite the paper that cite author y
A-P←P←P-A	The reverse relation of the above

(2) Measures of each relation: The four measures are the follows: path count (PC) is the number of

shortest path between two authors, normalized path count (NPC) is to discount PC by their overall connectivity, random walk (RW) and symmetric random walk (SRW) (Sun & Han, 2013).
(3) Evaluation of similarities based on link prediction: The relations and their measures combine the similarities, so there are 44 similarity indicators combined by 11 relations with four measures. We evaluate all the similarity indicators based on link prediction with precision and area under the curve (AUC).

Results

The three comparison perspectives are: (1) from the horizontal axis, compare which relation is best appropriate to the mechanism. (2) From the longitudinal axis, compare which measure is best to describe the mechanism. (3) Comparison between combined-relations-based and single-relation-based mechanisms.

The evolution mechanisms based on single-relation-based similarities

In Table 2 and Table 3, the entries emphasized in bold and italic corresponding to the highest accuracies from the horizontal axis.
In precision, the APAPA with NPC is the best appropriate and important mechanism in LIS where NPC plays the best in four measures, yet the APJPA with RW plays the worst. In AUC, the APAPA with SRW is the best mostly with little differences. There is lots of information loss in the projection from heterogeneous network to homogeneous network compared with CNs.

Table 2. The precision/AUC of single-relation-based similarities.

Relations	*PC(%)*	*NPC(%)*	*RW(%)*	*SRW(%)*
APAPA	38.4/87.5	***42.5***/87.5	31.7/87.7	41.4/***87.9***
APAPAPA	24.0/***86.2***	***32.9***/86	21.1/***86.2***	29.4/85.8
APJPA	3.2/76.8	***3.9***/77.2	0.9/76.7	2.6/***77.4***
APKPA	7.6/81.4	***20.4***/82.1	9.4/81.8	16.3/***82.3***
APKPKPA	2.2/70.8	***4.9/72.5***	2.5/70.9	4.3/72
CNs	23.4/84.1			

Comparison between combined-relations-based and single-relation-based mechanisms

The paper designs five combination strategies for comparison: (1) CR1: Combination of all relations without weights. (2) CR2: Combine all relations except APJPA. (3) CR3: Combination of all relations with weights denote by precision in Table 2. (4) CR4: the combination formed via just authors which is APAPA+APAPAPA. (5) CR5: the combination formed via just keywords, which is APKPA+APKPKPA. The precision and AUC are listed in Table 3.
In precision, the CR3 with NPC is the most appropriate and important mechanism in LIS where NPC plays the best in four measures, yet the CR5 with PC plays the worst. The AUC is consistent with the precision result mostly and others with little differences. The CR2 and CR3 with each measure are all outperformed the single-relation-based mechanisms. The CR4 performs much better than CR5 proves that in co-authorship formation the author is more important than research interest.

Table 3. The precision/AUC of different combinations of relations.

Relations	*PC(%)*	*NPC(%)*	*RW(%)*	*SRW(%)*
CR1	28.6/86.4	**40.8/88.6**	26.3/88.4	36/88.3
CR2	38.6/84.8	**43.7/87.4**	32.4/86.4	43.6/86.8
CR3	45.1/89.1	**49.2**/89.3	39.8/89.0	47.2/**89.5**
CR4	24.2/86	**38.6/86.4**	27.1/86.2	35.3/86.1
CR5	2.2/80.6	**16.7**/82.8	6.6/**83.1**	12/82.7

Conclusion and discussion

This paper uncovers the mechanisms of co-authorship network evolution by multirelations-based link prediction in LIS. In the next, we will consider other factors that influence research collaborations, all relations especially related to references to enhance the accuracy and validation in two or more different areas with different article types (e.g., journal and conference).

Acknowledgments

Our work is supported by the Ministry of Education of China Project of Humanities and Social Sciences (Grant No. 14YJC870025), the Fundamental Research Funds for the Central Universities (Grant No. 30915013101) and the National Natural Science Foundation of China (Grant No. 71173211).

References

Lu, L. & Zhou, T. (2010). Link prediction in complex networks: A survey. *Arxiv preprint arXiv:1010.0725*.

Ortega, J. L. (2014). Influence of co-authorship networks in the research impact: Ego network analyses from Microsoft Academic Search. *Journal of Informetrics, 8*(3), 728-737.

Sun, Y., Barber, R., Gupta, M., Aggarwal, C. C. & Han, J. (2011). Co-author Relationship Prediction in Heterogeneous Bibliographic Networks. *Proc. ASONAM*.

Sun, Y. & Han, J. (2013). Mining heterogeneous information networks: a structural analysis approach. *ACM SIGKDD Explorations Newsletter, 14*(2), 20-28.

Yan, E. & Ding, Y. (2009). Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis. *Journal of the American Society for Information Science and Technology, 60*(10), 2107-2118.

JOURNALS, DATABASES AND ELECTRONIC PUBLICATIONS

DATA ACCURACY AND DISAMBIGUATION

MAPPING AND VISUALIZATION

Citing e-prints on arXiv
A study of cited references in WoS-indexed journals from 1991-2013

Valeria Aman[1]

[1] *aman@forschungsinfo.de*
iFQ - Institute for Research Information and Quality Assurance, Schützenstraße 6a, 10117 Berlin (Germany)

Abstract

This study deals with the analysis of cited references in Web of Science (WoS) to e-prints on arXiv. Created in 1991, arXiv accelerated the scholarly communication and developed into a well-established e-print repository that functions as an essential access point to the latest research in physics, astrophysics, mathematics, computer science and related fields. Authors evidently rely on arXiv full texts and refer to them in their own research papers. These cited references to arXiv that represent the acceptance of e-prints in journals and series indexed in WoS are tackled in this paper. A total of 900,000 cited references to arXiv have been identified for the 1991-2013 period. Object of investigation is on the one hand the set of cited references to arXiv, and on the other hand the set of papers in WoS that cite arXiv. Among other things, the paper illustrates that citations to arXiv peak in the year after submission and drop rapidly. The geographical distribution of authorship citing arXiv in their papers shows that authors from the US, Germany, GB, France and Italy rely heavily on arXiv. The paper identifies "arXiv-friendly" journals where the majority of articles refer to arXiv.

Conference Topic
Journals, databases and electronic publications

Introduction

The arXiv is a convenient vehicle to disseminate research results prior to the publication of peer-reviewed articles. It is also common to submit postprints for reasons of wide availability and archiving. There is no doubt that e-prints are read by a wide community and are regarded to be of good quality. Thus, it is of interest to learn more about the perception of arXiv as a source of relevant information that supports researchers' ideas and discoveries. The study sets out to answer the following questions: 1) Do authors publishing in journals covered by Web of Science (WoS) cite e-prints on arXiv? 2) What characteristics in citations can be observed? 3) In which countries are authors situated that rely on e-prints in arXiv? 4) What are the journals that include the highest rate of articles with cited references to arXiv?

Background

The rise of preprints, e-prints and arXiv

There are several definitions for the term "preprint". Lim (1996) defines a "preprint" as a manuscript that has been reviewed and accepted for publication, a manuscript that has been submitted for publication, but for which a decision to publish has not been made yet, or a manuscript that is intended for publication, but is being circulated for comments among peers prior to journal submission. Electronic prints (e-prints) refer both to preprints and post-prints (peer-reviewed published papers), and other documents that are made available on the Internet. The "preprint culture" dates back to the 1960ies, when high-energy physicists were eager to disseminate their results by printing and mailing copies of their manuscripts simultaneously to journal submission (Goldschmidt-Clermont, 1965). The time consuming process of peer-review was hence effectively bypassed. With the advent of the World Wide Web in the early 1990ies, the emergence of new methods of scientific discourse were encouraged, altering the traditional channels of scholarly communication (Brown, 2001).

In summer 1991, Paul Ginsparg conceived the repository arXiv at the Los Alamos National Laboratoy (LANL) in New Mexico. Ginsparg (1994, p.157) stated that "the realization of arXiv was facilitated by a pre-existing 'preprint culture', in which the irrelevance of refereed journals to ongoing research has long been recognized". Ginsparg (1994, p.159) designed arXiv (formerly xxx.lanl.org) as a fully automated system, where users could maintain a database to disseminate information without outside intervention.

Originally, arXiv was intended for the High-Energy Physics (HEP) community, but expanded rapidly to cover all of Physics, Astrophysics, Mathematics and Computer Science. Since September 2003 arXiv covers Quantitative Biology. In April 2007 Statistics was included, followed by Quantitative Finance in December 2008. Today, arXiv is hosted at Cornell University in New York with seven mirror sites all over the world. It contains more than 1,000,000 full-text e-prints, receiving about 9,000 new submissions each month.[1] Researchers can check arXiv for new information, search for relevant papers, post their own papers and cite references by arXiv ID. It is a self-organizing publication mode that costs the users nothing (Langer, 2000). Another reason for arXiv's popularity is its democracy, because scientists "can post their research results without being hassled by grumpy editors and referees" (ibid., p.35). According to Ginsparg (1994, p.157) physicists have learned to determine from the author, title and abstract whether to read a paper "rather than rely on the alleged verification of overworked or otherwise careless referees".

Nowadays, researchers still regard it as valuable to publish their work in peer-reviewed journals. Prior to formal publication, the findings may be spread as conference proceedings, reports, working papers or preprints. As Heuer, Holtkamp and Mele (2008, p.2) point out "scientists expect unrestricted access to comprehensive scientific information in their field, state-of-the-art information venues to optimize their research workflow and quality assurance at the parallel existence of traditional peer-review and the immediacy of dissemination and feedback". A publication delay of several months between the completion of a work and its appearance in a peer-reviewed journal is simply a "negative phenomenon in scientific information dissemination" (Amat, 2008, p.379). Amat (ibid.) found that the publication delay depends primarily on the peer-review process (see also Luwel, 1998). ArXiv serves to overcome this delay and helps to circulate results upon realization.

Previous work

The citation behaviour of e-prints available through arXiv has been studied extensively. Youngen (1998) identified the growing importance of e-prints in the published literature. He found that e-prints became the first choice among physicists and astronomers for finding current research and keeping up with colleagues and competitors at other institutions. Brown (2001) studied citations of e-prints on arXiv in astronomy and physics journals from 1998 to 1999. The citation analysis showed that the peak of citations to e-prints is reached after three years, which is comparable to papers in print journals. Garner, Horwood & Sullivan (2001) determined the place of e-prints in the scholarly information delivery, concluding that rapid dissemination of results in form of preprints establishes priority and enables rapid feedback. Brown (2003) asked for the opinion of chemists about citing e-prints in the articles they author. Fifty-two percent said they would cite e-prints whenever possible, whereas 48% stated that they would not. Reasons for avoiding to cite the Chemistry Preprint Server (CPS) are the lack of relevant articles, the lack of customary to cite, and the lacking awareness of CPS (ibid., p.365). The study of infiltration of CPS e-prints into the literature of chemistry revealed that "no citations to e-prints were found in the journal literature using ISI's Web of Science from 2000 to 2001" (ibid., p.366). Prakasan & Kalyane (2004) focused on the

[1] http://arxiv.org/stats/monthly_submissions / [Last visited January 06, 2015]

citations in Science Citation Index to e-prints on arXiv, submitted under the four categories hep-ex, hep-lat, hep-ph and hep-th[2], providing a broad insight into citation habits.

Several studies focused on the citation impact of e-prints on arXiv, also within the Open Access debate (see Harnad & Brody, 2004; Antelman, 2004). Schwarz & Kennicutt (2004) analyzed articles published in the Astrophysical Journal in 1999 and 2002 and reported that papers posted to the astro-ph-section on arXiv were cited more than twice as often as those without a version on arXiv. In accordance, Metcalfes (2005) findings show that astronomy papers in the highly-cited journals Science and Nature received higher citation rates when their authors posted their papers on arXiv's astro-ph. Metcalfe (2006) studied the field of solar physics with the result that papers posted to arXiv are on average 2.6 times as often cited as papers not being posted. He concludes that higher citation rates are not a result of self-selection of outstanding papers, since conference proceedings reveal the same result. Moed (2007) analyzed how the citation impact of articles deposited in the Condensed Matter section in arXiv and subsequently published in a journal compares to that of articles not deposited on arXiv. He concluded that arXiv accelerates citations, because it makes papers earlier available. Davis & Fromerth (2007) examined whether mathematics journals from 1997 to 2005 with a previous preprint version on arXiv receive more citations than non-deposited. Their findings show that articles in arXiv receive on average 35% more citations, which translates to 1.1 citations per article. They explain the citation advantage with the Open Access, the Early View, and the Quality postulates, which are non-exclusive.

Henneken et al. (2007) analyzed whether e-prints on arXiv are preferred over the journal articles in four core journals in astrophysics. They found that as soon as an article is published, the community prefers to read and cite it, so that the usage in the NASA Astrophysics Data System (e-print system) drops to zero. They also showed that the half-life (the time at which the use of an article is half the use of a newly published article) for an e-print is shorter than for a journal article. Gentil-Beccot, Mele & Brooks (2009) investigate whether HEP scientists still read journals or rather prefer digital repositories. Their citation analysis shows that free and immediate dissemination of preprints results in a citation advantage for HEP journals. Furthermore, their analysis of clickstreams reveals that high-energy physicists prefer preprints and seldom read journals.

Some of the studies suggest that articles with a previous preprint on arXiv receive more citations than articles without. Other studies report no such effect. Gentil-Beccot, Mele & Brooks (2009) did not detect any citation advantage from publishing in Open Access HEP journals. Their finding is similar to that of Moed (2007) in Condensed Matter, Davis (2007) in Mathematics and Kurtz & Henneken in Astrophysics (2007).

Brody, Harnad & Carr (2006) examined the correlation of the number of article downloads and the number of citations. On the basis of arXiv they show that the short-term Web usage impact of e-prints predicts a medium-term citation impact of the final article. Haque and Ginsparg (2009; 2010) found that e-prints posted to arXiv at the beginning and end of a day reach a wider readership and receive higher citation rates over the course of ensuing years than posting in the middle of day. Shuai, Pepe & Bollen (2012) analyzed the online response to preprint publications on arXiv, studying the delay of article downloads and Twitter mentions following submission.

Larivière et al. (2014) analyzed the proportion of papers across all disciplines on arXiv for the 1991-2012 period, just as the proportion of arXiv papers that are published in WoS-indexed journals. They determine the time between arXiv submission and journal publication, ageing characteristics and impact of arXiv e-prints and their published alter ego. They also focus on

[2] High energy physics - experiment (hep-ex), high energy physics - lattice (hep-lat), high energy physics - phenomenology (hep-ph), and high energy physics - theory (hep-th).

the proportion of cited references in WoS to arXiv e-prints by discipline. Working with percentages, they quantify that journals in nuclear and particle physics have 6.6% of their references to arXiv e-prints, whereas in mathematics this share is below 1.5% (ibid., p.1163). Stimulated by the work of Larivière et al. (2014), this study sets out to quantify the number of cited references in WoS to arXiv manuscripts, and to provide a broader view on characteristics of cited references and the papers that include them.

Data and methods

Database

The study builds upon the bibliometric database at the "Competence Center for Bibliometrics for the German Science System" that is hosted at the iFQ.[3] It consists of data from Thomson Reuter's Web of Science. Peer-reviewed journal articles are the primary mode of communication of scientific research. Researchers write reviews or articles with discoveries, theories and results. To relate their work they cite other articles if they know the article and believe it to be relevant to their own work. They might also provide negative citations in order to disagree or to say that a paper has flaws (see Brody, Harnad & Carr, 2006). Citations can be therefore used as a measure of influence and importance of preceding articles.

The identification of references to arXiv depends on the quality of the bibliographic information (e.g. the presence of the reference to arXiv) and the extent to which WoS was able to parse the references of the citing articles. Identifying cited references to arXiv can lead to false positives, when a reference looks like an arXiv identifier but is actually not, or where authors make mistakes. A linking by bibliographic data is more precise as it builds upon author names, journal title, volume, page number, year of publication etc.

Data collection

Different from Youngen (1998), who analyzed those cited references that state explicitly "preprint" in ISI's SciSearch (p.451), this study also includes postprints. Hence, all manuscripts on arXiv are in the following referred to as "e-prints". The e-print identifier assigned by arXiv provides a standardized number that allows each e-print to be uniquely identified. This uniqueness is required for correct citing of the work. ArXiv has established a subject grouping and numbering system for submitted e-prints. Examples are Astrophysics (astro-ph), Condensed Matter (cond-mat), High-Energy Physics-Theory (hep-th) or Nuclear-Experiment (nucl-ex), followed by a numerical string, indicating the year and month of submission, and an increasing accession number. A typical example is quant-ph/95002, where quant-ph stands for Quantum Physics, "95" for the year 1995 and "002" for the accession number. Up to March 2007 this ID enabled a broad subject categorization. In April 2007, the arXiv-ID was changed and no longer contains subject categories. It consists of eight digits, of which the first four represent the year and month of submission. Divided by a period, they are followed by a four-digit long accession number, e.g.: arXiv: 0705.0002. We can infer that this e-print was loaded in May 2007. Since the accession number will soon reach its capacity, the length of the accession number has been extended by one digit in January 2015.[4]

The search for arXiv e-prints in the cited reference field in WoS was approached in several steps. E-prints up to 2007 were identified on the basis of an alphanumeric string that contains the subject category followed by the year of submission and the accession number.[5] E-prints published in 2007 or later were identified by the string "arXiv" followed by a numerical string. This led to an overall satisfying result, since the string "arXiv" is unique and causes

[3] http://www.bibliometrie.info/ [Last visited January 06, 2015]
[4] http://arxiv.org/new#dec19_2014 [Last visited January 06, 2015]
[5] The categories in bold print were used for the matching: http://arxiv.org/ [Last visited January 06, 2015]

almost no confusion. A low number of false positives cited references were deleted manually. Only one in four cited references had a publication year assigned, which is indeed not necessary, since it is part of the arXiv ID. With the application of Regular Expressions in SQL the year of e-print publication was deduced for more than 99% of cited references. A publication year was not deducible, where authors cited arXiv simply in this fashion: "arXiv". The search strategy may not include citations to works that technically have to be considered as arXiv e-prints. According to Youngen (1998, p.451) authors may have cited preprints as "submitted to...", "to be published in...", "in press" or "unpublished", depending on their state in the publication cycle. Thus, in reality, the number of citations to e-prints on arXiv may be much higher than presented here.

Data corpus[6]

With the search strategy described, 892,867 cited references to arXiv were identified for the 1991-2013 period, of which 357,557 have a distinct character string. Due to multiple subject categorizations in arXiv, author typos, or erroneous data parsing in WoS, one and the same e-print can be referred to in different spelling variants. Hence, the actual number of arXiv e-prints cited in the 1991-2013 period by papers in WoS is lower. At the same time 289,145 distinct papers were identified in WoS that constitute these 892,867 cited references. To relate these figures, Brown (2001) found 35,928 citations to arXiv e-prints (posted between 1991 and 1999) in astronomy and physics journals published in 1998-1999. In the following, analyses are based on the cited references to arXiv and the WoS-papers that include them.

Results and discussion

Figure 1 provides an overview of the data collected. The number of e-prints submitted to arXiv has been gradually rising from 303 in 1991 to 92,641 in 2013.[7] The number of papers in WoS citing at least one e-print on arXiv has steadily increased and comprises around 28,000 papers in 2013. In addition, we can see the number of cited references to e-prints on arXiv with the publication year of the citing paper as indicated on the x-axis. We can derive that a paper citing arXiv includes on average more than one citation to e-prints on arXiv. Most of the citations to e-prints were provided in 2012 (ca. 76,000).

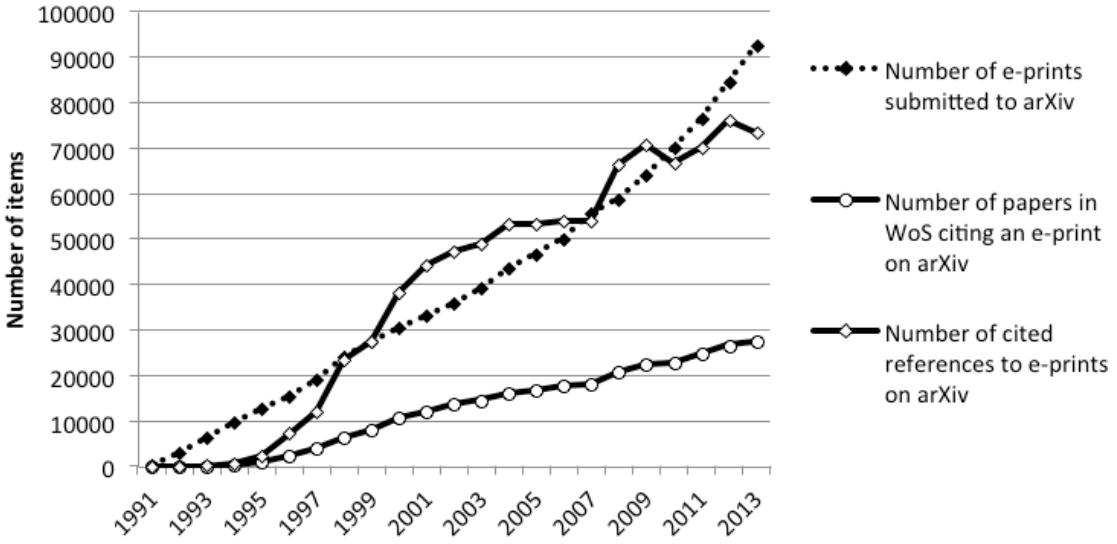


Figure 1. Overview of the yearly growth of submissions to arXiv, the number of papers in WoS citing arXiv e-prints according to their publication year, and the number of cited references.

[6] The data corpus can be requested on demand.
[7] http://arxiv.org/stats/monthly_submissions [Last visited January 06, 2015]

The analysis of document types shows that articles rank first with 96.0% of all WoS documents from 1991-2013 that cite arXiv. Reviews (3.2%) refer to arXiv as well, in order to provide a broad or up-to-date state of research. Editorials, Letters, Corrections and Notes also reference arXiv.

In the following, it does make a difference whether cited references are analysed or the WoS-papers that include those. Due to different citation habits, even within a broad field such as physics, it appears more suitable to consider primarily the citing papers. Table 1 provides an overview of the subject areas that constitute most of the citations to arXiv. The first column lists the Subject Categories[8] (SC) in WoS in a descendant order, regarding the number of arXiv citing papers assigned to this SC. We can see that Particle Physics ranks first (21%), followed by Astronomy and Astrophysics. In total, these 12 SC cover more than 90% of all citing papers that refer to arXiv between 1991 and 2013. The percentages and order of the SC changes when we have a look on the number of cited references to arXiv. Particle Physics still ranks first, claiming almost one-third of all cited references to arXiv. The results suggests that papers in Particle Physics have on average a higher number of cited references to arXiv than those in other SC.

Table 1. Overview of Subject Categories in WoS that contribute to the majority of papers that cite arXiv and their number of cited references. The data is based on 289,145 arXiv-citing papers in WoS that provide 892,867 cited references in 1991-2013.

Subject Category	No. of papers citing arXiv	Share in %	No. of cited references	Share in %
Physics, Particles & Fields	88,757	21.0	398,022	30.5
Physics, Multidisciplinary	70,383	16.7	248,091	19.0
Astronomy & Astrophysics	68,805	16.3	225,326	17.3
Physics, Mathematical	28,073	6.7	82,490	6.3
Physics, Condensed Matter	25,658	6.1	49,852	3.8
Mathematics	23,894	5.7	46,952	3.6
Physics, Nuclear	22,838	5.4	83,712	6.4
Optics	13,602	3.2	27,414	2.1
Physics, Atomic, Molecular & Chemical	12,754	3.0	25,625	2.0
Mathematics, Applied	10,976	2.6	20,169	1.5
Physics, Applied	9,223	2.2	17,099	1.3
Physics, Fluids & Plasmas	5,704	1.4	9,488	0.7

This leads us to the analysis of the distribution of cited references among the papers in WoS that cite arXiv. Table 2 illustrates the frequency of citing papers in WoS that include as many cited references as stated in the left column. We can see that six papers in WoS have more than 200 references to arXiv in their list of references. Every eleventh paper, out of the set of arXiv citing papers, includes 6 to 10 references to arXiv. Nevertheless, around 46% of citing papers provide a single reference to arXiv. A closer look on the paper with the highest number of cited references to arXiv shows that it is a review article from 2000 on String Theory and Gravity, where a link to arXiv was set additionally to the journal article reference. This brings us to the analysis of characteristics in citations to arXiv. Are e-prints on arXiv immediately cited when there is no corresponding journal article or are they also used in future and even preferred over the corresponding journal article?

[8] The 260 SC in WoS are assigned to journals on the basis of their scope and citation links.

Table 2. Distribution of cited references among WoS-papers that cite e-prints on arXiv.

Number of references to arXiv in a single paper	Number of papers citing arXiv	%
more than 200	6	0.00
151 to 200	8	0.00
101 to 150	29	0.01
51 to 100	222	0.08
21 to 50	2,567	0.89
11 to 20	9,375	3.24
6 to 10	25,859	8.94
5	12,544	4.34
4	18,939	6.55
3	30,969	10.71
2	56,204	19.44
1	132,423	45.80
Total	289,145	100.00

Figure 2 shows on the one hand the line graph of all citations to e-prints on arXiv up to 2013. Different from Figure 1 the x-axis signifies the year of e-print publication. Thus, the sudden decrease of cited e-prints from 2008 on is due to the fact that they had less time to be referenced than those posted in earlier years. In addition, Figure 2 provides bars indicating the years in which these e-prints were cited by WoS papers. Each bar represents the number of cited references to arXiv in the same year as the e-print was published, the subsequent year and two and three years respectively after publication of the e-print. The space between the line graph and the bars represents the cited references to e-prints that were provided more than three years after e-print publication. Since e-prints from recent years did not have much time to be cited, the bars coincide with the line graph of the total number of cited e-prints.

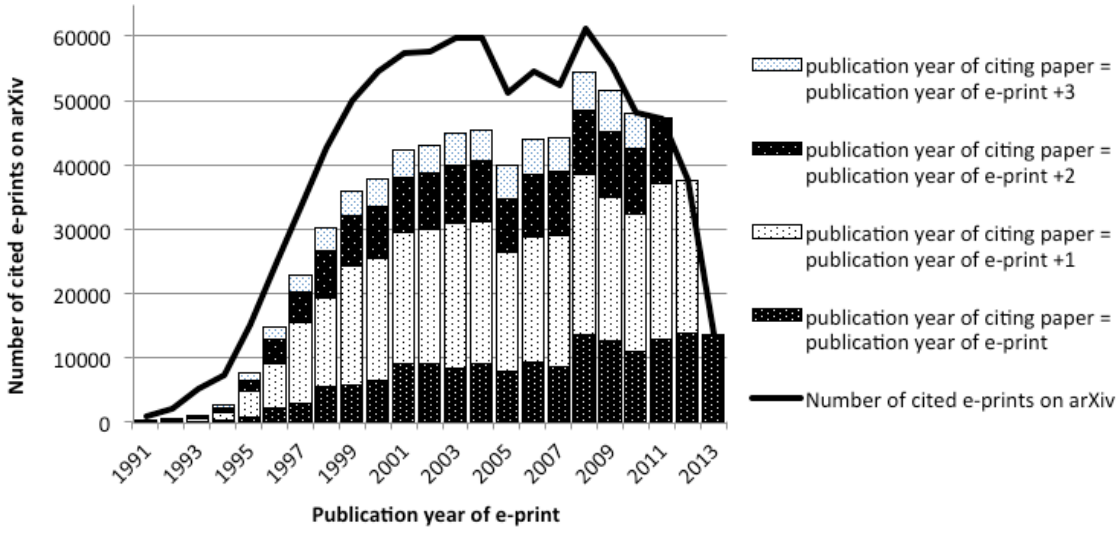


Figure 2. Time series of citation distribution. Illustrated are citations that equal the year of e-print submission, citations to e-prints that are one year old, up to the age of three years. The line graph signifies the total number of e-prints cited, published in the year as indicated.

It becomes evident that e-prints on arXiv are mostly cited in the subsequent year of e-print post. Almost half of all cited references in a year relate to e-prints that were placed on arXiv the preceding year. This is in accordance with Larivière et al. (2014, p.1166), who found that citations to e-prints on arXiv peak the year following submission. The figure also indicates that e-prints are cited immediately in the same year of posting. Only a small share of cited

references points to three-year old e-prints. On the contrary, Brown's (2011) analysis in astronomy and physics showed that the peak of citations to e-prints is reached after three years. The results in Figure 2 are in little accordance with Henneken et al. (2007, p.19) who showed that the usage of e-prints drops to zero as soon as the journal article has appeared, suggesting that authors have access to subscribed journals and prefer to cite the refereed version. Garner, Horwood & Sullivan (2001, p.251) quantified that 90% of papers on arXiv are later published in journals so that a corresponding article can be found and cited properly.

Nevertheless, there are many reasons that underscore the high citation rates of e-prints. Davis & Fromerth (2007) write that the arXiv copy is sufficient for the purpose of citing it in one's own work. They found that articles that are also accessible on arXiv receive 23% fewer downloads from the publisher's web site two years after publication (ibid., p.23). Gentil-Beccot, Mele & Brooks (2009) found that citations start before publication, because scientists in HEP do not wait for an article to be published. Even in the first few months after journal publication authors read and cite the preprint (ibid., p.6). According to Moed (2007) colleagues start to read a paper and cite it in their own articles earlier if it is deposited on arXiv. The following Figure 3 illustrates the relation between the publication year of a WoS-paper citing arXiv, and the publication year of the cited e-print. The whole bar in each year (y-axis) represents the total number of cited references to e-prints on arXiv from this year (cf. Figure 1). The cited references from each year are grouped by the publication year of the cited e-print. Each bar indicates the share of e-prints, according to their year of publication. For the year 2013 we can see that 13,000 cited references (top black part of the 2013-bar) refer to e-prints published in the same year. The lion's share of cited references in 2013 (24,000) is to e-prints published in 2012. In general, we can conclude from Figure 3 that the majority of references in each year points to e-prints published in the preceding year.

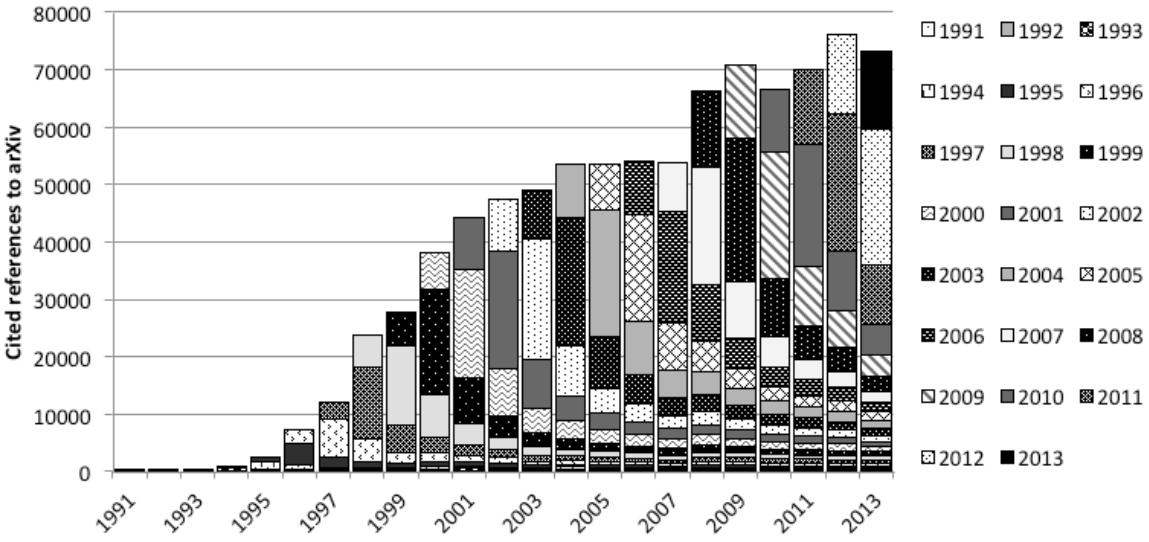


Figure 3. Time series of cited references to e-prints on arXiv. The x-axis represents the publication years of WoS-paper citing an e-print, whereas each bar represents the share of the years a cited e-print was published in.

To see where the authors that frequently cite arXiv are from, Table **3** provides a ranking of countries according to the highest number of papers in WoS with at least one cited reference to arXiv. USA rank first with one-third of all papers that cite arXiv. They are followed by Germany and Great Britain. Note that the percentages do not add up to 100, since co-authored papers can be attributed to multiple countries.

Table 3. Overview of countries that most frequently cite arXiv e-prints. The percentages are calculated on the basis of the total number of citing papers (289,145).

Rank	Country	No. of WoS-papers citing e-prints	%	Rank	Country	No. of WoS-papers citing e-prints	%
1	USA	97,085	33.6	11	Switzerland	14,489	5.0
2	Germany	45,842	15.9	12	India	11,764	4.1
3	GB	30,776	10.6	13	Poland	9,332	3.2
4	France	28,159	9.7	14	Brazil	9,004	3.1
5	Italy	27,896	9.6	15	Netherlands	8,361	2.9
6	China	25,467	8.8	16	South Korea	8,271	2.8
7	Japan	25,196	8.7	17	Australia	7,296	2.5
8	Russia	22,772	7.9	18	Israel	7,019	2.4
9	Spain	15,902	5.5	19	Sweden	5,402	1.9
10	Canada	14,879	5.1	20	Belgium	4,709	1.6

The journals whose articles most often cite e-prints on arXiv are identified in Table 4. On the left of the table, journals are ranked according to their number of citing papers in the 1991-2013 period. On the right of the table journals are ranked according to their number of cited references to arXiv. Evidently, most of the journals carry a majority of HEP content. Among these are *Physical Review D*, *Journal of High Energy Physics* (JHEP), *Physics Letters B* and *Nuclear Physics B*. Striking are also the astrophysical journals, among which we can find the *Astrophysical Journal, Monthly Notices of the Royal Astronomical Society* and *Journal of Cosmology and Astrophysical Physics*.

Table 4. Overview of journals in WoS with the highest number of papers citing arXiv and journals with most of the cited references to arXiv in the 1991-2013 period.

Journal	Citing papers	%	Journal	Cited ref.	%
Physical Review D	30,287	10.5	Physical Review D	112,261	12.6
Physical Review B	15,080	5.2	Journal of High Energy Physics	77,431	8.7
Journal of High Energy Physics	14,881	5.1	Physical Review B	66,750	7.5
Physical Review Letters	13,816	4.8	Nuclear Physics B	50,757	5.7
Physics Letters B	13,707	4.7	Physics Letters B	29,195	3.3
Physical Review A	9,599	3.3	Physical Review Letters	28,873	3.2
Astrophysical Journal	8,428	2.9	Classical and Quantum Gravity	22,969	2.6
Nuclear Physics B	8,033	2.8	Physical Review A	20,480	2.3
Monthly Notices of the Royal Astronomical Society	6,256	2.2	Journal of Cosmology and Astrophysical Physics	19,559	2.2
Physical Review E	5,081	1.8	International Journal of Modern Physics A	18,685	2.1
Sum	125,168	43.3	Sum	446,960	50.1

Youngen (1998) could not find firm rules for citing preprints, with the exception of the Astrophysical Journal, which stated that "References to private communications, papers in preparation, preprints, or other sources generally not available to readers should be avoided" (p.453). Nevertheless, it ranks seventh among the most active journals citing e-prints on arXiv. This restriction must have been eased over the years, as can be seen in Figure 4. Depicted are time series of percentages of papers in a journal that cite arXiv, for the ten

journals with the highest number of arXiv-citing papers (see Table 4). We can observe that up to 1997 the *Astrophysical Journal* had less than 10% of their papers citing e-prints on arXiv. This share was growing in the following years to reach approx. 25%.

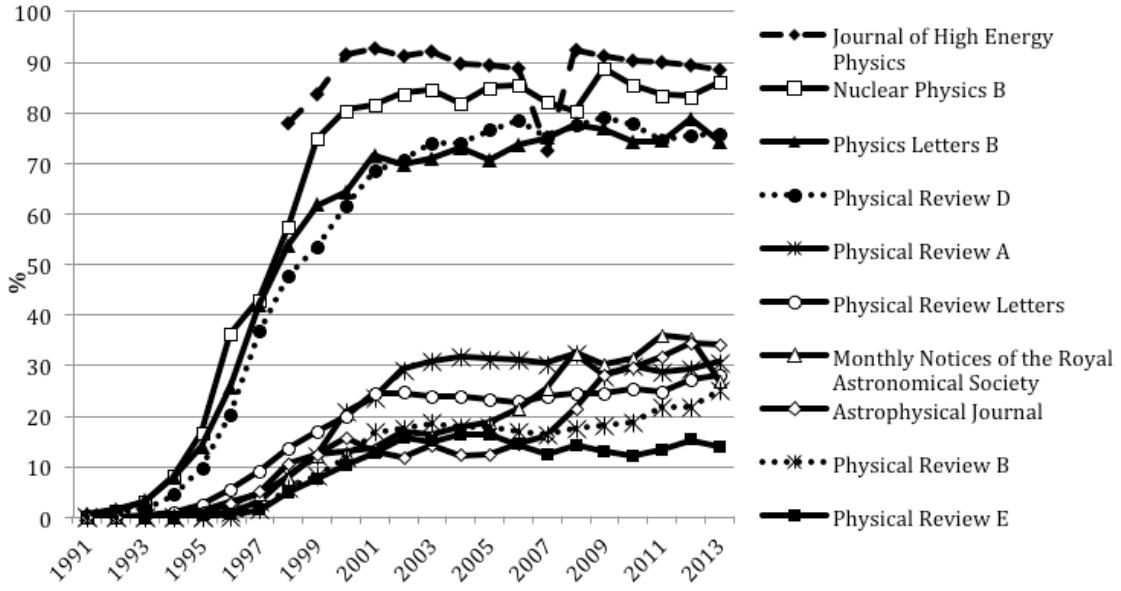


Figure 4: Time series of the percentages of papers in a journal that cite arXiv. Displayed are the 10 journals that most actively cite arXiv.

Striking is the decline of the share of papers in JHEP with references to arXiv in 2007, for which no explanation can be given. Overall, the shape of the line graphs suggests a rapid growth of arXiv's acceptance in the 1990ies and a constant reliance on arXiv in the past 15 years. The following table identifies other "arXiv-friendly" journals, where the majority of papers rely on arXiv. Since the number of papers published in a journal can differ immensely, Table 5 indicates percentages of the number of a journal's papers that cite arXiv. To provide an up-to-date view, only papers published between 2004 and 2013 are considered.

Table 5: Journals in WoS with the highest share of papers citing arXiv. Analyzed are only citing papers that were published between 2004 and 2013.

Journal	%	Journal	%
Journal of Cosmology and Astroparticle Physics	89.9	Journal of Physics G-Nuclear and Particle Physics	59.0
Advances in Theoretical and Mathematical Physics	81.7	International Journal of Modern Physics A	59.0
Annual Review of Nuclear and Particle Science	80.7	International Journal of Modern Physics D	57.5
Communications in Number Theory and Physics	79.8	Progress of Theoretical and Experimental Physics	56.2
European Physical Journal C	70.9	Physics Reports-Review Section of Physics Letters	55.5
Fortschritte der Physik-Progress of Physics	70.4	General Relativity and Gravitation	54.0
Quantum Information & Computation	69.3	Gravitation & Cosmology	54.0

Modern Physics Letters A	62.3	Journal of Sympletic Geometry	53.6
Progress in Particle and Nuclear Physics	61.5	Reviews of Modern Physics	52.8
Acta Physica Hungarica A-Heavy Ion Physics	60.4	Algebraic and Geometric Topology	52.2
Geometry & Topology	60.3	Progress of Theoretical Physics	51.7
Classical and Quantum Gravity	60.0	Astroparticle Physics	51.2

Ranking the journals on the basis of percentages instead of absolute numbers enables us to spot mathematics journals. The 24 journals listed prove that the circle of users coincides with the target group of arXiv that consists mainly of high-energy physicists. In HEP it is usual practice to submit papers to arXiv prior to journal submission. According to Gentil-Beccot, Mele & Brooks (2009) the arXiv often presents a version very similar to the published one. Finally, the arXiv version is freely available, while the journal versions require subscription.

Conclusions

The rapid dissemination of research results enabled by arXiv has accelerated the read-and-cite process (see Brody, Harnad & Carr, 2006). The identified number of cited references to arXiv and the rapid citation of e-prints in WoS-indexed journals indicate that e-prints are accepted within certain communities as well as among journal editors. Taking citation counts as a proxy for quality, e-prints on arXiv can be regarded as of good quality. They are valued, read and used within the scientific community, mainly because they present results upon finalization, circumventing the publication delay. To refer to these most up-to-date findings, authors evidently do not hesitate to cite arXiv e-prints in their research papers. The high number of cited references presented in this study suggests the usage of e-prints over the journal articles, as it was also found by Davis & Fromerth (2007). One reason for the preference of arXiv e-prints is the free availability of full text, especially if readers do not have access to the journal. Besides, the arXiv version is often similar to the formal journal article and can be easily cited by ID. An obvious reason to cite arXiv full texts even years after publication might be simply that the e-print does not have a published alter ego to be cited. Furthermore, the results showed that citations to e-prints peak in the year after publication and drop rapidly in the following years. Authors may still rely on the e-print but cite the formal publication, so the decline in citations does not necessarily indicate a decline in use. This could be proved in a future study with download data of arXiv e-prints over time.

Whereas this initial study is mostly exploratory, future work will link arXiv data to the data in WoS to examine, whether the cited e-prints have a journal version or not. So far, Larivière et al. (2014, p.1161) found that 64% of all arXiv e-prints are published in a WoS-indexed journal. An improved unification in our bibliometric database of institution names will allow analysing reasons why certain institutions rely on arXiv. Is it due to the presence of large physics departments, research centres, outstanding and highly-active researchers, collaboration or cutting-edge research? Moreover, a qualitative study of authors and their reasons to cite arXiv instead of the journal article would provide valuable information on the recent scholarly communication process.

References

Amat, C. B. (2008) Editorial and publication delay of papers submitted to 14 selected Food Research journals. Influence of online posting. *Scientometrics* 74, 3, 379-389.

Antelman, K. (2004). Do open-access articles have a greater research impact? *College and Research Libraries*, 65 (2004) 372 – 382.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science & Technology*, 57(8), 1060–1072.

Brown, C.M. (2001). The E-volution of preprints in the scholarly communication of physicists and astronomers. *Journal of the American Society for Information Science and Technology*, 52(3), 187–200.

Brown, C. (2003). The role of electronic preprints in chemical communication: Analysis of citation, usage, and acceptance in the journal literature. *Journal of the American Society for Information Science & Technology*, 54(5), 362–371.

Davis, P.M., & Fromerth, M.J. (2007). Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2), 203–215.

Garner, J., Horwood, L., & Sullivan, S. (2001). The place of eprints in scholarly information delivery. *Online Information Review*, 25(4), 250–253.

Gentil-Beccot et al. (2009). Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course. *Journal of the American Society for Information Science and Technology*, 60 (2009) 150–160.

Gentil-Beccot, A., Mele, S., & Brooks, T.C. (2009). Citing and reading behaviours in high-energy physics. How a community stopped worrying about journals and learned to love repositories. Retrieved January 6, 2015 from: arXiv:0906.5418.

Ginsparg, P. (1994). First steps towards electronic research communication. *Los Alamos Science,* 22, 156-165.

Goldschmidt-Clermont, L. (1965). Communication Patterns in High-Energy Physics. Retrieved January 6, 2015 from: http://eprints.rclis.org/archive/00000445/02/communication_patterns.pdf

Haque, A., & Ginsparg, P. (2009). Positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science and Technology*, 60(11), 2203–2218

Haque, A., & Ginsparg, P. (2010). Last but not least: Additional positional effects on citation and readership in arXiv. *Journal of the American Society for Information Science & Technology*, 61(12), 2381-2388.

Harnad, S. & Brody, T (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals*, D-Lib Magazine* 10.

Henneken, E.A. et al. (2007). E-prints and journal articles in astronomy: A productive co-existence. *Learned Publishing*, 20, 16-22.

Heuer, R.-D., Holtkamp, A., Mele, S. (2008). Innovation in Scholarly Communication: Vision and Projects from High-Energy Physics. pp.1-15. DESY-08-054. Retrieved January 6, 2015 from: https://bib-pubdb1.desy.de/record/86123/files/getfulltext.pdf

Kurtz, M & Henneken, E. (2007). Open Access does not increase citations for research articles from The Astrophysical Journal, Retrieved January 6, 2015 from: arXiv:0709.0896

Langer, James. (2000). "Physicists in the new era of electronic publishing." *Physics Today Online*, 53(8):35-38.

Larivière, V., Sugimoto, C.R, Macaluso, B., Milojevic', S., Cronin, B. & Thelwall, M. (2014). arXiv E-Prints and the Journal of Record: An analysis of Roles and Relationships. *Journal of the American Society for Information Science & Technology*, 65(6):1157–1169.

Lim, D. (1996). Preprint Servers: A New Model for Scholarly Publishing? *Australian Academic and Research Libraries (AARL)* 27 (1), 21–30.

Luwel, M. (1998). Publication delays in the science field and their relationship to the ageing of scientific literature. *Scientometrics*, 41, 29-40.

Metcalfe, T.S. (2005). The rise and citation impact of astro-ph in major journals. *Bulletin of the American Astronomical Society,* 37, 555–557. Retrieved January 6, 2015 from: http://arXiv.org/abs/astro-ph/0503519

Metcalfe, T.S. (2006). The citation impact of digital preprint archives for solar physics papers. *Solar Physics,* 239, 549-553.

Moed, H.F. (2007). The effect of 'Open Access' on citation impact: An analysis of arXiv's condensed matter section. *Journal of the American Society for Information Science & Technology*, 58(13), 2047–2054.

Prakasan, E.R. & Kalyane, V.L. (2004). Citation analysis of lanl high energy physics e-prints through Science Citation Index (1991-2002). Retrieved January 6, 2015 from: http://eprints.rclis.org/archive/00002200/

Schwarz, G.J., Kennicutt, R. C. J. (2004). Demographic and citation trends in astrophysical journal papers and preprints. *Bulletin of the American Astronomical Society,* 36 (2004), 1654–1663.

Shuai, X., Pepe, A. & Bollen, J. (2012). How the Scientific Community Reacts to Newly Submitted Preprints: Article Downloads, Twitter Mentions, and Citations. *PLoS ONE* 7(11): e47523.

Youngen, G.K. (1998). Citation patterns to traditional and electronic preprints in the published literature. *College & Research Libraries*, 59(5), 448–456.

Evolutionary Analysis of Collaboration Networks in *Scientometrics*

Yuehua Zhao[1], Rongying Zhao[2]

[1]*yuehua@uwm.edu*
University of Wisconsin-Milwaukee Milwaukee, School of Information Studies, P. O. Box 413 Milwaukee, WI 53201 (United States)

[2]*zhaorongying@126.com*
Wuhan University, School of Information Management, Research Center for China Science Evaluation, The Center for the Studies of Information Resources, Wuhan, Hubei 430072 (China)

Abstract
The research area of scientometrics began during the second half of the 19th century. After decades of growth, the international field of scientometrics has become increasingly mature. The present study intends to understand the evolution of the collaboration network in *Scientometrics*. The growth of the discipline is divided into three stages: the first time period (1978-1990), the second period (1991-2002), and the third period (2003-2014). Both macro-level and micro-level network measures between the studied time periods were compared. Macro-level analyses show that the degree distribution of the collaboration in each timespan are consistent with power-law, and both the average degree and average distance steadily increase with time. Micro-level structure analyses illustrate the authors with high performance in raw degree measure, degree centrality measure, and betweenness measure are dynamic in different timespans. From three dimensions (raw degree, degree centrality, and betweenness centrality), the collaboration dominators are identified in each time span. In addition, the visualization methods are applied to display the evolution of the collaboration networks for each of the three stages of scientometrics' development.

Conference Topic
Journals, databases and electronic publications

Introduction
Scientometrics is an interdisciplinary field that uses mathematical, statistical, and data-analytical methods and techniques to perform a variety of quantitative studies of science and technology (Chen, Börner, & Fang, 2013). In short, it can be defined as the science of science. The term "Scientometrics" has been first used as a translation of the Russian term "naukometriya" (measurement of science) coined by Nalimov and Mulchenko (1969). The research area of scientometrics began during the second half of the 19th century. This paper proposed a macro- and micro-level overview of the author collaboration patterns in journal *Scientometrics* to study the evolution of the field of scientometrics. The present study intends to understand the evolution of the collaboration network in *Scientometrics*. In this study, social network analysis methods are employed to describe the evolution of scientometrics over nearly 40 years after entering the development stage of this field. Both macro-level and micro-level network measures between the studied time periods were compared. Then, visualization methods were applied to display the evolution of the collaboration networks in three periods: the first time period (1978-1990), the second period (1991-2002), and the third period (2003-2014).

Related Works and Research Questions
Scientometrics has been studied for more than 100 years. Over the past years, scientists' studies of scientometrics shifted from the unconscious to consciousness, from qualitative research to quantitative research, and from external description to detailed study revealing the inherent properties of scientific production. Previous scholars (Pang, 2002; Yuan, 2010) tend to divide the development of scientometrics into three stages: embryonic period (from the

second half of the 19th century to early 20th century), the founding period (from the beginning of the 20th century to the 1960s), and development period (after the 1970s). In order to study the development period of scientometrics, Schubert (2002) indicated that as the representative communication channel of its field, the journal *Scientometrics* reflects the characteristic trends and patterns of the past decades in scientometric research. Therefore, in this study, we employed the publications in *Scientometrics* over the past 37 years to detect the evolution of the scientific collaboration networks in this field.

Previous research has provided some insight into the author collaboration network analysis in different disciplines. Barabasi et al. (2002) investigated the collaboration network in mathematics and neuroscience articles published between 1991 and 1998. Newman (2001) compared the co-authorship networks of in physics, biomedical research, and computer science, and found the differences of the collaboration networks between experimental and theoretical disciplines. By using the bibliometric methods, Ardanuy (2012) analyzed the level of co-authorship of Spanish research in Library and Information Science (LIS) until 2009, and found a significant increase in international collaboration. Given the advanced visualization techniques, Franceschet (2011) represented a collaboration picture of computer science collaboration including all papers published in the field since 1936.

These studies have investigated the collaboration networks in different disciplines and compared their differences. However, few studies investigated the field of scientometrics over the past 37 years. There is a need for researchers to identify and compare both the macro-level and micro-level characteristics of the scientific collaboration network in *Scientometrics* through different time periods.

This paper intended to address the following two research questions:

RQ1. What are the macro-level features of the collaboration networks in *Scientometrics* in each time period?

RQ2. What are the micro-level features of the collaboration networks in *Scientometrics* in each time period?

Method

Data collection

For the development period of scientometrics, the foundation of the journal *Scientometrics* (in September, 1978) is a landmark event. Following some of the predecessors (Schoepflin & Glänzel, 2001; Hou, 2006), this study used the journal as a representative model of scientometrics research. The research data involves 3627 documents published in *Scientometrics* during 1987 to 2014 retrieved from the Web of Science on December 10th, 2014, and the other 347 articles published from 1978 to 1986 retrieved on April 20th, 2013. The total of 37 years were divided into three periods: the first time period (1978-1990), the second period (1991-2002), and the third period (2003-2014).

The raw data extracted from Web of Science database that consisted of the bibliometric information of each paper. Microsoft Excel was applied to build the 2-mode author-to-paper matrices for each time period. In order to produce the collaboration networks, the 2-mode author-to-paper matrices were transferred to 1-mode author-to-author matrices based on the formula proposed by Breiger (1974): $P=A(A^T)$. In this case, the matrix A was the 2-mode author-to-paper matrix and the matrix AT was the transposition of the matrix A, and the 1-mode author-to-author matrix was generated by multiplying these two 2-mode matrices. In the produced author-to-author matrix, each row and column represented an author, the intersection cells contained the cumulative number of the co-authored papers by two authors, and the diagonal cells demonstrated the total number of papers written by each author.

Data analysis

Two social network analysis software packages (Ucinet and Netdraw) (Borgatti, Everett, & Freeman, 2002) were adopted in the data analysis to calculate the network measures and draw the networks. Ucinet is a software package which mainly deals with the social network analysis, and Netdraw, the network visualization tool, can be used to display the networks generated by Ucinet.

Results and Discussion

An overview

Over the 37 years, a total of 4,211 authors published 3,974 papers in *Scientometrics*. Figure 1 indicates the distribution of the number of articles and the number of scholars in each time period. In Figure 1, the *X*-axial represented the 3 time periods, and the *Y*-axial represented the frequencies, and the 2 bars in each period showed the number of authors and articles separately, and the line showed the trend of the differences between the two bars. Separately, 626 papers were contributed to by 435 authors from 1978 to 1990, 1,106 papers were published by 1,029 authors from 1997 to 2005, and 2,242 papers were written by 3,102 authors from 2006 to 2014. Based on Figure 1, both the number of articles and the number of authors increased over the three time spans. When we compared the two frequencies in each period, the number of articles was greater than the number of authors at the first two stages, but the number of authors boomed at the third stage which resulted in the number of authors being much greater than that of the authors. The increases of the total number of articles and authors suggested the rises of the collaboration opportunities through the three time periods.

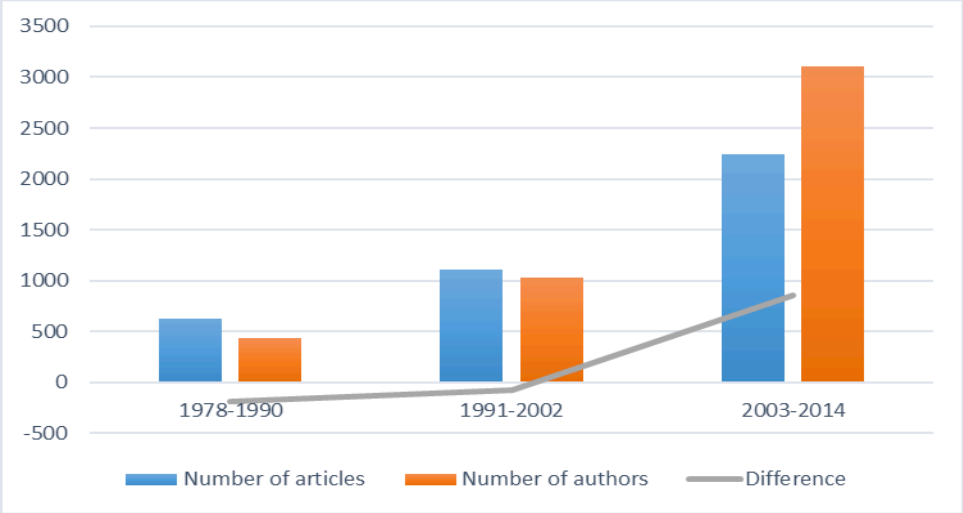


Figure 1. Distribution of the number of articles and authors in three time periods.

Macro-level structure analysis

In order to study the evolution of the scientific collaborations through three time periods, three 1-mode author-to-author matrices were plugged in Ucinet to calculate a variety of network measurements. There are a number of measures which can be used to evaluate the structure of a network. In this study, we will mainly focus on four elements to approach: degree distribution, average degree, average distance, and cluster coefficient.

The number of collaborators that each author has in a collaboration network is the degree of a node (Ding, Rousseau, & Wolfram, 2014). In Figure 2, three lines illustrated the distributions

of the node degree in each time span, respectively. The *X*-axial represented the number of authors, and the *Y*-axial represented the degree of the authors. From Figure 2, it can be seen that most authors held the low degree in all three periods. Based on the locations of three distribution lines, more authors tended to join more collaborations from 1978 to 2014 with the increase of the number of total authors published on the journal.

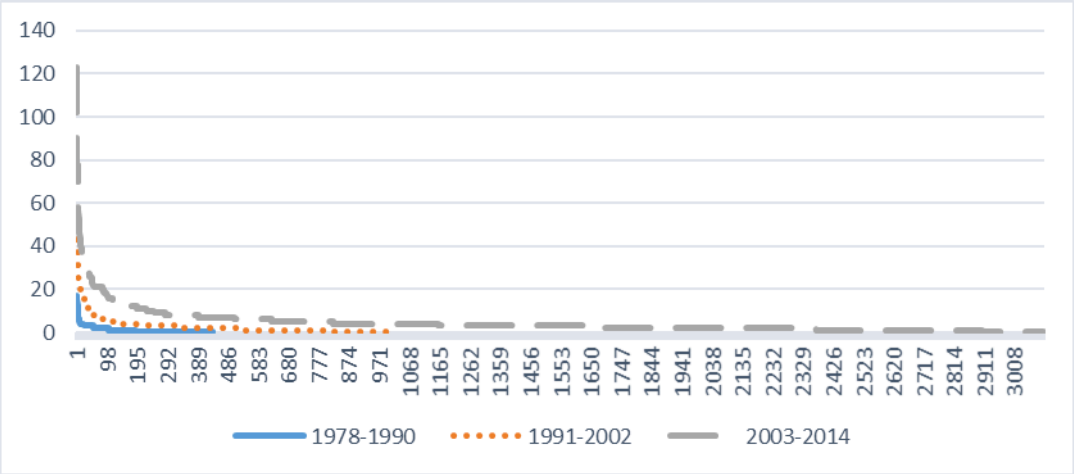
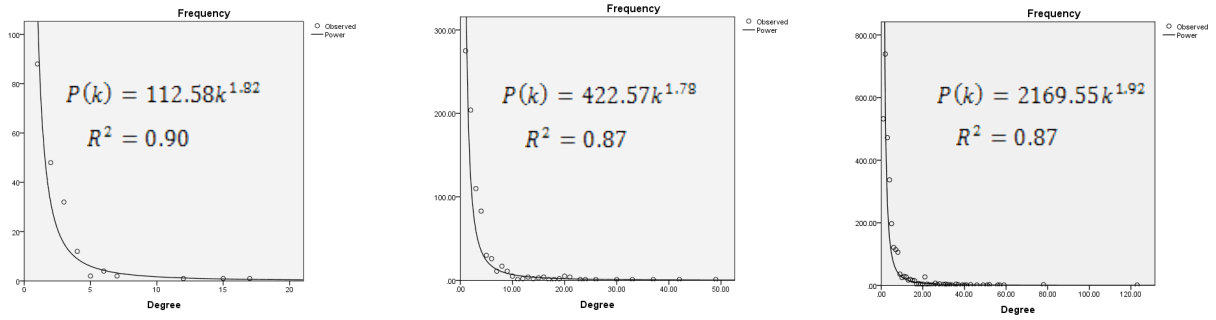


Figure 2. Degree distribution for authors in three time periods.

The degree distribution characterizes the spread of the edges each node has in a network. Although the degree distribution of a random graph is a Poisson distribution, Albert and Barabási (2002) have discovered that, for most large networks, the degree distribution has a power-law tail: $P(k) \sim k^{-\gamma}$, where $P(k)$ is the distribution function. In this study, the distributions of the collaboration network in each period were calculated and drawn in Figure 3. Power-law regression model was used to detect the degree distribution patterns in different timespans (Albert & Barabási 2002). Figure 3 illustrated the modeling results for the three periods, and the x-axis plots low degree nodes on the left and high degree nodes on the right; the y-axis indicates their probability. In both cases, power-law model performed the good fits to the observed data. In relationship between the degree of the authors and the corresponding frequencies can be estimated by: $P(k) = 112.58k^{1.82}$ with $R^2 = 0.90$ in 1978-1990, $P(k) = 422.57k^{1.78}$ with $R^2 = 0.87$ in 1991-2002, and $P(k) = 2169.55k^{1.92}$ with $R^2 = 0.87$ in 2003-2014. As discussed by Albert and Barabási (2002), the degree distribution of the collaboration network of high-energy physicists reach the almost perfect power-law



1978-1990 1991-2002 2003-2014

Figure 3. Degree distribution plots for collaboration networks.

with an exponent of 1.2, while the collaboration networks of mathematicians and neuroscientists between 1991 and 1998 held the degree exponents 2.1 and 2.5 (Barabasi et al., 2002). Comparing with those previous studies in different disciplines, the degree distribution of the collaboration of *Sicentometrics* in each timespan were consistent with power-law with degree exponents 1.82, 1.78, and 1.92, respectively. In addition to degree distribution, previous studies proved that there were several other useful indicators to feature a social network. Table 1 represented the four key measures for each time periods. Figure 3 describes the changes of each measure between 1978 and 2014.

Table 1. Four key measures of the collaboration networks in each time periods.

	1978-1990	*1991-2002*	*2003-2014*
Average Degree	0.794	2.101	3.435
Average Distance	1.412	4.673	7.106
Clustering Coefficient	0.941	0.873	9.014
Components	309	420	701
Diameter	4	11	19

Average degree is calculated by counting the average number of links per author (Barabasi et al., 2002). In the collaboration network, the average degree characterizes the interconnectedness between authors. Yin, Kretschmer, Hanneman, and Liu (2006) identified that the higher the average degree, the tighter the network. From Table 1, we can see that the average degree steadily increased with time, which demonstrated that authors cooperated more often. This results confirmed Barabasi et al.'s (2002) observations in Mathematics and Neuroscience. One possible reason might be the sharp increase of the total number of authors led to more possible connections between the new authors and also between the new authors and the existing authors.

The distance between two nodes is measured by the length of the shortest path between those two nodes. Average distance in a network is calculated by the average length of the geodesic paths between all reachable pairs of nodes (Borgatti, Everett, & Freeman, 2002). From Table 1, the average distance of the collaboration networks started form 1.412 (in 1978-1990), grew to 4.673 (in 1991-2002), and finally reached 7.106 (in 2003-2014). Watts and Strogatz (1998) examined that many social networks show a “small world” phenomenon that have small characteristic path lengths. According to Yin et al. (2006), short average distance allows authors to share information more rapidly. In this case, the average distance of the collaboration network enlarged with time, but actors were still able to reach the others within short paths in all periods. The cluster coefficient for the co-authorship network in *Scientometrics* appeared to have increased sharply: rising from 0.941 in 1978-1990 to 9.014 in 2003-2014.

Micro-level structure analysis

Micro-level structure analysis was adopted to measure the individual authors. One of the main purpose of social network analysis is to identify the core actors in a network. We applied four measures (raw degree, degree centrality, betweenness centrality, and closeness centrality) to investigate the structural characteristics of each author in each timespan.

Table 2 summarized the top 10 authors with highest degrees in each time period. Freeman (1978) defined the degree of a point as the number of other points to which a given point is adjacent. In the collaboration networks, the degree of an author represents the number of authors a given author co-authored with before. Schubert A held the highest degree with 17 in the first period, which showed he cooperated with 17 authors between 1978 and 1990. In both

second and third timespan, Glänzel W. achieved the first place with 49 and 123 collaborators in 1991-2002 and 2003-2014, respectively.

Table 2. Raw degree (top 10 authors) in each time period.

1978-1990		*1991-2002*		*2003-2014*	
Schubert, A	17	Glänzel, W	49	Glänzel, W	123
Braun, T	15	Schubert, A	42	Chen, DZ	78
Zsindely, S	12	Braun, T	37	Huang, MH	78
Moed, HF	7	Moed, HF	33	Debackere, K	59
Vanraan, AFJ	7	Gupta, BM	30	Zhang, X	57
Burger, WJM	6	Gomez, I	26	Rousseau, R	56
Courtial, JP	6	Courtial, JP	24	Gorraiz, J	52
Frankfort, JG	6	Rivas, AL	23	Thijs, B	52
Lepair, C	6	Dore, JC	21	Abramo, G	51
Lancaster, FW	5	Miquel, JF	21	D'Angelo, CA	49

Apart from the raw degree of the actors, the centrality is one of the most important structural attributes of social networks (Freeman, 1978). Over the past years, a number of centrality measures have been proposed by sociologists. In the case of co-authorship network, each centrality measure demonstrate special characteristics of the author cooperation. The centrality indicators are designed to identify the "core" authors from different perspectives. The degree centrality can be seen as an index of its potential communication activity. For the co-authorship network, the authors with high degree centrality may result in the status of "elite" (Yin et al., 2006). Freeman's (1978) betweenness centrality is based upon the frequency with which a point falls between pairs of other points on the shortest or geodesic paths connecting them. Regarding to the collaboration, betweenness centrality can be used to assess the potential of an author for control of communication in the knowledge flow network. Tables 3 and 4 summarized the top 10 authors with the highest degree and betweenness centralities in each time period, respectively.

From Table 3, we can see that authors with high degree centrality were dynamic in different timespans. New authors arrived in a field and gathered more collaborations, whereas the existing authors decayed, to some extent, with time. No author ranked in the top 10 in all three time periods. From the perspective of potential communication ability, the "star" of the collaboration networks changed over time. When it comes to the betweenness centrality, Glänzel W was no doubt the core author in both the second and third time periods. Interestingly, from both dimensions (degree centrality and betweenness centrality), Glänzel W occupied the genuine dominator (or "star") position from 2003 to 2014, which suggests that he possesses potential communication ability as well as the possible ability to control the communication between other authors in recent years.

Collaboration network visualization

Figures 4 to 6 present the evolution of the collaboration network in the three stages. Clearly, both the number of the authors and the collaborations boosted, which also illustrated the expansion of this field. With the time advanced, the collaborations between authors were strengthened. To highlight the changes in collaboration, we removed removed isolated nodes in the network in both Figures and displayed only the collaborating authors and their connections. The size of both the nodes and the labels indicated the degree of the authors. The strength of the collaboration was shown by the thickness of the ties between nodes. The authors with high degree in Table 2 were outstanding in the networks.

Table 3. Degree centrality (top 10 authors) in each time period.

1978-1990		*1991-2002*		*2003-2014*	
Courtial, JP	1.379	Moed, HF	1.846	Glänzel, W	1.419
Lepair, C	1.379	Courtial, JP	1.652	Rousseau, R	1.387
Lancaster, FW	1.149	Gupta, BM	1.458	De Moya-Anegon, F	0.967
Braun, T	0.92	Rousseau, R	1.458	Ho, YS	0.935
Dobrov, GM	0.92	Tijssen, RJW	1.458	Borner, K	0.903
Krebs, M	0.92	Glänzel, W	1.361	Park, HW	0.838
Nagy, JI	0.92	Gomez, I	1.263	Thelwall, M	0.838
Plagenz, K	0.92	Rivas, AL	1.263	Chen, DZ	0.838
Porta, MA	0.92	Deshler, JD	1.166	Wu, YS	0.806
Schubert, A	0.92	Gonzalez, RN	1.069	Debackere, K	0.806

Table 4. Betweenness centrality (top 10 authors) in each time period.

1978-1990		*1991-2002*		*2003-2014*	
Braun, T	0.017	Glänzel, W	1.408	Glänzel, W	5.478
Nagy, JI	0.016	Kretschmer, H	1.1	Rousseau, R	3.918
Courtial, JP	0.012	Moed, HF	1.017	Park, HW	2.17
Lepair, C	0.01	Gupta, BM	0.855	Leydesdorff, L	1.661
Schubert, A	0.007	Rousseau, R	0.489	Kretschmer, H	1.478
Dobrov, GM	0.005	Tijssen, RJW	0.397	Ho, YS	1.423
Inhaber, H	0.005	Gomez, I	0.351	Chen, J	1.374
Narin, F	0.005	Luwel, M	0.262	Meyer, M	1.284
Lancaster, FW	0.004	Braun, T	0.261	Huang, JS	1.219
Studer, KE	0.004	Schubert, A	0.259	Aguillo, IF	1.218

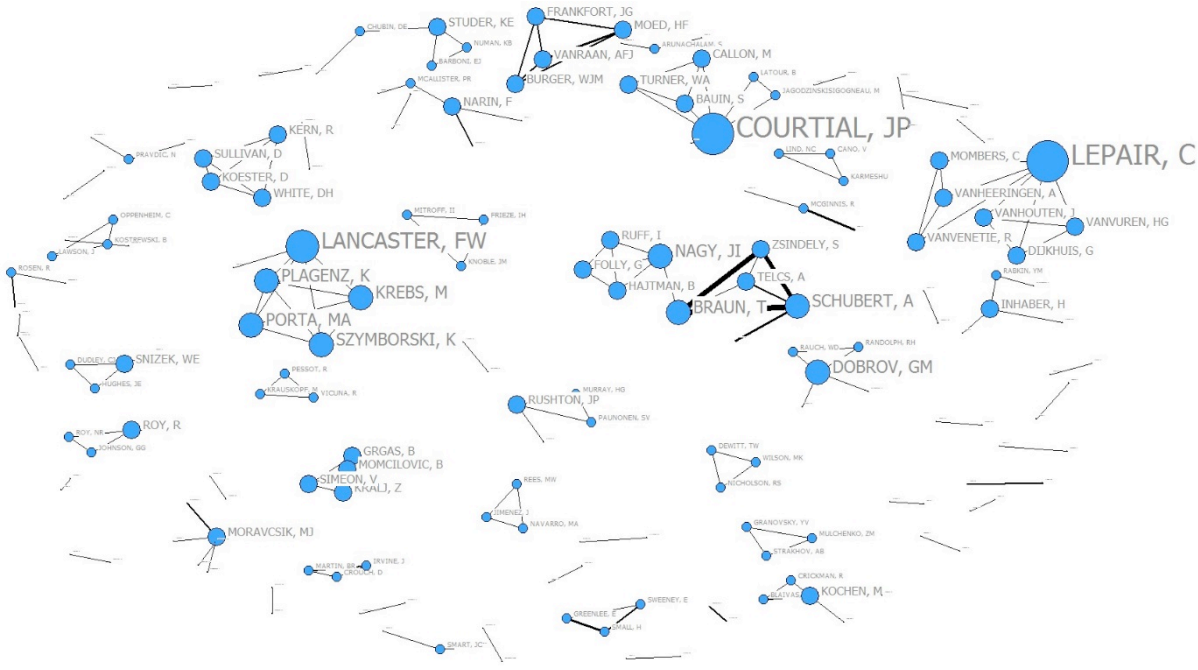


Figure 4. The collaboration networks in 1978-1990.

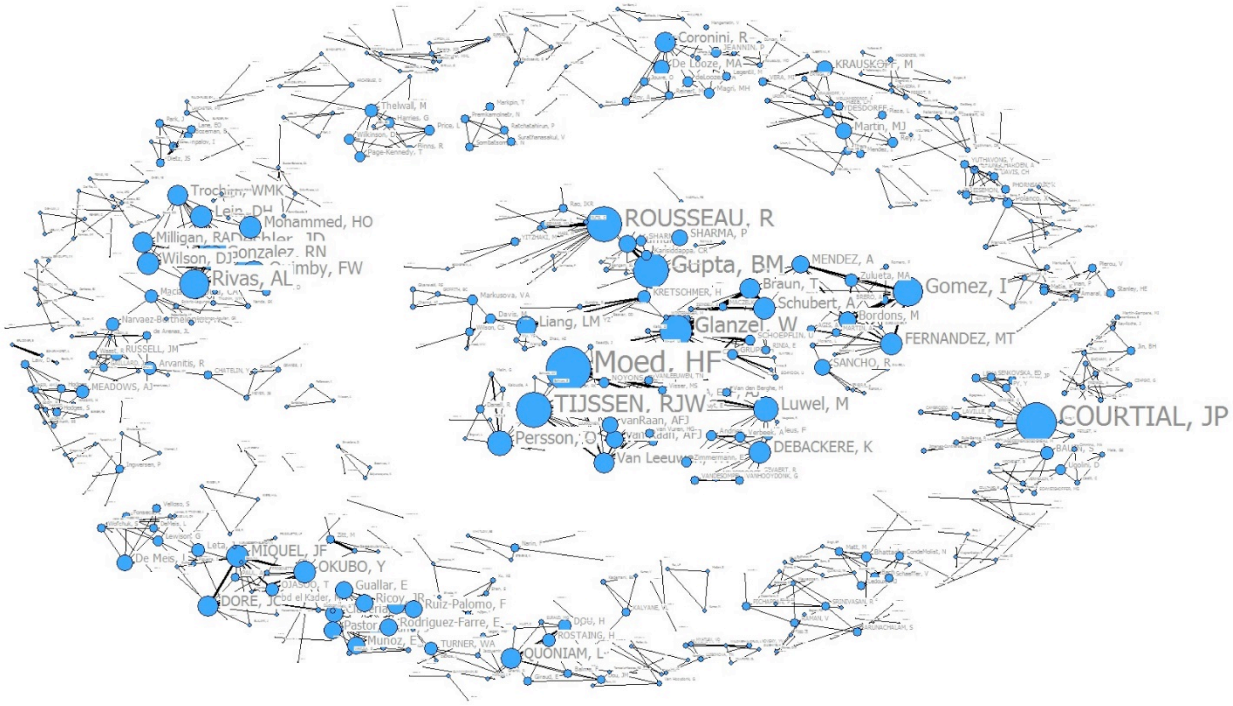


Figure 5. The collaboration networks in 1991-2002.

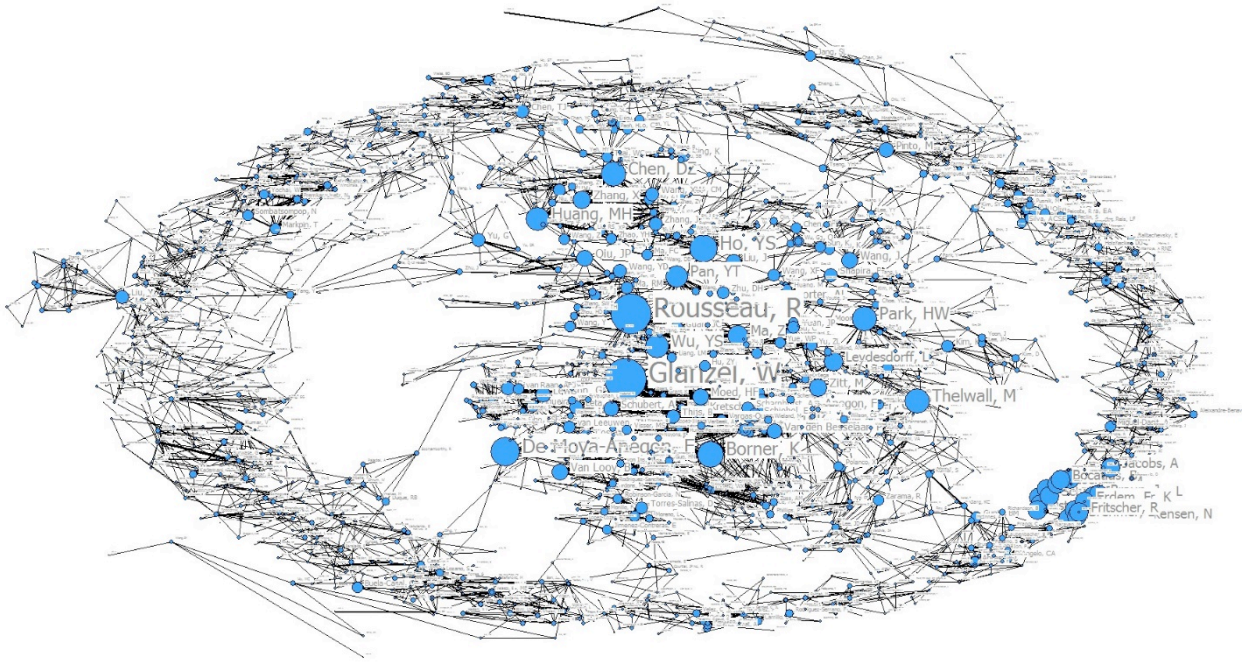


Figure 6. The collaboration networks in 2003-2014.

Conclusion

This paper approached the evolution of the scientific collaboration networks of scientometrics based on the publications in *Scientometrics*. The past 37 years were divided into three timespans: the first time period (1978-1990), the second period (1991-2002), and the third period (2003-2014). Based on the macro-level structure analyses, the degree distribution of the collaboration of *Scientometrics* in each timespan were consistent with power-law, and both the average degree and average distance steadily increased with time, which

demonstrated that the cooperation between authors was getting more frequent. Micro-level structure analyses illustrated the authors with high performance in raw degree measure, degree centrality measure, and betweenness measure were dynamic in different timespans. Interestingly, on each dimension, Glänzel W became the genuine dominator (or "star") in the most recent period: 2003-2014. Finally, the visualization of the evolution of the collaboration network in three stages was presented, and the boosts of the number of authors and their collaborators were displayed in the network graphs.

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97. doi:10.1103/RevModPhys.74.47

Ardanuy, J. (2012). Scientific collaboration in Library and Information Science viewed through the Web of Knowledge: the Spanish case. *Scientometrics*, *90*(3), 877–890.

Barabasi, A. L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and Its Applications*, *311*(3-4), 590–614.

Borgatti, S.P., Everett, M.G. & Freeman, L.C. (2002). *Ucinet 6 for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.

Breiger, R. L. (1974). The duality of persons and groups. *Social Forces*, *53*(2), 181–190.

Chen, Y., Börner, K., & Fang, S. (2013). Evolving collaboration networks in Scientometrics in 1978–2010: a micro–macro analysis. *Scientometrics*, *95*(3), 1051–1070.

Ding, Y., Rousseau, R., & Wolfram, D. (Eds.). (2014). *Measuring Scholarly Impact: Methods and Practice*. New York: Springer.

Franceschet, M. (2011). Collaboration in computer science: A network science approach. *Journal of the American Society for Information Science and Technology*, *62*(10), 1992–2012. doi:10.1002/asi.21614

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Hou, H. (2006). Study on the evolution of scientometrics based on the scientific map. Retrieved from http://www.cnki.net/

Nalimov, V.V. & Mulchenko, B.M. (1969). *Scientometrics*. Moscow: Nauka (in Russian).

Newman, M. (2001). Scientific collaboration networks. Network construction and fundamental results. *Physical Review E*, *64*(1), 016131.

Pang, J. (2002). *The Research Methodology of Scientometrics*. Beijing, China: Scientific and Technical Documentation Press.

Schoepflin, U., & Glänzel, W. (2001). Little Scientometrics, Big Scientometrics…and Beyond? *Scientometrics*, 30: 375-384.

Schubert, A. (2002). The Web of Scientometrics: A statistical overview of the first 50 volumes of the journal. *Scientometics*, 53(1):3-20.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, *393*(6684), 440–442.

Yin, L., Kretschmer, H., Hanneman, R. A., & Liu, Z. (2006). Connection and stratification in research collaboration: An analysis of the COLLNET network. *Information Processing & Management*, *42*(6), 1599–1613.

Yuan, J. (2010). *The Advanced Tutorial of Scientometrics*. Beijing, China: Scientific and Technical Documentation Press.

Open Access Publishing and Citation Impact - An International Study

Thed van Leeuwen, Clifford Tatum, and Paul Wouters

leeuwen@cwts.nl, c.c.tatum@cwts.leidenuniv.nl, p.f.wouters@cwts.leidenuniv.nl
CWTS, Leiden University, Wassenaarseweg 62a, Leiden (the Netherlands)

Abstract
This paper describes the analysis of open access (OA) publishing in the Netherlands in an international comparison. As OA publishing is now actively stimulated by Dutch science policy, similar to the UK, a bibliometric baseline measurement is conducted to assess the current situation, to be able to measure developments over time. For the study we collected data from various sources, and for three different smaller European countries (the Netherlands, Denmark, and Switzerland). Not all of the analyses for this baseline measurement are included here; the analysis presented in this paper mainly focuses on the various ways OA can be defined while using Web of Science, and the problems with interpreting these results. From the data we collected, we can conclude that the way OA is currently registered in various electronic bibliographic databases is quite unclear, and various methods applied deliver results that are different, although the impact scores point in the same direction.

Conference Topic
Journals, databases, and electronic publications

Introduction
Acceleration of open access goals in the Netherlands coincides with implementation of new current research information systems (CRIS) at Dutch universities and research institutes. This deployment of institutional CRIS systems provides an opportunity for national level tracking of open access through coordinated metadata schemes and common registration practices. As open access is notoriously difficult to measure, contemporary analyses often employ random sampling techniques (Archambault et al., 2014; Björk et al., 2010). All publication records in a given sample are tested to determine the proportion of full texts that are open access publications. National level coordination of research information provides an opportunity for improved, more precise assessment of open access publishing. In this study we use bibliographic data to establish a baseline analysis of the proportion of open access publishing in the Netherlands.

Assessment of open access publishing is complicated by a growing diversity of what counts as open access, the copyright restrictions for when a publication can be made openly accessible, and the lack of clear and consistent identification of open access publications in bibliographic data. To examine these challenges we begin with a definition from the Budapest open access Initiative (BOAI):

> Free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited. (BOAI 2002)

This definition highlights two distinct channels of access: (1) human access to read, download, and reuse the full text of published articles; and (2) machine access to crawl, index, or analyze the content of articles. The BOAI also proposes two operational paths to access through open access journals and self-archiving in repositories, subsequently referred to as

Gold open access and Green open access (Bailey, 2005). Hybrid open access generally refers to the situation whereby authors can pay to make their articles in subscription journals openly accessible on the Web (Björk, 2012).

In addition to the broad categories of Gold, Green, and Hybrid modes of open access, multiple versions of a manuscript may exist due to variations in publishers' licensing agreements. These agreements typically specify how, when, and under which conditions a manuscript may be openly accessible on the web. For example, a publisher may allow Green open access through self-archiving in an institutional repository. However, publishers' copyright restrictions differ on the stage of manuscript development that may be openly accessible, thus assigning different rights to different versions of the text. Commonly specified version types include the submitted manuscript (before peer review), the accepted manuscript (peer-reviewed but not formatted), and an exact copy of the published manuscript (Björk et al., 2013). This creates the possibility that the open access version of a manuscript is substantively different from the published version. In such instances, it is unclear whether the open access version has been sufficiently validated through the quality control measures such as peer review.

Another variation is delayed access, which is applied as an embargo period, after which a copy of the publication may be self-archived or the publisher may remove access restrictions on the journal website. Embargo periods are generally specified as a delay of 6, 12, 18, or 24 months after publication, with 12 months being the most common embargo period (Laakso & Björk, 2013). For Green open access, it is thus left to authors and institutions to track and manage a variety of self-archiving policies, which in itself has been shown to be a barrier to open access (Davis & Connolly, 2007). However, this kind of administrative overhead is largely absent from subscription journals that convert articles to open access after a specified delay (e.g. 12 months). In addition, a bibliometric analysis of 'delayed access' journals found journal and article impact factors higher than comparable averages from both subscription journals and direct (no delay) open access journals (Laakso & Björk, 2013).

A common refrain among proponents of open access is that open access publishing yields increased citation impact. While there are conflicting reports regarding an open access citation advantage (OACA), heightened attention to this issue has increased our understanding about citation behaviour more generally. Numerous bibliometric studies claim that open access publishing results in a significant increase in citations. In these studies the size of advantage varies widely based on a variety of issues, such as disciplinary differences, methodological approaches, variation in how open access is defined, and difficulty in determining when an article is made openly accessible (Swan, 2010). In addition, a number of confounding factors have been shown to influence citation frequency such as early exposure to draft versions of a manuscript (Moed, 2007), self-selection bias whereby an author may choose open access for only her best publications (Kurtz et al., 2007), the availability at multiple access points (Xia, Myers & Wilhoite, 2011), and physical proximity of researchers (Lee et al., 2010).

To control for these factors, Davis et al. (2008) employ randomized controlled trial methods, whereby randomly selected articles in subscription based journals are switched to open access. The resulting configuration is similar to hybrid open access, such that the article is made to be openly accessible and is listed among the non-open access articles on the journal's website. In the Davis et al. (2008) study a citation advantage was not present. However, the research design used to control for confounding variables (randomized controlled trial) also limited applicability of the findings to the hybrid model of open access. More recently, Archambault et al. (2014) show variation in the accumulation of citations associated with the different modes of open access. The authors find a citation *advantage* most prominently associated with the self-archiving mode of open access (Green OA) and a citation

disadvantage associated with full and immediate open access journals (Gold OA). This study also establishes a general ranking of citation accumulation on the bases of open access, listed in order of most to least: Green OA, Other OA, Not OA, and Gold OA." (Archambault et al., 2014, pp. 20, 24)

To address the variability of circumstances associated with open access publishing, recent studies invert the research design from top-down queries of bibliometric datasets to bottom-up testing whether a publication is an open access publication. This approach involves random sampling of a given publishing domain, harvesting full-texts from the Internet, and analysis of available metadata from harvested manuscripts (Björk et al., 2010). While this approach circumvents much of the variability noted above, it is nevertheless dependent on the presence and quality of metadata. (The potential for improved metadata practices is addressed in the discussion section below.)

The objective of our analysis is to show the challenges of bibliometrically analysing OA publications and associated impact scores. We use Web of Science (WoS) data, either directly retrieved from the database, or combined with article-level data extracted from journals listed in the Directory of Open Access Journals (DOAJ). As both data sources are incomplete with respect to open access publications, the analysis is focused on comparison of relative output and relative impact among three European countries of similar size and scientific production: the Netherlands, Denmark, and Switzerland, in order to show developments in time, as well as differences resulting from both approaches. It is important to note that Green OA articles are excluded from our analysis. While the Netherlands maintains a robust national repository for Green OA (NARCIS), there is not yet a reliable system of identifying the self-archived state of publications within bibliometric datasets. As such, the proportion of open access and associated impact comparisons are limited to the available data on Gold OA.

Data collection

In the study we make use of data from various sources. The Web of Science (WoS) database is used in its internet version, available to most Dutch researchers. We also used the CWTS version of the WoS, a tailor-made database based upon state-of-the-art bibliometric techniques and indicators. In this version, the functionality to search for OA output is not yet available. Finally, we make use of the journals and the publications listed in the Directory of Open Access Journals (DOAJ). From this data source, we will further focus on the digital object identifiers (DOIs), while leaving out other elements (such as the license types, as this information is unclearly defined as well as unclearly linked to the publications).

Method I: The first way of data collection from WoS starts from the desktop interface of the WoS database. The functionality to collect this information is not yet available in the in-house WoS database at CWTS, so therefore we had to collect these data from the internet version directly. This approach involved the following steps:

1) Collect the output of one of the selected countries for a particular year;
2) Within that set, further distinguish the OA part of that selected output;
3) Download these publications from the WoS database (including the so-called UT-code, a unique identifier within WoS that allows for linking to the CWTS WoS database);
4) Select within the CWTS database the output for the three countries;
5) Match the selected output from the Internet version of the WoS with the in-house CWTS version;
6) Create two sets within the CWTS database, an OA formatted set of publications, and a non OA formatted set of publications.

These steps were taken for all three countries, collecting publications from 2000-2013.

The definition of how the publications were defined as OA is based upon the following statement on the WoS database' website: "The Thomson Reuters Links open access Journal

Title List includes free journal content that are available for linking from the Web of Science."

Method II: The second method started from the Directory of Open Access Journals (DOAJ). This list contains journals that have implemented the Gold open access business model. CWTS has downloaded the complete list, and all publications published in the journals on the DOAJ list. By making use of this dataset, we could use a second approach to the OA output of the three countries taking the following steps:

1) First select within the CWTS database the output for the three countries;
2) Collect their Digital Object Identifiers (doi);
3) Match these with the doi's of the publications downloaded from the DOAJ list;
4) Create two sets within the CWTS database, an OA formatted set of publications, and a non OA formatted set of publications.

We focused on articles, letters and reviews only, excluding other types of documents such as editorials, meeting abstracts, book reviews, etc. The choice for these types is based upon the importance of these three types in communicating scientific findings among peers, and their relative homogeneity within the system.

Methods

In the study we present a number of indicators. In cases we present numbers of publications, this is indicated with a P. In case citation data are presented, we use MNCS (Mean Normalized Citation Score), as well as the MNJS, the field normalized journal impact indicator, to indicate the normalized impact scores in the study (Waltman et al., 2011a; Waltman et al., 2011b). While the output indicator can be used for the various electronic systems we use in the study, and P can relate to various document types analysed, the citation impact indicators are used only within the context of the WoS database. In case of the impact indicators, the length of the citation window is one year longer than the presented year block (so in case of the last block, 2009-2012, the citation impact is measured up until 2013, currently the last year fully covered in the CWTS WoS database).

Results

First we present the results from Method I, described above. The output numbers of the three countries according to the methodology I are found in Table 1 along with the two separate parts of the output, distinguished by openness. The analysis covers the period 2000 up until 2012 for publication data, and up until 2013 for citation impact data. In this analysis we use moving publication year windows, in order to create more solid and stable trend lines, as we are more interested in the trends than in variation from year to year.

The data presented in Table 1 clearly show that OA publishing is becoming increasingly important, in all three selected countries. The Netherlands is lagging somewhat behind Denmark and Switzerland, albeit with only a small part of the total output.

In Figure 1, we have distinguished between the open access format output of the three countries (indicated by the 'Ex OA' label to the country names). What we observe are increasing trends for the parts of the output not published in OA format, which is also visible for the OA format of the output of these three countries, and as shown above in Table 1, increases somewhat faster for Denmark and Switzerland as compared to the Netherlands.

Table 1. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output, 2000-2012.

	NL Ex OA	NL OA	Share OA	DK Ex OA	DK OA	Share OA	CH Ex OA	CH OA	Share OA
2000 - 2003	75607	712	1%	30616	452	1%	53283	995	2%
2001 - 2004	78087	858	1%	31262	557	2%	54793	1220	2%
2002 - 2005	81849	1180	1%	31972	728	2%	56982	1836	3%
2003 - 2006	85386	1663	2%	33024	949	3%	60319	2217	4%
2004 - 2007	88745	2349	3%	34082	1244	4%	63205	2790	4%
2005 - 2008	92349	3265	4%	35273	1631	5%	65920	3517	5%
2006 - 2009	96278	4269	4%	36672	1997	5%	69518	3912	6%
2007 - 2010	101270	5587	6%	38726	2554	7%	72687	4981	7%
2008 - 2011	106560	7299	7%	41417	3264	8%	76658	6354	8%
2009 - 2012	111990	9504	8%	44264	4420	10%	80786	7990	10%

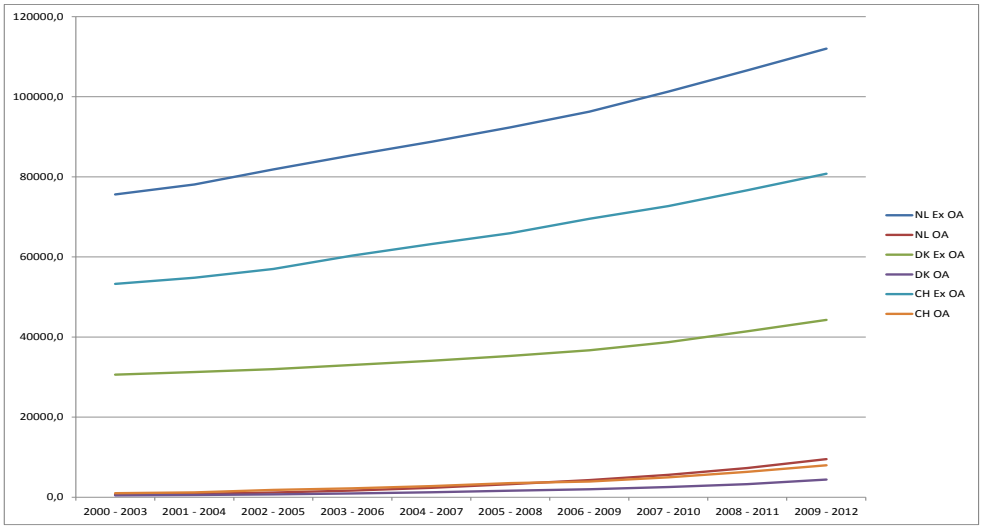


Figure 1. Output development (P) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013.

In Table 2, we present the citation impact scores as represented by the MNCS indicator, the field normalized impact of the outputs of the three countries, again separated by the two types of publication output: open access and non-open access publications.

Figure 2 shows that for all three countries the non-OA part of the output has a citation impact well above world average, with Switzerland topping the other two countries, which have a nearly equal field normalized impact score. The impact of OA publications is lower for all three countries. The impact of the OA part of the national outputs of Denmark and Switzerland were initially well above world average. This is also the case for Swiss publications, as the OA format published output is lower on MNCS only from 2007-2010/2011 onwards. In case of Denmark, this drop started somewhat earlier, while in the case of the Netherlands, the OA output never got an impact higher than that of the non-OA format output. Another interesting phenomenon is the increase of the gap between the impact of OA and non-OA output. This is particularly the case for Switzerland and Denmark, where we observe a clear drop of the impact of OA format output compared to their non-OA formatted output, and to a lesser extent for the Netherlands, where the two impact lines are more slowly diverging.

Table 2. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output, 2000-2012.

	NL Ex OA	NL OA	DK Ex OA	DK OA	CH Ex OA	CH OA
2000 – 2003	1,29	0,99	1,30	1,03	1,37	1,11
2001 - 2004	1,30	0,95	1,29	1,31	1,35	1,21
2002 - 2005	1,30	0,99	1,29	1,39	1,36	1,36
2003 - 2006	1,31	1,07	1,31	1,34	1,36	1,46
2004 - 2007	1,30	1,12	1,31	1,30	1,38	1,47
2005 - 2008	1,31	1,13	1,32	1,30	1,39	1,48
2006 - 2009	1,35	1,15	1,34	1,26	1,39	1,39
2007 - 2010	1,38	1,17	1,37	1,26	1,42	1,37
2008 - 2011	1,40	1,18	1,40	1,25	1,46	1,36
2009 - 2012	1,44	1,18	1,44	1,18	1,50	1,33

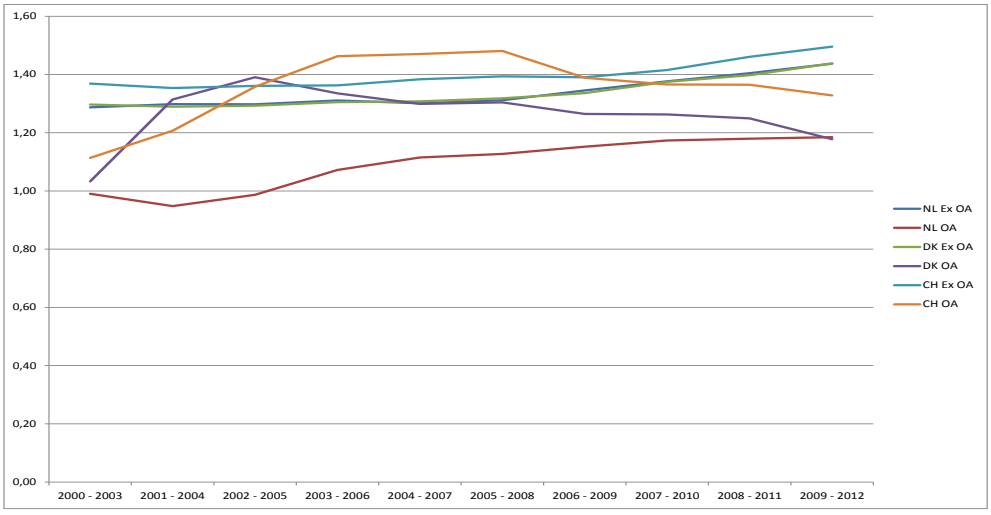


Figure 2. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013.

If we shift our focus towards the journal impact analysis (see Table 3 and Figure 3), for which we use the indicator MNJS, we see an even more interesting phenomenon. While the output in non-OA format published journals shows a choice for journals with increasing impact scores, the OA format published outputs end up in journals with decreasing field normalized impact scores. We even notice a diverging trend in these two clusters of trend lines: non-OA format published journals tend to show increasing impact scores, while OA format published journals show decreasing impact trends. This is striking since these are three of the 'scientifically stronger' nations, as far as can be measured with bibliometric instruments.

Here we start with the results from methodology II. The results of the output analysis are shown in Table 4, which again covers a similar distinction between OA and non-OA format output, but now according to the definition described above under Method II. We combined the DOIs of journals on the DOAJ list with the DOIs available in the WoS. From the total set of 787,611 DOIs in the DOAJ list, we matched 226,641 publications in WoS on the basis of available DOIs. The reason for this seemingly low recall is twofold. In the first place, not all journals covered by the DOAJ list are processed for the WoS database, and secondly, not all publications in journals covered in WoS do contain DOIs. This means that for some journals that are both covered in the DOAJ list as well as in WoS, a match is impossible, particularly

for the earlier years in the analysis. Like the first methodology we followed, we separated the OA format published output from the Netherlands, Denmark, and Switzerland from the total set of publications for the three countries under study.

Table 3. Journal-to-field citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output, 2000-2012

	NL Ex OA	NL OA	DK Ex OA	DK OA	CH Ex OA	CH OA
2000 - 2003	1,18	0,95	1,15	0,84	1,19	1,06
2001 - 2004	1,19	0,97	1,16	1,02	1,20	1,03
2002 - 2005	1,19	1,00	1,16	1,08	1,20	1,19
2003 - 2006	1,20	1,06	1,16	1,11	1,20	1,20
2004 - 2007	1,22	1,09	1,18	1,12	1,22	1,11
2005 - 2008	1,24	1,09	1,20	1,10	1,24	1,14
2006 - 2009	1,26	1,11	1,22	1,07	1,26	1,11
2007 - 2010	1,29	1,11	1,25	1,06	1,29	1,11
2008 - 2011	1,30	1,10	1,26	1,05	1,31	1,11
2009 - 2012	1,32	1,09	1,28	1,00	1,33	1,09

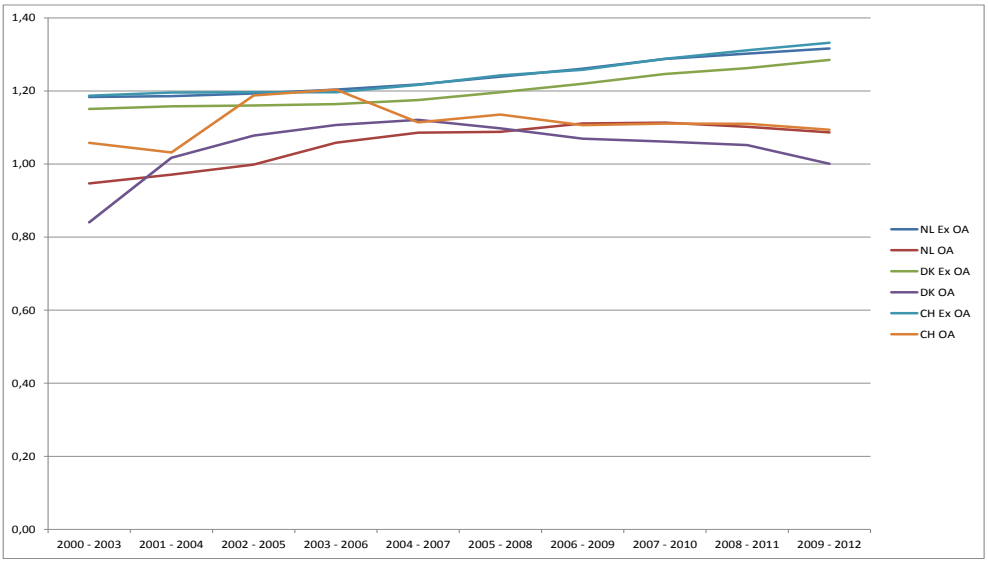


Figure 3: Journal impact development (MNJS) of Denmark, the Netherlands, and Switzerland, 2000-2012/2013.

First of all, we observe that the overlap between the DOAJ list/WoS combinations with Dutch/Danish/Swiss publications in WoS is much smaller compared to the previous analysis on Dutch/Danish/Swiss output in OA format, which is most likely the result of the missing DOIs in the WoS database. If we compare the results of Table 1 with those presented in Table 4, we find much lower shares of OA output compared to the overall output of the three countries. This is further underlined by Figure 4, in which the OA format output of the three countries is at the low end of the graph, while we simultaneously observe a strong increase in the output of the non-OA format output of the three countries.

Table 4. Output (P) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output (based on DOI-matching), 2000-2012

	NL Ex OA	NL OA	Share OA	DK Ex OA	DK OA	Share OA	CH Ex OA	CH OA	Share OA
2000 - 2003	75607	10	0%	30616	4	0%	53283	2	0%
2001 - 2004	78087	35	0%	31262	25	0%	54793	30	0%
2002 - 2005	81849	136	0%	31972	83	0%	56982	97	0%
2003 - 2006	85386	344	0%	33024	170	1%	60319	232	0%
2004 - 2007	88745	648	1%	34082	312	1%	63205	420	1%
2005 - 2008	92349	1068	1%	35273	486	1%	65920	690	1%
2006 - 2009	96278	1531	2%	36672	664	2%	69518	972	1%
2007 - 2010	101270	2207	2%	38726	924	2%	72687	1461	2%
2008 - 2011	106560	3036	3%	41417	1231	3%	76658	2062	3%
2009 - 2012	111990	3896	3%	44264	1595	4%	80786	2608	3%

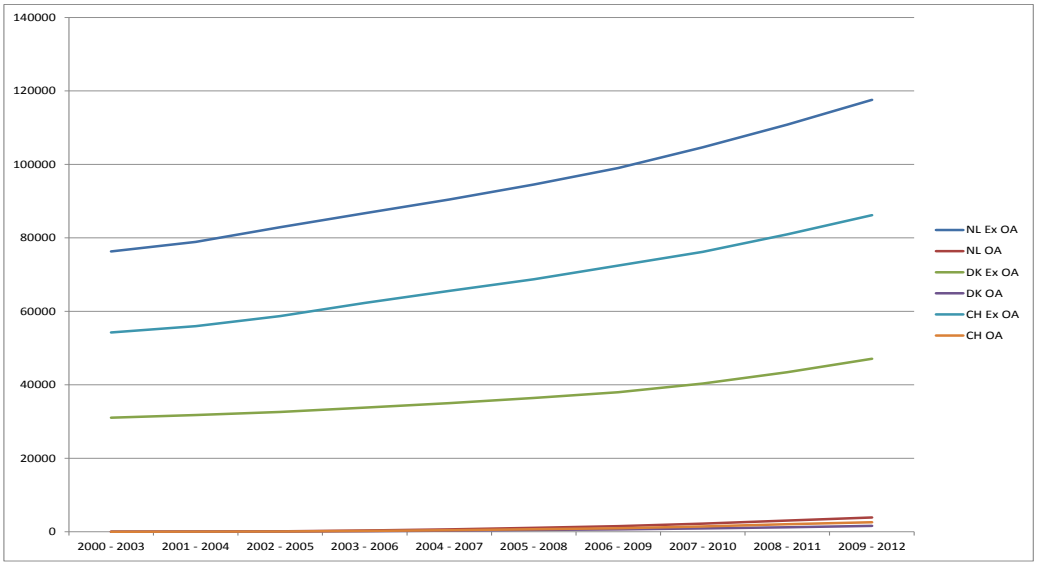


Figure 4. Output development (P) of Denmark, the Netherlands, and Switzerland, based on matching of DOI's, 2000-2012/2013.

In Table 5, we present the impact scores of the three countries, again distinguishing OA format output and non-OA format output. Again we observe lower impact scores for the OA format output of the three countries, except for the starting block of the analysis (please note that the output numbers are extremely low in this part of the analysis for the Netherlands and Denmark, respectively 10 and 4 papers). From the second year block onwards, we observe increasing trends in the impact of the OA format of the three countries, although we must stress that this is also the case for the non-OA format output of the three countries.

Figure 5 shows this stable development of both sets of publications in time, whereby the impact scores are increasing on both sets, although the 'difference' remains more or less the same between the two sets of scores.

In Table 6 we present the outcomes of the analysis on the journal impact scores, based upon methodology II. Here we observe, similar to the previous outcomes, fluctuations in the initials years of the analysis for the OA format output, followed by a more stable situation from 2005-2008 onwards. This finding is even more visible in the graphical representation of Table 6, as in Figure 6.

Table 5. Citation impact (MNCS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output (based on DOI-matching), 2000-2012

	NL ex OA	NL OA	DK ex OA	DK OA	CH ex OA	CH OA
2000 - 2003	1,28	1,65	1,29	1,32	1,36	
2001 - 2004	1,29	0,87	1,29	0,91	1,35	1,03
2002 - 2005	1,29	0,87	1,30	0,98	1,36	1,18
2003 - 2006	1,31	0,87	1,31	0,78	1,37	0,95
2004 - 2007	1,30	0,75	1,31	0,72	1,39	0,96
2005 - 2008	1,31	0,83	1,32	0,86	1,40	0,91
2006 - 2009	1,35	0,85	1,34	0,89	1,40	0,92
2007 - 2010	1,38	0,90	1,38	0,96	1,42	0,97
2008 - 2011	1,40	0,97	1,40	1,00	1,46	1,07
2009 - 2012	1,43	1,03	1,43	0,96	1,49	1,06

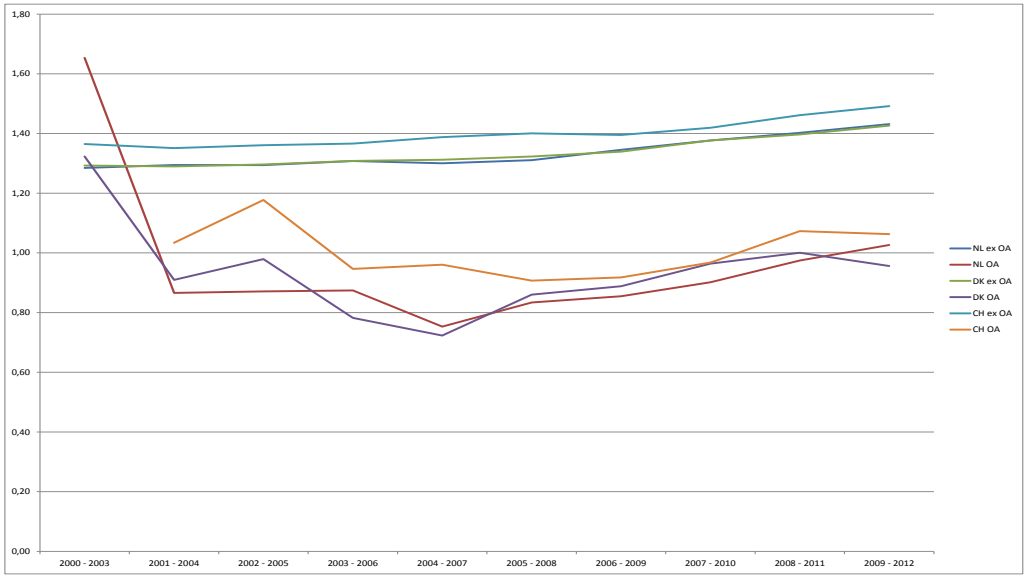


Figure 5. Impact development (MNCS) of Denmark, the Netherlands, and Switzerland, based on matching of DOIs, 2000-2012/2013.

Table 6. Journal-to-field citation impact (MNJS) of Denmark, the Netherlands, and Switzerland, distinguishing OA and non-OA output (based on DOI-matching), 2000-2012

	NL ex OA	NL OA	DKex OA	DK OA	CH ex OA	CH OA
2000 - 2003	1,18	0,54	1,15	1,28	1,19	0,24
2001 - 2004	1,18	0,84	1,16	0,92	1,19	1,22
2002 - 2005	1,19	0,77	1,16	0,84	1,20	1,00
2003 - 2006	1,20	0,84	1,16	0,79	1,20	0,90
2004 - 2007	1,22	0,86	1,18	0,83	1,22	0,88
2005 - 2008	1,24	0,88	1,20	0,86	1,24	0,86
2006 - 2009	1,26	0,90	1,22	0,87	1,26	0,87
2007 - 2010	1,29	0,94	1,24	0,91	1,29	0,91
2008 - 2011	1,30	0,97	1,26	0,93	1,31	0,96
2009 - 2012	1,31	0,97	1,27	0,92	1,32	0,97

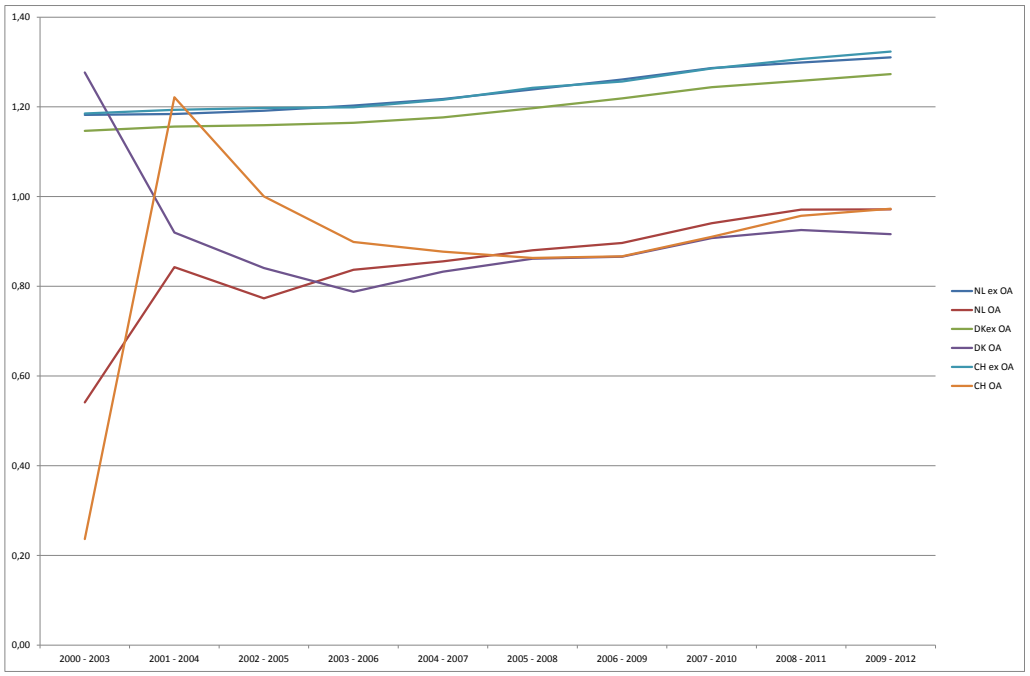


Figure 6: Journal impact development (MNJS) of Denmark, the Netherlands, and Switzerland, based on matching of DOI's, 2000-2012/2013

Conclusion and Discussion

In this final part of the paper, we will summarize the main bibliometric findings, and then move towards limitations in the ways OA is now disclosed in electronic systems supporting bibliometric analyses. Finally, we will discuss the need to improve identification of open access publications and the use of bibliometric techniques to measure OA.

Please note that our conclusions are mainly related to the domains in which journal publishing is the dominant way of communication (the natural, life and medical sciences, and to a lesser extent the social sciences and humanities (van Leeuwen, 2013). We observe for the three countries that the share in output in OA journals is lagging behind as compared to the journals that maintain the non-OA format. We observe a divergence in the development of citation impact for (Gold) OA and non-OA publications with consistently lower impact for the OA publications.

Second, we observe that OA journals have lower journal impact scores than non-OA journals. This may mean that they still struggle to find their position within the total 'reputational hierarchy' of the domain, and as such also within the WoS database. This is a common problem for new journals, and OA journals are no exception. It should be noted however, that our findings associated with OA impact are consistent with what others have found: Gold OA is associated with no citation advantage or a disadvantage (e.g. Archambault et al., 2014). With the inclusion of the various forms of Green OA, we would expect to find a larger proportion of open access articles and a more nuanced outcome related to impact. That Green OA has been found to have increased accumulation of citations (Archambault et al., 2014), may be associated with the circumstances identified above as confounding factors (e.g. early exposure, multiple access points, and proximity of researchers).

Third, we may need to worry about the role of peer review in the journals that are part of the expansion of the WoS database in the last couple of years, many of which are in the OA segment of the database. The Institute for Scientific Information, the predecessor of the current owner of the WoS database Thomson Reuters, always clearly indicated that a properly functioning peer review system within a journal was one of the conditions for a journal to be included in the system (next to other criteria, such as international focus, regular appearance,

preferably in the English language, etc.). We do not know whether this is still such a strong criterion, particularly given the fact that so many new journals appeared around the OA development.

A fourth conclusion relates to the messy situation around the various manners by which open access is defined in electronic databases. The two different ways open access can be operationalized within the world of WoS is an example of this unclear and somewhat messy situation. The fact that the Scopus database did not have the functionality to clearly define open access for users of the system is another instance of the situation around open access. Further examples of this lack of clarity are the various ways open access is operationalized by the publishing industry. There is no clear way of operationalizing in the larger databases of the various business models (such as Gold, Green, and Hybrid open access). Yet another example relates to the various license types related to open access.

A recently published metadata standard for open access holds some promise for improving both human and machine identification of open access publications (Carpenter, 2013). Here, too, stakeholders involved in the new standard were unable to agree on a precise definition of open access. Instead, the standard specifies metadata elements for *free to read* and *license reference*, the latter of which should point to copyright information publicly accessible on the Web (NISO 2015). Increased attention to national research assessment and increased use of institutional CRIS systems together provide a potentially welcoming context for implementing new metadata practices. This would ideally include the possibility of tracking open access among the diversity of research outputs maintained by CRIS systems and considered in assessment events. In this context, it becomes important to assign openly accessible, persistent identifiers to all research objects (Tatum & Wouters 2014). This would increase the potential use of institutional research information for tracking open access as part of regular research assessment practices, rather than relying solely on estimation derived from random sampling of commercial datasets.

References

Archambault, E., Amyot, D., Deschamps, Nicol, A., Provencher, F., Rebout, L. & Roberge, G. (2014). Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013. *Rapport, Commission Européenne DG Recherche & Innovation; RTD-B6-PP-2011-2: Study to Develop a Set of Indicators to Measure Open Access.*

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T. & Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. *PLoS ONE,* 5 (6): e11273. doi:10.1371/journal.pone.0011273.

Björk, B.-C. 2012. The Hybrid Model for Open Access Publication of Scholarly Articles: A Failed Experiment? *Journal of the American Society for Information Science and Technology,* 63 (8): 1496–1504. doi:10.1002/asi.22709.

Björk, B-C., Laakso, M., Welling, P. & Paetau, P. (2013). Anatomy of Green Open Access. *Journal of the Association for Information Science and Technology,* 65 (2): 237–50. doi:10.1002/asi.22963.

BOAI. (2002). Budapest Open Access Initiative. *The Open Society Foundations*. http://www.opensocietyfoundations.org/openaccess.

Carpenter, T. (2013). Progress Toward Open Access Metadata. *Serials Review,* 39 (1): 1–2. doi:10.1016/j.serrev.2013.02.001.

Davis, P.M., & Connolly, M.J. L. (2007 March). Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace. http://hdl.handle.net/1813/5195.

Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E. & Murray, S.S. (2005). The Effect of Use and Access on Citations. *Information Processing & Management*, Special Issue on Infometrics, 41 (6): 1395–1402. doi:10.1016/j.ipm.2005.03.010.

Laakso, M., & Björk, B.-C. (2013). Delayed Open Access: An Overlooked High-impact Category of Openly Available Scientific Literature. *Journal of the American Society for Information Science and Technology,* 64 (7): 1323–29. doi:10.1002/asi.22856.

Lee, K., Brownstein, J.S., Mills, R.G. & Kohane, I.S. (2010). Does Collocation Inform the Impact of Collaboration? *PLoS ONE,* 5 (12): e14279. doi:10.1371/journal.pone.0014279.

van Leeuwen, T.N. (2013). Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship? *Bibliometrie - Praxis und Forschung*, 1-18. http://www.bibliometrie-pf.de/article/viewFile/173/215

Moed, H.F. (2007). The effect of "Open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society of Information Science & Technology*, 58 (13), 2047-2054

NISO. 2015. *Access License and Indicators NISO RP-22-2015*. National Information Standards Organization.

Swan, A. (2010). The Open Access Citation Advantage: Studies and Results to Date. Technical Report. http://eprints.ecs.soton.ac.uk/18516/.

Tatum, C. & Wouters, P.F. (2014). Next Generation Research Evaluation: The ACUMEN Portfolio and Web Based Information Tools. *OpenAIRE-COAR Conference: Open Access Movement to Reality - Putting the Pieces Together*. Athens. doi:10.6084/m9.figshare.1033681.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. & van Raan, A.F.J. (2011a). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, 5(1), 37-47.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. & van Raan, A.F.J. (2011b). Towards a new crown indicator: An empirical analysis. *Scientometrics*, 87 (3), 467-481.

Xia, J., Myers, R.L. & Wilhoite. S.K. (2011). Multiple Open Access Availability and Citation Impact. *Journal of Information Science,* 37 (1): 19–28. doi:10.1177/0165551510389358.

Measuring the Competitive Pressure of Academic Journals and the Competitive Intensity within Subjects

Ma Zheng[1], Pan Yuntao[2], Wu Yishan[2], Yu Zhenglu[2] and Su Cheng[2]

[1] *mazheng@istic.ac.cn*
Institute of Scientific and Technical Information of China (ISTIC), 15 Fuxing Rd. 100038 Beijing (China); and Nanjing University, School of information Management, 22 Hankou Rd. 210093 Nanjing (China)

[2] *panyt@istic.ac.cn, wuyishan@istic.ac.cn*, *luluyu@istic.ac.cn, sucheng@istic.ac.cn*
Institute of Scientific and Technical Information of China (ISTIC), 15 Fuxing Rd. 100038 Beijing (China)

Abstract
A journal's impact and similarity with rivals is closely related to its competitive intensity. A subject area can be considered as an ecological system of journals, and can then be measured using the competitive intensity concept from plant systems. Based on Journal Citation Reports data from 1997, 2000, 2005, 2010, and 2013, we calculated the mutual citation, cosine similarity, and competitive relationship matrices for mycology journals. We derived the mutual citation network for mycology according to Journal Citation Reports data from 2013. We calculated each journal's competitive pressure, and the competitive intensity for the subject. We found that competitive pressures are very variable among journals. Differences between a journal's absolute and relative influence are related to the competitive pressure. A more powerful journal has lower competitive pressure. New journals have more competitive pressure. If there are no other influences, the competition intensity of a subject will continue to increase. Furthermore, we found that if a subject has more journals, its competitive intensity decreases.

Conference Topic
Journals, databases, electronic publications

Introduction
Scientific and technical (S&T) journals have an important role in science and knowledge dissemination. Journals that are focussed on the same subject are at competition with each other. We must build a favourable competitive environment to realize the optimal allocation of limited resources. At the same time, the "survival of the fittest" mechanism boosts the development of S&T journals.

To build a sustainable environment and competition mechanism, we must analyse and measure the present environment of S&T journals, especially in terms of competition. Many researchers have investigated the competitive environment of S&T journals.

Reaching a consensus on the relationship between the journal environment and competition

Scholars began to study the competitive relationship of journals in the 1920s. Competition is mainly related to the resources of subeditors, editors, and authors. Studies found that competitive power is related to a journals' impact factor (IF) (Campanario 1996). Zhu (1999) discussed the relationship between an S&T journal's quality and competitive spirit. A few years later, scholars proposed that competition is a basic attribute of science and noted the differences between different journals' abilities to secure resources. Powerful journals typically attract more attention, which results in a Matthew effect on the journal's development. Scholars have attempted to measure competition between journals using quantitative indexes (Manfred & Scharnhorst, 2001). Researchers have generally accepted that S&T journals develop within a competitive environment. They have explored definitions of the competition between S&T journals (Cai, 2003), how to increase a journal's core competitive strength (Chen 2005), and how to take advantage of market competition (Gao,

2004). Recently, Leydesdorff, Wagner and Bornmann (2014) focused on competition between highly cited journals dependent on the proportions of most-frequently cited publications in the European Union, China, and the United States, which are represented differently because they use different databases.

Determining the competitive relationship between journals using quantitative methods

Leydesdorff noted that Pearson correlations could be used as similarity measures for citation patterns based on bi-connected graphs (Leydesdorff, 2004). He then used principal component analysis and factor analysis to design indicators for the position of the cited journals in the dimensions of the database (Leydesdorff, 2006). Yang analysed the relationship between a journal's value chain and competitive edge using value chain theory (Yang, 2006). As a whole, these ideas and methods for quantitatively measuring a journal's competitive relationship have not been generally accepted, and are not fully developed.

Applying research ideas from ecological competition

Recently, ideas related to competition and competitive intensity in ecology have been applied to research related to S&T journals. Scholars such as Tao, Daoping and Gaoming (2007) have attempted to consider the survival and development of S&T journals from an ecological perspective. Xinyan (2008) researched the concentration ratio of an S&T journal's market share and its competition. She also analysed the index model of competitive intensity in ecology, and applied it to measure a journal's competitive intensity (CI). This was a meaningful exploration, but did not result in a proper index for measuring a journal's distance in terms of the ecological system of S&T journals (Xinyan, 2008).

The competitive environment of S&T journals has been extensively analysed. Progress has been made in terms of the quantitative analysis. Although the CI concept from ecology is useful, we do not know how to define and measure the "distance" between journals. The institute of Scientific and Technical Information of China has measured journal similarity using the mutual citation matrix and cosine similarity method since 2011 (ISTIC, 2011). This provides a measurement of the distance between journals.

In this study, we considered a journal's absolute impact value and similarity as parameters based on the *Journal Citation Reports*. We measured the competitive pressures of mycology journals and the CI for the entire subject using scientometrics and the CI.

Methodology

In this study, we used the concept of CI from the field of ecological research to define the "competitive pressure" among S&T journals. The following design scheme illustrates how we calculate the relevant values.

Main factors that influence the competitive relationship between S&T journals

In a relatively closed ecological environment, the CI mainly depends on the differences between plant diameters and the distance between plants. In this closed environment, the competitive relationships between plants can indicate the strength of the overall competition within the ecological environment.

If we consider journals that focus on one subject, we are investigating a relatively closed ecological environment. Then, all the individual journals can be viewed as separate plants. As shown in Figure 1, the respective "diameters" (D_i and D_j) of journals i and j, and the "distance" (L_{ij}) between them are the major factors of the competitive relationship.

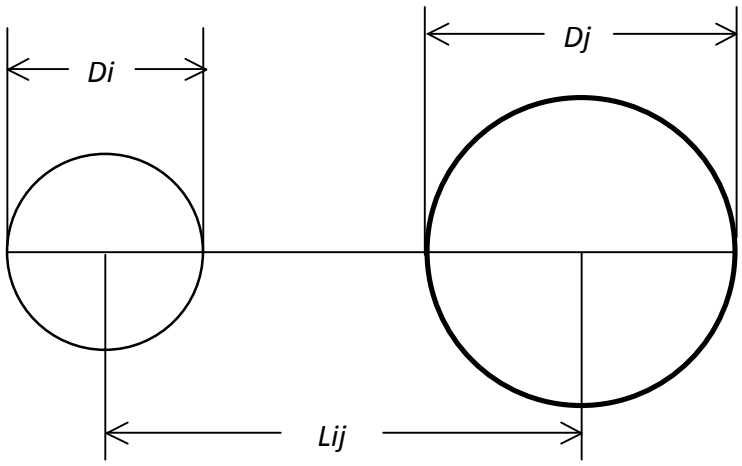


Figure 1. Main factors influencing the competitive relationship between S&T journals.

The number of total citations can be used as an alternative indicator to reflect the influence of the journal

The absolute influence of the journal can be seen as the plant thickness (diameter). Typically, a thicker plant is more capable of competing for resources and fighting rivals. Similarly, more influential journals are generally stronger in terms of their access to excellent manuscripts, funding, and attention. Journals with weaker influences are under more pressure from competitors.

The absolute influence of journals can be quantified using three main indicators: total citations (TC), IF, and the number of published papers.

Among these indicators, the IF is more likely to fluctuate. The number of papers is more vulnerable to subjective factors and can sometimes change dramatically. For example, a change to the journal's publishing cycle from bimonthly to monthly will lead to a sudden increase in the number of papers, and an accordingly sharp drop in the IF (because of a doubled denominator). Compared with the IF and paper number, the total citation indicator is relatively more stable and objective. It visually reflects the influence of journals, is less effected by other factors, and has a distinct advantage in terms of long term monitoring.

Additionally, the IF depends on the average number of citations of paper in a journal, so the total citation is equal to the IF multiplied by the number of papers. From this point of view, the total citation is monotonic in the mathematical sense.

Considering the above discussion, the total citation can be used as an alternative indicator of the influence of a journal. Therefore, in this study, we use the total citation as the diameter (D_i) of journal *i*. That is,

$$D_i = TC_i, \tag{1}$$

where TC_i is the total citation of journal *i.*

The similarity of two journals can be compared using the "distance" between them

It is widely accepted within the ecological community that competition is most intense when the same species live in the same environment (Clements, 1905). The similarity between two journals is also an important factor in their competitive relationship. In other words, a greater similarity between two journals leads to more intense competition. The similarity between two journals can be compared using the "distance" between them (L_{ij}).

Zheng, Na & Guozhen (2012) calculated a citation matrix for a sample of Chinese journals, which is classified into 61 subjects. They calculated the similarities for each journal in a specific subject area, and then constructed the similarity matrix for the journals. We used the same definition, and calculated the distance between periodicals using

$$\mathrm{Lij} = \frac{1}{Sij} - 1, \quad (2)$$

where S_{ij} is the cosine similarity indicator between i and j. S_{ij} is in the range of [0,1], and l_{ij} is in the range of [0,∞]. A S_{ij} value that is closer to 1 means that journals i and j are more similar. Accordingly, the distance L_{ij} is closer to zero. Conversely, if S_{ij} is closer to zero, i and j are less similar and the distance L_{ij} is closer to infinity.

Calculating the competition pressure between S&T journals

We used Hegyi's quantitative measurement for plant competition in ecology (Hegyi, 1974). Suppose that there are n journals for a subject, the target journal is called i and is set as the "basic journal", and the other is called j and considered a "rival journal". Then, $CRij$ is the competitive pressure on journal i from rival j. It is calculated using

$$\mathrm{CRij} = \frac{Dj}{Di \cdot Lij}. \quad (3)$$

We can assume that the competitive pressure on i from j is inversely proportional to the absolute influence of i, is directly proportional to the absolute influence of the rival, and is inversely proportional to the distance between the journals. This assumption is consistent with an intuitive understanding of the competitive relationship.

Combining Equations (1), (2), and (3), we get

$$\mathrm{CRij} = \frac{TCj}{TCi \cdot \left(\frac{1}{Sij} - 1\right)}, \quad (4)$$

where TC_i and TC_j represent the TC for i and j, and S_{ij} is the cosine similarity between periodicals.

CR_{ij} and CR_{ji} represent the competitive relationship between i and j. The cosine similarity S_{ij} measures the angular distance between a journal and its rival, so S_{ij} and S_{ji} are equal. However, CR_{ij} and CR_{ji} are not equal if TC_i is not equal to TC_j. Equation (4) implies that C_{ij} and C_{ji} have a mutually reciprocal relationship.

We can conclude from the definition that the basic journal is under less competitive pressure if it has a higher total citation value than its competitor, and vice versa. The more similar the journals are, the greater the competitive pressure. A journal does not compete with itself, so CR_{ii} is zero.

Calculating the competitive pressure on basic journal i

Suppose that, within its discipline, basic journal i has n-1 rival journals. Then, CI_i is the total competitive pressure on journal i from all of its rivals,

$$\mathrm{CIi} = \sum_{n}^{j=1} CRij. \quad (5)$$

Overall competitive strength for a specific subject

The number of competing journals depends on the subject classification. To compare disciplines, we define the overall competitive strength as CIS. It is the average competitive pressure for all journals, i.e.,

$$\mathrm{CIS} = \frac{1}{n}\sum_{n}^{i=1} CIi. \quad (6)$$

Analysis and Results

We calculated the mutual citation, similarity, competitive relationship, and competitive pressure matrices for the journals, and the CI for mycology using Journal Citation Reports (*JCR*) data from 1997, 2000, 2005, 2010, and 2003.

The inter-citation matrices for the target subject, and the similarity and competitive relationships

We used journals focussed on mycology to demonstrate how to calculate and analyse inter-citations within the target subject, and the similarities and competitive relationships between journals.

There are 23 journals indexed in the *JCR 2013* for mycology (*n=23*). The inter-citation matrix (*C*) was constructed by calculating the inter-citations of each pair of journals. We used the cosine similarity method to transform the inter-citation matrix to the similarity matrix, *R*. The cosine similarity is calculated using

$$\text{Cosine}(x,y) = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}. \quad (7)$$

We transformed *R* into a net document and used Pajek to produce Figure 2, which shows the mutual citation network for mycology according to *JCR 2013*. Each node represents a journal, and a node's area represents the journal's TC. The location of the journal and the thickness of the link represent its similarity with its rivals.

From another perspective, we considered the whole subject area as an ecological space. Then, the 23 journals are independent plants. Figure 2 can be regarded as an ecological system with 23 plants, as viewed from above. The differences between the plant diameters and distances between plants determine the CI and the state of the journals.

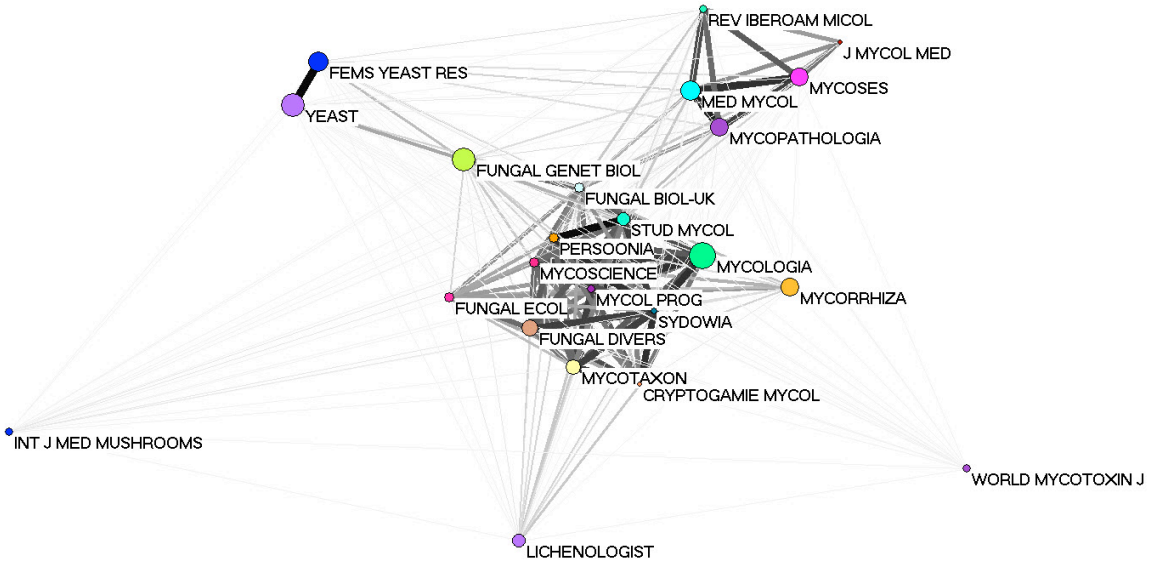


Figure 2. Mutual citation network of journal focussed on mycology, according to *JCR 2013.*

We applied Equation (4) to construct the competitive pressure matrix (*CR*) for the 23 journals, by considering each journal's TC and the cosine similarities between each journal pair.

Competitive pressure for a journal (CI)

Equation (5) shows that the CI of a journal is a combination of the competitive pressure from all of each rivals. We measured the competitive pressure of the all journals using competitive relationship matrices for mycology at five time points.

Table 1 shows that there were large differences in the competitive pressures of the rival journals. The maximum was 408.198 and the minimum was 0.022. In *JCR 2013,* two journals had competitive pressures over 100, 15 were between 10 and 100, and six were under 10.

Table 1. Competitive intensity (CI) for mycology journals.

Title	*1997*	*2000*	*2005*	*2010*	*2013*
CRYPTOGAMIE MYCOL	79.15	278.326	37.227	90.551	140.329
EXP MYCOL	13.81				
FEMS YEAST RES				17.673	32.585
FUNGAL BIOL-UK				81.125	48.575
FUNGAL DIVERS			28.170	8.875	14.402
FUNGAL ECOL				16.954	23.032
FUNGAL GENET BIOL	4.394	14.820	2.985	1.929	3.222
INT J MED MUSHROOMS				0.341	2.175
J MED VET MYCOL	13.572				
J MYCOL MED	42.521	18.324	31.853	17.819	41.412
LICHENOLOGIST			3.753	3.057	3.249
MED MYCOL		28.391	5.748	7.315	18.067
MIKOL FITOPATOL	3.280	1.854	2.389		
MYCOL PROG				189.149	98.921
MYCOL RES	3.751	6.649	11.217	11.919	
MYCOLOGIA	4.663	7.341	12.558	5.09	6.046
MYCOPATHOLOGIA	11.130	4.616	5.069	6.109	17.724
MYCORRHIZA	4.993	8.529	4.174	2.036	2.292
MYCOSCIENCE				30.886	53.764
MYCOSES	10.392	3.991	3.422	12.211	18.333
MYCOTAXON	16.890	20.216	18.220	15.182	16.865
PERSOONIA	94.223	84.520	408.198		92.237
REV IBEROAM MICOL				31.666	35.185
STUD MYCOL	139.528	69.935	51.901	31.591	36.342
SYDOWIA			116.148	298.986	230.812
WORLD MYCOTOXIN J					0.095
YEAST	0.031	0.022	0.318	5.028	15.638

Table 2 shows the competitive intensities compared with the IF and TC, for mycology journals in 2013. The rankings based on the IF and TC is different from the CI rankings. Some journals are ranked in the top 10 in terms of TC and IF but have low CIs, and some are ranked in the bottom five in terms of TC and IF but have higher CI*s*. Therefore, a more powerful journal has lower competitive pressure. We have only listed the results based on the 2013 data, but they were similar for 1997, 2000, 2005, and 2010. The difference between a journals' absolute and relative influence is related to its competitive pressure.

There are certainly some exceptions. Journals that are extremely similar have a significant influence on the competitive pressure. For example, some journals have TCs that are greater than one thousand and are very similar to other journals with the same mass influence, so they also have high competitive pressures. However, some journals are focused on narrow fields and have distinctive characteristics, and therefore do not have much competition because there are not many similar journals, although their TC may be high.

Table 2. Competitive intensity (CI) compared with impact factor (IF) and total citations (TC), for mycology journals in 2013.

Title	*CI 2013*	*rank*	*IF 2013*	*rank*	*TC 2013*	*rank*
CRYPTOGAMIE MYCOL	140.329	2	1.153	18	254	22
FEMS YEAST RES	32.585	10	2.436	7	2935	5
FUNGAL BIOL-UK	48.575	6	2.139	10	790	14
FUNGAL DIVERS	14.402	17	6.938	2	2120	9
FUNGAL ECOL	23.032	11	2.992	5	701	15
FUNGAL GENET BIOL	3.222	20	3.262	4	4298	2
INT J MED MUSHROOMS	2.175	22	1.123	19	554	19
J MYCOL MED	41.412	7	0.4	22	247	23
LICHENOLOGIST	3.249	19	1.613	14	1285	12
MED MYCOL	18.067	13	2.261	9	3132	4
MYCOL PROG	98.921	3	1.543	16	623	18
MYCOLOGIA	6.046	18	2.128	11	5754	1
MYCOPATHOLOGIA	17.724	14	1.545	15	2913	6
MYCORRHIZA	2.292	21	2.985	6	2650	7
MYCOSCIENCE	53.764	5	1.288	17	926	13
MYCOSES	18.333	12	1.805	12	2451	8
MYCOTAXON	16.865	15	0.643	21	1959	10
PERSOONIA	92.237	4	4.225	3	669	16
REV IBEROAM MICOL	35.185	9	0.971	20	649	17
STUD MYCOL	36.342	8	9.296	1	1461	11
SYDOWIA	230.812	1	0.213	23	355	21
WORLD MYCOTOXIN J	0.095	23	2.38	8	454	20
YEAST	15.638	16	1.742	13	4268	3

Figure 3 shows the difference between the CI rankings for a set of journals between 1997 and 2000, and a second set of journals between 2005 and 2013. For the first set, the CI rankings for most of the 14 journals decreased from 1997 to 2013, and only four were in the top ten. This typically means that the competitive pressures of traditional journals (with a longer publishing history) were declining. At the same time, most of the second set started in a high competitive pressure situation, and approximately half of them remained in the top ten of the CI ranking. This means these new journals had to face more challenges.

Competitive intensity for a subject

Equation (6) shows that the CI for a subject is the average competitive pressure of all the journals. We calculated the CIs for mycology in 1997, 2000, 2005, 2010, and 2013.

Table 3 shows that the competitive intensity for a subject (CIS) increased from 1997 to 2005, but the number of journals only increased from 15 to 17. We can see that the CIS decreased between 2005 and 2010 because the number of journals increased from 17 to 23 (by approximately 35%). By analysing the relationship between the subject's scale and CIS, we can see that more journals correspond to low CIs. From 2010 to 2013, the number of journals was stable at 23 so the CIS increased. In the absence of any other influences, the CIS will continue to increase.

By analysing the competitive pressure on each journal and the CIS, we can determine the state of the competitive environment using a quantitative method, and compare the competitive

relationships of different journals and subjects. Through a comparative analysis, we can research reasons for any differences and provide S&T publications with scientific data and tools. Additionally, the data can be used to monitor the S&T journals environment at a macro level, and help decision makers with regard to administration.

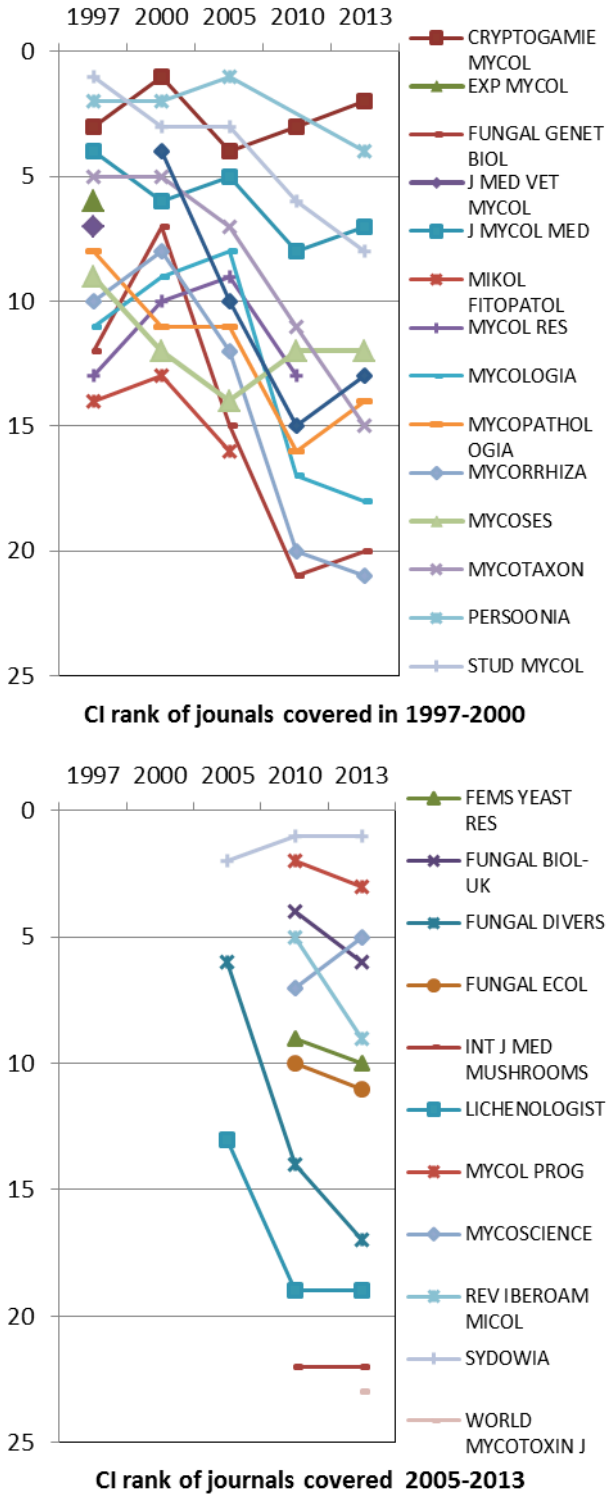


Figure 3. Relationship between competitive intensity (CI) and time.

Table 3. Competition intensity (CI) and number of journals for mycology

	1997	*2000*	*2005*	*2010*	*2013*
number of journal	15	14	17	23	23
CIS	29.489	39.110	43.726	38.500	41.361

Conclusions

There is vast difference in the CIs between subjects and competition pressures between journals.

We have measured journals' competition pressures and the CIS using quantitative methods. The differences between journals' competitive environments may be caused by many related factors. Different journal attributes are related to competitive pressure. For example, the competitive environment and resources vary among multidisciplinary, ordinary professional, and specialized professional journals. Fundamental research or academic journals and engineering or application journals have different competitive features. Chinese journals are obviously different to English language journals. So the factors that influence competitive pressure and intensity, measurements of these related factors, and mechanisms that influence journals' competitive environments must be studied further.

The competitive pressure from a powerful rival may be equal to the pressure from several weakly similar journals.

The ecological concept of CI is a combination of all kinds of competitive pressure. So the competitive pressure on a journal is a combination of the competitive pressure from all of its rivals. The competitive pressure from a powerful rival may be equal to the pressure from several weakly similar journals. The combination of competitive pressure for each journal may be different, which can lead to a high competitive pressure and number of rivals. It can be used as reference when analysing a target journal's competition.

A journal's homogeneity is important when developing S&T journals. Using our quantitative method, we found that homogeneity is obvious in some fields, especially journals that lack "personality". Such journals have higher competitive pressures. The homogeneity of a journal increases its competitive pressure, and the homogeneity of a subject hinders a favourable competitive environment. There is typically fierce competition between two journals that are very similar. Abnormal cooperative relationships exist between some journals, who adopt inter-citation journal group models. These very similar journals pursue high IFs and cited rates. The academic misconduct phenomenon is one problem that results from a journal's homogeneity.

More study is required for multidisciplinary or interdisciplinary journals.

In our method, each journal only belongs to one subject. However, developments in science and technology have led to fusions and evolutions in subject areas. Most articles belong to more than one subject area. At the same time, some journals are multidisciplinary, so it can be difficult to define their subject. We measured a journal's competitive pressure in terms of only one subject. Future research is required to determine how to measure and compare competitive pressure and similarities for multidisciplinary or interdisciplinary subjects.

A favourable competitive environment is only possible at the proper scale

The scale of the subject (number of journals) is related to its competitive pressure and intensity. A favourable competitive environment is only possible at the proper scale. If there

are too many or too few journals the CI decreases. In S&T journal administration, the distribution and trends of the CIs can be used as a reference to promote the development of favourable and sustainable environments.

The research findings in this study can be used as a reference for a new journal when choosing a subject and field.

In management science, there are "red ocean" and "blue ocean" strategies when facing competitive environments. The red ocean strategy directly reacts to competition, whereas the blue ocean strategy avoids direct competition and exploits new markets (Chan & Mauborgne, 2005). When facing competition from rivals, S&T journals must choose an optimal path based on the current environment and future positioning. Journals with relative advantages tend to use red ocean strategies, proactively consolidating and extending their advantages. Relatively weak journals use blue ocean strategies, seeking paths that reduce homogeneity problems and competitive pressures. The findings of this study can be used as a reference for a new journal when choosing a subject and field. In a fiercely competitive fields, it is difficult to successfully launch a new journal without obvious diversity.

Acknowledgments

This research was supported by National Key Technology Support Program of China. (Project Number: 2015BAH25F01)

References

Bonitz, M. & Scharnhorst, A. (2001). Competition in Science and the Matthew Core Journals. *Scientometrics*, 51, 37-54.

Cai, Y. (2003). On Market Competitiveness of Sci-tech Journals. *Chinese Journal of Scientific and Technical Periodicals*. 14, 345-348

Campanario, J.M. (1996). The Competition for Journal Space among Referees, Editors, and Other Authors and Its Influence on Journals' Impact Factors. *Journal of the American Society for Information Science*, 47,184-192

Clements, F.E. (1905). *Research Methods in Ecology*. Lincoln Nebraska: University Publishing Company.

Chen, X. (2005). On sets of core competitiveness of scientific and technical journal. *Media*. 2005, 9, 55

Gao, J. (2004). Market competitive game of sci-tech periodicals. *Acta Editologica*, 16, 319-320.

Hegyi, F. (1974). A simulation model for managing jack-pine stands. In *Growth models for tree and stand simulation*. J. Fries (Ed.). Stockholm: Royal College of Forestry.

Institute of Scientific and Technical Informaiton of China. (2011). *Chinese S&T Journal Citation Reports 2011*. Beijing: Scientific and Technical Documention Press.

Kim, W.C. & Mauborgne, R. (2005). *Blue Ocean Strategy*. Beijing: The Commercial Press.

Leydesdorff, L., Wagner, C.S., & Bornmann, L. (2014). The European Union, China, and the United States in the top-1% and top-10% layers of most-frequently cited publications: Competition and collaborations. *Journal of Informetrics*, 8, 606-617

Leydesdorff, L. (2006). Can Scientific Journals be Classified in terms of Aggregated Journal-Journal Citation Relations using the Journal Citation Reports? *Journal of the American Society for Information Science and Technology*, 57, 601-613.

Leydesdorff, L. (2004). Clusters and Maps of Science Journals Based on Bi-connected Graphs in the Journal Citation Reports. *Journal of Documentation*, 60, 317-427.

Tao, Y., Daoping, W. & Gaoming. Z. (2007). Think deeply in ecology of sci-tech periodicals' survival and development. *Acta Editologica.*, 19, 3-5.

Xinyan, L. (2008). Study on Competition Intensity in Scientific and Technical Journals Publishing Industry. (Unpublished Master's Dissertation) Institute of Scientific and Technical Information of China.

Zheng, M., Na, W. & Guozhen, Z. (2012). The Analysis of Mutual Citation Network on Patterns of Chinese S&T Core Journal Groups. *Studies in Science of Science*. 2012, 30, 983-991.

Zhu, J., & Mei, H. (1999). Relation between competitiveness and quality for S&T journal. *Science Technology and Publication*, 5, 27-29

SciELO Citation Index and *Web of Science*: Distinctions in the Visibility of Regional Science

Diana Lucio-Arias[1], Gabriel Velez-Cuartas[2] and Loet Leydesdorff [3]

[1] *dlucioarias@gmail.com*
Colombian Observatory of Science and Technology, Bogota (Colombia)

[2] *gabrielvelezcuartas@gmail.com*
Grupo de Investigación Redes y Actores Sociales, Departamento de Sociología, Facultad de Ciencias Sociales y Humanas, Universidad de Antioquia; Calle 70 No. 52-21, Medellín (Colombia)

[3] *loet@leydesdorff.net*
University of Amsterdam, Amsterdam School of Communication Research, Amsterdam (The Netherlands)

Abstract
In this study we compare the visibility and performance of Latin American and Caribbean (LAC) Science in terms of its presence in the core collection indexes included in the *Web of Science* (WoS) —*Science Citation Index Expanded*, *Social Sciences Citation Index*, and *Arts & Humanities Citation Index*—and the *Scielo Citation Index* (SciELO CI)—which was recently integrated into the WoS platform. The purpose of this comparison is to provide some inputs to reconstruct the role of SciELO as a communication platform for science produced in Latin America and the Caribbean, and to provide some reflections on the potential impacts—in terms of a better understanding of the global scientific scenery—of the articulation of SciELO CI into WoS: Are there significant differences in the region´s scientific results when studied from publications included in SciELO CI versus those included in the traditional core collection of the WoS? Are regional exercises, such as SciELO, successful in enhancing the visibility of regional scientific production?

Conference Topic

Journals, databases and electronic publications

Introduction

Although the participation of Latin American and Caribbean (LAC)-edited journals in WoS has increased over time, this growth is not comparable to the growth in the participation of scientific articles with at least one author affiliated to an institution in LAC. This increase in participation has been interpreted as a successful integration of LAC science into the world repertoires despite a persistent and notorious gap in the making of good scientific journals (Meneghini, Mugnaini & Packer, 2006). The difference in the nature and characteristics of the journals considered and included in each of the indices justifies our expectation of finding significant differences in the science produced in LAC and communicated through WoS or SciELO CI indexed journals: while the inclusion policy of WoS targets the top quality journals by discipline, the program SciELO has had an inclusive policy aimed at increasing visibility and circulation of LAC journals and their content.[1]

[1] SciELO (Scientific Library on Line) was a program that was initiated in Brazil in 1997 with the purpose of offering a core of Brazilian scientific journals in an open access mode through internet. The program had a successful expansion in the region and now includes, in addition to Brazilian, journals from Chile, Cuba, Spain, Venezuela, Colombia, Argentina, Costa Rica, Mexico, Portugal, Peru, and Uruguay. It is important to note that the SciELO program transcends the SciELO citation index which is the subject of this study. Not all the scientific journals that belong to the SciELO collection and whose content has been made available through SciELO´s program belong to ScieLO´s citation index.

Another difference in the origins of SciELO and WoS that might be helpful in explaining the differences in regional scientific communication is related to the disciplinary context of each of the indexes. A lot has been written about the "natural" or hard sciences origin of WoS, which derived from the Science Citation Index (Garfielfd, 1971), but was expanded to include a broader range of journals and then accompanied by the Social Science Citation Index and later on by the Arts & Humanities Citation Index. The three indexes have been operative since 1978. SciELO, on the other hand, resulted from cooperation of the Fundacao de Amparo a Pesquisa do Estado do Sao Paulo (FASPEP) and the Latin American and Caribbean Center for Health Sciences Information (Bireme) of Panamerican and World Health Organization (PHO/WHO).

We believe that SciELO´s contribution to global science relies on its impact in the circulation of LAC scientific production and therefore the visibility of this production. In the last 15 years, SciELO played an important role in the development of capabilities in LAC to produce world-class scientific results, particularly though the consolidation of a regional base of high-quality scientific journals. The financial requirements to maintain such an exercise updated, expanding and relevant (Aguillo, 2014), together with the potential of SciELO indexed journals to provide a representation of LAC science, might explain the interest behind the inclusion of the regional exercise in the Thomson Reuters owned databases.

The inclusion of SciELO into WoS has had a mixed reception in the LAC scientific community. In 2007, an alliance between Scopus and SciELO raised expectations of all SciELO information to be included in Scopus (Elsevier, 2007). The potential impacts of the inclusion of the journals, and the ambiguity of whether all SciELO journals would be included in Scopus raised some concerns in the LAC scientific community. The negotiations behind SciELO´s inclusion either in Scopus or WoS, was perceived by some editors of LAC journals as a "sell-out" of SciELO´s principles and allowed uncertainty in the future of the regional journal structure that SciELO had aimed to consolidate.

With this paper we expect to contribute on the relevance of both indexes and the complementarities between them as they represent different styles of scientific communication that transcend the center-periphery debate on scientific production. This section is followed by a section in which we introduce the data and methods employed for this study. The results section will focus on the differences between the indices; specifically in the geographical, collaborative aspects, and cognitive characteristics of the communications in each. We finish this contribution with some reflections on the challenges and opportunities of the integration of SciELO into WoS.

Data and Methods

We downloaded all the bibliographical information from the core collection of the WoS (SCI expanded, SSCI, A&HCI) for 79,924 documents that responded to the search query for affiliation country to any LAC countries AND publication year 2012. The same information was downloaded for 30,518 documents that responded to the same search query in the SciELO CI available through WoS. While participation of LAC authors explains 73% of the total publications in SciELO CI, in WoS, this participation is lower than 5%.[2] The organization of the information into relational databases was possible through dedicated routines available at http://www.leydesdorff.net/scielo and http://www.leydesdorff.net/software/isi/index.htm.

[2]In January 2015, a total of 1,899,805 documents were included in WoS with publication year 2012, and 41,621 in SciELO CI.

In order to assess some of the differences in the sets of data considered in this analysis, we provide some descriptive statistics in Table 1. We include the mean and the standard deviation to provide some order of magnitude and dispersion among attributes.

From Table 1, differences among the types of communications included in each set are evident. The mean (μ), represents the average number of authors, addresses, citations, cited references and subject categories per document and the standard deviation (σ) is included to illustrate dispersion in these data. The documents in journals indexed in WoS have more citations, and more frequently result from collaborations among larger number of authors in European or American institutions. These documents are more codified (in terms of the cited references used) as well, and, in general, have a significantly larger impact (in terms of citations received). The mean and standard deviation of the journals are included to represent the average number of LAC documents per journal. Although fewer journals concentrate LAC scientific production in SciELO CI than that in WoS, dispersion among different titles is greater; as can be expected, SciELO CI indexed journals have a larger participation of LAC authors compared with authors from other countries. A total of 163 journals are indexed in both WoS and SciELO CI.

Table 1. Differences in the sets of LAC publications from SciELO CI and WoS Core collection.

LAC publications	**SciELO CI**			**WoS Core Collection**		
Records	30,518			79,924		
Statistics	**N**	**μ**	**σ**	**N**	**μ**	**σ**
Authors	91,269	3.8	2.4	306,560	14	144,3
Addresses	11,858	2.3	1.5	168,390	3.9	14.3
Times cited	7,733	0.3	0.7	274,225	3.4	18.6
Cited references	681,151	26.2	19.1	1,969,653	37	29
Subject Categories	186	1.2	0.7	246	1.5	0.8
Journals	750	40.7	44.5	7,268	10.9	28.0

We use the Overlay maps Toolkit available at http://www.leydesdorff.net/overlaytoolkit (Rafols, Porter & Leydesdorff, 2012) to provide the different visualizations of the relations among disciplines in each of the document sets (SciELO CI and WoS core collection). We rely on these visualizations to suggest disciplinary differences in each of the sets of documents. We expect some of these differences to reflect on diverse goals and interests in the management of each of the indices and which were shortly introduced above.

To reflect upon the distinctions in the collaborative nature of the communications in each index, we build a collaboration network between countries using Pajek.

Results

In this section we provide some results on the differences between communications in the Core Collection of WoS and the recently integrated SciELO CI, focusing on the regional, collaborative and cognitive aspects underlying these communications. In Table 2, we provide the number of records in each of the sets by country of origin of the authors. To normalize for documents with a high number of co-authorships we include a fractional counting of documents considering the total number of signing authors.

The divergence in the countries' participation in the scientific production of LAC can result from (a) the degree in which the specific country has become articulated in the SciELO program and the efforts in increasing the SciELO journal list of each country. As can be expected, the most important SciELO journal collection is from Brazil and it includes 337 journal titles, Colombia follows with a total of 184 journal titles, Mexico has 149, Argentina

and Chile 107 and 106 journal titles each. Another explanation is (b) the specific country´s treatment and importance of national scientific journals.

The policy effort supporting national scientific journals varies in the region where some countries privilege international publication while others aim at balancing international visibility with support to local journals and local publishers (Vessuri, Guédon & Cetto, 2014). Different publication strategies are also evident from Table 2 where the effect of fractional counting seems to be more drastic for communications in journals indexed in WoS Core collection than in SciELO CI. Colombia, for example, has relied on collaborating with international peers to increase their participation in international journals and databases (Lucio-Arias, 2013).

Table 2. Regional distribution of papers in WoS Core collection and SciELO CI.

Country	SciELO CI		WoS	
	Records	Fractional	Records	Fractional
Brazil	19,537	11,929.5	44,812	21,844.1
Colombia	3,065	2,312.2	4,007	1,734.9
Chile	2,409	1,754.3	7,277	3,562.0
Mexico	2,336	1,529.2	13,041	5,879.3
Cuba	1,979	1,053.5	966	320.8
Argentina	1,625	1,223.8	9,975	4,953.8
Venezuela	526	340.8	1,240	543.9
Peru	480	344.0	975	336.1
Costa Rica	284	189.4	514	310.8
Uruguay	99	51.8	868	195.3
Ecuador	53	25.0	465	153.4
Bolivia	42	20.0	85	17.0
Guatemala	23	11.4	52	8.0
Panama	22	8.0	416	120.7
Puerto Rico	22	8.0	N/A	N/A
Paraguay	27	10.7	43	6.1
El Salvador	11	5.1	24	3.1
Jamaica	10	3.1	9	1.8
Nicaragua	20	8.4	31	4.3
Honduras	3	1.0	25	2.8
Dominica	1	0.2	2	0.4
Dominican Republic	1	0.2	33	4.4

The alliances and collaborations reflect important differences in the networks of collaboration that emerge from LAC scientific communications in each of the indices considered (See Figures 1 and 2).

Collaborations in WoS suggest the importance of North America and Europe as allies in the production of scientific knowledge in the region. Collaboration of LAC countries with peers "from the north" dominates scientific communications where LAC participate. Regional collaboration seems not very relevant and in fact not as important as collaboration with Asia, Africa and Oceania. South-South collaboration has received a lot of attention (Arunachalam & Doss, 2000; Chandiwana & Ornbjerg, 2003) and has become an important issue in the

development policy agenda.[3] We believe, nevertheless, that South-South collaboration depicted in Figure 1 is mostly mediated by developed countries and does not represent necessarily a transfer and exchange of resources and knowledge.

The resulting map of collaborations in LAC scientific communications in journals indexed in SciELO CI, suggest a more pronounced strategy based on the regional conjugation of research efforts. Collaboration with Europe is mainly oriented towards Spain and Portugal, suggesting language and cultural similarities as a strong motivation to collaborate. Collaboration with North America and particularly with the United States might rely on geographic proximity as this is stronger in the case of Mexico.

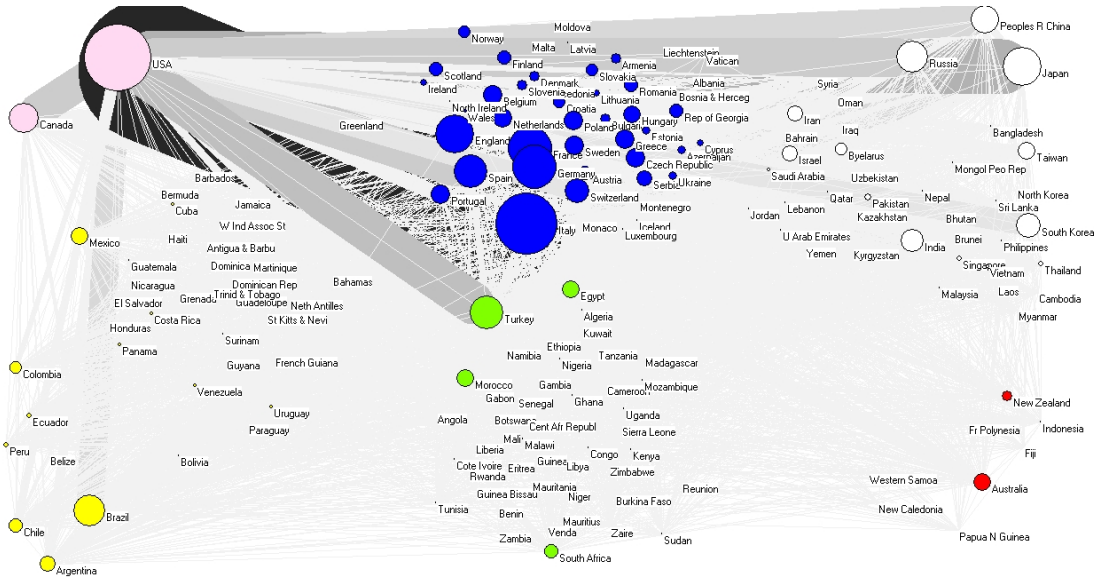


Figure 1. International Collaboration from LAC communications in WoS Core Collection.

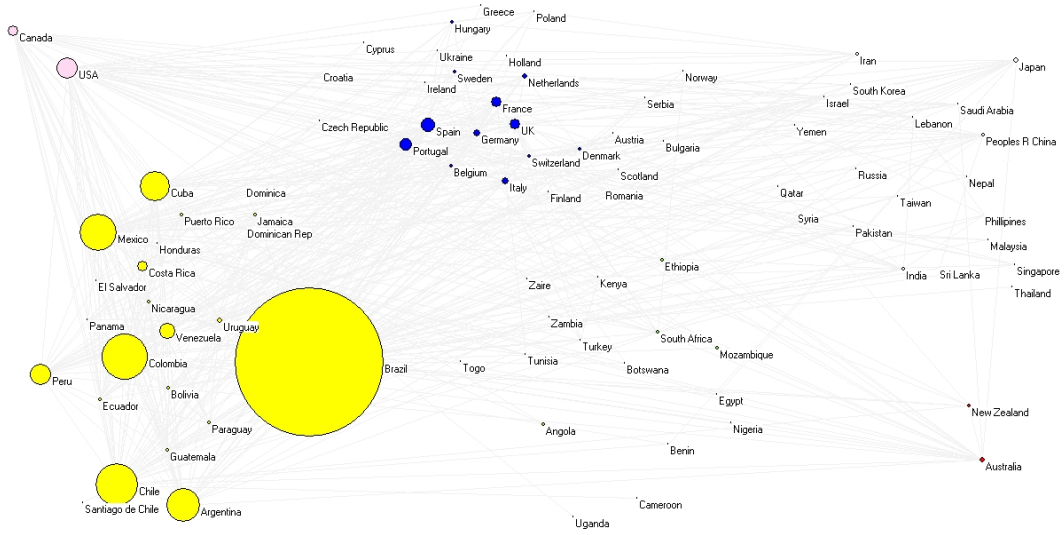


Figure 2. International Collaboration from LAC communications in SciELO CI.

Although it deserves further research, we expect collaborations in SciELO to be a better representation of South-South cooperation, which implies an exchange of resources and ideas within developing countries to solve similar development problems. Collaboration in Figure 2

[3] There is a United Nations Office for South-South cooperation with a website at http://ssc.undp.org/content/ssc.html.

within LAC, Africa and Asia might be a better representation of South-South cooperation. We expect less mediation of the North in the South-South collaboration for the case of SciELO CI indexed communications.

In summary, the differences between Figures 1 and 2 suggest distinct communication practices when (a) aiming at results with international visibility than when the main goal is (b) regional or local diffusion of scientific results through regional journals. While for WoS (Figure 1) strong ties can be indicated with North America and Europe, regional collaboration seems dominant in Figure 2. The participation of the USA in Figure 1 and Brazil in Figure 2 should be interpreted considering that these countries have the highest numbers of indexed journals in each of the respective databases.

This can also result from the different disciplines represented in each index. While WoS has some dominance of "hard" sciences, which are more prone to be published in English and in collaboration, for SciELO CI the disciplinary participation seems to favor the social sciences (see Figure 3 and 4).

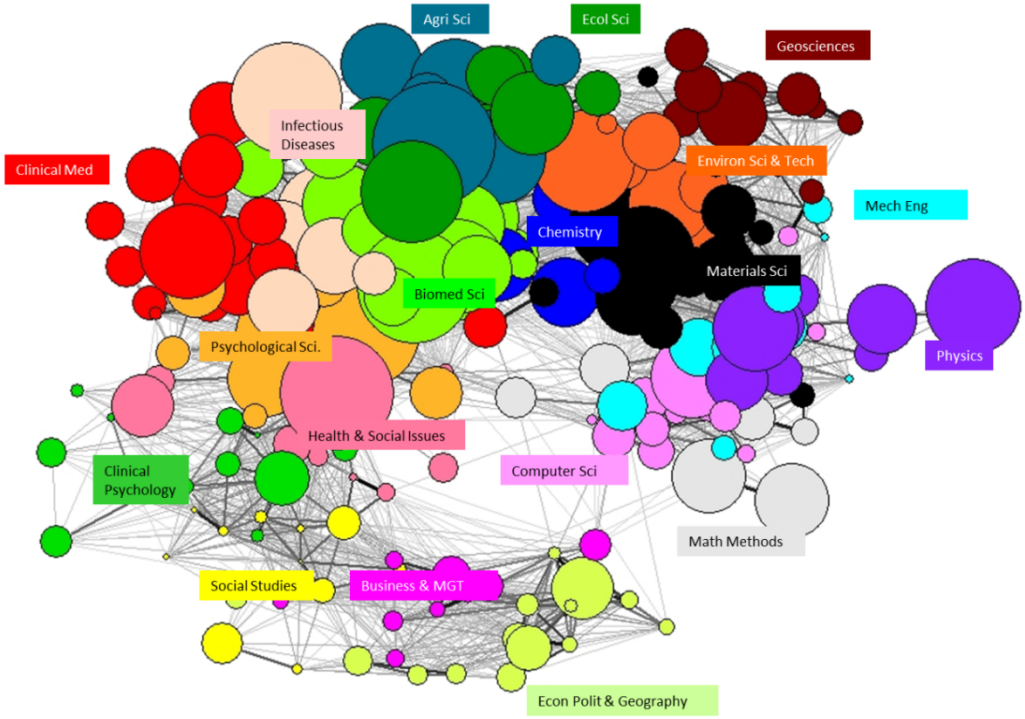


Figure 3. LAC map of Science, WoS Core Collection; 224 Web of Science Categories.

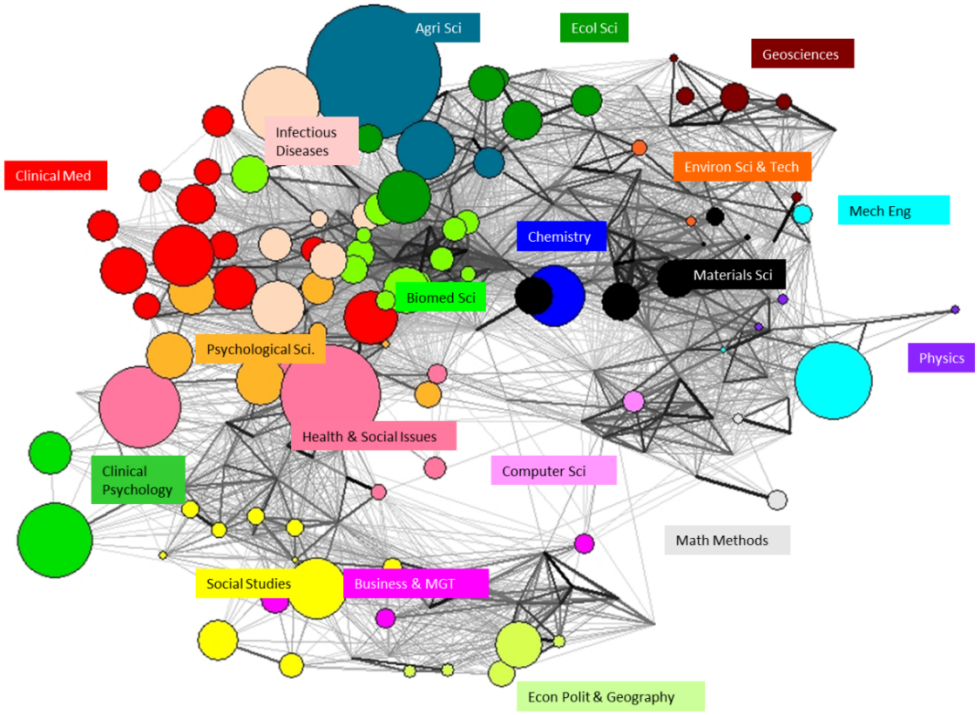


Figure 4. LAC map of Science, SciELO CI.; 224 Web of Science Categories.

Figures 3 and 4 suggest differences in the thematic orientation of the communications in each index. Contributions from the natural sciences are better represented in WoS Core Collection; nevertheless, SciELO CI provides a valuable insight into the regional scientific production in the social and health sciences (where social aspects of the health and medical sciences like research in public health has a better representation), and agriculture. Our expectation is that in-depth analysis of the subjects addressed by the communications would exhibit differences in the sets; communications in SciELO CI will address topics of regional relevance.

Reflections and Further Work

In the last twenty years, scientific development together with technological change and productive innovation have raised interest in the LAC countries, and as a consequence been targeted on the public-policy agenda. Important aspects in the institutionalization of scientific research, such as the consolidation of public institutions for the promotion of science technology and innovation, strengthening of public research institutes, the growth of PhD programs, and the formation and formalization of a journal structure, to socialize scientific results obtained in the region, have also characterized these last decades.

Although growth in the participation of LAC scientific production in traditional databases, such as Web of Science and Scopus, has also been the norm in this period, a common concern in the community has been the challenges to properly socialize scientific results when they are of little interest for mainstream scientific journals. The perseverance in LAC scientific communications of Spanish and Portuguese, as the main languages for communication, particularly in sciences with an important social component, demands alternative means of communication outside international journals as they might have their own structures. Leydesdorff and Bornmann (in press), for example, found a specific citation pattern of Spanish and Portuguese journals in library and information sciences (LIS).

This demand has been acknowledged and as a consequence, most LAC countries have an important structure of national journals. This poses other types of challenges in terms of

research assessment and evaluation. While rankings of international journals and measures based on citations allow researchers and librarians to make informed decisions on the expected quality of a scientific journal´s content, this distinction is more difficult and in occasions impossible when considering national publications. The proliferation of local journals edited by faculties or departments for the diffusion of mainly their own researchers' findings makes the distinction among journals harder.

The need to assess and monitor research results comes together with the demand for a transparent classification among scientific communications. How to assess scientific communications included in international journals versus regional or national journals? In part as a response to this need, different LAC countries have joined the SciELO program. SciELO, in our perspective, has had a positive impact on the consolidation of regional research capabilities and in providing a proper infrastructure for regional exchange and communication.

As was suggested in the collaboration networks analyzed, the SciELO program seems to have transcended the LAC region and includes authorships from Africa and Asia suggesting a platform for South-South collaboration. Other causes for the dominance of the international collaborations in scientific communications in WoS are the cognitive dominance of the biomedical and natural sciences, where collaboration among geographical dispersed groups of individuals is very common. The type of research that results in publications indexed in WoS Core Collection might also cause the dominance of international collaboration in WoS when compared to SciELO CI. Researchers from LAC countries might have a marginal participation in these collaboration networks. This position results of a collaboration among many authors and contributions in the form of data processing instead of cognitive contributions and argumentations. Successful collaborations in the region should hold the researchers in leadership positions (Moya Anegón et al., 2013).

From a cognitive perspective, the inclusion of SciELO CI into WoS offers new opportunities of coverage of disciplines and specialties where the particularities of the territory and the social context are important. Public health, social sciences and agriculture are relevant in SciELO CI; the participation of the LAC scientific communications in these disciplines in the core collection of the WoS has traditionally been low. In this sense, the 15% overlap of Scielo CI journals in both indexes suggests that the inclusion of SciELO CI in the WoS benefits WoS in terms of coverage of regional scientific advances, particularly of communications that have a local object of study and where communication is more original and responds to regional capabilities, but also regional issues and problems.

The inclusion of SciELO CI has raised some concerns among the editors of Spanish[4] and Portuguese journals that have benefitted from a special treatment and inclusion in WoS but that do not have an important position in SciELO CI. Editors of these journals fear that the policy of articulation of SciELO CI into the WoS might result in exclusion of their journals from WoS.

Inclusion of SciELO CI into WoS, responds to the need for a more inclusive representation of scientific results despite regional constrains and conditions. This has resulted from the competition of services offered by Thomson Reuters and Elsevier. The strategies aimed at improving regional visibility are different in Scopus and in the Web of Science. While Scopus has aimed at increasing coverage by increasing their base of regional journals, the globalization of the Web of Science (Testa, 2011) has meant the articulation of regional exercises. The Chinese Journal Database has been hosted in the WoS since 2008, the

[4] FECyT (Spain´s foundation for science and technology) has had an important role in certifying quality of its quality journals in order to support their inclusion in the WoS after an alliance with Thomson Reuters around 2007 (FECyT, 2011)

inclusion of SciELO CI and the Korean Journal Database has been operative since 2014. We believe that the strategy followed by Thomson Reuters provides the cumulative expertise of circulation and visibility promoted regionally, by programs similar to SciELO. We would like to explore this issue further in the future to understand how the inclusion of SciELO CI might put the WoS back in the competition for visibility of regional results.

References

Aguillo, I. (2014). Políticas de información y publicación científica. *El Profesional de la Información*, 23 (2), 113-118.

Arunachalam, S. & Doss, M.J. (2000). Mapping international collaboration in science in Asia through co-authorship analysis. *Current Science*, 79 (5), 621-628

Chandiwana, S. & Ornbjerg, N. (2003). Review of North South and South South cooperation and conditions necessary to sustain research capability in developing countries. *Journal of Health Population and Nutrition*, 21(3), 288-97.

Elsevier. (2007). Elsevier News America Latina. Retrieved on January 10, 2015 from: http://www.elsevier.com.br/bibliotecadigital/news_dez07/pdf/edicao_03_esp_ok.pdf

FECYT. (2011). Análisis de la presencia de las revistas científicas españolas en el JCR de 2010. Retrieved on January 10, 2015from: http://icono.fecyt.es/informesypublicaciones/Documents/2011_07_27RevEspanolasJCR2010.pdf

Garfield, E. (1971). The mystery of the transposed journal lists—wherein Bradford's Law of Scattering is generalized according to Garfield's Law of Concentration. *Current Contents*, 3(33), 5–6.

Lucio-Arias, D. (2013). Colaboraciones en Colombia, un análisis de las coautorías en el Web of Science 2001-2010. In, J. Lucio (Ed.). Observando el sistema nacional de ciencia y tecnología, sus actores y sus productos. Bogotá: OCyT.

Leydesdorff, L., & Bornmann, L. (in press). The Operationalization of "Fields" as WoS Subject Categories (WCs) in Evaluative Bibliometrics: The cases of "Library and Information Science" and "Science & Technology Studies". *Journal of the Association for Information Science and Technology*. http://arxiv.org/abs/1407.7849.

Meneghini, R., Mugnaini, R. & Packer, A.L. (2006). International versus national oriented Brazilian scientific journals. A scientometric analysis based on SciELO and JCR-ISI databases. *Scientometrics*, 69(3), 529-538.

Rafols, I., Porter, A. L. & Leydesdorff, L. (2010). Overlay science maps: a new tool for research policy and library management. *Journal of the American Society for Information Science and Technology*, 61(9), 871–1887.

Vessuri, H., Guédon, J.C., & Cetto, A.M. (2014). Excellence or quality? Impact of the current competition regime on science and scientific publishing in Latin America and its implications for development. *Current Sociology*, 62(5), 647-665.

Testa, J. (2011). The globalization of the Web of Science. http://wokinfo.com/media/pdf/globalwos-essay.pdf.

Book Bibliometrics – A New Perspective and Challenge in Indicator Building Based on the Book Citation Index

Pei-Shan Chi[1], Wouter Jeuris[1], Bart Thijs[1] and Wolfgang Glänzel[1,2]

peishan.chi@kuleuven.be, wouter.jeuris@kuleuven.be, bart.thijs@kuleuven.be, wolfgang.glanzel@kuleuven.be
[1]KU Leuven, ECOOM and Dept. MSI, Leuven (Belgium)
[2]Library of the Hungarian Academy of Sciences, Dept. Science Policy & Scientometrics, Budapest (Hungary)

Abstract
This study aims to gain a better understanding of communication patterns in different publication types and the applicability of the Book Citation Index (BKCI) for building indicators for use in both informetrics studies and research evaluation. The authors investigate the differences not only in citation impact between journal and book literature, but also in citation patterns between edited books and their monographic authored counterparts. The complete 2005 volume of the Web of Science Core collection database including the three journal databases and the BKCI has been processed as source documents. Annual cumulative citation rates in a three-year (x3) and a nine-year (x9) citation window are applied to compute the citation impact of different types of publications. The ratio x3/x9 is utilized as a kind of prospective Price index to examine the extent of ageing. The results of this study show that books are more heterogeneous information sources and addressed to more heterogeneous target groups than journals. Comparatively, the differences between edited and authored books in terme:s of the citation impact are not so impressive as books vs. journals. Humanities have the most different citation impact between books and journals, whereas life sciences have the most similar impact between two groups.

Conference Topic
Journals, databases and electronic publications; Citation and co-citation analysis

Introduction

Some consequences of the absence of books in bibliometric analyses

In contrast to the natural and life sciences, social scientists and humanists publish in different formats, specifically, they rather produce books and contributions to edited volumes and monographs than journal articles (Bourke & Butler, 1996; Pestaña, Gómez, Fernández, Zulueta & Méndez, 1995; Nederhof, 2006; Sivertsen & Larsen, 2012). Books should not be ignored by bibliometrics, not only because they are a major output type but also due to their high impact. Hicks (1999) states that the best social science is often found in books, which is reflected in their citation rates. The danger of ignoring books is illustrated by research, which explores the differences between the worlds of book and journal publishing (e.g., Nederhof, van Leeuwen & van Raan, 2010; Butler & Visser, 2006; Amez, 2013; Clemens, Powell, Mcllwaine & Okamoto, 1995; Hicks & Potter, 1991; Bourke & Butler, 1996; Chi, 2014a). Furthermore, citations to and from books are distributed differently from those to and from journal articles, and often originate from outside the cited work's specialty (Broadus, 1971). Some studies show that books reference more books than articles, and journal articles refer to more articles than books (Larivière, Archambault, Gingras & Vignola-Gagné, 2006; Line, 1979), indicating that citations from journal articles are not the largest source of citations obtained by book publications.

Even though the importance of books in scholarly communication, notably in the social sciences and humanities, was proved by previous studies, only few and small-scale case studies investigating the characteristics of books were conducted by bibliometricians due to the lack of a reliable and comprehensive data source providing citation links. These studies either investigate the citations of so-called non-source items in the references of Web of

Science (WoS) journal papers (Butler & Visser, 2006; Hammarfelt, 2011; Amez, 2013; Chi, 2014a) or analyse citations in other alternative databases such as Google Books or Google Scholar (Kousha & Thelwall, 2009; Kousha, Thelwall & Rezaie, 2011; Samuels, 2011, 2013). All in all, large-scale bibliometric studies analysing the citation patterns of book literature have not been conducted in the past decade.

A new approach to explore citation patterns of books and its limitations

In 2011, Thomson Reuters released a new collection in the WoS, Book Citation Index (BKCI), to allow users to discover book literature and trace its comprehensive citation links alongside journal literature (Adams & Testa, 2011). BKCI covers over 60,000 editorially selected books starting from 2005 with an additional 10,000 new titles each year (Book Citation Index, 2015).

Even though the BKCI broadens the coverage of WoS and allows researchers to tackle studies based on numerous and qualified bibliographic data of books and book chapters in different aspects, the new database is not fully developed yet (Leydesdorff & Felt, 2012; Torres-Salinas, Robinson-García, Jiménez-Contreras & Delgado López-Cózar, 2012; Gorraiz, Purnell & Glänzel, 2013; Torres-Salinas, Robinson-García, Campanario & Delgado López-Cózar, 2013a; Torres-Salinas, Rodríguez-Sánchez, Robinson-García, Fdez-Valdivia & García, 2013b; Torres-Salinas, Robinson-García, Cabezas-Clavijo & Jiménez-Contreras, 2014). Some limitations mentioned in previous studies include:

- Coverage
  BKCI indexes 61% of 60,000 books in the social sciences and humanities (in November 2014, see Book Citation Index, 2015), which is not too arguable due to the nature of the publication behavior of scholars in different fields. However, its indexing bias in terms of language, country, and publisher is large. For example, 96% of the indexed books are written in English (Torres-Salinas et al., 2014) and the United States and England account for 35% of all publications and 75% of publishers in BKCI (Gorraiz et al., 2013; Torres-Salinas et al., 2014). Furthermore, Springer, Palgrave and Routledge alone account for 50% of the total database (Torres-Salinas et al., 2014) evincing a rather high concentration of publishers.
- Completeness of records
  Gorraiz et al. (2013) report the absence of affiliation data in BKCI but it has been confirmed by Torres-Salinas et al. (2014) that their later downloaded data does include affiliation information which could be used to analyse research units such as countries or institutions. Moreover, the low share of BKCI indexed items with references data (<30%, see Chi, 2014b) would also limit the validity of relevant studies.
- Document type classification
  A further limitation of the BKCI comes from the lack of a clear distinction of document types due to the different forms of book literature.
  - *Books*
    Gorraiz et al. (2013) argue that 'book' might be considered to be at a higher hierarchical level as 'journal' instead of being treated as a document type, and consequently point out the lack of cumulative citation counts from different hierarchies in BKCI. It is in line with the warning raised by Leydesdorff and Felt (2012) that monographs may be underrated in terms of citation impact or overrated using publication performance indicators. Furthermore, Gorraiz et al. (2013) question the fuzzy boundaries of subtypes of book and how to treat new editions.

  - *Monographs and edited volumes*

It was discovered that edited books usually have a greater impact than non-edited books (Leydesdorff & Felt 2012, Torres-Salinas et al., 2014, Chi, 2014a; Amez, 2013). This may be because of the effects of working collectively with a more diverse content and the higher average number of book chapters per book (Torres-Salinas et al., 2014). However, a global consensus on how to cite the book editor(s), the book author(s) or the author(s) of the book chapter is lacking (Gorraiz et al., 2013). Even though it is possible to distinguish bibliometrically between monographs and edited volumes among the type 'book', a normalization for the credit of a monograph is required (Leydesdorff & Felt, 2012).

- *Book series and annual series*
  BKCI covers annual series, which are part of the journal and series literature and indexed by other collections of WoS as well. They are assigned to the pubtype 'Journal' in BKCI (the other two pubtypes are 'Books' and 'Books in series'), and all are published by the publisher Annual Reviews. Leydesdorff and Felt (2012) indicate the problems from ignoring differences between book series and annual series. As noticed by Torres-Salinas et al. (2012, 2013b), this publisher presents an outlier pattern showing a behavior more closely linked to journals rather than monographs.

The research purposes of this study

In this study, we analyse and compare BKCI items jointly with journals literature to answer the following open questions based on the revealed limitations of using the database. Some of these questions have already been addressed but not yet answered by, e.g., Adams & Testa (2011) and Gorraiz et al. (2013). These issues apply to differences in citation impact between journal and book literature but also to the question whether edited books with different contributors for each chapter essentially deviate in their citation patterns from their monographic authored counterparts.

1. What is the feature of books in the sciences (including life sciences, natural sciences, technical sciences), social sciences and humanities through the lens of the BKCI?
2. Is there any difference between the ageing of periodical and monographic literature?
3. Is there a difference in citation patterns of edited and authored books?

The findings are expected to allow a better understanding of communication patterns in different publication types and the applicability of the BKCI for building indicators for use in both informetrics studies and research evaluation.

Methodology

Data sources

The complete 2005 volume of the Web of Science Core collection database including the three journal databases Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) as well as the Book Citation Index (BKCI) has been processed as source documents. The two proceedings editions of the core collection have been excluded because of the large overlap among the book, proceedings and journal databases (cf. Gorraiz et al., 2013). The choice of volume 2005 was made for two reasons, particularly, because 2005 was the first BKCI volume and this allowed us to trace citations till end of 2013, i.e., for a full period of nine years.

In addition, we have split up the BKCI database into two parts, namely those books that could be identified as edited books and the rest, which was considered to refer to authored books. Overlap with proceedings and journals were removed to obtain a correct dataset for the

analysis. Only so-called citable document types have been taken into account, that is, articles, letters and reviews for journals, books and citable book chapters for the BKCI. All documents extracted from the BKCI have been analysed both individually and aggregated to the book level.

Subject classification

All items extracted from the database have been assigned to the 74 individual subfields according to the *modified* Leuven-Budapest classification system. Multiple assignments are quite frequent at this level of granularity. The original scheme was introduced by Glänzel and Schubert (2003) and has been recently modified to provide a better categorisation for the social sciences and humanities. The modified version has been developed for the use with the BKCI but is also fully compatible with the journal and proceedings editions of the WoS Core Collection as it is based on the WoS and Journal Citation Reports (JCR) subject categories. Major fields and subfields in the sciences of the previous version have not been changed. The modified classification scheme is presented in Figure 1.

THE LEUVEN – BUDAPEST CLASSIFICATION SCHEME FOR THE SCIENCES, SOCIAL SCIENCES AND HUMANITIES

0. **MULTIDISCIPLINARY SCIENCES**
   - X0 multidisciplinary sciences

1. **AGRICULTURE & ENVIRONMENT**
   - A1 agricultural science & technology
   - A2 plant & soil science & technology
   - A3 environmental science & technology
   - A4 food & animal science & technology

2. **BIOLOGY (ORGANISMIC & SUPRAORGANISMIC LEVEL)**
   - Z1 animal sciences
   - Z2 aquatic sciences
   - Z3 microbiology
   - Z4 plant sciences
   - Z5 pure & applied ecology
   - Z6 veterinary sciences

3. **BIOSCIENCES (GENERAL, CELLULAR & SUBCELLULAR BIOLOGY; GENETICS)**
   - B0 multidisciplinary biology
   - B1 biochemistry/biophysics/molecular biology
   - B2 cell biology
   - B3 genetics & developmental biology

4. **BIOMEDICAL RESEARCH**
   - R1 anatomy & pathology
   - R2 biomaterials & bioengineering
   - R3 experimental/laboratory medicine
   - R4 pharmacology & toxicology
   - R5 physiology

5. **CLINICAL AND EXPERIMENTAL MEDICINE I (GENERAL & INTERNAL MEDICINE)**
   - I1 cardiovascular & respiratory medicine
   - I2 endocrinology & metabolism
   - I3 general & internal medicine
   - I4 hematology & oncology
   - I5 immunology

6. **CLINICAL AND EXPERIMENTAL MEDICINE II (NON-INTERNAL MEDICINE SPECIALTIES)**
   - M1 age & gender related medicine
   - M2 dentistry
   - M3 dermatology/urogenital system
   - M4 ophthalmology/otolaryngology
   - M5 paramedicine
   - M6 psychiatry & neurology
   - M7 radiology & nuclear medicine
   - M8 rheumatology/orthopedics
   - M9 surgery

7. **NEUROSCIENCE & BEHAVIOR**
   - N1 neurosciences & psychopharmacology
   - N2 psychology & behavioral sciences

8. **CHEMISTRY**
   - C0 multidisciplinary chemistry
   - C1 analytical, inorganic & nuclear chemistry
   - C2 applied chemistry & chemical engineering
   - C3 organic & medicinal chemistry
   - C4 physical chemistry
   - C5 polymer science
   - C6 materials science

9. **PHYSICS**
   - P0 multidisciplinary physics
   - P1 applied physics
   - P2 atomic, molecular & chemical physics
   - P3 classical physics
   - P4 mathematical & theoretical physics
   - P5 particle & nuclear physics
   - P6 physics of solids, fluids and plasmas

10. **GEOSCIENCES & SPACE SCIENCES**
    - G1 astronomy & astrophysics
    - G2 geosciences & technology
    - G3 hydrology/oceanography
    - G4 meteorology/atmospheric & aerospace science & technology
    - G5 mineralogy & petrology

11. **ENGINEERING**
    - E1 computer science/information technology
    - E2 electrical & electronic engineering
    - E3 energy & fuels
    - E4 general & traditional engineering

12. **MATHEMATICS**
    - H1 applied mathematics
    - H2 pure mathematics

13. **SOCIAL SCIENCES I (GENERAL, REGIONAL & COMMUNITY ISSUES)**
    - Y1 education, media & information science
    - Y2 sociology & anthropology
    - Y3 community & social issues

14. **SOCIAL SCIENCES II (ECONOMIC, POLITICAL & LEGAL SCIENCES)**
    - L1 business, economics, planning
    - L2 political science & administration
    - L3 law

15. **ARTS & HUMANITIES**
    - K0 multidisciplinary
    - K1 arts & design
    - K2 architecture
    - K3 history & archaeology
    - K4 philosophy & religion
    - K5 linguistics
    - K6 literature

Figure 1. The modified version of the Leuven-Budapest classification scheme for the WoS.

Data processing

In order to analyse citation impact and ageing patterns over subfields, we have calculated the following statistics:

- Annual citation rates (both increments and cumulated) for the year of publication 2005 (1) till 2013 (9). In this study, however, we only use cumulative citation impact in a three-year (x_3) and a nine-year (x_9) citation window.
- The ratio x_3/x_9 as a kind of prospective Price index and an indicator of ageing.

We have calculated all statistics on the basis of both individual book chapters, where available, and for the complete books. Chapters were considered the equivalent of journal articles in terms of the aggregation level. Unfortunately, chapter-based citation statistics proved not to be reliable since citations to individual chapters could not be identified in many cases as they were assigned to the book in the database. This is not necessarily due to the database producer: often the authors of the citing documents are responsible for this uncertainty. In order to avoid biased indicators or otherwise incomplete or distorted results we decided to use only citation indicators for complete books, which, of course, results in a serious loss of information and a more intricate interpretation. This applies above all to edited books, where chapters are authored by different contributors, and a distinction between different chapters would be of paramount importance.

A further issue is the small size of the publication set resulting from this restriction. We have found many subfields with fewer than 30 books each: This threshold might be critical for the interpretation and reliability of statistics like mean values and shares (e.g., Glänzel & Moed, 2013). Furthermore, we have not assigned books to corporate addresses of authors/editors because the availability of author affiliation in books is rather low (see, e.g., Gorraiz et al., 2013).

Results

It is not the aim of the present paper to study the subject coverage of the BKCI database since, on one hand, we can refer to the study by Adams and Testa (2011) in the context of broader subject areas and, on the other hand, a subject analysis at the level of subject categories can easily be conducted using the analyse tool of the web version of Thomson Reuters WoS Core Collection. Nevertheless we would just like to mention in passing that we can confirm that subfields in the social sciences and humanities have a better representation in the BKCI than in the other databases of the WoS.

Ten subfields had a share larger than 5% in the 2005 volume of the BKCI: Among those 10 subfields applied mathematics was the only representative of the sciences. Slightly more than 12% of all books could be assigned each to business, economics, planning and political science & administration, respectively. All books in the humanities (except for multidisciplinary and arts & design) as well as education, media & information science and sociology & anthropology in the social sciences were among the top ten in terms of subject representation.

In the first step we looked at citation patterns of book and journals literature by disciplines in a nine-year citation window. What we intended to do was not to compare citation impact over across fields but to compare subject-specific citation patterns between journals and books. It is a well-known fact that the subject is one of the factors influencing citation impact; the document type is another one (cf. Glänzel, 2013). Thus the publication type such as journal, proceeding, or monograph is expected to play a role in this context as well. Figure 2 plots the mean citation rates of subfields based on the nine-year citation window of books against the corresponding journal indicators. The volume year of the source items was 2005. Only subfields have been chosen in which at least 30 books have been published in that year. Subfields are ranked according the subfield impact in the BKCI. The results are somewhat unexpected here: Not the life sciences – as expected from journal literature – exhibit the highest citation impact for books but disciplines in chemistry and the geosciences. Consequently, the correlation between the corresponding x_9 values is medium ($r = 0.420$). In this respect, there are no dramatic differences between edited and authored books. The correlation between these two book types with $r = 0.762$ is relatively strong.

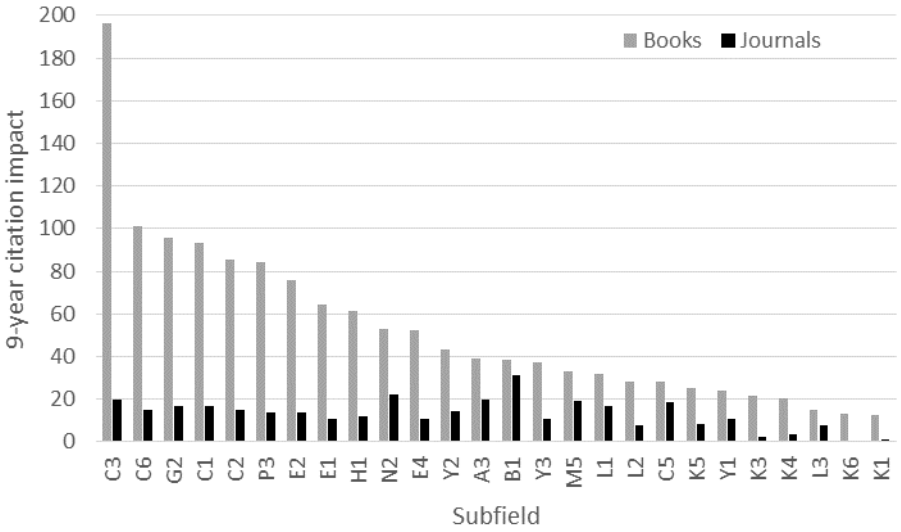


Figure 2. Most cited subfields in the mirror of the BKCI vs. SCIE/SSCI/AHCI. [Data sourced from Thomson Reuters Web of Science Core Collection].

It is known from journal literature that ageing is the fastest in the life and the natural sciences, followed by applied sciences, mathematics, social sciences and humanities (see Glänzel & Schoepflin, 1999). Ageing patterns can be characterised as a combination of phases of maturing and decline in citation processes (Glänzel & Schoepflin, 1995; Moed, van Leeuwen & Reedijk, 1998). The transition from the first to the second phase is marked by a peak in the annual increments of citation impact. This peak ranges according to the ageing of the discipline under study typically between the second and the fifth year beginning with the date of publication. The ratio (x_3/x_9) can thus serve as a proxy for literature ageing in the mirror of citation processes.

The plot of the prospective 'Price Index' (x_3/x_9) of books indexed in the 2005 volume of the BKCI against the corresponding journal indicators for the same volume is shown in Figure 3. The x_3/x_9 ratios are ranked in descending order according to the journal database editions of the WoS. At the left-hand side the disciplines with the fastest aging (highest ratios) can be found, while the low end is formed by slow-ageing subfields (cf. black bars in Figure 3). The grey bars representing the subfields in the BKCI show a rather subject-balanced situation. High (between 20% and 25%) as well as low (between 10% and 15%) shares can be found in both science and SSH subfields. The correlation between the x_3/x_9 ratios for books and journals is practically zero. This is illustrated in Figure 4. We just mention in passing that also the correlation between the corresponding ratios of edited and authored books is low ($r = 0.110$) as well. This substantiates that citation processes of books are more complex as these apparently depend on more factors than in the case of journal literature. Notably ageing seems not to be principally characterised by subject-specific peculiarities. Books are thus more heterogeneous information sources and addressed to more heterogeneous target groups than journals (and possibly proceedings).

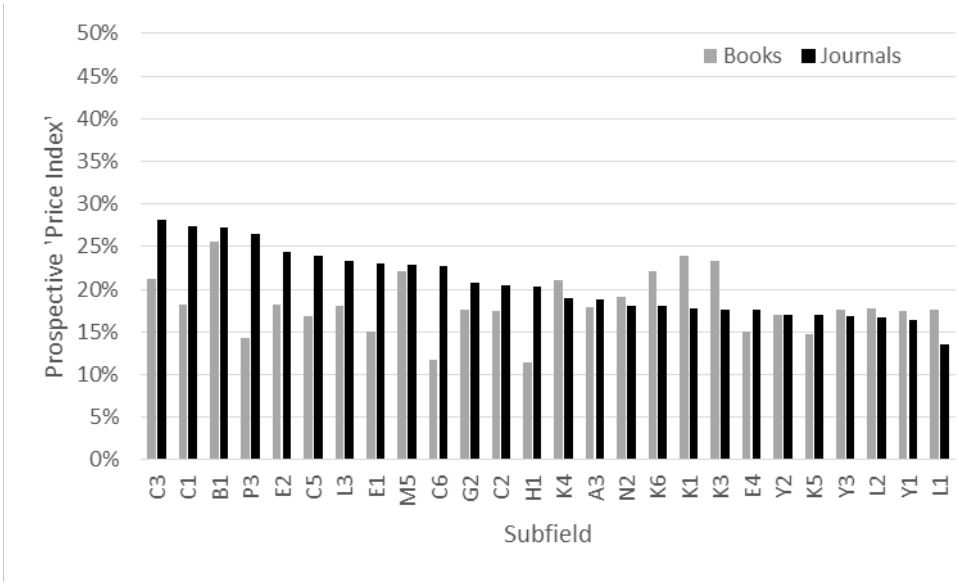


Figure 3. Prospective 'Price Index' of subfields in the BKCI vs. SCIE/SSCI/AHCI. [Data sourced from Thomson Reuters Web of Science Core Collection].

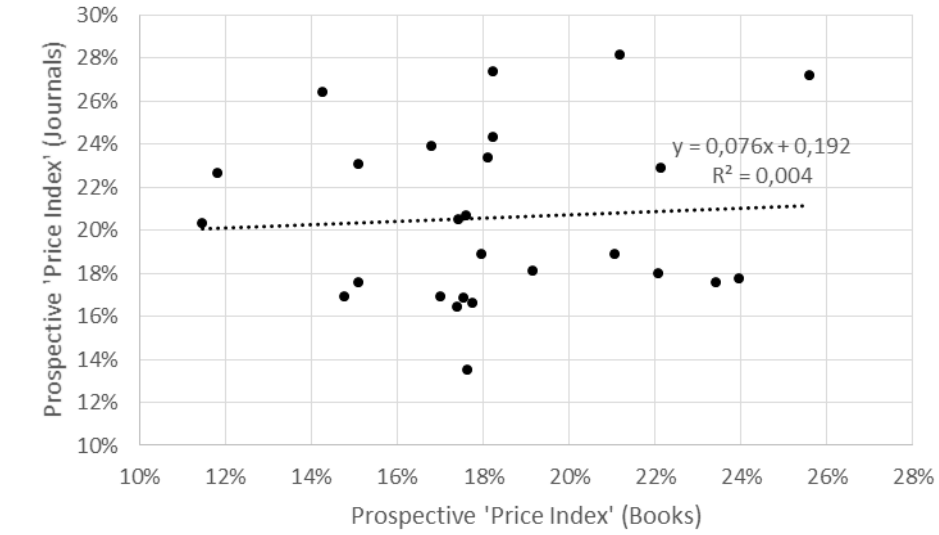


Figure 4. Scatter plot of prospective 'Price Index' of subfields in the BKCI vs. SCIE/SSCI/AHCI. [Data sourced from Thomson Reuters Web of Science Core Collection].

Conclusion

It is confirmed in this study that subfields in the social sciences and humanities have a higher representation in the BKCI (59%) than they have in the other databases of the WoS (12%). Disciplines in chemistry and the geosciences, instead of life sciences, have the highest citation impact for books. Humanities is the field having the highest difference between citation impact of books and journals. In contrast, life sciences have the most similar impact in books and journals. Compared to other sciences, technical sciences have relatively moderate characteristics in different perspectives.

It is not surprising to see that the social sciences and humanities have the largest increase of both the coverage and citation impact in the BKCI compared to journal literature in the other databases of the WoS. The BKCI could be an initial approach to explore wider targets of bibliometric analyses in the social sciences and humanities. The books in the basic sciences have unexpectedly high citation impact, whereas books in the life sciences do not reflect the

dominant position in journal literature but have been found to be on a relatively similar scale of citation counts as journals. This may imply that using BKCI data for bibliometric analyses in basic sciences would be a powerful approach to drag in more citation information.

For the ageing of periodical and monographic literature, the results of this study indicate a clear boundary between the two groups. The differences between books and journals are obvious, but the ageing of books is balanced between subjects. The differences between edited and authored books in terms of the 9-year citation impact are not so impressive as the other group books and journals. However, their disparities in ageing ratios are more evident than those of citation impact. The more complex citation processes of books, compared to journal literature, are shown in this study, the more heterogeneous characteristics of books should therefore be addressed.

The different ageing patterns of book and journal literature, i.e., books do not have as strong discipline specific patterns as journals, may lead to a universal condition for applying or building indicators in the collections of BKCI. It especially needs to be taken into account while designing indicators that are sensitive to the observed citation period. Moreover, the heterogeneous characteristics of books from their different formats such as edited or authored volumes result in more complex citation patterns than journals. These findings on the differences between periodical and monographic literature are worth further studies of indicator design to take into account.

References

Adams, J. & Testa, J. (2011). Thomson Reuters Book Citation Index. In E. Noyons, P. Ngulube, J. Leta (Eds.), *The 13th Conference of the International Society for Scientometrics and Informetrics* (Volume I, 13–18). Durban, South Africa: ISSI, Leiden University and the University of Zululand.

Amez, L. (2013). Citation patterns for social sciences and humanities publications. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger & H. Moed (Eds.), *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference* (Volume II, 1891–1893). Vienna: AIT GmbH.

Book Citation Index (2015). Retrieved January 16, 2015 from: http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/

Bourke, P. & Butler, L. (1996). Publication types, citation rates and evaluation. *Scientometrics*, 37(3), 473–494.

Broadus, R. N. (1971). The literature of the social sciences: A survey of citation studies. *International Social Sciences Journal*, 23(2), 236–243.

Butler, L. & Visser, M. S. (2006). Extending citation analysis to non–source items. *Scientometrics*, 66(2), 327–343.

Chi, P. S. (2014a). Which role do non-source items play in the social sciences? A case study in political science in Germany. *Scientometrics*, 101(2), 1195–1213.

Chi, P. S. (2014b). T*he Characteristics and Impact of Non-Source Items in the Social Sciences – A Pilot Study of Two Political Science Departments in Germany*. Berlin: Humboldt-Universität zu Berlin.

Clemens, E. S., Powell, W. W., Mcllwaine, K. & Okamoto, D. (1995). Careers in print: Books, journals, and scholarly reputations. *American Journal of Sociology*, 101(2), 433–494.

Glänzel, W. (2013), *Bibliometrics as a research field.* Course script, KU Leuven, 3rd Edition.

Glänzel, W. & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, 21(1), 37–53.

Glänzel, W. & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing and Management*, 35(3), 31–44.

Glänzel, W. & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56(3), 357–367.

Glänzel, W. & Moed, H. F. (2013). Opinion paper: thoughts and facts on bibliometric indicators. *Scientometrics*, 96(1), 381–394.

Gorraiz, J., Purnell, P. & Glänzel, W. (2013). Opportunities and limitations of the book citation index. *Journal of the American Society for Information Science and Technology*, 64(7), 1388–1398.

Hammarfelt, B. (2011). Interdisciplinarity and the intellectual base of literature studies: Citation analysis of highly cited monographs. *Scientometrics*, 86(3), 705–725.

Hicks, D. (1999). The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. *Scientometrics*, 44(2), 193–215.

Hicks, D. & Potter, J. (1991). Sociology of scientific knowledge: A reflexive citation analysis or science disciplines and disciplining science. *Social Studies of Science*, 21(3), 459–501.

Kousha, K. & Thelwall, M. (2009). Google book search: Citation analysis for social science and the humanities. *Journal of the American Society of Information Science and Technology*, 60(8), 1537–1549.

Kousha, K., Thelwall, M. & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, 62(11), 2147–2164.

Larivière, V., Archambault, E., Gingras, Y. & Vignola-Gagné, E. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57(8), 997–1004.

Leydesdorff, L. & Felt, U. (2012), Edited volumes, monographs and book chapters in the Book Citation Index. *Journal of Scientometric Research*, 1(1), 28–34.

Line, M. B. (1979). The influence of the type of sources used on the results of citation analyses. *Journal of Documentation*, 35(4), 265–284.

Moed, H. F., van Leeuwen, T. N. & Reedijk, J. (1998). A new classification system to describe the ageing of scientific journals and their impact factors. *Journal of Documentation*, 54 (4), 387–419.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1), 81–100.

Nederhof, A. J, van Leeuwen, T. N. & van Raan, A. F. J. (2010). Highly cited non-journal publications in political science, economics and psychology: A first exploration. *Scientometrics*, 83(2), 363–374.

Pestaña, A., Gómez, I., Fernández, M. T., Zulueta, M. A. & Méndez, A. (1995). Scientometric evaluation of R&D activities in medium-size institutions: A case study based on the Spanish Scientific Research Council. In M. Koenig & A. Bookstein (Eds.), *Proceedings of the Fifth Biennial International Conference of the International Society for Scientometrics and Infometrics* (425–434). Medford: Learned Information, Inc.

Samuels, D. J. (2011). The modal number of citations to a political science article is greater than zero: Accounting for citations in articles and books. *PS: Political Science and Politics*, 44(4), 783–792.

Samuels, D. J. (2013). Book citations count. *PS: Political Science and Politics*, 46(4), 785–790.

Sivertsen, G. & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: an empirical analysis of the potential. *Scientometrics*, 91(2), 567–575.

Torres-Salinas, D., Robinson-García, N., Jiménez-Contreras, E. & Delgado López-Cózar, E. (2012). Towards a 'Book Publishers Citation Reports'. First approach using the 'Book Citation Index'. *Revista Española de Documentación Científica*, 35(4), 615–620.

Torres-Salinas, D., Robinson-García, N., Campanario, J. M. & Delgado López-Cózar, E. (2013a). Coverage, specialization and impact of scientific publishers in the Book Citation Index. *Online Information Review*, 38(1), 24–42.

Torres-Salinas, D., Rodríguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J. & García, J. A. (2013b). Mapping citation patterns of book chapters in the book citation index. *Journal of Informetrics*, 7(2), 412–424.

Torres-Salinas, D., Robinson-García, N., Cabezas-Clavijo, Á. & Jiménez-Contreras, E. (2014). Analyzing the citation characteristics of books: edited books, book series and publisher types in the book citation index. *Scientometrics*, 98(3), 2113–2127.

When is an Article Actually Published? An Analysis of Online Availability, Publication, and Indexation Dates

Stefanie Haustein[1], Timothy D. Bowman[1] and Rodrigo Costas[2]

[1] *stefanie.haustein@umontreal.ca, timothy.bowman@umontreal.ca*
École de bibliothéconomie et des sciences de l'information, Université de Montréal, Montréal (Canada)

[2] *rcostas@cwts.leidenuniv.nl*
Center for Science and Technology Studies, Leiden University, Wassenaarseweg 62A, 2333 AL Leiden (The Netherlands)

Abstract
With the acceleration of scholarly communication in the digital era, the publication year is no longer a sufficient level of time aggregation for bibliometric and social media indicators. Papers are increasingly cited before they have been officially published in a journal issue and mentioned on Twitter within days of online availability. In order to find a suitable proxy for the day of online publication allowing for the computation of more accurate benchmarks and fine-grained citation and social media event windows, various dates are compared for a set of 58,896 papers published by Nature Publishing Group, PLOS, Springer and Wiley-Blackwell in 2012. Dates include the online date provided by the publishers, the month of the journal issue, the Web of Science indexing date, the date of the first tweet mentioning the paper as well as the Altmetric.com publication and first seen dates. Comparing these dates, the analysis reveals that large differences exist between publishers, leading to the conclusion that more transparency and standardization is needed in the reporting of publication dates. The date on which the fixed journal article (Version of Record) is first made available on the publisher's website is proposed as a consistent definition of the online date.

Conference Topic
Journals, databases and electronic publications

Introduction
The process of scholarly communication, which usually begins with the formulation of a research idea and hypothesis and ends with publishing results to share them with the scientific community (Garvey & Griffith, 1964), has been sped up by means of electronic publishing (Dong, Loh, & Mondry, 2006; Wills & Wills, 1996). The publication delay, which Amat (2008, p. 382) defined as the "chronological distance between the stated date of reception of a manuscript by a given journal and its appearance on any print issue of that journal", has been accelerated by email and online manuscript handling systems as well as online publication (Wills & Wills, 1996). The delay period consists of the review process, which constitutes the main delay and ends with the acceptance of the manuscript, followed by technical delays of journal production and paper backlog.

Various studies have analyzed publication delays and found differences between scientific fields, journals, and publishers (e.g., Abt, 1992; Amat, 2008; Björk & Solomon, 2013; Das & Das, 2006; Diospatonyi, Horvai, & Braun, 2001; Dong et al., 2006). Since long delays interfere with priority claims and slow down scientific discourse, publication speed plays an important role for authors and scholarly communication (Rowlands & Nicholas, 2006; Schauder, 1994; Tenopir & King, 2000). Short publication delays can therefore be considered as a quality indicator reflecting the up-to-dateness of scientific journals (Haustein, 2012). Publishers have begun to reduce delays by making so-called *early view*, *in press*, *ahead of print* or *online first* versions of accepted papers available before they appear in an (print) issue. It has been shown for food research journals that online ahead of print publication has reduced publication delay by 29% (Amat, 2008), while Das and Das (2006) reported for 127 journals in 2005 average lags of three months between online and print issues publications

with particular differences between publishers. Tort, Targino, and Amaral (2012) showed that this lag increased significantly over time for six neuroscience journals. Online dates are now being recorded in bibliometric databases like Scopus, which impacts bibliometric analyses (Gorraiz, Gumpenberger, & Schlögl, 2014; Heneberg, 2013). Together with the increasing popularity of preprint servers (such as arXiv and SSRN) and institutional repositories, such *in press* versions have helped to speed up the read-cite-read cycle. As a result manuscripts increasingly cite papers that have not been officially published in a journal issue. Although scholarly communication has always involved sharing different versions of a manuscript with colleagues before, during, and after formal publication—such as exchanging drafts for feedback before submission or diffusing preprints after acceptance—, the electronic era makes these versions 'public', searchable, and (often) permanently retrievable on the web. To define and distinguish between various versions, the National Information Standards Organization (NISO) agreed upon the following versions of a journal article (NISO/ALPSP Working Group, 2008):

- Author's Original (AO) – manuscript ready to submit.
- Submitted Version Under Review (SMUR) – manuscript under formal peer review.
- Accepted Manuscript (AM) – version of journal article accepted for publication.
- Proof (P) – copy-edited version of accepted article.
- Version of Record (VoR) – fixed version of journal article formally published.
- Corrected Version of Record (CVoR) – VoR in which errors have been corrected.
- Enhanced Version of Record (EVoR) – VoR updated or enhanced with supplementary material.

It is important to note that by the NISO definition, the VoR is defined as a "fixed version of a journal article that has been made available by any organization that acts as a publisher by formally and exclusively declaring the article 'published'" (NISO/ALPSP Working Group, 2008, p. 3). This definition includes early views and in press articles without information on volume and issue or other identifiers as long as the content and layout of the article are fixed.

When it comes to bibliometric indicators, the acceleration of the publication process has been reflected in obsolescence patterns (Egghe & Rousseau, 2000) as well as citing half-lives (Luwel & Moed, 1998). These increasing online-to-print lags were shown to artificially increase citation rates including the immediacy index and impact factor (Heneberg, 2013; Seglen, 1997; Tort et al., 2012). The speed of scholarly communication becomes particularly visible in the context of social media metrics (the so-called altmetrics); for example, mentions of scientific documents on Twitter happen within hours (and sometimes within minutes) of online availability (Shuai, Pepe, & Bollen, 2012).

We argue that in the fast-moving digital era, the use of the publication *year* of the journal issue as the smallest level of time aggregation for bibliometric indicators is becoming insufficient, particularly in research evaluation contexts, due to the following factors:

a. acceleration of the read-cite-read cycle due to electronic publishing;
b. commonplace of online publication before publication of the journal issue; and
c. increasing online-to-print lags.

Following NISO's terminology, we suggest that the date of the first public online appearance of the VoR is the most relevant and should be used as the basic time unit to determine the official publication date of a paper. This would allow for the construction of more accurate citation and social media event windows, for example, citation windows of equal length (in days or months) for papers published in January or December, as well as the construction of more exact benchmarks by aggregating citations and social media events per week (e.g., tweets and Facebook shares) or month (citation rates) depending on the evaluation context.

Although many publishers now report online publication dates, many different dates are presented and the information provided varies between publishers, as no official standards

exist on publication dates. This paper explores and aims to verify various 'publication' dates in order to find a good proxy for the actual date of online availability. Thus, the paper aims to answer the following research questions:

1. Which publishers specify online dates and how do they provide them?
2. How reliable are dates provided by the publishers and how do they compare to each other?
3. What other existing dates can be used as a proxy of the online publication date of the VoR?

Methods and Materials

The dataset of this study was retrieved from the Web of Science (WoS) (as the major citation database) and is restricted to the publication year 2012 to limit effects of changes over time. To validate the publication dates provided by the publishers, the dates of the first tweet mentioning the particular paper were obtained from Altmetric.com. We argue that a tweet cannot link to a paper before it exists, thus the first tweet cannot have appeared before the online publication date. Tweets captured by Altmetric.com are linked to the documents via the DOI resulting in 313,301 WoS 2012 papers with at least one event captured by Altmetric.com (Haustein, Costas, & Larivière, 2015). Altmetric records that contained an arXiv ID or Astrophysics Data System (ADS) ID were removed to exclude tweets to preprints, which could have been made public before the online publication of the VoR. Twitter mentions are thus restricted to the mentions or links to the publisher's website, DOI, or PubMed ID.

Table 6. Top 10 publishers according to number of papers with types of dates available according to data provided by the publisher via API (a), in the metadata (m) of the webpage, on the webpage only (w), or as dynamic content only (d). Publishers selected for this study are highlighted in grey.

Publisher	Papers	Received	Revised	Accepted	Version of Record	Online	Publication	Date	Journal Issue	Journal Issue Online
Elsevier	51,292	d	d	d		d	a		w	
Wiley-Blackwell	47,958	w		w		m,w[i]	m		w,m	w
Lippincott	21,944							m	w,m	
Springer	19,225					m	m,a	m	w,m,a	
PLOS	16,208	w		w			a,m		a,m	
BMC	11,930	w		w			w,m		w,m	
NPG	11,181	w,m		w,m		m,a	w,m,a		w,m,a	
ACS	11,024							m,w	w	
Oxford	10,368	w		w		w		m	w,m	
Sage	8,776				w	w		m	w,m	

[i] Wiley provides two online dates "article published online" as well as "online date". See explanations below.

The top 10 publishers[1] of papers in the WoS-Altmetric dataset can be found in Table 1 together with the date information provided via API, in the metadata, in the webpage only, or as dynamic content of the webpage. It can be seen (in the headings of the table) that multiple terms exist to describe the online publication date and that multiple types of dates are made available on the website, in the metadata, or via the API; these include received, revised, accepted, version of record, online, publication, and date. Based on checking samples of articles for each of the publishers, we assume that the dates provided as *Version of Record*, *Online*, *Publication* and *Date* (Table 1) refer to (first) online appearances of the VoR required

[1] Publisher names from WoS were cleaned searching for name variants, but mergers and acquisitions were not accounted for. For example, BMC is considered an independent publisher, although it was acquired by Springer in 2008.

for this study. Wiley-Blackwell, Springer, PLOS, and Nature Publishing Group (NPG) were chosen due to their coverage and the technical feasibility of retrieving online date information. While Elsevier was the most represented publisher in this sample, it was difficult to obtain the required date information for their articles using PHP because this information is inserted dynamically into the webpage using JavaScript; Elsevier offers an API, but when queried[2] it was found to provide access to only the issue date and not to the online publication dates required for this study.

Using the DOI, the respective publishers' web platforms were queried to retrieve online dates. PLOS, Springer, and NPG each offer an API, but it was found that in some instances additional date information was only made available by searching the web page. In order to obtain the dates for Wiley, Springer and NPG, a PHP script was written that retrieved the HTML of the page. The HTML was then searched for metadata containing date information (e.g. <meta name="prism.publicationDate" content="2012-01-05"/>). When date information was found, it was saved to a relational database for evaluation. In instances where the article website had no (or missing) metadata available, the HTML was parsed and the contents of specific HTML tags found to contain date information was extracted and saved to a relational database; for the Wiley articles, a second script was written to retrieve dates not found in the metadata.

To compare different dates available and test in how far they can be used as proxies for online publication dates, other date information was obtained from WoS and Altmetric, so that together with the information from publishers the following dates were available:

- *online date*: retrieved from the publishers websites as part of the article metadata. For NPG ("Advance Online Publication"[3]), Springer ("Online First"[4]), and Wiley-Blackwell ("Early View"[5]) this date marks when the VoR was made publicly available on the publisher's website. For PLOS the online date equals the publication date because there is no difference between online and issue dates.
- *journal issue date*: the date from the journal issue as recorded by WoS. Since only a minority of papers provided the day of the month, the journal issue date was converted to the first of each month. Based on all 1.3 million papers in WoS published in 2012, 3.2% were published in issues spanning several months (such as JAN-FEB for a double issue). These were converted to the first day of the first month. A small percentage (0.5%) of papers appeared in seasonal issues (SPR, SUM, FAL, WIN). Since the data indicates that these are published at the beginning, middle, as well as the end of the particular season, these dates were disregarded. An additional 11.3% of all 2012 papers did not provide any issue date. Figure 1 provides an overview of the distribution of the 1.3 million WoS 2012 papers per journal issue date information.
- *Altmetric publication date*: the publication date as recorded by Altmetric.com, which is a mix of the journal issue date and online date (personal communication with Euan Adie and Jean Liu) as retrieved from the publisher. This is also the date Altmetric.com uses to compute the Altmetric score and provide benchmarks for papers of the same age. As shown in Figure 2, particular peaks can be observed for January 1 of each year as well as the first or last of each month. This might reflect common publishing practices, but could also be caused by aggregating data without actual day (and month) information. It was found that 15.1% of Altmetric.com records[6] did not have any publication date or they had incorrect dates (e.g. dates up to 2037).

[2] Using the http://api.elsevier.com/content/abstract/doi/{doi} API call
[3] http://www.nature.com/authors/author_resources/about_aop.html
[4] http://www.springer.com/authors/journal+authors/helpdesk?SGWID=0-1723213-12-817311-0
[5] http://olabout.wiley.com/WileyCDA/Section/id-404512.html#ev
[6] Based on 2.1 million Altmetric.com records collected in August 2014.

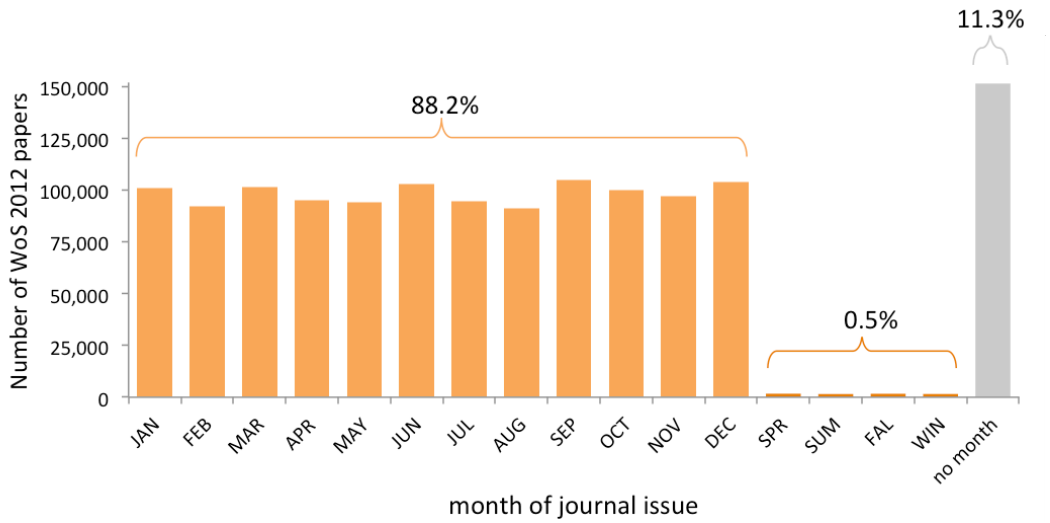


Figure 1. Number of WoS 2012 papers per months of journal issue.

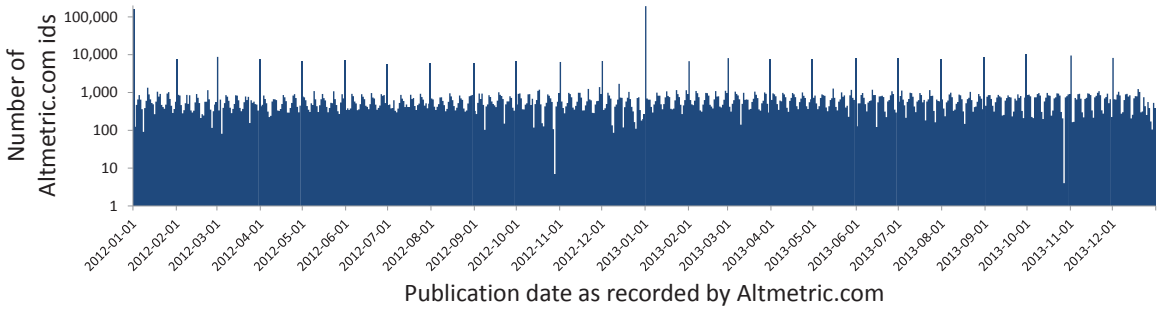


Figure 2. Number of Altmetric.com ids per Altmetric.com publication date from January 2013 to December 2014.

- *Altmetric first seen date*: the datestamp when Altmetric.com captured the first event for a particular document, which is missing for 4% of all records.[7]
- *First tweet date*: the datestamp of the first tweet[8] captured by Altmetric.com (excluding all papers with links to arXiv IDs or ADS IDs to ensure that the tweet did not refer to a preprint).
- *WoS indexing date*: the day when the document was indexed by WoS, which for 2012 papers was mostly during (37.7%) or in the month before (11.5%) or after (29.4%) the journal issue month.

In addition to the dates above we were also able to retrieve the following information for the papers published by Wiley-Blackwell:

- *Manuscript received*: the date the AO was submitted.
- *Manuscript accepted*: the date the AM was accepted.
- *Article first published online*: we could not determine the exact meaning of this date; for 95.6% of the total 34,507 Wiley-Blackwell documents it was identical with the *online date* and for 1.6% it was missing. For 2.3% of papers the *article first published online date* occurred before the *online date* by, on average, 35 days, which suggest

[7] Based on 2.1 million Altmetric.com records collected in August 2014.
[8] Twitter is the most common source covered by Altmetric.com (Robinson-García, Torres-Salinas, Zahedi, & Costas, 2014), so it makes sense to work with this date and not from other less common sources (e.g. Facebook or blogs).

that it marks the publication of the AM. However, in 137 cases (0.4%), it followed the *online date* by, on average, 52 days.

The final dataset—that is, the match of WoS, Altmetric.com, and papers with online dates retrieved from the four publishers—included 71,175 papers. For better comparison, it was restricted to papers for which all five dates tested as proxies for online publication (i.e., journal issue, Altmetric publication and first seen date, first tweet and WoS indexing date) were available. This amounted to a total of 58,896 papers, 12.5% NPG, 16.3% PLOS, 24.6% Springer and 46.6% Wiley-Blackwell.

Results and Discussion

Descriptive statistics comparing the online date to the five potential proxies are presented in Table 2, highlighting particular differences for the four publishers. Based on the assumption that the online date provided by the publishers were correct, the Altmetric publication date, first seen date, as well as the first tweet date seem to be the best proxies for online publication, while the journal issue and WoS indexing date show the largest deviations from the online publication dates. These differences reflect the nature of these dates. For example, Altmetric collects its publication dates from the publishers websites and while first tweets are known to happen shortly after publication (Shuai et al., 2012), WoS processing takes more time, namely, on average between 39 days for PLOS or 163 days for Springer papers. The 61 (NPG), 84 (Wiley-Blackwell), and 146 (Springer) days between online and journal issue date mostly reflect the backlog between online availability and publication of the journal issue. Although the (print) issue is generally assumed to follow online publication chronologically, results in Table 2 show that for 3.47% of Springer, 9.09% of Wiley-Blackwell, and 20.04% of NPG papers analyzed the online date came after the journal issue date, which is considered negative delay (Das & Das, 2006).

Although Altmetric and Twitter dates work better than journal issue and WoS indexing, none of the dates seem to reflect the online date well and large differences can be observed between publishers, in particular for Wiley-Blackwell, which questions the validity of any of the five dates as a reliable proxy of the publication of the VoR across publishers. The Altmetric publication date, which overall shows the smallest difference compared to the online date provided by the publishers—on average, 9 days for Springer, 12 days for NPG, 27 days for PLOS, and 121 for Wiley-Blackwell—is also problematic, because it is set to a date prior to online publication in 43.37% of Springer, 55.38% of NPG, 63.83% of Wiley-Blackwell, and 66.49% of PLOS papers. The variance between publishers affects Altmetric scores (but arguably also citation scores) when benchmarking a paper's scores against that of papers of the same reported age.

Based on the assumption that a tweet cannot mention a paper before it exists in the online space it links to, the online dates provided by Wiley-Blackwell seem to be the most problematic (Figure 3), as 14.52%[9] of the 27,432 analyzed papers had tweets linking to them before the date that the publisher identifies as the online publication date. On the other hand, none of the PLOS papers and few of the Springer (0.08%) articles were mentioned on Twitter before the online publication date. Although all of the papers analyzed have been tweeted, the mean number of days between online date and first tweet was higher than expected, ranging from 15 days for PLOS to 92 days for Springer. Moreover, the first mention on Twitter happened on the day of online publication for 1.06% (Springer) and 34.47% (NPG) sampled papers, which—particularly considering that about 80% of recent papers are never tweeted

[9] Results change only slightly when using the *article first published online* date, i.e. 14.61% of Wiley-Blackwell papers had a tweet appear before this date.

(Haustein, Costas, & Larivière, 2015)—limits the usefulness of the first tweet date as a proxy for online publication.

Table 2. Statistics for chronological distance (in number of days) of the journal issue month, Altmetric publication and first seen date, first tweet date and WoS indexing date with the online date for NPG, PLOS, Springer and Wiley-Blackwell.

Chronological distance to online date in number of days		NPG *n=7,391*	PLOS *n=9,600*	Springer *n=14,473*	Wiley-Blackwell *n=27,432*
Journal issue month[i]	% before	20.04%	n/a[ii]	3.47%	9.09%
	% identical	5.47%		0.11%	0.29%
	% after	74.50%		96.42%	90.62%
	mean	61		146	84
	standard deviation	78		111	93
	min	-330		-269	-423
	max	548		1,850	1,032
Altmetric publication date	% before	55.38%	66.49%	43.37%	63.83%
	% identical	39.35%	31.41%	34.11%	2.81%
	% after	5.28%	4.44%	22.52%	33.36%
	mean	12	27	9	121
	standard deviation	68	79	48	322
	min	-3,013	-697	-519	-16,761
	max	411	526	1,850	5,016
Altmetric first seen date	% before	3.48%	0.00%	0.08%	14.59%
	% identical	32.88%	36.64%	1.04%	14.26%
	% after	63.64%	63.36%	98.89%	71.15%
	mean	35	12	90	63
	standard deviation	87	49	164	122
	min	-459	0	-257	-533
	max	890	602	1,843	1,228
First tweet date	% before	3.52%	0.00%	0.08%	14.52%
	% identical	34.37%	37.23%	1.06%	15.21%
	% after	62.21%	62.77%	98.85%	70.27%
	mean	37	15	92	65
	standard deviation	92	59	169	127
	min	-459	0	-257	-533
	max	890	811	1,843	1,393
WoS indexing date	% before	2.72%	0.00%	0.10%	0.05%
	% identical	0.01%	0.00%	0.00%	0.00%
	% after	97.27%	100.00%	99.90%	99.95%
	mean	83	39	163	97
	standard deviation	81	20	113	94
	min	-302	9	-252	-359
	max	576	262	1,866	1,049

[i] First of the journal issue month as recorded by WoS.
[ii] PLOS does not distinguish between online and issue date, so that the two dates are actually identical.

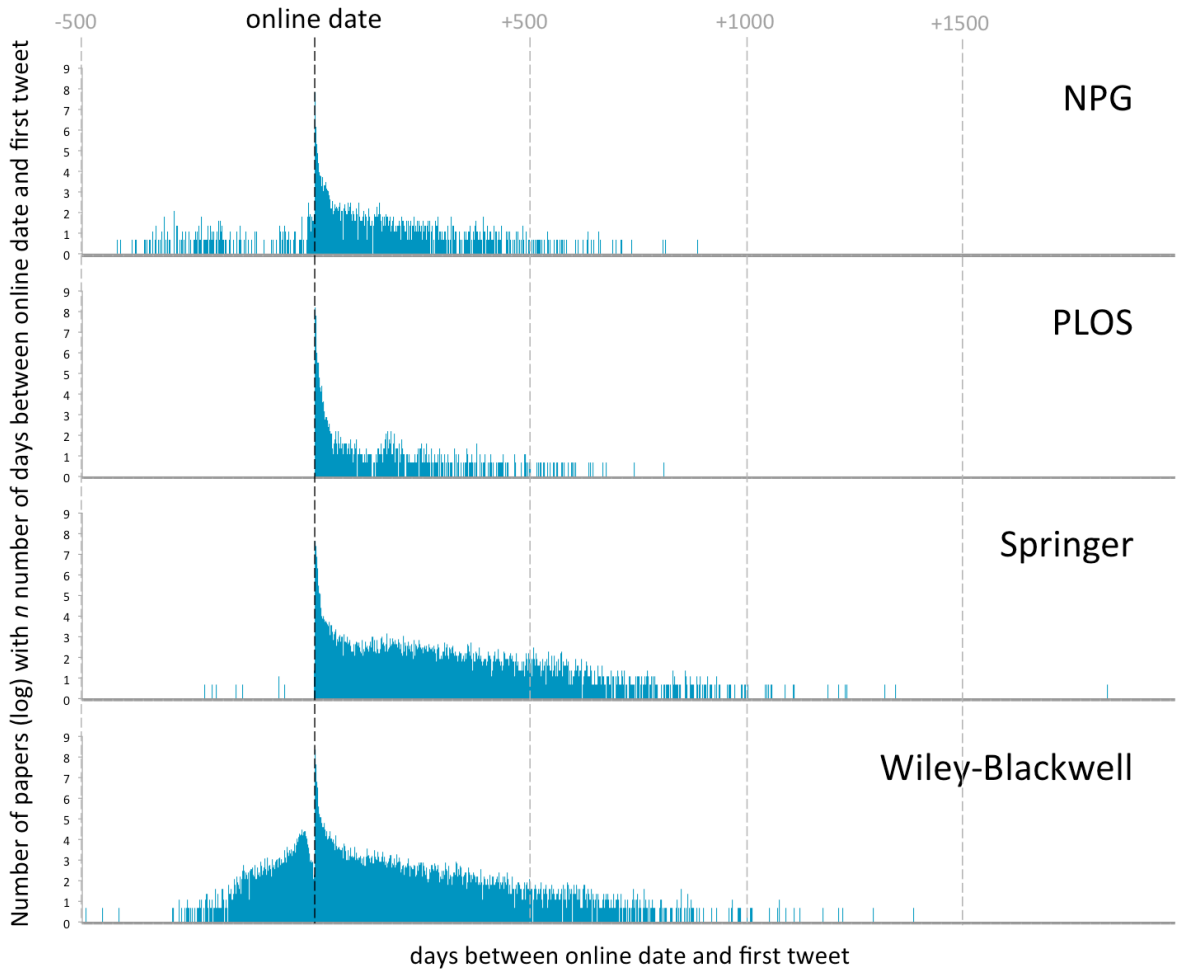


Figure 3. Number of papers (log) with *n* days between online date and first tweet per publisher.

Conclusions and Outlook

Currently none of the investigated dates represent a good proxy for the date a journal article was actually available online. In particular, the finding that a considerable amount of Wiley-Blackwell papers had been mentioned on Twitter before the online date, suggests that inconsistencies exist in terms of how publishers report online dates. This applies to the technical aspects as well as to actual content and vocabulary used. Thus, even when online dates can be retrieved from the publishers' websites or via API, they do not seem to always (and in a similar way for every publisher) mark the actual point in time when something was made accessible online. There is, thus, an urgent need for transparency and standardization of various dates reported by publishers in order to assure comparability of online dates across publishers. Adopting the vocabulary developed by NISO, specific dates could be reported for each version of the journal article, and the first appearance of the VoR would thus mark the date the fixed version of the document appeared online. A standardized vocabulary and a common definition of what various publication dates mean would not only improve benchmarking in the context of research evaluation but would also help to accurately determine the start of open access embargo periods required by certain funders, such as the NIH in the United States or the European Research Council. Currently these embargo periods, delaying green open access by a couple of months to years to protect publishers' revenue, are supposed to begin with publication of the article, which can refer to either journal issue or

online date.[10] Setting the start date of the embargo to the online publication date of the VoR would remove a potential loophole that allows the publishers to increase the embargo period during which they have the exclusivity of access.

Until such a standard is implemented, research on metrics should focus on obtaining more publisher-independent date information. One potential proxy for online publication could be the date when a DOI resolved successfully for the first time. Recently CrossRef has implemented the DOI Chronograph, a tool which tracks various deposits of metadata by the publisher as well as the first day of successful DOI resolution (Wass, 2015). Future work will investigate in how far these dates can be used to create fine-grained benchmarks needed in the context of social media metrics. Regarding citations, where monthly proxies are sufficient, the WoS Indexing date should be further investigated.

Acknowledgments

The authors would like to thank Euan Adie and Altmetric.com for access to their data and acknowledge funding from the Alfred P. Sloan Foundation, grant no. 2014-3-25.

References

Abt, H. A. (1992). Publication practices in various sciences. *Scientometrics*, *24*(3), 441–447. doi:10.1007/BF02051040

Amat, C. B. (2008). Editorial and publication delay of papers submitted to 14 selected Food Research journals. Influence of online posting. *Scientometrics*, *74*(3), 379–389. doi:10.1007/s11192-007-1823-8

Björk, B.-C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914–923. doi:10.1016/j.joi.2013.09.001

Das, A., & Das, P. (2006). Delay between online and offline issue of journals: A critical analysis. *Library & Information Science Research*, *28*(3), 453–459. doi:10.1016/j.lisr.2006.03.019

Diospatonyi, I., Horvai, G., & Braun, T. (2001). Publication speed in analytical chemistry journals. *Journal of Chemical Information and Modeling*, *41*(6), 1452–1456. doi:10.1021/ci010033d

Dong, P., Loh, M., & Mondry, A. (2006). Publication lag in biomedical journals varies due to the periodical's publishing model. *Scientometrics*, *69*(2), 271–286. doi:10.1007/s11192-006-0148-3

Egghe, L., & Rousseau, R. (2000). The influence of publication delays on the observed aging distribution of scientific literature. *Journal of the American Society for Information Science*, *51*(2), 158–165. doi:10.1002/(SICI)1097-4571(2000)51:2<158::AID-ASI7>3.0.CO;2-X

Garvey, W. D., & Griffith, B. C. (1964). Scientific information exchange in psychology: The immediate dissemination of research findings is described for one science. *Science*, *146*(3652), 1655–1659. doi:10.1126/science.146.3652.1655

Gorraiz, J., Gumpenberger, C., & Schlögl, C. (2014). Usage versus citation behaviours in four subject areas. *Scientometrics*, *101*(2), 1077–1095. doi:10.1007/s11192-014-1271-1

Haustein, S. (2012). *Multidimensional Journal Evaluation. Analyzing Scientific Periodicals beyond the Impact Factor*. Berlin / Boston: De Gruyter Saur. doi:10.1515/9783110255553

Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing social media metrics of scholarly papers: the effect of document properties and collaboration patterns. *PLoS ONE*, *10*(3), e0120495. doi:10.1371/journal.pone.0120495

Heneberg, P. (2013). Effects of print publication lag in dual format journals on scientometric indicators. *PLOS ONE*, *8*(4), e59877. doi:10.1371/journal.pone.0059877

Luwel, M., & Moed, H. F. (1998). Publication delays in the scientific field and the their relationship to the ageing of scientific literature. *Scientometrics*, *41*(1-2), 29–40.

NISO/ALPSP Journal Article Versions (JAV) Technical Working Group. (2008). *Journal Article Versions (JAV): Recommendations of the NISO/ALPSP JAV Technical Working Group*. Baltimore. Retrieved from http://www.niso.org/publications/rp/RP-8-2008.pdf

Robinson-García, N., Torres-Salinas, D., Zahedi, Z., & Costas, R. (2014). New data, new possibilities: exploring the insides of Altmetric.com. *El Profesional de La Informacion*, *23*(4), 359–366. doi:10.3145/epi.2014.jul.03

Rowlands, I., & Nicholas, D. (2006). The changing scholarly communication landscape: An international survey of senior researchers. *Learned Publishing*, *19*, 31–55. doi:10.1087/095315106775122493

[10] http://authorservices.wiley.com/bauthor/faqs_fundingbodyrequirements.asp

Schauder, D. (1994). Electronic publishing of professional articles: Attitudes of academics and implications for the scholarly communication industry. *Journal of the American Society for Information Science*, *45*(2), 73–100. doi:10.1002/(SICI)1097-4571(199403)45:2<73::AID-ASI2>3.0.CO;2-5

Seglen, P. O. (1997). Citations and journal impact factors: questionable indicators of research quality. *Allergy*, *52*(11), 1050.

Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: article downloads, Twitter mentions, and citations. *PLoS ONE*, *7*(11), e47523. doi:10.1371/journal.pone.0047523

Tenopir, C., & King, D. W. (2000). *Towards Electronic Journals: Realities for Scientists, Librarians, and Publishers*. Washington, DC: Special Libraries Association.

Tort, A. B. L., Targino, Z. H., & Amaral, O. B. (2012). Rising publication delays inflate journal impact factors. *PLoS ONE*, *7*(12), e53374. doi:10.1371/journal.pone.0053374

Wass, J. (2015). Introducing the CrossRef Labs DOI Chronograph. *Crosstech blog post 12 January 2015*. Retrieved January 21, 2015, from http://crosstech.crossref.org/2015/01/introducing-chronograph.html

Wills, M., & Wills, G. (1996). The ins and the outs of electronic publishing. *Internet Research*, *6*(1), 10–21. doi:10.1108/10662249610123647

Analysis of the Obsolescence of Citations and Access in Electronic Journals at University Libraries

Chizuko Takei[1], Fuyuki Yoshikane[2] and Hiroshi Itsumura[3]

[1] *naoe.chizuko@ynu.ac.jp*
University of Tsukuba, Graduate School of Library, Information and Media Studies, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

[2] *fuyuki@slis.tsukuba.ac.jp,* [3]*hits@slis.tsukuba.ac.jp*
University of Tsukuba, Faculty of Library, Information and Media Science, 1-2 Kasuga, Tsukuba, Ibaraki (Japan)

Abstract

This study analyzes the correlation between the obsolescence of citations and access concerning a broad range of subjects, including fields that have not been dealt with in previous studies, shedding light on the differences between these two types of obsolescence and the characteristics for each field. The analysis investigates approximately 1,200 journals that were randomly sampled from 11 subject fields in SpringerLink and 20 subject fields in ScienceDirect. Metrics such as cited half-life and download half-life are employed to examine the relationship between the rate of obsolescence of citations and access. As a result, no strong correlation between citations and access is observed in most fields with regard to the short-term obsolescence. As for the long-term obsolescence, on the other hand, comparatively strong and significant correlations are seen in natural sciences other than medicine-related fields ($p < 0.05$).

Conference Topic

Journals, databases and electronic publications

Introduction

This study analyzes the relationship between the obsolescence of citations and access for usage of electronic journals in Japanese university libraries. The Big Deal, which is a package contract for electronic journals, has been rapidly adopted among Japanese university libraries. Irrespective of the university's size, the Big Deal drastically increased the number of accessible titles of journals at contract universities. However, with ongoing budget cuts and increasing journal prices, price hikes for the Big Deal are putting pressure on library budgets. This situation makes it difficult for libraries not only to subscribe to new journals but also to maintain existing subscriptions. As withdrawal from the Big Deal results in a drastic decrease in the number of accessible titles of journals, and thereby a collapse of the library's academic information framework, collection building of journal backfiles is necessary to alleviate the impact of these losses.

The collection development of journal backfiles differs from that of current files, which have a strong tendency to become fixed owing to budgetary considerations. This is because library staffs at many universities select and propose journal backfiles to be introduced under their own direction, for example, by utilizing special proposals received from publishers shortly before the accounting period. However, few Japanese universities have sought to implement a planned introduction of journal backfiles by scrutinizing the level of on-campus demand and the effectiveness of such an introduction.

As Takei, Yoshikane, and Itsumura (2013) pointed out, effective methods of collecting journal backfiles have rarely been studied in the literature. Investigating the development of backfiles requires perspectives focusing on the articles that fall into disuse, that is, obsolescence. Slower obsolescence represents stronger demand of researchers for older articles in the concerned field. Obsolescence analysis has been performed on library

collections to evaluate a decrease in the use of documents over time. The obsolescence of books is assessed on the basis of the number of times a book is used by lending year and accession year. In contrast, obsolescence of journals is based on citations and access to documents. Understanding the relationship between the obsolescence of citations and access will make it possible to estimate the obsolescence of access on the basis of information regarding the obsolescence of citations. This relationship has already been examined in certain fields, such as chemistry, and for specific journals, as will be described in the next section. However, the nature of documental use (citations and access) varies by field, and trends in the differences between the obsolescence of citations and access may also differ by field. Thus, this study employs several indices of obsolescence, some of which had not been adopted before our previous study (Takei, Yoshikane & Itsumura 2013), and analyzes obsolescence of access and citations for a wide range of subjects, including fields that have not previously been examined. We shed light on the differences between both types of obsolescence and their characteristics in each field.

Related Research

There are some indices for analyzing the relationship between citations and downloads (access). Impact Factor (IF), Immediacy Index (II), and Cited Half-life (CHL) are major indices of citations, while Download Impact Factor (DIF), Download Immediacy Index (DII), Download Half-life (DHL), and Usage Half-life (UHL), which is used as a synonym of DHL, are indices of downloads. According to the definition of Journal Citation Reports (JCR), IF is "the average number of times articles from the journal published in the past two years have been cited in the JCR year," II is "the average number of times an article is cited in the year it is published," and CHL is "the median age of the articles that were cited in the JCR year." IF and II indicate how frequently articles in the journal are cited within several years after publication and immediately after publication, respectively. CHL shows the degree of demand for older articles in the journal. In contrast, DIF and DII analogically apply the definitions of IF and II to downloads, respectively, and both DHL and UHL replicate the definition of CHL to access. Using these indices, many studies have been conducted on the relationship between citations and downloads to evaluate journal collections. For instance, Duy and Vaughan (2006) analyzed local citation data and IF with journal usage in the fields of chemistry and biochemistry. Good correlations were seen between local citation data and journal usage, whereas no significant correlation was observed between IF and journal usage. Other examples can be found in Chu and Krichel (2007), McDonald (2007), Bollen and van de Sompel (2008), and Watson (2009). In particular, there are some studies on obsolescence of access and citations related to electronic journals. For instance, Nicholas et al. (2005) surveyed synchronous obsolescence of access, revealing that over half of all usage was accounted for by items published within the last 15 months. Moreover, several studies have analyzed the relationship between obsolescence of citations and access by calculating and comparing the densities of citations and access (e.g., Kurtz et al., 2005; Moed, 2005; Brody et al., 2006).

In recent years, Schloegl and Gorraiz (2010; 2011) conducted more multifaceted studies related to oncology and pharmacology, using indices such as IF, II, and CHL. In the case of oncology journals in 2006, the results indicated that the means of UHL and CHL were 1.7 years and 5.6 years, respectively. Similar results were found in the case of pharmacology journals in the same year. Furthermore, they calculated CHL and found a medium-sized correlation between CHL and UHL in pharmacology ($r = 0.42$). Wan et al. (2010) examined the relationship between DII and citation indicators using the Chinese full-text database, the Chinese National Knowledge Infrastructure (CNKI).They found that DII had the potential to be a predictor for other indices such as h-index. While a moderate correlation between DII

and II was observed in the field of agriculture and forestry (r = 0.57), a strong correlation was found in psychology (r = 0.8). In addition, Gorraiz, Gumpenberger and Schloegl (2013) investigated the differences in obsolescence between citations and downloads in five fields in ScienceDirect, and Guerrero-Bote and Moya-Anegon (2013) observed the influence of language on the relationship between citations and downloads.

However, these analyses have only been performed for limited fields, including organic chemistry, astronomy, and astrophysics, and for selected journals in those fields. Although our previous work analyzed the obsolescence of citations and access with regard to all fields in Springer's SpringerLink and suggested the predictability of the long-term obsolescence of access on the basis of that of citations (Takei et al., 2013), its sample size for each field was small and insufficient for generalizing the results for the whole field.

Therefore, this study examines Elsevier's ScienceDirect in addition to SpringerLink to increase the sample size. SpringerLink is a collection comprising 11 fields focusing on Science, Technology, and Medicine (STM), whereas ScienceDirect is a collection comprising 23 fields including social sciences as well as STM. Analyzing both collections will enable a survey for a wider range of fields; besides, as for the fields included in both, it will facilitate an analysis based on more samples. It is assumed that indices of obsolescence that are effective for predicting the effects of backfiles will differ by field. Utilizing data of the two collections, we clarify the relationship in obsolescence between citations and downloads for each field.

Methodology

This study targeted Yokohama National University (YNU) in Japan, a medium-sized national university without a medical school. YNU consists of four undergraduate colleges (Education and Human Sciences, Economics, Business Administration, and Engineering Science) and five graduate schools (Education, International Social Sciences, Engineering, Environment and Information Sciences, and Urban Innovation). The university comprises around 600 full-time teaching staff and 10,000 students (around 2,600 graduate and 7,500 undergraduate students).

The survey employed the 2009–2012 editions of JCR as citation data, and statistics on the use of full text by publication year in the style of COUNTER Journal Report 5 for SpringerLink (2010–2012) and ScienceDirect (2001–2012) as access data. COUNTER Journal Report 5 defines the number of downloads, the number of times accessed, and the number of times used as the number of times the "full text" of an article is used. As with many studies, we employed this definition and referred to it as access count. COUNTER report has some limitations, for example, it does not reflect all of researchers' activities or could not distinguish the number of access by unique users. However, it reflects a certain amount of user's needs and it is useful to evaluate journal collections. We examined all the 11 fields in SpringerLink and 20 of the 23 fields in ScienceDirect (excluding Decision Science, Nursing and Health Professions, and Veterinary Science and Veterinary Medicine, for which the number of journals suitable for our analysis was less than 10). Because, for both collections, statistics contained sections in which the access count for multiple publication years had been summed up, the access count was divided by the number of years in the section to calculate the access count for each year.

The main concern of this study is to examine the practical predictability of local usage (i.e., access count in a given university) for each field based on global citation data, which is easily available from JCR, for collection management. Although local data does not always correspond with global data as shown in earlier studies (e.g., Duy & Vaughan, 2006; Bollen & van de Sompel, 2008), there may be a certain relationship between them because the

former is a part of the latter and the former partly reflects the latter. Thus, we compared local access data to global citation data in order to reveal the predictability of local access.

The sampling procedure was as follows. First, from all 2,782 journals in SpringerLink and all 1,792 journals in ScienceDirect, we extracted the journals whose fields could be identified on the basis of the title lists of publishers, excluding journals whose full text had never been accessed at YNU. As for ScienceDirect, where journals are classified into multiple fields, this study employed the fields first listed in Web of Science to ensure the same analysis conditions as for SpringerLink. Consequently, 1,567 and 1,657 journals were selected from SpringerLink and ScienceDirect, respectively.

Next, journals with index values listed in the relevant edition of JCR were sampled and rearranged in descending order of cumulative ratio of access counts for each field. These journals were separated into three layers according to the cumulative ratio of access counts as illustrated in Figure 1, i.e., less than 70%, 70% up to (not including) 90%, and 90% and above.

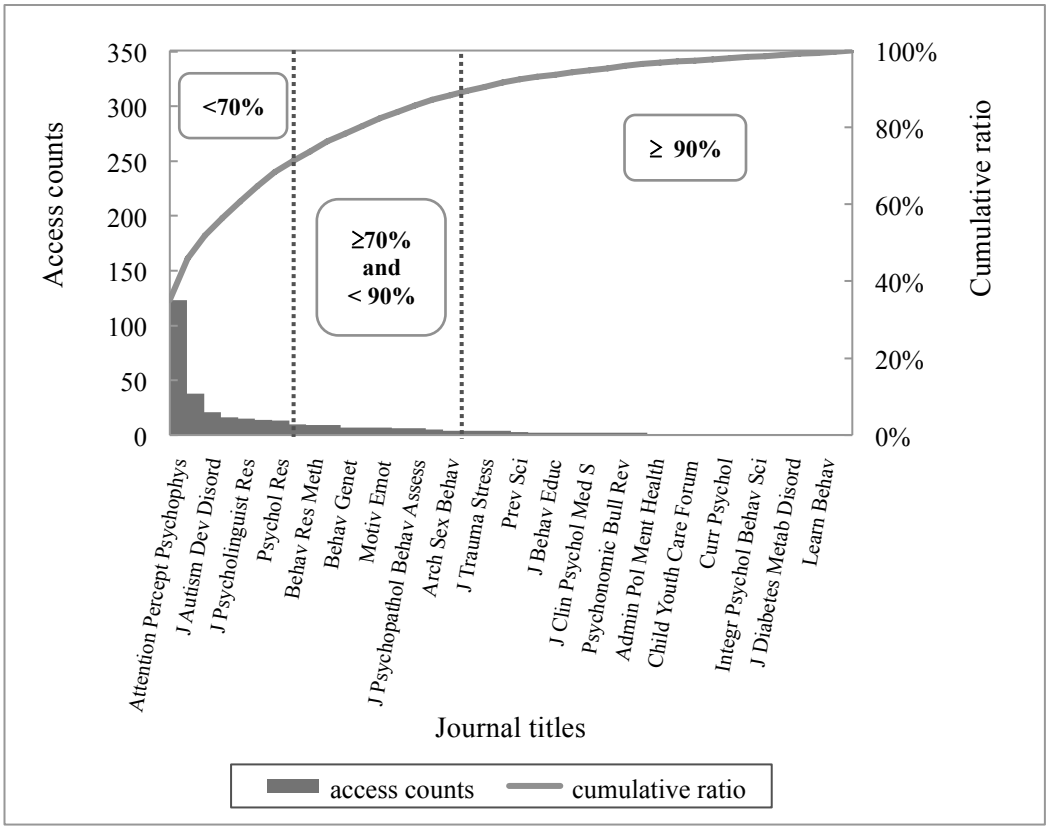


Figure 1. An example of 3 layers according to the cumulative ratio of access counts (Behavioral Science in SpringerLink).

To examine overall trends in each field, 15 journals were then randomly sampled from each of the layers in each field other than the three fields of ScienceDirect described above; for layers with less than 15 journals, all journals were considered. On this occasion, we sampled the journals that fulfilled the following conditions to obtain data for calculating the indices regarding obsolescence as of 2011 and 2012:

(a) Journals whose access count in 2011 and 2012 is not zero to analyze long-term obsolescence.

(b) Journals included in collections from 2011 to 2012 to analyze short-term obsolescence.

(c) Journals that fulfill the conditions of both (a) and (b) to examine the relationship between the two types of obsolescence.

As a result, the number of titles that became the targets of research was as follows:

SpringerLink: (a) 417, (b) 469, (c) 135

ScienceDirect: (a) 773, (b) 752, (c) 571

Tables 1 and 2 show the number of titles by field in the collections of SpringerLink and ScienceDirect, respectively. With regard to the sampling condition (c), we excluded 6 fields of SpringerLink (Behavioral Science; Business and Economics; Computer Science; Humanities, Social Sciences and Law; Mathematics and Statistics; and Medicine) and one field of ScienceDirect (Psychology) for which we obtained only 10 samples or less.

Table 1. Number of titles by field in SpringerLink

Subject	Sampling condition (a)	Sampling condition (b)	Sampling condition (c)
Behavioral Science (BS)	17	30	N/A
Biomedical and Life Sciences (BL)	45	45	32
Business and Economics (BE)	29	40	N/A
Chemistry and Materials Science (CM)	45	45	35
Computer Science (CS)	40	45	N/A
Earth and Environmental Science (EE)	45	45	30
Engineering (EG)	42	42	16
Humanities, Social Sciences and Law (HS)	30	42	N/A
Mathematics and Statistics (MS)	45	45	N/A
Medicine (MD)	34	45	N/A
Physics and Astronomy (PA)	45	45	22
Whole	417	469	135

Table 2. Number of titles by field in ScienceDirect

Subject	Sampling condition (a)	Sampling condition (b)	Sampling condition (c)
Agricultural and Biological Sciences (AB)	41	41	41
Biochemistry, Genetics and Molecular Biology (BG)	45	45	45
Business, Management and Accounting (BM)	36	34	20
Chemical Engineering (CE)	40	40	40
Chemistry (CH)	36	35	35
Computer Science (CS)	45	45	35
Earth and Planetary Sciences (EP)	45	45	43
Economics, Econometrics and Finance (EF)	45	45	30
Energy (EN)	22	21	16
Engineering (EG)	45	45	45
Environmental Science (ES)	36	36	35
Health Sciences (HE)	45	43	20
Immunology and Microbiology (IM)	37	37	17
Materials Science (MT)	43	42	43
Mathematics (MA)	36	36	21
Neuroscience (NS)	38	34	12
Pharmacology, Toxicology and Pharmaceutical Science (PT)	30	29	18
Physics and Astronomy (PA)	33	33	32
Psychology (PC)	36	29	N/A
Social Sciences (SS)	39	37	23
Whole	773	752	579

Sampling conditions: (a) Journals whose access count in 2011 and 2012 is not zero to analyze long-term obsolescence; (b) Journals included in collections from 2011 to 2012 to analyze short-term obsolescence; (c) Journals that fulfill the conditions of both (a) and (b) to examine the relationship between the two types of obsolescence.

This study employs the following indices as measures of obsolescence:

(1) Obsolescence of citations:
  (1A) Cited Half-life (CHL)
  (1B) Immediacy Index/Impact Factor (II/IF), i.e., ratio between II and IF
(2) Obsolescence of access:
  (2A) Download Half-life (DHL)
  (2B) Download Immediacy Index/Download Impact Factor (DII/DIF), i.e., ratio between DII and DIF

CHL and DHL express slower obsolescence, while II/IF and DII/DIF express faster obsolescence, as values become higher. In addition, whereas CHL and DHL are indices of obsolescence of use that take into consideration long periods of time, II/IF and DII/DIF particularly focus on the change in usage during several years after publication. DII/DIF, the ratio between DII and DIF, had not been used in obsolescence analysis before our previous study (Takei et al., 2013). However, given that the use of journals is generally concentrated at the time immediately after publication, it seems that DII/DIF would also prove useful as an index representing the nature of documental use in each field. For example, as for 2012, DII/DIF of Medicine is 5368.33 whereas DII/DIF of Earth and Environmental Science is 41.17 in SpringerLink. This means that the former field tends to progress quickly and the "latest" findings attract a lot of attention in the field whereas the latter field is inclined to emphasize not only the "latest" results but also previous ones. Therefore, DII/DIF was used in combination with II/IF in this study. The survey examined the degree of accordance—that is, correlation—of obsolescence between citations and access for each field with respect to the long-term (CHL and DHL) and the short-term (II/IF and DII/DIF). First, the values of these indices were calculated as of 2012. Data for CHL, II, and IF was obtained from the JCR of 2012. DHL, DII, and DIF analogically apply the definitions of CHL, II, and IF in JCR, respectively, to access count. To compute these indices, we set the sampling conditions (a) and (b) described above. In the analysis of short-term obsolescence based on the sampling condition (b), DII and DIF were used with the addition of one to avoid division by zero. Furthermore, to compare the tendencies in 2012 with those in the preceding year (i.e., to observe changes in documental use), the values as of 2011 were also obtained in the same manner.

If good correlations are found between the indices of citations and access in some fields, the information of CHL or II/IF obtained from JCR greatly helps us to determine the strategy to collect journal backfiles for these fields. That is, the correlations suggest the predictability of the use of journal backfiles by the information that can be obtained before introducing them.

Results

First, to determine the degree of accordance of obsolescence of citations and access, correlations between each pair of indices were observed: (A) between CHL and DHL; and (B) between II/IF and DII/DIF. The samples for analyzing (A) and (B) were extracted on the sampling conditions (a) and (b), respectively. The distributions of II/IF and DII/DIF had high values of skewness (2.71–12.97). Moreover, we cannot obtain exact values for CHL from JCR, in which the maximum value of CHL is 10, that is, even if its true value is greater than 10, CHL is described as 10. Thus, Spearman's rank correlation coefficient ρ was employed instead of Pearson's product-moment correlation coefficient r, which should be applied to interval or ratio scale data following a normal distribution.

Table 3 shows the correlation coefficients for (A) CHL and CHL and those for (B) II/IF and DII/DIF by field. There are differences between SpringerLink and ScienceDirect, both in the number and scope of fields. Therefore, to make it easier to compare the results of both collections, we reclassified all fields into the following 6 fields: Humanities and Social Sciences, Medicine, Chemistry and Engineering, Mathematics and Computer Science, Agricultural and Environmental Science, and Physics, as shown in Table 3.

As for 2012, the correlation coefficients for all fields were (A): $\rho = 0.50$ ($p < 0.05$) and (B): $\rho = 0.04$ ($p < 0.05$) in SpringerLink; (A): $\rho = 0.30$ ($p < 0.05$) and (B): $\rho = 0.03$ in ScienceDirect. While a moderate correlation was observed for (A), almost no correlation was found for (B). With regard to individual fields, in the case of (A), the strongest and statistically significant correlation was seen for Physics and Astronomy ($\rho = 0.59$, $p < 0.05$) in SpringerLink and for Energy ($\rho = 0.62$, $p < 0.05$) in ScienceDirect.

Table 3. Rank correlation ρ of obsolescence between citations and access.

Subject		2012 (A)		2012 (B)		2011 (A)		2011 (B)	
Humanities and Social Sciences	BS (S)	0.25		0.04		0.11		−0.10	
	BE (S)	0.46	*	0.07	*	0.32		−0.10	
	HS (S)	0.33		0.13		0.04		0.14	
	BM (E)	0.09		−0.27		−0.31		0.28	
	EF (E)	0.26		0.01		0.13		0.08	
	PC (E)	0.16		0.22		−0.04		0.00	
	SS (E)	0.05		−0.07		0.36	*	−0.04	
Medicine	BL (S)	0.51	*	0.28		0.29		0.40	*
	MD (S)	0.32		0.19		0.40	*	0.39	*
	HE (E)	0.09		−0.06		0.22		0.17	
	IM (E)	0.05		0.06		0.18		0.24	
	NS (E)	0.30		−0.31		0.18		0.08	*
	PT (E)	0.08		0.05		0.27		0.04	
Chemistry and Engineering	CM (S)	0.57	*	0.09		0.62	*	0.00	
	EG (S)	0.50	*	0.04	*	0.72	*	0.26	
	BG (E)	0.26		0.15		0.50	*	0.22	
	CE (E)	0.60	*	0.32	*	0.57	*	0.28	
	CH (E)	0.30	*	0.05		0.66	*	0.10	*
	EG (E)	0.34	*	0.04		0.42	*	0.26	
	MT (E)	0.56	*	0.07		0.56	*	0.03	
Mathematics and Computer Science	CS (S)	0.43	*	−0.06		0.45	*	0.09	
	MS (S)	0.43	*	0.07		0.52	*	−0.11	
	CS (E)	0.25		0.13		0.23		0.17	
	MA (E)	0.36	*	0.05		0.41	*	−0.20	
Agricultural and Environmental Science	EE (S)	0.47	*	0.02		0.53	*	0.03	
	AB (E)	0.15		0.04		0.36	*	0.18	
	ES (E)	0.46	*	−0.24		0.39	*	0.18	
Physics	PA (S)	0.59	*	0.08		0.39	*	−0.12	
	EP (E)	0.32	*	0.27		0.32	*	−0.21	
	EN (E)	0.62	*	0.11		0.73	*	0.23	
	PA (E)	0.35	*	0.10		0.33		−0.30	
Whole	(S)	0.50	*	0.04	*	0.45	*	0.01	
	(E)	0.30	*	0.03		0.37	*	0.08	*

(A): correlations between the indices of long-term obsolescence (CHL and DHL) on the sampling condition (a).
(B): correlations between the indices of short-term obsolescence (II/IF and DII/DIF) on the sampling condition (b)
(S): fields in SpringerLink. (E): fields in ScienceDirect. *Significant ($p < 0.05$)

In the case of (B), the correlation was significant and stronger in Chemical Engineering (ρ = 0.32, $p < 0.05$) in ScienceDirect than in other fields, and negative correlations were witnessed in some fields unlike in the case of (A). Meanwhile, as for 2011, the correlation coefficients for all fields were (A): $\rho = 0.45$ ($p < 0.05$) and (B): $\rho = 0.01$ in SpringerLink; (A): $\rho = 0.37$ ($p < 0.05$) and (B): $\rho = 0.08$ ($p < 0.05$) in ScienceDirect. With regard to individual fields, the correlation between indices changed according to the base years of observation. In the case of (A), for example, while Energy showed the strongest significant correlation both in 2012: ρ = 0.62 ($p < 0.05$) and in 2011: $\rho = 0.73$ ($p < 0.05$), the correlation for Chemistry varied from ρ = 0.66 ($p < 0.05$) in 2011 to 0.30 ($p < 0.05$) in 2012 in ScienceDirect. In the case of (B), for example, the correlation for Medicine varied from $\rho = 0.39$ ($p < 0.05$) in 2011 to 0.19 in 2012 in SpringerLink.

Concerning the 6 fields after reclassification, somewhat strong and significant correlations were seen between the indices of long-term obsolescence (CHL and DHL) in natural sciences other than Medicine, particularly in Physics and in Chemistry and Engineering.

Engineering (EG), Computer Science (CS), and Physics and Astronomy (PA) are included in both SpringerLink and ScienceDirect. Comparing SpringerLink and ScienceDirect, we find differences in the degree of correlation for these fields. The access count of the latter fluctuated considerably by year compared to that of the former in YNU. The gap between global data and unrepresentative local data might result in these differences.

Furthermore, we examined the correlations of pairs of indices for journal usage, including pairs other than (A) and (B), based on the sampling condition (c). To enable comparison with the results of previous studies and to take into account the strength of raw values, Pearson's product-moment correlation r was also studied along with Spearman's rank correlation ρ. When calculating the product-moment correlations, the data was logarithmically transformed to reduce skewness of distribution. As examples, Tables 4 and 5 show the correlation coefficients for SpringerLink (in 2012). Similar results were also obtained for SpringerLink (in 2011) and ScienceDirect (in 2011 and 2012). An example of these was shown in Table 6. The gray-colored cells in the tables indicate the correlations between the indices for citations and access, and moreover, the cells enclosed in boxes indicate the correlations between the indices relating to the obsolescence of citations and access. Little difference exists between the results of the three types of correlations, i.e., the rank correlation and the product-moment correlations before and after logarithmic transformation.

Table 4. Rank correlation ρ between indices for all 6 fields in 2012 in SpringerLink on the sampling condition (c).

	II	IF	DII	DIF	CHL	DHL	II/IF	DII/DIF
II	1	0.81 *	0.17 *	0.24 *	−0.04	−0.01	0.53 *	0.00
IF		1	0.05	0.20 *	−0.01	0.07	0.01	−0.15
DII			1	0.55 *	0.07	−0.19 *	0.10	0.57 *
DIF				1	0.21 *	0.01	0.05	−0.30 *
CHL					1	0.53 *	−0.03	−0.11
DHL						1	−0.10	−0.20 *
II/IF							1	0.12
DII/DIF								1

*Significant ($p < 0.05$)

Among pairs of the indices relating to obsolescence, while the strongest significant correlation (around 0.5, $p < 0.05$) was observed between CHL and DHL, which are the indices corresponding to (A), only weak correlations were found in the remaining pairs. However, an exception was found for Energy (ScienceDirect in 2011): a strong and positive

correlation was also seen between II/IF and DII/DIF, the indices corresponding to (B), as shown in Table 7.

Table 5. Product-moment correlation *r* after logarithmic transformation between indices for all 6 fields in 2012 in SpringerLink on the sampling condition (c).

	II	IF	DII	DIF	CHL	DHL	II/IF	DII/DIF
II	1	0.82 *	0.09	0.18 *	−0.03	0.05	0.57 *	−0.08
IF		1	0.04	0.19 *	−0.01	0.08	0.00	−0.15
DII			1	0.63 *	0.07	−0.21 *	0.10	0.57 *
DIF				1	0.19 *	0.01	0.03	−0.28 *
CHL					1	0.56 *	−0.04	−0.11
DHL						1	−0.03	−0.27 *
II/IF							1	0.08
DII/DIF								1

*Significant ($p < 0.05$)

Table 6. Rank correlation *ρ* between indices for all 6 fields in 2011 in SpringerLink on the sampling condition (c).

	II	IF	DII	DIF	CHL	DHL	II/IF	DII/DIF
II	1	0.81 *	0.11	0.02	0.00	0.20 *	0.59 *	0.07
IF		1	0.16	0.13	0.08	0.19 *	0.08	0.04
DII			1	0.58 *	−0.04	−0.22 *	−0.09	0.58 *
DIF				1	0.07	−0.14	−0.22 *	−0.27 *
CHL					1	0.54 *	−0.05	−0.08
DHL						1	0.15	−0.12
II/IF							1	0.10
DII/DIF								1

*Significant ($p < 0.05$)

Table 7. Rank correlation *ρ* between indices for Energy in 2011 in ScienceDirect on the sampling condition (c).

	II	IF	DII	DIF	CHL	DHL	II/IF	DII/DIF
II	1	0.86 *	0.73 *	0.62 *	−0.12	−0.30	0.71 *	0.33
IF		1	0.49	0.69 *	−0.30	−0.37	0.36	0.05
DII			1	0.55 *	−0.01	−0.19	0.74 *	0.71 *
DIF				1	−0.06	−0.07	0.29	−0.08
CHL					1	0.77 *	0.23	0.15
DHL						1	0.01	0.02
II/IF							1	0.64 *
DII/DIF								1

*Significant ($p < 0.05$)

Discussion and Conclusions

Results of the analysis indicated that, for 8 fields of SpringerLink and 7 fields of ScienceDirect, statistically significant positive correlations of over 0.4 were observed between CHL and DHL, which are the indices of long-term obsolescence, in both or either year. Furthermore, having reclassified all fields of both collections into 6 fields, comparatively strong and significant correlations were seen between CHL and DHL in natural sciences other

than Medicine, particularly in Physics and in Chemistry and Engineering. This result suggests that, to a certain degree, it is possible to predict the long-term obsolescence of access on the basis of the value of CHL obtained from JCR with regard to natural sciences.

In addition to Spearman's rank correlation coefficients ρ, we also examined the correlations between indices for all fields using Pearson's product-moment correlation coefficients r, and no major differences were observed between both types of correlations. Comparing with previous studies such as Schloegl and Gorraiz (2010; 2011) and Wan et al. (2010), our results indicated the same tendency regarding the indices of long-term obsolescence (CHL and DHL). However, in the case of other indices, a different tendency was observed. Wan et al. (2010), for example, investigated many indices and reported the following correlations between indices: DII and II showing $\rho = 0.24$ ($p = 0.0964$), DII and IF showing $\rho = 0.41$ ($p = 0.0034$), II and IF showing $\rho = 0.59$ ($p < 0.0001$) in agriculture and forestry; DII and II showing $r = 0.8$ in psychology. Meanwhile, in this study, almost no correlations were witnessed between DII and II and between DII and IF in most fields, whereas strong and significant correlations were observed between II and IF ($\rho = 0.81$, $r = 0.82$) as indicated in Tables 4 and 5. This is thought to be partly due to the characteristics of local use along with differences in the fields and databases. For example, citation speed in YNU may be slower than that of global trends, or research areas of researchers in YNU may be specific and narrow, i.e., a large proportion of the journals that they read may not be core journals for their research and thus their research activities (citations) may not correspond to global trends. If one focuses on this issue, the relationship between local access and local citation should be investigated. In addition to this, citation age may also influence the results. Citation age is larger than publication time lag of the citing article, which is mostly around one year. In contrast, downloads (access) tend to be concentrated in the publication year, that is to say, there is little time lag. This might cause different tendencies of downloads and citations in the short-term (e.g., weak correlation between DII and II in Tables 4–6).

Furthermore, the results of 2011 and 2012 for both collections indicate that the degree of correlation in several fields such as Chemistry may vary considerably by year, and the indices with a strong correlation differ depending on the field. Regarding the variation in the indices of short-term obsolescence (II/IF and DII/DIF), we can guess that it would be easily influenced by such factors as the change in the number of papers, the frequency of publication, and special issues of journals. In contrast, regarding the variation in the indices of long-term obsolescence (CHL and DHL), factors such as the transfer to another publisher, title change, and discontinuation of publication may exert influence.

This study focused on the relationship between the obsolescence in local access and global citation for the purpose of grasping the predictability of the former based on the latter. Although one should take into consideration various ways such as cost-effectiveness (e.g., Bergstrom et al., 2014) when introducing journal backfiles efficiently, our approach would also be useful for making a decision.

In future research, aiming to clarify the characteristics themselves of document use by researchers in Japan, we will investigate the citation data in Japanese universities, including YNU, and compare it with the corresponding access data. Moreover, we would like to observe the obsolescence of access and citation for a longer period for further examination of the tendency concerning the variation in the relationship between them.

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research (C) 23500294 (2013) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and we would like to show our gratitude to the support.

References

Bergstrom, T. C., Courant, P. N., McAfee, R. P. & Williams, M. A. (2014). Evaluating big deal journal bundles. *Proceedings of the National Academy of Sciences*, 111(26), 9425–9430.

Bollen, J. & van de Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1), 136-149.

Brody, T., Harnad, S. & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8), 1060-1072.

Chu, H. & Krichel, T. (2007). Downloads vs. citations in economics: Relationships, contributing factors & beyond. In: *Proceedings of the 11th International Society for Scientometrics and Informetrics Conference* (pp. 207-215). Madrid, Spain.

Duy, J., & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *The Journal of Academic Librarianship*, 32(5), 512-517.

Gorraiz, J., Gumpenberger, C., & Schloegl, C. (2013). Differences and similarities in usage versus citation behaviours observed for five subject areas. In: *Proceedings of the 14th International conference of the international Society for Scientometrics and Informetrics (ISSI2013)* (pp. 519-535). Vienna: University of Wien.

Guerrero-Bote, V. P., & Moya-Anegon, F. (2013). Relationship between downloads and citation and the influence of language. In: *Proceedings of the 14th International conference of the international Society for Scientometrics and Informetrics (ISSI 2013)* (pp. 1469-1484). Vienna: University of Wien.

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. S., Martimbeau, N., & Elwell, B. (2005). The bibliometric properties of article readership information. *Journal of the American Society for Information Science and Technology*, 56(2), 111-128.

McDonald, J. D. (2007). Understanding journal usage: A statistical analysis of citation and use. *Journal of the American Society for Information Science and Technology*, 58(1), 39-50.

Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56(10), 1088-1097.

Nicholas, D., Huntington, P., Dobrowolski, T., Rowlands, I., Jamali, M. H. R., & Polydoratou, P. (2005). Revisiting 'obsolescence' and journal article 'decay' through usage data: An analysis of digital journal use by year of publication. *Information Processing and Management,* 41(6), 1441-1461.

Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of Oncology journals. *Scientometrics*, 82(3), 567-580.

Schloegl, C., & Gorraiz, J. (2011). Global usage versus global citation metrics: The case of Pharmacology journals. *Journal of the American Society for Information Science and Technology,* 62(1), 161-170.

Takei, C., Yoshikane, F., & Itsumura, H. (2013). Use of electronic journals in university libraries: an analysis of obsolescence regarding citations and access. In: *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics (ISSI 2013)* (pp. 1772-1783). Vienna: University of Wien.

Wan, J.-K., Hua, P.-H., Rousseau, R., & Sun, X.-K. (2010). The journal download immediacy index (DII): Experiences using a Chinese full-text database. *Scientometrics*, 82(3), 555-566.

Watson, A. B. (2009). Comparing citations and downloads for individual articles. *Journal of Vision*, 9(4), 1-4.

Dynamics between National Assessment Policy and Domestic Academic Journals

Eleonora Dagienė[1] and Ulf Sandström[2]

[2] *eleonora.dagiene@vgtu.lt*
Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 Vilnius (Lithuania)

[2] *ulf.sandstrom@indek.kth.se*
KTH, Indek — Department of Industrial Economics and Management,
Lindstedtsvägen 30, 10044 Stockholm (Sweden)

Introduction

Normally research assessment methodologies assume that the highest scores should be given to articles published in recognised high impact journals. While these high impact journals are mostly published in the US and UK, lower citation rates are particular to journals published in other countries. Subsequent to expansion of the Web of Science in 2007–2009, the research platform was generously augmented with scientific journals issued by local publishers of non-English speaking countries (Leeuwen et al., 2001; van Raan, van Leeuwen, & Visser, 2011). Analysts agree that papers in national journals are usually less frequently cited in comparison to articles published in English (Haiqi & Yamazaki, 1998; Meneghini & Packer, 2007; Moed, 2002; Ponomariov & Toivanen, 2014; Russell, 1998; Tijssen et al., 2006). Research evaluations in several Eastern European countries largely build on data from Thomson Reuters and Elsevier databases. An overview provided by Dejan Pajić (Pajić, 2014) demonstrates that methodologies of most countries award papers in leading international journals rather than national ones. In some countries, articles published in national journals either receive a lower score or are given no score. The Lithuanian methodology is but an illustration of this.

The way a journal reflects the internationalized nature of science may be determined by many methods, one of which is based on the distribution of authoring and citing countries (Zitt & Bassecoulard, 1998).

The aim of the paper is to analyse the impact of the national assessment policy on the development of research journals published in the same country.

Lithuanian Assessment Methodologies and Journal Publishing in Lithuania 2005–2013

Five Lithuanian research assessment methodologies were designed in the period 2005–2010. It should be underlined that there is a great difference between assessment of papers in Sciences and papers in Social Sciences & Humanities. While in Social Sciences and Humanities, researchers have to be published in peer-reviewed journals only, papers in the Sciences have especially high requirements: to gain a score, they have to be published in journals indexed by Web of Science and have an impact factor. The methodology of 2010 was grossly disadvantageous to most Lithuanian journals as it was centred on papers published in high ranking journals (Maskeliūnas, 2011). Lithuanian research journal publishing and other quantitative indicators as well as technical publishing issues have already been analysed in several papers (Dagiene, 2011, 2013). In 2006, Thomson Reuters Web of Science database had only 5 indexed Lithuanian journals; while in 2007, it had 21; and since 2008, there were 29 journals in WoS with Lithuania as the publishing country. One supplementary journal—BALT J OF MANAGEMENT—has been added to this list although its country of origin is England and it is published by Emerald, the Editor-in-Chief and the Managing Editor are from Lithuania.

Data and Methodology

All data analysed in this research has been retrieved from the Web of Science databases: SCIE, SSCI and A&HCI. All indicators employed in this research and listed below have been analysed for two periods: 2008–2010 and 2011–2013. This is done because Lithuanian methodology was changed in 2010, using not only journal impact factors but also JCR data with thresholds measuring the "citation quality" of journals. The main quantitative and qualitative indicators of the Lithuanian journals are presented in the appendix. NJCS – Normalized journal citation score is the impact of the journal set normalized in relation to its sub-fields (average=1.00) (Sandström, 2009).

Citation indicators showed an improvement over the recent years: in 2011–2013, the number of cites by foreign researchers increased by 10% compared to 2008–2010; besides, citation from core journals increased by 19%, which confirms the growing internationalization of Lithuanian journals.

Figure 1 presents dynamics of internationalization indicators of Lithuanian journals.

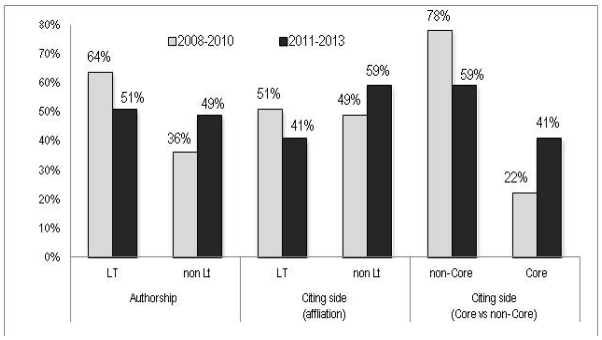


Figure 1. Dynamics of internationalization indicators of Lithuanian journals.

Authorship: from period I to period II, there's an overall drop in LT share and growth of foreign researchers from 36% to 49% if we count averages of all LT journals.

Conclusions

National policy has an influence on scholarly communication and puts the pressure on the national journals. There is some tension but also a response from the journals; thus, over a short period of time we see rather substantial changes.

Firstly, from 2008–2010 to 2011–2013, the relative share of the Lithuanian authors in authorship became smaller; secondly, papers published in Lithuanian journals are more often cited by researchers affiliated to non-Lithuanian institutions; thirdly, papers published in Lithuanian journals are more often cited by papers published in core journals defined as such by Leiden (CWTS 2014).

References

CWTS Leiden Ranking (2014) Retrieved, March 3, 2014, from: http://www.leidenranking.com/methodology/indicators

Dagiene, E. (2011). Changes in Lithuanian research journal publishing in 2009–2010. *Sciecominfo, 7*(1). Retrieved from http://nile.lub.lu.se/ojs/index.php/sciecominfo/article/view/4906

Dagiene, E. (2013). Progressive Opportunities for Research Journal Publishing. *Proceedings of the 5th Belgrade International Open Access Conference 2012: Journal Publishing in Developing, Transition and Emerging Countries* (pp. 11–23). Centre for Evaluation in Education and Science. doi:10.5937/BIOAC-94

Haiqi, Z., & Yamazaki, S. (1998). Citation indicators of Japanese journals. *Journal of the American Society for Information Science, 49*(4), 375–379. doi:10.1002/(SICI)1097-4571(19980401)49:4<375::AID-ASI7>3.0.CO;2-X

Leeuwen, T. N. Van, Moed, H. F., Tijssen, R. J. W., Visser, M. S., & Raan, A. F. J. Van. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics, 51*(1), 335–346. doi:10.1023/A:1010549719484

Meneghini, R., & Packer, A. L. (2007). Is there science beyond English? Initiatives to increase the quality and visibility of non-English publications might help to break down language barriers in scientific communication. EMBO Reports, 8(2), 112–6. doi:10.1038/sj.embor.7400906

Moed, H. F. (2002). Measuring China's research performance using the Science Citation Index. *Scientometrics, 53*(3), 281–296. doi:10.1023/A:1014812810602

Pajić, D. (2014). Globalization of the social sciences in Eastern Europe: genuine breakthrough or a slippery slope of the research evaluation practice? *Scientometrics*. doi:10.1007/s11192-014-1510-5

Ponomariov, B., & Toivanen, H. (2014). Knowledge flows and bases in emerging economy innovation systems: Brazilian research 2005–2009. *Research Policy, 43*(3), 588–596. doi:10.1016/j.respol.2013.09.002

Russell, J. M. (1998). Publishing patterns of Mexican scientists: Differences between national and international papers. *Scientometrics, 41*(1-2), 113–124. doi:10.1007/BF02457972

Sandström, U. (2009). Bibliometric evaluation of research programs. The Swedish Environmental Protection Agency, 81 p.

Tijssen, R. J. W., Mouton, J., van Leeuwen, T. N., & Boshoff, N. (2006). How relevant are local scholarly journals in global science? A case study of South Africa. *Research Evaluation, 15*(3), 163–174. doi:10.3152/147154406781775904

Van Raan, A. F. J., van Leeuwen, T. N., & Visser, M. S. (2011). Severe language effect in university rankings: particularly Germany and France are wronged in citation-based rankings. *Scientometrics, 88*(2), 495–498. doi:10.1007/s11192-011-0382-1

Zitt, M., & Bassecoulard, E. (1998). Internationalization of scientific journals: A measurement based on publication and citation scope. *Scientometrics, 41*(1-2), 255–271. doi:10.1007/BF02457982

Appendix. The main quantitative and qualitative indicators of the Lithuanian journals.

Journal title	Period I – 2008-10 II – 2011-13	THREE MOST FREQUENT COUNTRIES (TOP3) in the authors' affiliations	LT Authorship	TOP3 Authorship	Shift Towards International	NJCS 1=Global avg.
Included in Science Citation Index Expanded (SCI-EXPANDED) – Web of Science Core Collection						
BALT ASTRON	I	LITHUANIA | CZECH REPUBLIC | USA	22.17%	46.95%		0.11
	II	LITHUANIA | ESTONIA | USA	6.95%	34.89%	25.7%	0.07
BALT FOR	I	LITHUANIA | ESTONIA | FINLAND	35.96%	77.34%		0.21
	II	LITHUANIA | ESTONIA | FINLAND	30.54%	62.29%	19.5%	0.19
BALT J ROAD BRIDGE E	I	LITHUANIA | SOUTH KOREA | ITALY	62.95%	77.07%		0.65
	II	LITHUANIA | POLAND | ITALY	45.74%	66.60%	13.6%	0.68
BALTICA	I	LITHUANIA | ESTONIA | LATVIA	36.47%	70.20%		0.29
	II	LITHUANIA | ESTONIA | RUSSIA	74.93%	85.57%	-21.9%	0.12
CHEMIJA	I	LITHUANIA | IRAN | INDIA	94.01%	98.33%		0.14
	II	LITHUANIA | IRAN | BULGARIA	85.94%	91.06%	7.4%	0.08
ELEKTRON ELEKTROTECH	I	LITHUANIA | LATVIA | ROMANIA	61.67%	77.21%		0.25
	II	LITHUANIA | LATVIA | PEOPLES R CHINA	40.10%	58.08%	24.8%	0.21
INFORMATICA-LITHUAN	I	LITHUANIA | SLOVENIA | PEOPLES R CHINA	57.78%	74.81%		1.08
	II	LITHUANIA | PEOPLES R CHINA | TAIWAN	46.00%	62.77%	16.1%	1.04
INF TECHNOL CONTROL	I	LITHUANIA | POLAND | ALGERIA	81.15%	86.89%		0.34
	II	LITHUANIA | TAIWAN | PEOPLES R CHINA	61.51%	88.17%	-1.5%	0.56
J CIV ENG MANAG	I	LITHUANIA | POLAND | TURKEY	43.73%	69.33%		1.28
	II	LITHUANIA | POLAND | TAIWAN	30.03%	54.69%	21.1%	0.71
J ENVIRON ENG LANDSC	I	LITHUANIA | TURKEY | ESTONIA	70.28%	80.47%		0.47
	II	LITHUANIA | TURKEY | INDIA	71.68%	82.57%	-2.6%	0.26
J VIBROENG	I	LITHUANIA | LATVIA | POLAND	66.10%	82.03%		0.11
	II	LITHUANIA | PEOPLES R CHINA | POLAND	28.57%	84.18%	-2.6%	0.41
LITH J PHYS	I	LITHUANIA | UKRAINE | INDIA	88.91%	91.61%		0.12
	II	LITHUANIA | LATVIA | RUSSIA	69.43%	83.55%	8.8%	0.09
LITH MATH J	I	LITHUANIA | GERMANY | HUNGARY	72.27%	83.33%		0.42
	II	LITHUANIA | PEOPLES R CHINA | GERMANY	51.10%	75.64%	9.2%	0.31
MATER SCI-MEDZ	I	LITHUANIA | ESTONIA | CZECH REPUBLIC	83.44%	90.16%		0.18
	II	LITHUANIA | ESTONIA | LATVIA	64.50%	79.20%	12.2%	0.22
MATH MODEL ANAL	I	LATVIA | ESTONIA | LITHUANIA	20.61%	59.02%		0.51
	II	LATVIA | LITHUANIA | PEOPLES R CHINA	18.28%	55.28%	6.3%	0.51
MECHANIKA	I	LITHUANIA | ROMANIA | ALGERIA	71.28%	83.67%		0.51
	II	LITHUANIA | PEOPLES R CHINA | IRAN	48.57%	76.89%	8.1%	0.41
MED LITH	I	LITHUANIA | ESTONIA | USA	92.33%	94.77%		0.11
	II	LITHUANIA | LATVIA | ESTONIA	67.40%	84.24%	11.1%	0.17
NONLINEAR ANAL-MODEL	I	LITHUANIA | INDIA | BANGLADESH	64.86%	82.97%		0.50
	II	LITHUANIA | INDIA | PEOPLES R CHINA	47.62%	75.62%	8.9%	0.61
TRANSPORT-VILNIUS	I	LITHUANIA | PEOPLES R CHINA | TURKEY	56.83%	67.51%		1.19
	II	LITHUANIA | PEOPLES R CHINA | SERBIA	43.10%	65.38%	3.2%	0.56
VET ZOOTECH-LITH	I	LITHUANIA | POLAND | ESTONIA	82.13%	91.88%		0.13
	II	LITHUANIA | POLAND | ESTONIA	69.36%	83.67%	8.9%	0.11
ZEMDIRBYSTE	I	LITHUANIA | ITALY | POLAND	73.74%	86.59%		0.19
	II	LITHUANIA | TURKEY | POLAND	59.79%	80.30%	7.3%	0.35
Included in Social Sciences Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) – Web of Science Core Collection						
BALT J OF MANAGEMENT	I	ESTONIA | LITHUANIA | USA	17.30%	62.89%		0.29
	II	ESTONIA | LITHUANIA | FINLAND	16.34%	67.91%	-8.0%	0.35
FILOS-SOCIOL	I	LITHUANIA | POLAND | NETHERLANDS	88.31%	96.10%		0.41
	II	LITHUANIA | POLAND | LATVIA	90.57%	96.60%	-0.5%	0.41
INT J STRATEG PROP M	I	LITHUANIA | FINLAND | ENGLAND	25.71%	58.57%		0.80
	II	LITHUANIA | PEOPLES R CHINA | ENGLAND	24.27%	59.75%	-2.0%	0.86
INZ EKON	I	LITHUANIA | ESTONIA | POLAND	93.03%	97.23%		0.92
	II	LITHUANIA | CZECH REPUBLIC | SPAIN	65.78%	77.47%	20.3%	0.77
J BALT SCI EDUC	I	TURKEY | USA | SLOVAKIA	3.92%	60.10%		0.09
	II	TURKEY | SLOVENIA | FINLAND	2.25%	74.36%	-23.7%	0.43
J BUS ECON MANAG	I	LITHUANIA | TURKEY | ESTONIA	52.07%	65.70%		1.52
	II	LITHUANIA | TURKEY | SPAIN	20.11%	49.84%	24.1%	0.99
LOGOS-VILNIUS	I	LITHUANIA | FRANCE	99.32%	100%		0.14
	II	LITHUANIA | POLAND | FRANCE	99.44%	100%	0.0%	0.35
PROBLEMOS	I	LITHUANIA | BYELARUS | POLAND	92.64%	96.93%		0.52
	II	LITHUANIA | ESTONIA | USA	82.81%	93.75%	3.3%	n.a.
TECHNOL ECON DEV ECO	I	LITHUANIA | POLAND | LATVIA	64.55%	80.43%		1.81
	II	LITHUANIA | PEOPLES R CHINA | POLAND	37.85%	62.22%	22.6%	2.46
TRANSFORM BUS ECON	I	LITHUANIA | POLAND | ROMANIA	42.41%	76.70%		0.51
	II	LITHUANIA | POLAND | ROMANIA	39.89%	79.45%	-3.6%	0.14

Correlation between Impact Factor and Public Availability of Published Research Data in Information Science & Library Science Journals

Rafael Aleixandre-Benavent[1], Luz Moreno-Solano[2], Antonia Ferrer Sapena[3], Enrique Alfonso Sánchez Pérez[4]

[1] *rafael.aleixandre@uv.es*
INGENIO (CSIC-Universitatd Politècnica de València) & UISYS-Universitat de València, Palacio Cerveró - Pza. Cisneros 4, 46003 Valencia (Spain)

[2] *luz.moreno@cchs.csic.es*
IFS, Centro de Ciencias Humanas y Sociales (CCHS). CSIC. Albasanz 26-28. 28037 Madrid (Spain)

[3] *anfersa@upv.es*
DCADHA. Universitat Politècnica de València. Spain. Camino de Vera s/n, 46022 Valencia (Spain)

[4] *easancpe@mat.upv.es*
Instituto Universitario de Matemática Pura y Aplicada. Universitat Politècnica de València. Camino de Vera s/n, 46022 Valencia (Spain)

Introduction

Scientists continuously generate research data but only a few part of them are published. If these data were accessible and reusable, researchers could examine them and generate new knowledge. Currently, the barriers to data sharing are phased out and public research organizations are demanding ever more insistently that publications resulting from publicly funded projects and data that support them should be published in open (Savage & Vickers, 2009). The purpose of this work is: a) to analyse policies concerning open availability of raw research data in journals in the Information Science & Library Science (ISLS); and b) to determine whether there is a correlation between the impact factor and policies of these journals concerning storage and reuse of scientific data.

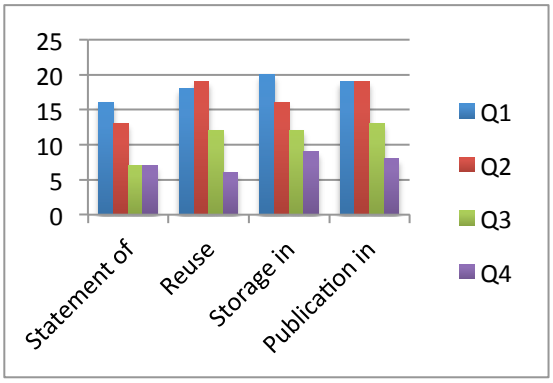
Method

We reviewed the policies related to public availability of papers and data sharing in the 85 journals included in the ISLS category of Journal Citation Reports, 2012 edition. We reported information about the statement of policy regarding: a) complementary material; b) reuse; c) storage in repositories; d) publication on a website; e) journal impact factor; and f) quartile (Q). We have performed a statistical analysis using Chi-square test of the difference regarding each point considered.

Results

The results obtained after analysing the four main variables are presented in Table 1. The variable "Statement of complementary material" was accepted in 50% of the journals. The results were quite similar between the first and second Q and between the third and fourth Q. Regarding the reuse of data, 65% of the journals support this possibility. The highest percentage of response was in the journals of the first Q that accept the reuse of data (86%). The variable "Storage in thematic or institutional repositories", 67% of the journals specified that it was possible. The percentage of journals that accepts storage in institutional repositories decreases by the quartile of journals (e.g., journals in lower quartiles are less supportive). For publication of the manuscript in a website, 69% of the journals accepted it (Figure 1).

Figure 1. Journals supporting each variable by quartile (Q).



Statistical analysis:
Chi-square tests suggest that there is a strong correlation between being a top quartile journal and allowing (a) complementary material (χ^2=11.318, p <.001); (b) reuse of research data (χ^2=19.888, p <.001); (c) storage in thematic and institutional repositories (χ^2=13.080, p <.001); and (d) in personal websites (χ^2=17.350, p <.001).

Conclusions

Our results show that, of the four variables analysed, three have an acceptance rate close to 70% (reuse, publication of the manuscript in a website and storage in thematic or institutional repositories), while the percentage of journals that include the ability to deposit data as supplementary material is lower (50%). These percentages are somewhat higher than those found in a previous study that analysed public availability of published research data in Substance abuse journals (Aleixandre-Benavent et al., 2014). In another study that analysed the same variable in high-impact journals (Alsheikh-Ali et al., 2011), 88% had a statement in their instructions to authors related to public availability and sharing of data, which is 38 percentage points above the average found in the LSIS journals (50%). We found a positive correlation between being a top journal in JCR and having an open policy. A previous paper pointed out that, despite the willingness of some journals to accept supplementary materials, policies, when present, were weak (Borrego & Garcia, 2013). As future research, it would be interesting to raise the question whether journals having high impact factor and open research data is related to the fact that these journals are often owned by rich publishers that are more open for new developments and also have the financial capacities to support such developments.

Acknowledgments

This work has benefited from assistance by the National R+D+I of the Ministry of Economy and Competitiveness of the Spanish Government (CSO2012-39632-C02-01) and Prometeo Program for excellent research groups of Generalitat Valenciana (GVPROMETEO2013-041).

References

Aleixandre-Benavent, R., Vidal-Infer, A., Alonso-Arroyo, A., Valderrama-Zurián, J.C., Bueno-Cañigral, F., & Ferrer-Sapena A. (2014). Public availability of published research data in substance abuse journals. *International Journal of Drug Policy*, *25*, 1143–1146.

Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., & Ioannidis, J.P.A. (2011). Public Availability of Published Research Data in High-Impact Journals. *PLoS ONE*, *6*(9): e24357.

Borrego, A., & Garcia, F. (2013). Provision of supplementary materials in library and information science scholarly journals. *Aslib Proceedings*, *65*(5): 503-514.

Savage, C.J., & Vickers, A.J. (2009). Empirical study of data sharing by authors publishing in PLOS journals. *PLoS ONE*, *4*(9), e7078.

Table 1. Results from main variables analysed in the 85 ISLS journals.

Quartile *	Statement of complementary material			Reuse			Storage in thematic or institutional repositories			Publication in website		
	A	NA	NS	A	NA	NS	A	NA	NS	A	NA	NS
	n (%)	n (%)	N (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
1	16 (76%)	-	5 (24%)	18 (86%)	-	3 (14%)	20 (95%)	-	1 (5%)	19 (90%)	-	2 (%)
2	13 (62%)	-	8 (38%)	19 (90%)	1 (5%)	1 (5%)	16 (76%)	-	5 (24%)	19 (90%)	1 (5%)	1 (5%)
3	7 (33%)	2 (10%)	12 (57%)	12 (57%)	3 (14%)	6 (29%)	12 (57%)	-	9 (43%)	13 (61%)	2 (10)	6 (29%)
4	7 (32%)	2 (9%)	13 (59%)	6 (27%)	1 (5%)	15 (68%)	9 (40 %)	1 (5%)	12 (55%)	8 (36%)	1 (5%)	13 (59%)
Total	43 (50%)	4 (5%)	38 (45%)	55(65%)	5 (6%)	25 (29%)	57 (67%)	1 (1%)	27(32%)	59 (69%)	4 (5%)	22 (26%)
	85			85			85			85		

Quartile on ISLS journals in JCR-2012. A: Accepted. NA: Not Accepted. NS: Not Specified

Use of CrossRef and OAI-PMH to Enrich Bibliographical Databases

Mehmet Ali Abdulhayoglu[1] and Bart Thijs[2]

[1]*Mehmetali.abdulhayoglu@kuleuven.be*
Centre For R&D Monitoring (ECOOM), K.U. Leuven, Waaistraat 6, B-3000 Leuven (Belgium)

[2]*Bart.thijs@kuleuven.be*
Centre For R&D Monitoring (ECOOM), K.U. Leuven, Waaistraat 6, B-3000 Leuven (Belgium)

Introduction

Today prominent and comprehensive databases such as Thomson Reuters' Web of Science (WoS) or Elsevier's Scopus are highly in use for bibliometric research. However, these databases do not index full texts hindering researchers to carry out more detailed analyses. Besides, it is possible that some indexed publications do not have DOI numbers playing an important role to access full texts. This paper focuses on how these above-mentioned deficiencies might be overcome by harnessing the Web sources CrossRef and OAI-PMH. Glenisson, Glänzel, Janssens, & De Moor (2005) and Alexandrov, Gelbukh, & Rosso (2005) stated and showed that full text can have an added value in comparison to abstract and title combination when mapping or clustering disciplines and subfields are in question. Therefore, automatic, rapid and free access to full texts of scientific publications might yield a significant contribution to bibliometric research.

Sources

CrossRef

CrossRef provides, besides its other valuable services, a Text and Data Mining (TDM) service enabling researchers to access full-texts of scientific papers for free (Lammey, 2014). This initiative might be a good alternative when considering the policies of the publishers over TDM hindering or retarding the scientific initiatives (Van Noorden, 2012). In this context, by means of a CrossRef REST API, which is free to be used by the public, the developer can access the metadata that CrossRef assembles from more than 4,400 publishers. Besides the metadata such as title, source (e.g. journal, book chapter etc.) name, co-author names, volume year, volume, issue, subject category, two additional important items might be given. These records are *license* and *links* where *link* gives the related full text link and *license* presents an URL link to the license which must be accepted when a GET request is triggered to access the full text. Figure 1 depicts how to access a full text through CrossRef for a given sample digital object identifier (DOI) and a java GET request. In CrossRef's web site, other methods are given to access full text. Since it is not mentioned in the site, we opt to give a java sample through a snippet.

Figure 1. Process of accessing a full text presented by CrossRef by applying *license* and *link* information.

As of 22/12/14, CrossRef has thousands of publications metadata having both full text and license info from the publishers using creative commons license (CC-BY) which encourages the reuse and distribution of content. These publishers are given in Figure 2.

Publisher	CrossRef Number	CrossRef & WoS Number
HINDAWI PUBLISHING CORPORATION	123552	30737
PENSOFT PUBLISHERS	2233	1712
AIP PUBLISHING	273	5
AMERICAN ASSOCIATION OF PHYSICISTS IN MEDICINE (AAPM)	39	11
AMERICAN VACUUM SOCIETY	4	1
ACOUSTICAL SOCIETY OF AMERICA (ASA)	1	0

Figure 2. Number of publications according to publishers using creative commons license (CC-BY) with full text info within CrossRef and within CrossRef-WoS DOI combination.

On the figure's last column, the number of publications, which appear in both CrossRef and WoS, is given for those WoS records only having a DOI. Even though only a few publishers are willing to allow their contents to be mined, we believe that this number will increase over time as also stated by Van Noorden (2014).

Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)

OAI-PMH emerged aiming at enabling e-print archives to be interoperated (Van de Sompel & Lagoze, 2000). The content of the metadata depends on data provider, for example, while BMC

is providing full texts as well as other metadata, most of data providers such as arXiv do not provide full text or they just mention the URL link not guaranteeing that the full text can be freely downloaded. Below, some example links are given from arXiv and BMC which can be applied to harvest data.

http://www.pubmedcentral.nih.gov/oai/oai.cgi?verb=ListRecords&from=2014-01-01&metadataPrefix=pmc&set=bmcbiology (1)

http://export.arxiv.org/oai2?verb=ListRecords&metadataPrefix=arXiv&set=cs (2)

While former link gives the results only for the journal BMC Biology and those recorded in the repository later than 2014/01/01, later link invokes all the data from computer science discipline in arXiv repository without any date limitation. Note that both results will be invoked in accordance with their own XML schema.

Application

Combining WoS - arXiv - CrossRef

Leveraging arXiv repository, we harvested their OAI-PMH compatible data (See (2)) to combine with our WoS database by matching titles through a character N-Gram text matching process (Abdulhayoglu, Thijs, & Jeuris, 2014). In particular, from arXiv we retrieved title and DOI information for only the *computer science(cs)* discipline to deal with a relatively small data set. There were about 60,000 arXiv records while we have, in WoS, more than 35 million records indexed between 1991 and 2014. We searched for arXiv records within WoS and we found around 18,000 matches having a Salton similarity score higher than 0.90.

Besides 10,000 matches having identical titles, there were more than 7,000 matches having both Salton and Kondrak scores higher than 0.90. Finally, there were only about 200 matches having lower similarity Kondrak scores which can be re-checked manually or simply removed.

We examined the matches having very high similarity scores around 0.90-0.99 and saw that the small character corruptions might appear both on the database or repository side. Additionally, some terms might be given as a text string while it might appear as a symbol in the other source for exp. alpha and α. As a result a similarity score higher than 0.90, especially for Kondrak, can be applied for string matches. So, considering the observations just mentioned, we retained about 6,000 matches having both Salton and Kondrak scores higher than 0.90 and DOI information from the arXiv side.

The retrieved DOI numbers were supposed to be used for accessing full texts through CrossRef. However, a few accessed records have a CC-BY license and we could only grab 286 publications and download their full texts in pdf format. We controlled each full text whether they are correct by checking titles. During this optional process we applied a java pdf parser (*itextpdf*) and correctly extract the title information of those 286 publications. Besides *itextpdf*, CrossRef has its own tool named *pdfextract,* however, it is only applied on Linux environment. Lipinski, Yao, Breitinger, Beel, & Gipp (2013) compare some other extractors.

Conclusions and Discussions

Employing CrossRef and OAI-PMH, a process of accessing full texts of scientific publications indexed in WoS database is explained. Computer science articles from arXiv repository are matched with whole WoS database. Despite a high number of matches, the number of publications appearing within CrossRef repository having creative commons license is quite low. Though a small number of publications has creative commons license, CrossRef seems to ease the issue of accessing full texts freely in time (Van Noorden, 2014).

Acknowledgments

Authors would like to thank Rachael Lammey and Karl Ward from CrossRef, Meshna Koren from Elsevier, Mikail Shaikh from Springer and IT admins from arXiv for their valuable guidance and helps for their TDM systems.

References

Abdulhayoglu, M. A., Thijs, B., & Jeuris, W. (2014). Matching bibliographic data from publication lists with large databases using N-Grams. *Available at SSRN 2464065*.

Alexandrov, M., Gelbukh, A., & Rosso, P. (2005). An approach to clustering abstracts. In *Natural Language Processing and Information Systems* (pp. 275-285). Berlin: Springer.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management*, *41*(6), 1548-1572.

Lammey, R. (2014). CrossRef's Text and Data Mining Services. *Learned Publishing*, 27(4), 245-250.

Lipinski, M., Yao, K., Breitinger, C., Beel, J., & Gipp, B. (2013, July). Evaluation of header metadata extraction approaches and tools for scientific PDF documents. In *Proc. 13th ACM/IEEE-CS Joint Conf. on Digital Libraries* (pp. 385-386). ACM.

Van de Sompel, H., & Lagoze, C. (2000). The Santa Fe convention of the open archives initiative. *D-Lib Magazine*, (2), 2011-10.

Van Noorden, R. (2014). Elsevier opens its papers to text-mining. *Nature*, *506*(7486), 17-17.

Van Noorden, R. (2012). Trouble at the text mine. *Nature*, *483*(7388), 134-135.

Does Scopus Put its Own Journal Selection Criteria into Practice?

Zehra Taşkın[1], Güleda Doğan[1], Sümeyye Akça[1], İpek Şencan[1], and Müge Akbulut[2]

[1] *ztaskin@hacettepe.edu.tr, gduzyol@hacettepe.edu.tr, sumeyyeakca@hacettepe.edu.tr, ipeksencan@hacettepe.edu.tr*
Hacettepe University, Department of Information Management, Ankara (Turkey)

[2] *mugeakbulut@gmail.com*
Yıldırım Beyazıt University, Department of Information Management, Ankara (Turkey)

Introduction

Scopus has been one of the main abstract and citation databases introduced by Elsevier in 2004 to the scientific area. With the multidisciplinarity and international coverage aspects, it is one of the largest databases of peer-reviewed literature in the fields of science, technology, medicine, social sciences, arts, and humanities. There have been several literature studies assessing different aspects of Scopus since the very beginning. The following consists mainly of a description of Scopus, comparing it with the other databases, from the point of usability and accessibility, evaluations regarding the number of citations, and so on. Although there have been many studies about content evaluation and comparisons with other databases, to our knowledge no study has been published focusing on the journal selection criteria of Scopus. The main goal of this study is to evaluate Scopus journals and draw a picture regarding the quality of the journals indexed in Scopus. The two research questions of this study are:

- Do the journals indexed in Scopus match with the Scopus indexing criteria?
- Is there any contribution of the journals that does not fulfil the criteria of Scopus with respect to diversity of authors, institutions and countries as well as internationality of referees, editors and authors?

Methodology

The universe of the study consists of the 2013 Scopus journal list downloaded from SCImago Journal Rank (SJR) on September 18th, 2014. Two groups of countries that have more than 1,000 journals and less than 100 journals in Scopus were left out of the content of this study because of their projected effects on the sample. As a result, 6,151 journals from 23 countries constituting the sample frame were sampled with the systematic sampling method with a rate of 1:30 and 203 journals were chosen for the sample in proportion to 23 countries' journal counts in Scopus.

These 203 journals were evaluated according to the criteria outlined in Table 1, which is mainly based on Scopus journal selection criteria.[1] The contextual criteria were removed because of the requirement to have a comprehensive knowledge of related field. Furthermore, revised Scopus criteria and some new added criteria are marked with grey in Table 1.

Table 1. Criteria selected and used to evaluate Scopus journal content.

Criteria categories	Criteria
Minimum technical criteria (Pre-selection conditions)	Peer-review content and have a publicly available description of the peer review process
	Have an International electronic Standard Serial Number (eISSN) as registered with the ISSN International Centre
	English abstracts and titles
	Regular publication
	References in Roman script
	Publicly available publication ethics and malpractice statement
Journal policy	Editorial policy available
	Type of peer review
	Reviewer list available online
	Diversity in geographical distribution of editors
	Volume of editorial board
	Diversity in geographical distribution of authors
Journal standing	Citedness of journal articles in Scopus
Publishing regularity	No delays or interruption in the publication schedule
Online availability	Full journal content available online
	Journal website available
	English language journal website available
General information about journal	Country of the journal
	Number of issues per year
	First publishing year of the journal
	Journal back issues available on the journal website

Findings and Results

There are only 13 journals providing all of the minimum technical criteria of Scopus. The majority of the journals (190) did not meet at least one criterion. Six journals fulfilled only one criterion of Scopus. Journals and their fulfilment of evaluation criteria are shown in Figure 1. The baseline of the radar graphic (Fig. 1) was created by using "yes"

[1] http://www.elsevier.com/online-tools/scopus/content-overview#content-policy-and-selection

answers to the criteria. We found that 32% of journals did not have an International Electronic Standard Serial Number available (eISSN). Most of the journals (82% and 69% respectively) did not match the criteria of reviewers list being available online and having publicly available publication ethics and malpractice statement. Journals were successful about applying the criteria of available references in Roman script, regular publication and English abstracts and titles.

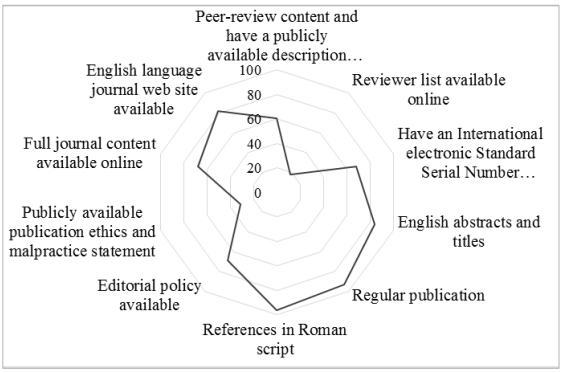


Figure 1. Radar graphic presentation of journals' fulfilment of evaluation criteria.

The evaluation criteria were divided into five classes in this study. These classes are accessibility, peer-review process, policy issues, internationalization and citation levels of journals. The detailed evaluation of each criterion is found in the following sections of this study.

We decided that accessibility on the web, regular publication and references in Roman script consist of the main components of the accessibility criteria in our study. Fifty-one percent of journals in our sample have had all the issues since the launch of their websites and had websites that included full contents of the issues (titles, abstracts, full texts, etc.). Almost all journals had references in Roman script (97%) and most of the journals had English titles/abstracts (84%) and English websites (82%).

The criteria of peer-review process consists of a journal having detailed information about how it is managed and its peer-review board list being available online. We found that 40% of the journals did not have any information on their websites about the peer-review process. Those that did, 73% did not have any information about how their peer-review processes were managed (e.g., double blind, single blind and so on). Only 18% of journals published a list of their reviewers. Under these circumstances, it was hard to determine the diversity of reviewers.

Having accessible publication policies and publicly available publication ethics and malpractice statements were regarded as policy issues. We found that 32% of the journals did not have any editorial policy on their websites. In addition, 68% of the journals did not have any publicly available publication ethics and malpractice statements. Because policy issues were parts of Scopus's minimum criteria, it was expected that journals without these policies would not have passed the preliminary evaluation. However, all these journals have been indexed in Scopus over the years.

The diversity of authors and the editorial board were important for Scopus' evaluation team. We evaluated the diversities as part of this study. Twenty-nine percent of the journals did not have a list of editorial board on their websites. The median for geographic diversity of editors was about 6 within the rest of journals. Eight journals had editors from more than 20 countries. A journal had editors from 53 different countries, while 21% had editors from only one country.

Author diversity is also important for internationalization of journals. We calculated the number of countries by using author affiliations of the last 10 published articles/reviews of each journal. Nine journals did not give any country information for their authors. The median for geographic diversity of authors was 4 within the rest of the journals. Authors were from only one country in 26% of the journals.

Citations are essential for indexed journals within citation databases, as almost all the performance evaluations rely on citations. We evaluated the citation levels of journals by using total cites (three years) indicator of SCImago database. The median number of citations was calculated as 26. Fourteen journals did not have any citations during the three-year period. Six journals had over 1,000 citations.

Conclusions

Citation databases are important for authors, decision-makers, institutions, countries and others. Therefore, it is vital to index high-quality journals for them. Citation databases have strict selection criteria to evaluate journals before indexing to achieve their aims. The criteria of databases are generally based on journal policy, regularity of publication, diversity and so on. We evaluated the journal selection criteria of Scopus and checked the extent of their implementation within this study.

According to the results of our study, the publishers, editors and Scopus should strive to enhance quality. On Scopus' side, Scopus must put the selection criteria into practice strictly and control indexed journals on the aspects of these criteria. Because of the huge competitive environment in the journal market recently, Scopus as well as other publishers of commercial citation databases should consider quality issues more importantly than commercial concerns. A comparative study on journal selection of citation databases may be the continuation of this study.

On the Correction of "Old" Omitted Citations by Bibliometric Databases

Fiorenzo Franceschini[1], Domenico Maisano[2] and Luca Mastrogiacomo[3]

[1] *fiorenzo.franceschini@polito.it,* [2]*domenico.maisano@polito.it,* [3]*luca.mastrogiacomo@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract
Omitted citations – i.e., missing links between a cited paper and the corresponding citing papers – are the main consequence of several bibliometric-database errors. To reduce these errors, databases may undertake two actions: (i) improving the control of the (new) papers to be indexed, i.e., limiting the introduction of "new" dirty data, and (ii) detecting and correcting errors in the papers already indexed by the database, i.e., cleaning "old" dirty data. The latter action is probably more complicated, as it requires the application of suitable error-detection procedures to a huge amount of data. Based on an extensive sample of scientific papers in the Engineering-Manufacturing field, this study focuses on old dirty data in the Scopus and WoS databases. To this purpose, a recent automated algorithm for estimating the omitted-citation rate of databases is applied to the same sample of papers, but in three different-time sessions. A database's ability to clean the old dirty data is evaluated considering the variations in the omitted-citation rate from session to session. The major outcomes of this study are that: (i) both databases slowly correct old omitted citations, and (ii) a small portion of initially corrected citations can surprisingly come off from databases over time.

Conference Topic
Data Accuracy and disambiguation

Introduction
An important branch of the bibliometric literature examines errors in bibliometric databases. Several studies show that the major consequence of database errors is represented by omitted citations, i.e., citations that should be ascribed to a certain (cited) paper but, for some reason, are lost (Moed, 2005; Buchanan, 2006; Jacsó, 2006, Li et al., 2010; Olensky, 2013).

Franceschini et al. (2013) proposed an automated algorithm for estimating the omitted-citation rate of bibliometric databases. This algorithm requires the combined use of two or more bibliometric databases and is based upon the hypothesis that the mismatch between the citations occurring in one database and another one is evidence of possible errors/omissions.

In a further study by Franceschini et al. (2014), this algorithm was applied to a relatively large set of publications, showing that, depending on the bibliometric database in use (Scopus or WoS), omitted citations are not distributed uniformly among publishers; e.g., regarding the publications in the Engineering-Manufacturing field, citations from papers published by Wiley-Blackwell are more likely to be omitted by Scopus, while those from papers published by ASME (American Society of Mechanical Engineers) are more likely to be omitted by WoS. A reason behind this result is that some editorial styles imposed by certain publishers can probably hamper the correct identification of the cited papers by some databases.

The presence of database errors, as well as journal coverage or author disambiguation, is probably one of the major concerns of database administrators. In the authors' opinion, database administrators may undertake two actions for reducing database errors:

1. Limiting the introduction of "new" dirty data in a database, i.e., errors concerning new papers to be indexed;
2. Cleaning "old" dirty data, i.e., errors concerning papers/journals already indexed by a database.

The recent effort by reviewers, publishers and database administrators in checking the cited article lists of new papers probably contributes to reducing "new" dirty data. This hypothesis is corroborated by a recent study by Franceschini et al. (2015), which shows that the databases' propensity to omit newer citations is generally lower than that to omit older citations.

Cleaning up old dirty data is certainly much more complicated because it requires the systematic application of suitable error-detection procedures to a huge amount of data. However, this effort would be essential for improving the quality of a database significantly.

This paper focuses on the ability of the major multidisciplinary bibliometric databases, i.e., Scopus and WoS, to clean up old dirty data. For this evaluation, we use a new procedure, derived from the automated algorithm by Franceschini et al. (2013). This procedure consists in (i) repeating the omitted-citation-rate analysis on the same sample of (cited and citing) articles, but in different-time sessions, and (ii) observing any variation in the results. A database's ability to clean old dirty data will be evaluated considering the variation in the omitted-citation rate from one session to another one.

The remainder of this paper is organized into four sections. The section "Automated algorithm for examining the omitted citations" briefly recalls the algorithm by Franceschini et al. (2013). The section "Methodology" describes the methodology used in our study, focusing on data collection and analysis. The section "Results" illustrates the results of the analysis, investigating similarities and differences between the two databases examined. Finally, the section "Conclusions" summarizes the original contributions of this paper, highlighting the major results, limitations and suggestions for future research.

Automated algorithm for analysing the omitted citations

Before recalling the algorithm, we present an introductory example to illustrate how it works. Let us consider a fictitious paper of interest, indexed by Scopus and WoS. The number of citations received by this paper is four in Scopus and six in WoS (see Table 1).

Table 1. Citation data relating to a fictitious article, according to Scopus and WoS. The union of the citations recorded by the two databases (see the first column) is a total of eight citations. Among the citations, only five come from sources officially covered by both databases (highlighted in grey).

Citation No.	Scopus	WoS
1	✓	
2		✓
3	Omitted	✓
4	✓	✓
5	✓	✓
6	Omitted	✓
7		✓
8	✓	Omitted
Total	4	6

The union of the citations recorded by the two databases is a total of eight citations. Among the citations, only five come from sources (i.e., journals or conference proceedings) officially covered by both databases (highlighted in grey in Table 1). Focusing on these five "theoretically overlapping" (TO) citations, two are omitted by Scopus (but not by WoS) and one is omitted by WoS (but not by Scopus). Therefore, from the perspective of the paper of interest, a rough estimate of the omitted-citation rate is $2/5 \approx 40\%$ in Scopus and $1/5 \approx 10\%$ in WoS. The same reasoning can be extended to multiple papers of interest and more than two bibliometric databases.

The automated algorithm, which is based on the combined use of two bibliometric databases (Scopus and WoS in this case), can be summarised in three steps:

1. Identify a set of (P) papers of interest, indexed by both the databases.
2. For each (i-th) paper of the set, identify the TO citations, defined as the portion of documents issued by journals officially covered by Scopus and WoS. The number of TO citations concerning the i-th paper of interest will be denoted as γ_i.
3. For each (i-th) paper of the set and for each database, determine the number (ω_i) of TO citations that do not occur in it and classify them as omitted citations. The omitted-citation rate (p) relating to the P papers of interest, according to a database, can be estimated as:

$$\hat{p} = \sum_{i=1}^{P} \omega_i / \sum_{i=1}^{P} \gamma_i \,. \qquad (1)$$

We emphasize that p is estimated on the basis of (i) a set of papers of interests and (ii) a portion of the total citations that they obtained (i.e., that ones related to citing articles purportedly covered by both the databases). For a more detailed description of the algorithm, we refer the reader to Franceschini et al. (2013).

The ability of bibliometric databases to clean old dirty data will be evaluated by applying this algorithm to the same sample of TO citations, in three different-time sessions.

Methodology

The study is based on the analysis of the citations obtained from a relatively large sample of papers of interest. The papers were issued by 33 scientific journals (i) included in the ISI Subject Category of Engineering-Manufacturing (by WoS) and (ii) covered by Scopus; Table 2 reports the list of these journals. For each journal, we considered the papers published in the time-window from 2006 to 2012 and the citations that they obtained from papers issued in the same period.

Data collection was repeated in three different-time sessions, spaced about seven months apart: i.e., session I on August 2013, session II on March 2014 and session III on September 2014. We remark that the duration of each data-collection session (i.e., a few days) is negligible with respect to the time period between two consecutive sessions.

To enable comparisons between data collected in different sessions, we adopted two measures:

1. Among the papers of interest (or cited papers) – i.e., those issued by the 33 Engineering-Manufacturing journals – we selected those indexed in each of the three sessions, by both the (Scopus and WoS) databases; in formal terms:

$$A = A^{(I)} \cap A^{(II)} \cap A^{(III)}, \qquad (2)$$

   A being the set of cited papers selected for our analysis and $A^{(I)}$, $A^{(II)}$ and $A^{(III)}$ the sets of papers indexed by both the databases, at the moment of session I, II and III respectively.

   Also, we excluded articles without DOI code or whose DOI code is not indexed by both databases, as they would be difficult to disambiguate.
2. Among the citations, we selected the so-called TO citations, i.e., those obtained from journals purportedly covered by both databases and issued in the 2006-to-2012 time-window. To avoid any misunderstanding, we excluded citations from journals covered in the 2006-to-2012 time-window, but later banned from the database[1]. The official lists of documents covered by the databases in use – which are essential for determining the TO

[1] A possible misunderstanding arises from the fact that, in some cases (mostly on Scopus), the expulsion of a journal from a database entails the entire removal of previously indexed papers, while in other cases (mostly on WoS), previously indexed papers are not necessarily removed.

citations – were retrieved from the databases' websites (Scopus Elsevier, 2015; Thomson Reuters, 2015).

Table 2. List of the Engineering-Manufacturing journals examined. For each journal, it is reported its title and ISSN code. Journals are sorted alphabetically according to their title

Journal title	ISSN
AI EDAM - Artificial Intelligence for Engineering Design Analysis and Manufacturing	0890-0604
Assembly Automation	0144-5154
CIRP Annals - Manufacturing Technology	0007-8506
Composites Part A - Applied Science and Manufacturing	1359-835X
Concurrent Engineering - Research and Applications	1063-293X
Design Studies	0142-694X
Flexible Services and Manufacturing Journal	1936-6582
Human Factors and Ergonomics in Manufacturing & Service Industries	1090-8471
IEEE Transaction on Components Packaging and Manufacturing Technology	2156-3950
IEEE Transactions on Semiconductor Manufacturing	0894-6507
IEEE-ASME Transactions on Mechatronics	1083-4435
International Journal of Advanced Manufacturing Technology	0268-3768
International Journal of Computer Integrated Manufacturing	0951-192X
International Journal of Crashworthiness	1358-8265
International Journal of Machine Tools & Manufacture	0890-6955
International Journal of Production Economics	0925-5273
Journal of Advances Mechanical Design Systems and Manufacturing	1881-3054
Journal of Computing and Information Science in Engineering - Transactions of the ASME	1530-9827
Journal of Intelligent Manufacturing	0956-5515
Journal of Manufacturing Science and Engineering - Transactions of the ASME	1087-1357
Journal of Manufacturing Systems	0278-6125
Journal of Materials Processing Technology	0924-0136
Journal of Scheduling	1094-6136
Machining Science and Technology	1091-0344
Materials and Manufacturing Processes	1042-6914
Proceedings of the Institution of Mechanical Engineers Part B - Journal of Engineering Manufacture	0954-4054
Packaging Technology and Science	0894-3214
Precision Engineering - Journal of the International Societies for Precision Engineering and Nanotechnology	0141-6359
Production and Operations Management	1059-1478
Production Planning & Control	0953-7287
Research in Engineering Design	0934-9839
Robotics and Computer-Integrated Manufacturing	0736-5845
Soldering & Surface Mount Technology	0954-0911

The sample of TO citations used in the analysis is the union of the TO citations (that meet the above requirements), collected in each of the three sessions. In formal terms, this sample of TO citations is:

$$B = B^{(I)} \cup B^{(II)} \cup B^{(III)}, \quad (3)$$

$B^{(I)}$, $B^{(II)}$ and $B^{(III)}$ being the TO citations collected during session I, II and III respectively.
This sample of TO citations will be used for estimating the omitted-citations rate of a certain database, in a certain session; the relationship in Eq. 1 can be used, being:

- $\hat{p}$ the estimate of the omitted-citation rate related to a certain session and a specific database;
- P the number of (cited) articles of interest;
- γ_i the number of TO citations relating to the i-th of the P articles of interest;
- ω_i the portion of the TO citations, collected in a certain session, which are omitted by a specific database.

Being $\hat{p}$ just an estimate of p – albeit the best possible – a relevant symmetrical $(1 - \alpha)$ confidence interval (*CI*) can be constructed as[2]:

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{\sum_{i=1}^{P} \gamma_i}}, \tag{4}$$

with:

α, the type-I error;

$z_{1-\alpha/2}$ the unit normal deviate corresponding to $1 - \alpha/2$.

In this case, we consider a symmetrical 95% *CI*, therefore $\alpha = 5\%$ and $z_{97.5\%} \approx 2$.

By adopting this procedure, we will obtain six different estimates of the omitted-citation rate, i.e., one for each of the three sessions and each of the two databases in use. The comparison of these estimates will tell us whether the databases examined are able to correct old omitted citations.

Results

The total number of papers of interest, i.e., those issued by the Engineering-Manufacturing journals examined, is $P = 23{,}806$. The corresponding TO citations are $\Sigma\gamma_i = 97{,}698$. Table 3 contains the $\hat{p}$ values and the relevant 95% *CI*s, relating to the three sessions and the two databases examined.

Table 3. Main results of the (repeated) analysis of the omitted-citation rate of databases. Citing and cited articles were issued from 2006 to 2012. Statistics concern each of the three sessions (i.e., session I, II and III) for Scopus and WoS respectively.

		(a) Scopus				(b) Wos			
Session	$\sum_{i=1}^{P} \gamma_i$	$\sum_{i=1}^{P} \omega_i$	$\hat{p}$	95% *CI*		$\sum_{i=1}^{P} \omega_i$	$\hat{p}$	95% *CI*	
I (August 2013)	97,698	5,183	5.3%	5.2%	5.4%	7,370	7.5%	7.4%	7.7%
II (March 2014)	97,698	4,607	4.7%	4.6%	4.8%	6,376	6.5%	6.4%	6.7%
III (October 2014)	97,698	4,473	4.6%	4.4%	4.7%	6,404	6.6%	6.4%	6.7%

$P = 97{,}698$ is the total number of (cited) articles, published by 33 Engineering-Manufacturing journals;

$\sum \gamma_i$ is the total number of TO citations (which is independent on the session);

$\sum \omega_i$ is the total number of omitted citations relating to each session and each database;

$\hat{p}$ is the estimate of the omitted-citation rate relating to each session and each database;

The 95% *CI* around $\hat{p}$ is obtained applying the approximated relationship in Eq. 4.

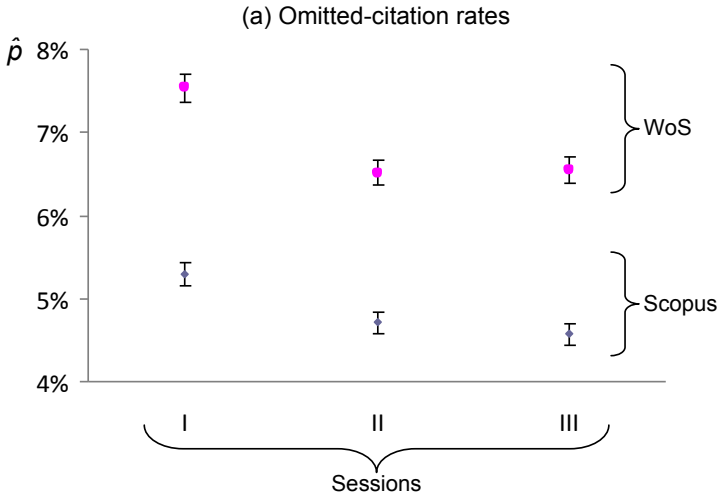
[2] The *CI* construction in Eq. 4 is grounded on the following considerations:

- For a generic sample consisting of $n = \Sigma\gamma_i$ TO citations, the number of omitted citations will be a binomially distributed variable with mean value $n \cdot p$ and variance $n \cdot p \cdot (1 - p)$;
- The aforesaid binomial distribution can be approximated by a normal distribution with the same mean value and variance. This approximation is acceptable in the case $n \cdot p \geq 5$ (Ross, 2009), which is generally satisfied when considering relatively large sets of TO citations.
- Based on the previous approximation, the percentage of omitted citations for a sample of n TO citations will be a normally distributed variable with mean value p and variance $p \cdot (1 - p)/n$. Since p is not known, it can be replaced by its best estimate $\hat{p}$.

In conclusion, Eq. 4 defines a symmetric *CI* around $\hat{p}$, which – with a probability $(1 - \alpha)$ – will include the "true" p value.

The $\hat{p}$ values of both databases tend to decrease over time, denoting that dirty data have been partially cleaned. Interestingly, the major reduction in the $\hat{p}$ values is between the session I and II for both databases; on the other hand, variations between session II and III are not significant, since the 95% *CI*s are partially overlapped (see Figure 1(a)); as regards WoS, we can even notice an imperceptible increase in the $\hat{p}$ value between session II and III.

The overall reduction in the number of omitted TO citations ($\Sigma\omega_i$) for WoS is greater than that for Scopus (i.e., 7,370 – 6,404 = 966 against 5,183 – 4,473 = 710); however, consistently with what observed in other studies (Franceschini et al., 2014; 2015), we note that the omitted-citation rates in Scopus are generally lower than those in WoS. Figure 1(b) shows that the overall percent variations in the $\hat{p}$ values between session I and III are very similar (i.e., -13.7% and -13.1%, for Scopus and WoS respectively).



(b) Relevant percent variations

$$\Delta p = \left(\frac{\hat{p}_{final} - \hat{p}_{initial}}{\hat{p}_{initial}} \right)$$

	Δp		
	I to II	II to III	I to III
Scopus	-11.1%	-2.9%	-13.7%
WoS	-13.5%	0.4%	-13.1%

Figure 1. (a) Graphical representation of the omitted-citation rate in the three sessions, for Scopus and WoS, and (b) relevant percent variations.

Having verified that both databases tend to slowly correct old omitted citations, we now investigate the possible differences in the indexing of individual TO citations, from one session to another one. Table 4 summarizes the eight possible events concerning the correct/missing indexing of individual TO citations. Since there are two possible indexing states (i.e., correct or missing indexing) for each of the three sessions, the total number of possible events is $2^3 = 8$; the file containing the complete list of individual TO citations, with the relevant cited papers, and their session-by-session indexing by the databases, is available under request to authors.

Not surprisingly, the most frequent events are those with no variation (i.e., the type 1 and 2 events in Table 4), in which the TO citations are indexed correctly ("✓") or incorrectly ("×") in all the three sessions; the portion of TO citations with no variation is 98.7% for Scopus and 98.5% for WoS). The type 3 and 4 events represent corrections in the TO-citation indexing, in session II and III respectively. The total number of corrections in WoS is basically larger to that in Scopus, probably due to the larger level of "initial dirt" in the former database, compared to that one in the latter. Moreover, we note that almost all of the corrections by WoS are concentrated in session II (i.e., 1193 out of 1215).

Despite these differences, the percentage of TO citations corrected by Scopus and WoS are pretty close to each other (i.e., roughly 1% and 1.2% respectively). This similarity is even more interesting if we consider the fact that, among the set of corrected TO citations, a relatively small subset is shared between the two databases (i.e., 392 citations out of (997 + 1,215 – 392) = 1,820, corresponding to about 21.5% of the set of corrected TO citations).

Table 4. Overall statistics concerning the indexing of the individual TO citations, in each session. Symbols "✓" and "×" respectively identify the TO citations correctly indexed or omitted in a certain session.

Type of event		Session			(a) Scopus				(b) Wos			
					Single event		Aggregated events		Single event		Aggregated events	
		I	II	III	TO citations	Percent	TO citations	Percent	TO citations	Percent	TO citations	Percent
No variation	1	✓	✓	✓	92,296	94.5%	96,411	98.7%	90,195	92.3%	96,214	98.5%
	2	×	×	×	4,115	4.2%			6,019	6.2%		
Correction	3	×	✓	✓	765	0.8%	997	1.0%	1,193	1.2%	1,215	1.2%
	4	×	×	✓	232	0.2%			22	0.0%		
Anomalous variation	5	✓	×	×	102	0.1%	290	0.3%	164	0.2%	269	0.3%
	6	✓	✓	×	112	0.1%			77	0.1%		
	7	×	✓	×	0	0.0%			0	0.0%		
	8	✓	×	✓	76	0.1%			28	0.0%		
				Total	97,698	100%	97,698	100%	97,698	100%	97,698	100%

The type 5 to 8 events are characterized by anomalous variations, in which some TO citations, which are correctly indexed in a certain session, are omitted in one (or more) subsequent sessions. It is surprising how citations, which were initially indexed correctly, can come off from a database over time; in other words, these events represent a form of generation of dirty data, which is independent of the introduction of new data in the database. Fortunately, the incidence of these abnormalities is rather low (coincidentally, about 0.3% for both Scopus and for WoS); in the future, we may conduct a thorough analysis of these anomalies, based on their manual examination.

Conclusions

The analysis presented in this paper shows that the two bibliometric database examined tend to gradually reduce the number of old omitted citations, although this reduction is relatively slow for both. It would be interesting to see to what extent these cleanings were due to error-correction campaigns structured by database administrators, or simply due to impromptu database-inaccuracy reports by authors and/or database users (even checking and cleaning up bibliometric data in personal research profiles, such as ResearcherID, Scopus Author ID, ORCID, etc.).

Results of this study show other interesting similarities/coincidences between the two databases examined:

1. Comparing the results related to session I and III (spaced about fourteen months apart), we noticed a 13-to-14% reduction in the *p* values for both Scopus and WoS.
2. For both databases, the greatest reduction in the omitted-citations rate was registered in session II and not in session III. This could be just a coincidence or it could denote a sort of "seasonality" of the two databases in cleaning up old dirty data.
3. The portion of TO citations whose indexing varies in the three sessions is roughly the same for both databases, i.e., roughly 1 to 1.5%. Apart from the previously omitted TO citations that have been justly corrected, they include a small portion of abnormal variations, i.e., TO citations correctly indexed in some session and subsequently omitted. Coincidentally, the percentage of abnormal variations is 0.3% for both databases.

The proposed analysis has several limitations. Even though the set of TO citations includes almost one-hundred thousand citations, the relevant cited papers are all confined within the Engineering-Manufacturing field. Also, the analysis was repeated in three sessions over a

total period of about 14 months; therefore, it reflects a database's ability to correct errors in short/middle-term period, but not in the long-term period.

In the future, we plan to extend the study to a longer time-scale (e.g., 2 or 3 years) and/or to scientific articles in other disciplines. Furthermore, the study will be expanded for investigating possible links between the omitted citations' propensity to be corrected and the publishers of the relevant citing papers.

References

Buchanan, R.A. (2006). Accuracy of Cited References: The Role of Citation Databases. *College & Research Libraries*, 67(4), 292-303.

Franceschini, F., Maisano & D., Mastrogiacomo, L. (2013). A novel approach for estimating the omitted-citation rate of bibliometric databases. *Journal of the American Society for Information Science and Technology*, 64(10), 2149-2156.

Franceschini, F., Maisano, & D., Mastrogiacomo, L. (2014). Scientific journal publishers and omitted citations in bibliometric databases: Any relationship? *Journal of Informetrics*, 8(3), 751-765.

Franceschini, F., Maisano, & D., Mastrogiacomo, L. (2015). Influence of omitted citations on the bibliometric statistics of the major Manufacturing journals. To appear in *Scientometrics*. A draft version is available at http://staff.polito.it/fiorenzo.franceschini/Pubblicazioni/Revised_IJPE-D-13-01272.pdf.

Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3), 297-309.

Li, J., Burnham, J.F., Lemley, T., & Britton, R.M. (2010). Citation analysis: comparison of Web of Science, Scopus, Scifinder, and Google Scholar. *Journal of Electronic Resources in Medical Libraries* 7(3), 196-217.

Moed, H.F. (2006). *Citation analysis in research evaluation* (Vol. 9). Springer.

Olensky, M. (2013). Accuracy Assessment for Bibliographic Data. *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, vol. 2, pp. 1850-1851, Vienna, Austria.

Ross, S.M. (2009). *Introduction to probability and statistics for engineers and scientists*. Academic Press.

Schenker, N., & Gentleman, J.F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3), 182-186.

Scopus Elsevier (2015). *Scopus Content Coverage*. Available at http://www.scopus.com [retrieved on August 2013, March 2014 and October 2014].

Thomson Reuters (2015). *Master Journal List*, http://ip-science.thomsonreuters.com/mjl/ [retrieved on August 2013, March 2014 and October 2014].

Can We Track the Geography of Surnames Based on Bibliographic Data?

Nicolas Robinson-Garcia[1], Ed Noyons [2] and Rodrigo Costas [2]

[1] *elrobin@ugr.es*
EC3Metrics spin-off and EC3 Research Group, Universidad de Granada, Granada (Spain)

[2]*noyons@cwts.leidenuniv.nl* [3] *rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (The Netherlands)

Abstract
In this paper we explore the possibility of using bibliographic databases for tracking the geographic origin of surnames. Surnames are used as a proxy to determine the ethnic, genetic or geographic origin of individuals in many fields such as Genetics or Demography; however they could also be used for bibliometric purposes such as the analysis of scientific migration flows. Here we present two relevant methodologies for determining the most probable country to which a surname could be assigned. The first methodology assigns surnames based on the most common country that can be assigned to a surname and the Kullback-Leibler divergence measure. The second method uses the Gini Index to evaluate the assignment of surnames to countries. We test both methodologies with control groups and conclude that, despite needing further analysis on its validity; these methodologies already show promising results.

Conference Topic
Data Accuracy and disambiguation

Introduction
Tracking the geographical origin of individuals has multiple applications and is of interest to many fields. For instance, in biomedical research it is used for racial and ethnic classification as this information is useful for identifying risk factors in epidemiological and clinical research (Burchard et al., 2003). It is also of interest in the field of Demography to analyse migration movements (e.g. Chen & Cavalli-Sforza, 1983) or migratory influences in a given country (Hatton & Wheatley Price, 1999). In the field of bibliometrics, scientific migration flows between countries has been a subject of study as they are considered beneficial for the exchange of new ideas and scientific knowledge between countries (Moed & Halevi, 2014) as well as to analyse case studies to identify the spread of researchers of a given nationality around the world (Costas & Noyons, 2013).

Surnames have been used as a proxy of geographic, ethnic and even genetic origin for some time now. According to Kissin (2011) "the use of surnames in human population biology dates back to 1875, when George Darwin used frequency of occurrences of the same surname in married couples to study in-breeding". Geographic information related to surnames may also be of use in the field of bibliometrics, especially with regard to collaboration and mobility studies. So far only few papers have been found using surname data for bibliometric purposes. Kissin and colleagues (Kissin & Bradley, 2013; Kissin, 2011) have performed several studies focused on the analysis of Jewish surnames in the database MEDLINE. Also Freeman and Huan (2014) recently analysed the effect of diversity of authorship in the impact of scientific publications.

Until recently, these studies relied on manually curated lists of surnames related to ethnic groups, languages or countries. In the last few years, surname research has been developed and many methodologies have been proposed to discern statistical approaches to geographically classify surnames (a good review on the subject can be found in Cheshire, 2014). In this regard, two types of approaches can be found: 1) probability and Bayesian

methods and 2) clustering techniques. For this, we can focus either on the concentration of surnames by areas or on tracking surnames to their original region (Cheshire, 2014).

So far the results reported are quite satisfactory (Mateos, 2007). While regional studies with large data sets offer relatively accurate results due to the skewness of the surnames distribution (Cheshire, 2014), there are still problems when applying these methodologies at a global level. Such limitations are due to migratory movements and data restrictions. For instance, the surname 'Lee' is considered in many studies as British. However, it is most common in the United States and at the same time in Asia. Also data availability may be an issue as most of it comes from census data and demography studies which usually come from different sources and present differences between them.

In this paper we suggest the use of a single data source to develop a methodology to track the geography of surnames worldwide. We propose using the authors' affiliation data from a scientific bibliographic database. For this purpose we analyse two different useful methodologies: one based on the application of information theoretic measures, and a second one based on the use of inequality indexes.

This paper is structured as follows. First we describe the data collection and processing. Then we describe each of the two methodologies proposed for assigning countries to names: one based on the Kullback-Leibler divergence (Kullback & Leibler, 1951) and a second one using the Gini Index, usually used in the field of Economics. In order to test the validity of each methodology, we compared our results with those from a list of surnames based on language origin for 11 different languages. Finally we conclude discussing the limitations of our methodologies, further developments and the potential use of this type of studies for the field of bibliometrics.

Data collection and processing

The goal of this paper is to develop a methodology to assign surnames to countries based on the bibliographic data offered by authors from a scientific database. For this we used the in-house CWTS version of the Web of Science database (not including the Conference Proceedings Citation Index or the Book Citation Index). This database covers all publications and authors for the 1980-2013 time period. The next step needed was to identify authors and relate them with their country of origin. Such approach assumes certain limitations:

- *Reliance on a single data source*. This means that errors or misrepresentations by countries derived from the Web of Science database will reflect on the quality of the result findings reported. Also, the surname information is restricted to the time period employed in the analysis, meaning that migration flows which have taken place before 1980 are not considered. This means that the origin of the surname is tracked according to a fixed image.
- *Limitations in the data*. We are working with a bibliographic database, implying that scholarly related patterns (e.g. migrations of scholars, mobility programs, issues related on how scholars use their name in publications, etc.) as well as database-coverage related problems (e.g. orientation of the database towards Anglo-Saxon countries, the lack of coverage of surnames that have never published, etc.) can play a role. Also, possible mistakes from the database (e.g., wrong linkage of authors to addresses, typos, transcription problems, lack of information, etc.) should be taken into account when interpreting the results.

In Figure 1 we offer an overview of the methodology followed. For all the surnames in all the publications covered in the Web of Science we detected all the 'trusted' linkages between authors and countries. By a trusted linkage we mean a surname-country relationship that is

unambiguously registered in a publication[1] based on linkages between authors and countries according to bibliographic data. This implies that only in those cases where there is strong evidence that an author is linked to a country, the link is created and the combination (surname-country) is taken into consideration for the statistical analysis. These trusted linkages were created based on the following author-country combinations:

- Authors and countries from the *reprint address* field in the Web of Science are directly linked to their affiliation (Costas & Iribarren-Maestro, 2007).
- *Registered combinations of author and affiliations* recorded in the Web of Science, as from 2008 onwards WoS registers the linkage between authors and countries as they appear in the publications.
- *First authors* are assigned to the *first address* in the publication. As Calero and colleagues (2006) show the linkage of the first author with the first address of the publication is quite reliable.
- *One country publications*. For all publications with only one address or only national collaboration all their authors can be assigned to this country.

As a result, a matrix distribution of surnames by countries was created. Based on this matrix, two approaches were considered to assign surnames to countries. The first one consisted on assigning surnames to the countries with the highest frequency (in terms of publications containing the surname-country trusted linkage) which complied certain levels of assurance. This level of assurance was obtained by means of the Kullback-Liebler divergence or information gain measure. The second approach was to assign surnames according to their relative concentration by countries. This was done by using the Gini Index. In the next two subsections we detail each of the two methods proposed and the results obtained for each of them.

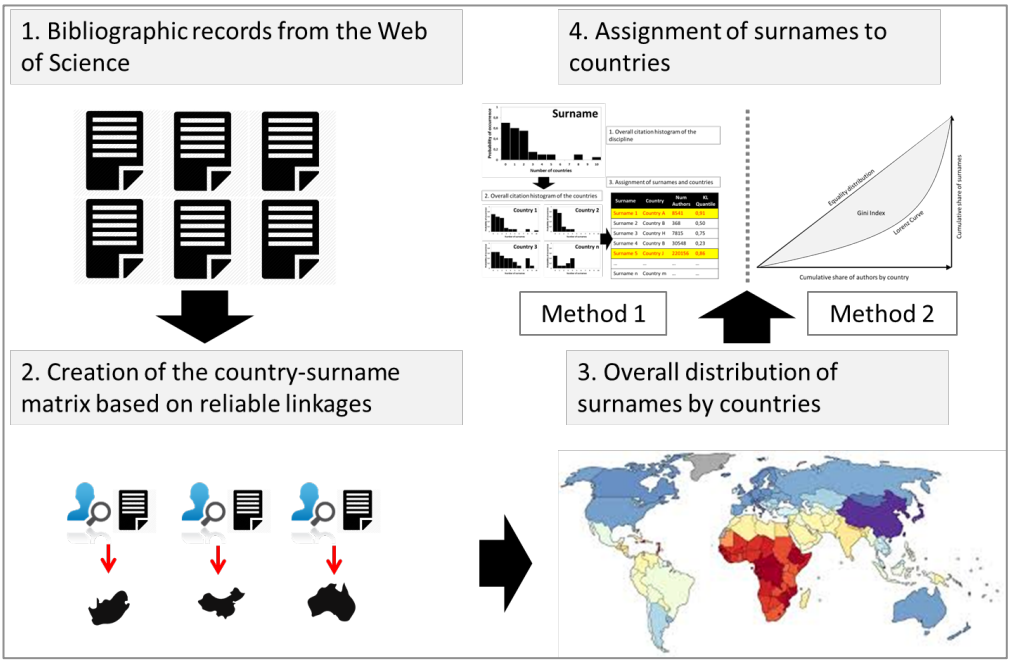


Figure 1. Overview of the methodology followed for assigning countries to surnames.

[1] For many publications in the Web of Science, not all the authors are directly linked to their affiliations in the paper, therefore sometimes it is very difficult to establish to which affiliation (and country) belongs every author.

Method 1: Kullback-Leibler divergence and distribution by country

When identifying the geographic origin of a surname one plausible approach is to consider that a surname will belong to the country with the largest number of occurrences. However, this assumption entails two problems that have to be solved. Firstly, while using raw data will benefit countries with a large presence in the database (e.g. Western and Anglo-Saxon countries), relative indicators will benefit smaller countries, preventing from a balance between countries. Secondly, some surnames may show similar numbers in various countries.

In order to overcome such limitations, we need a reasonable method to characterize the belonging of surnames to each country; and secondly, we have to be able to measure what is the amount of relative information between such characterizations. Here we propose the use of the information gain or Kullback-Leibler divergence measure (Kullback-Leibler, 1951). This measure allows us to select the country that contributes with more information to a given surname. It compares two distributions: a true probability distribution $p(x)$ and an arbitrary probability distribution $q(x)$, and indicates the difference between the probability of X if $q(x)$ is followed, and the probability of X if $p(x)$ is followed. Although it is sometimes used as a distance metric, information gain is not a true metric since it is not symmetric and does not satisfy the triangle inequality (making it a semi-quasimetric) (García et al., 2013).

In this paper, the true probability distribution $p(x)$ is represented by the authors' distribution of a given surname in the country with the highest number of such surname, while the arbitrary probability distribution $q(x)$ is represented by the frequency distribution of the surname in the rest of the countries. The objective is, on the one hand, to characterize the information gain between two probability distributions with a minimal number of properties, which are natural and thus desirable. Second, it aims to determine the form of all error functions satisfying these properties, which we have stated to be desirable for predicting surname-country dissimilarity. This analysis allows identifying similar and dissimilar distributions from a given one, but it does not explain the reasons for such dissimilarity. Such an approach has been previously used in the field of bibliometrics for very different purposes. For instance, Waltman and van Eck (2013) use it to identify national journals from international journals. García and colleagues (2013) use the Kullblack-Leibler divergence measure to determine similar academic institutions (García, et al., 2013). Finally, Torres-Salinas and colleagues (2013) apply it to characterize the field-specialization of publishers based on the citation patterns of book chapters (Torres-Salinas et al., 2013). In Figure 2 we summarize the main steps followed for assigning countries to surnames.

If we predict the similarity between the given surname and the country based on their information gain, then we can set a minimum value of information gain that should be reached in order to ensure that the assignment made is correct, thus relating the surname with the country that leads to the most alike assignment to the frequency distribution. In this case we have established a minimum value up to the percentile 0.8[2] of the overall distribution of surnames and main country by the Kullback-Leibler divergence measure in order to determine a good assurance in the surname-country association.

[2] In other words, we consider that up to 80% of the surname-country linkages based on the highest KL divergence measures are informative, and we disregard 20% of the combinations in which the surname and the country cannot be considered as a reliable linkage (as the surname could also reasonably belong to another country, based on the overall distribution of the surname across countries).

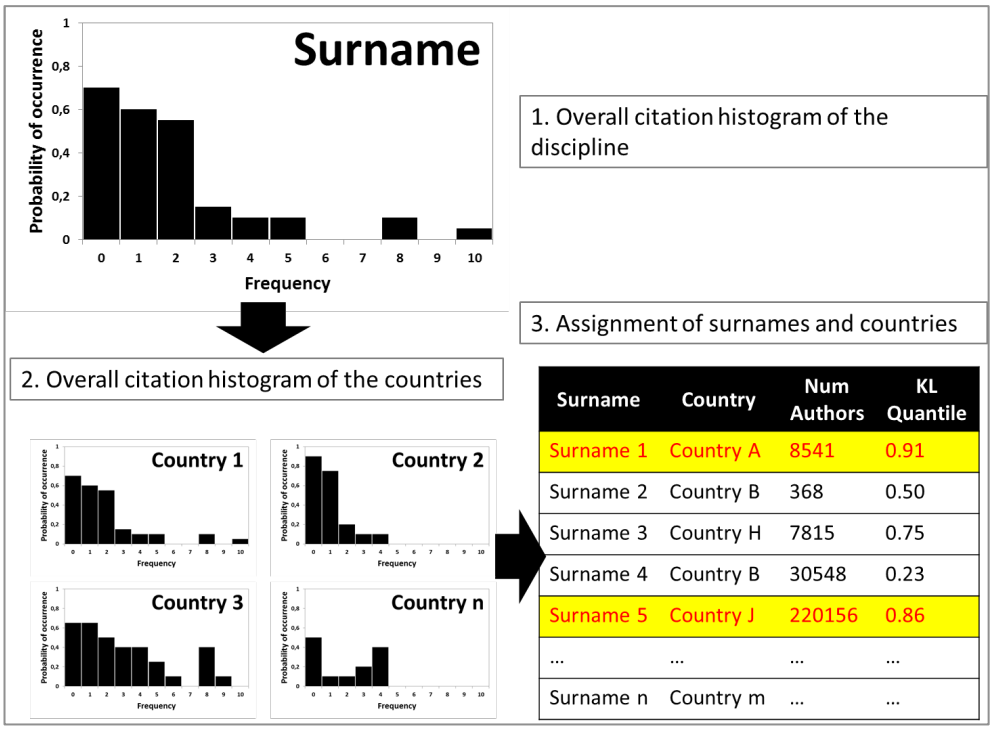


Figure 2. Overview of Method 1 employing the Kullback-Leibler divergence measure.

Table 1. Distribution of top 36 countries with the highest number of surnames according to Method 1. Kullback-Leibler Divergence.

Country	Surnames	Country	Surnames	Country	Surnames
FRANCE	138349	MEXICO	38367	FINLAND	15160
GERMANY	112445	BRAZIL	37198	UKRAINE	14582
RUSSIA	111716	GREECE	34917	CZECH REPUBLIC	14427
SPAIN	83529	IRAN	34235	NORWAY	12892
USA	76219	THAILAND	32426	DENMARK	12861
ITALY	69637	TURKEY	27671	ARGENTINA	11714
ENGLAND	63885	SWEDEN	26134	HUNGARY	10541
JAPAN	56345	ISRAEL	24482	PEOPLES R CHINA	10472
CANADA	49775	AUSTRALIA	24259	ROMANIA	9976
NETHERLANDS	41306	BELGIUM	22203	SOUTH AFRICA	9504
INDIA	41198	SWITZERLAND	21402	NIGERIA	9313
POLAND	40446	AUSTRIA	18048	EGYPT	8682

Results

A total of 1,568,052 surnames were assigned to 119 different countries. Table 1 shows the distribution by surnames of the 36 countries with the higher number of surnames assigned. As observed, the largest number of surnames is assigned to France (8.8%), followed by Germany (8.0%), Russia (7.1%) and Spain (4.9%).

As observed, some countries with the same language appear in this list, such as England and United States for English language or Spain and Mexico for Spanish language. Also some manual normalization of countries was required due to changes in the name of countries (i.e., USSR and Russia or Germany and Federal Republic of Germany).

Method 2: Gini inequality index and concentration by country

Another plausible approach to assigning countries to surnames is to consider the right country as the one where a given surname is more concentrated. For this, we suggest the use of inequality indexes such as the Gini Index. This indicator has already been used in the field of bibliometrics. For example, Torres-Salinas and colleagues (2014) employ it to determine the level of specialization of academic publishers indexed in the Book Citation Index. It is a measure of statistical dispersion. It is defined based on the Lorenz Curve, which plots the proportion of population (y axis, surnames in our case) that is cumulatively concentrated by the bottom *x*% of the population. In Figure 3 we represent its interpretation. The equality distribution is represented by a 45 degrees line. The Gini Index is defined as the ratio of the area that lies between the line of equality and the Lorenz Curve. Its value ranges between 0 and 1, 0 meaning total equality (or dispersion) and 1, total inequality (or concentration). The hypothesis we pose is that a surname can be assigned with certain levels of reliability to the country which shows a higher concentration of such surname, hence relativizing the presence of a given country in the database.

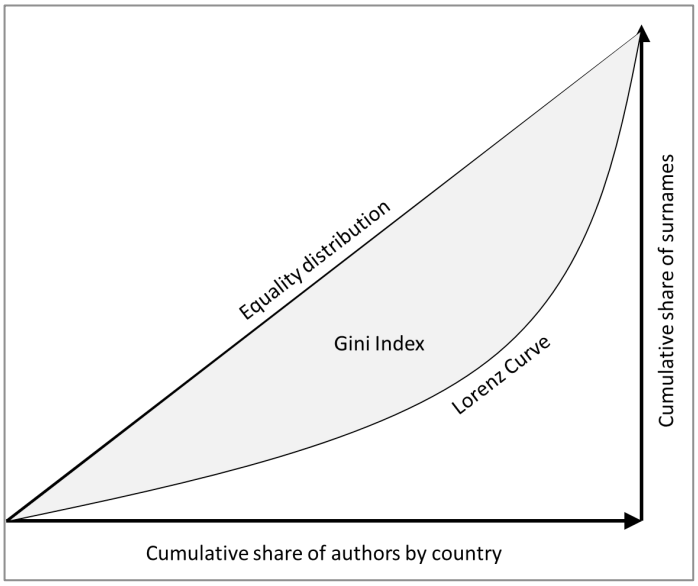


Figure 3. Interpretation of the Gini Index.

Table 2. Distribution of top 36 countries with the highest number of surnames according to Method 2. Gini Index

Country	Surnames	Country	Surnames	Country	Surnames
USA	310739	NETHERLANDS	40528	UKRAINE	17580
FRANCE	117938	BRAZIL	38386	ARGENTINA	16275
GERMANY	111375	GREECE	38034	FINLAND	16060
RUSSIA	94369	IRAN	37162	CZECH REPUBLIC	15166
SPAIN	77387	THAILAND	35090	NORWAY	15074
ITALY	65699	TURKEY	28473	DENMARK	14347
JAPAN	52399	ISRAEL	28360	HUNGARY	12291
ENGLAND	47521	SWEDEN	26051	ROMANIA	11767
CANADA	46146	SWITZERLAND	25029	SOUTH AFRICA	11018
POLAND	44087	BELGIUM	23863	NIGERIA	10619
INDIA	42897	AUSTRALIA	23396	CHINA	9531
MEXICO	41066	AUSTRIA	21609	EGYPT	9158

In Table 2 we show the distribution of surnames by countries for the top 36 countries with the highest number of surnames. A total of 1,885,782 surnames were matched to a list of 343 countries. The country with the largest number of surnames assigned is the United States, representing 16.5% of the total share, and followed by France (6.25%) and Germany (5.9%). In general terms we observe that this methodology distributes surnames among a larger number of countries, showing a less skewed distribution.

Validation

In order to validate the results of each method and determine their performance, we tried to compare them with a 'valid' list of surnames by countries. However, identifying such a list entails certain limitations. First, there is no 'perfect' and unique linkage between countries and surnames. Secondly, these linkages are not usually done for countries but rather for languages, cultures, ethnicities, etc. We decided to use a list of surnames by language provided from Wikipedia[3] and select a sample of languages.

Table 3. Control table of correspondences between countries and languages.

Normalized country	Languages	Countries
Denmark	Danish	Denmark; Greenland
England	Celtic; Anglo-Cornish; English; Scottish; Irish	Antigua & Barbuda; Australia; Bahamas; Barbados; Belize; Bermuda; Canada, England, Ghana; Gibraltar; Grenade; Guyana; Ireland; Jamaica; Liberia; Malawi; Mauritius; Micronesia; N Wales; Namibia, New Zealand, Nigeria; Scotland; Sierra Leone; Solomon Islands; South Africa, St. Kitts & Nevis; St. Lucia; St. Vincent; Trinidad & Tobago; USA; Wales; Zambia
Finland	Finnish	Finland
France	Breton; French	Benin; Burkina Faso; Congo; Côte Ivoire; Polynesia; France; French Guayana; Gabon; Guadeloupe; Guinea; Haiti; Ivory Coast; Mali; Martinique; Monaco; New Caledonia; Niger; Reunion; Senegal; Togo; Upper Volta
Germany	German	Austria; Germany; Liechtenstein
Greece	Greek	Greece
Iceland	Icelandic	Iceland
Italy	Italian	Italy; San Marino; Vatican
Japan	Japanese	Japan
Netherlands	Afrikaans; Dutch	Holland; Netherlands; Surinam
Portugal	Portuguese	Angola; Brazil; Cape Verde; Guinea Bissau; Mozambique; Portugal
Spain	Basque; Catalan; Galician;	Andorra; Argentina; Bolivia; Chile; Colombia; Costa Rica; Cuba; Dominican Republic; Ecuador; El Salvador; Guatemala; Honduras; Mexico; Nicaragua; Panama; Paraguay; Peru; Spain; Uruguay; Venezuela

We chose 20 different languages grouped in what we called 12 'normalized' countries, that is, the most representative countries of these 20 languages. Then we crossed our sample table with the surnames obtained from Web of Science and identified the countries to which each of the two methods proposed assigned these surnames. The list of countries was then processed in order to identify the 20 languages selected. We assigned to each retrieved country one of the selected language if one of the following premises was given (Table 3):

1. It was the official language of the country. For instance, French is the official language of countries such as Gabon, Haiti or Martinique.

[3] http://en.wikipedia.org/wiki/Category:Surnames_by_language

2. It is not the main language but it is only spoken in a given area. For instance, Galician, Basque and Catalan surnames were assigned to Spain, or Breton to France.
3. There is more than one official language (which is also used in other countries). This is the most important limitation noted from our validation method, as it excludes countries such as Switzerland, Belgium or Luxembourg (which have several languages spoken in more than one country). The only exception noted is Canada, which has been attributed to English language, acknowledging the important limitation towards French language.

Our validation list from Wikipedia contains a total of 8,239 surnames. After crossing this list with our list of surnames retrieved from Method 1, a total of 7,625 surnames were matched. In Table 4 we include the distribution of surnames by normalized countries according to our control list (Table 3), the coverage of 'valid' assignments made, that is, those surnames which could be assigned with certain levels of assurance according to their information gain; and the share of correct assignments.

Table 4. Distribution of surnames by countries of the control sample for 12 normalized countries according to their language, valid assignments and correct assignments according to the two methods proposed.

	METHOD 1*			METHOD 2**		
Countries	**Surnames**	**% coverage**	**% correct**	**Surnames**	**% coverage**	**% correct**
DENMARK	123	91.06%	68.75%	123	100%	60.16%
ENGLAND	932	28.76%	80.97%	929	100%	58.56%
FINLAND	225	99.11%	94.62%	224	100%	91.96%
FRANCE	562	88.08%	68.28%	560	100%	50.54%
GERMANY	2186	52.24%	69.00%	2170	100%	43.78%
GREECE	170	84.12%	78.32%	168	100%	78.57%
ICELAND	29	100.00%	65.52%	28	100%	100.00%
ITALY	972	87.65%	86.97%	968	100%	64.77%
JAPAN	1349	98.74%	98.95%	1347	100%	91.39%
NETHERLANDS	471	88.11%	60.96%	468	100%	41.67%
PORTUGAL	137	98.54%	92.59%	136	100%	91.91%
SPAIN	469	93.18%	48.74%	464	100%	54.74%
Total	**7625**	**73.22%**	**79.03%**	**7585**	**100%**	**61.29%**

*** Method 1:** Kullback-Leibler divergence; **** Method 2:** Gini Index

As observed, in general terms the coverage of 'reliable' assignments made was of 73.2% of the sample list. However, significant differences can be found by country. While in the case of Iceland all surnames were assigned with certain levels of assurance (>80 quartile of the Kullback-Leibler divergence distribution), in the case of England only 28.8% of the surnames were considered valid. Also the coverage figures are quite low for Germany (52.2%). From these surnames covered, around 80% of them were assigned to the correct country. The highest figures of correct assignments are observed for Japan (98.9%, also with a coverage of 98.5%), while the lowest figures go to Spanish surnames (48.7% of correct assignments with a coverage of 93.2%). In the case of England, although the coverage is low, 80.1% of the assignments were correct. In the case of Germany the share is lower (69%).

Regarding the methodology based on the Gini Index, a total of 7585 surnames were retrieved after crossing the list of surnames obtained with the control list. As observed, the coverage of 'reliable' assignments with this methodology is much higher (100%), however, many differences are observed on the share of correct assignments. In general terms this

methodology performs not as well as the first one, with 61.2% of all assignment correct. However, in some cases its share of correct assignments is higher. This is the case of Iceland where the 29 surnames of the control list were correctly assigned. Also the share of correct assignment for Spain increases (54.7%) but still shows low values.

Discussion and conclusions

In this paper we propose the identification of the geographic origin of surnames for bibliometric purposes. For this, we propose the use of scientific databases in order to work with data worldwide. In this way we overcome a major restriction of this type of studies regarding data availability (Cheshire, 2014). We propose two methodologies to assign countries to surnames. The first method is based on the number of surnames found in a given country when its Kullback-Leibler divergence measure is below the 80th percentile of all the combinations with the lowest Kullback-Leibler values. The second methodology is based on the concentration of a given surname in a country, using the Gini Index to calculate such concentration.

In this regard, a preliminary validation has been done comparing the coverage and correct assignments made with a sample list of 20 languages grouped into 12 'normalized countries'. The results reported are promising, especially for the first methodology. In fact, this has already been applied successfully elsewhere (Costas & Noyons, 2013). But the second methodology ensures a 100% coverage of all surnames. However, much research is still needed and further refinements in both methodologies. First, we believe that thresholds of minimum publications of a surname by country should be established in order to improve the methodology based on the Gini Index. Regarding the Kullback-Leibler divergence methodology, we considered reliable assignments those which were below the 80th percentile, however, different thresholds should be also tested. Finally, we will consider other validation lists as some questionable assignments were found in this control list (e.g., Pinto is assigned to Italian language, but it could also be assigned to Spanish or even Portuguese) which may blur the evaluation of the actual performance of each method. These methods should also be compared with those developed elsewhere.

The use of surnames to track demographic movements or analyse diversity in collaboration shows interesting opportunities for implementing these methodologies in bibliometric analyses. One example of such application is the recent work of Freeman and Huan (2014). However, frequently little attention to the methodology employed for assigning countries, languages or ethnicities to surnames is paid, something that may represent a challenge to results based on these data. Thus, understanding better the limitations and possibilities of these data is critical for a proper use. Although further research is still needed, we believe that applying methodologies such as the ones suggested here using bibliographic databases will lead to more reliable results.

References

Burchard, E., Elad, Z., Coyle, N., Gomez, S.L., Tang, H., Karter, A.J., Mountain, J.L., Pérez-Stable, E.J., Sheppard, D. & Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, 348(12), 1170-1175.

Calero, C., Buter, R., Cabello Valdés, C. & Noyons, E. (2006). How to identify research groups using publication analysis: An example in the field of nanotechnology. *Scientometrics*, 66(2), 365-376.

Chen, K.-H. & Cavalli-Sforza, L.L. (1983). Surnames in Taiwan: Interpretations based on geography and history. *Human Biology*, 55(2), 367-374.

Chesire, J. (2014). Analysing surnames as geographic data. *Journal of Anthropological Sciences*, 92, 99-117.

Costas, R. & Iribarren-Maestro, I. (2007). Variations in content and format of ISI databases in their different versions: The case of the Science Citation Index in CD-ROM and the Web of Science. *Scientometrics*, 72(2), 167-183.

Costas, R. & Noyons, E. (2013). "Detection of different types of 'talented' researchers in the Life Sciences through bibliometric indicators: Methodological outline". Retrieved from: http://hdl.handle.net/1887/22165

Freeman R.B. & Huang, W. (2014). Collaboration: Strength in diversity. *Nature*, 513(7518), 305.

García, J., Rodriguez-Sánchez, R., Fdez-Valdivia, J., Robinson-García, N. & Torres-Salinas, D. (2013). Benchmarking research performance at the university level with information theoretic measures. *Scientometrics*, 95(1), 438-452.

Hatton, T.J. & Wheatley Price, S. (2005). Migration, migrants and policy in the United Kingdom. In: Zimmermann, K.F. (ed.), *European migration: What do we know?* (pp. 113-172). Oxford University Press.

Kissin, I. (2011). A surname-based bibliometric indicator:publications in biomedical journal. *Scientometrics*, 89(1), 273-280.

Kissin, I. & Bradley, E.L.J. (2013). A surname-based patent-related indicator: The contribution of Jewish inventors to US patents. *Scientometrics*, 97(2), 357-368.

Kullback, S. & Leibler, R.A. (1951). On the information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86

Mateos, P. (2007). A review of name-based ethnicity classification methods and their potential in population studies. *Population, Space and Place*, 13(4), 243-263.

Moed, H.F. & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics*, 101(3), 1987-2001.

Torres-Salinas, D., Robinson-García, N., Campanario, J.M. & Delgado López-Cózar, E. (2014). Coverage, field specialisation and the impact of scientific publishers indexed in the Book Citation Index. *Online Information Review*, 38(1), 24-42.

Torres-Salinas, D., Rodriguez-Sánchez, R., Robinson-García, N., Fdez-Valdivia, J. & García, J.A. (2013). Mapping citation patterns of book chapters in the Book Citation Index. *Journal of Informetrics*, 7(2), 412-424.

Waltman, L. & van Eck, N.J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics*, 96(3), 699-716.

An 80/20 Data Quality Law for Professional Scientometrics?

Andreas Strotmann[1] and Dangzhi Zhao[2]

[1] *andreas.strotmann@gmail.com*
ScienceXplore, D-01814 Bad Schandau (Germany)

[2] *dzhao@ualberta.ca*
University of Alberta, School of Library and Information Studies, Edmonton, Alberta (Canada)

Scientometric network error consequences

Only very recently have researchers begun looking at what concrete effect the errors in a network model caused by name ambiguities in the data sources may have on the results of popular types of network analysis. The results that they report are quite alarming in the aggregate: not only do typical evaluative analyses of individuals (e.g., citation rankings) suffer significantly from these errors, but there is mounting evidence that even the most basic statistical features of realistic large-scale networks are hugely distorted by ambiguities. Strotmann et al. (2009), for example, document significant distortions in co-authorship network visualizations, and Diesner and Carley (2013) report that "minor changes in accuracy rates of [name disambiguation] lead to comparatively huge changes in network metrics, while the set [of] top-scoring key entities is highly robust. Co-occurrence based link formation entails a small chance of false negatives, but the rate of false positives is alarmingly high."

In fact, Fegley and Torvik (2013) go so far as to dismiss one of the most famous recent results in large-scale social network analysis, the exact power-law distribution from preferential attachment (Barabási & Albert, 1999), at least in the case of scientific collaboration networks (Barabási et al., 2002), as a mere artefact produced by a lack of name disambiguation in the underlying dataset! The ultimate irony here is that Fegley and Torvik's (2013) data are consistent with an interpretation that Barabási's cooperation network power may have been induced by a power law distribution of name ambiguities rather than co-authorships.

Similarly, Strotmann and Zhao (2013) find that even highly stable statistical analysis methods of author co-citation analysis fail in the face of large-scale ambiguity errors in the underlying dataset.

While for evaluative bibliometrics the most serious problem is generally the "splitting" of individuals, i.e., the failure to recognize each and every one of an individual's contributions correctly (especially of high-performing individuals), Fegley and Torvik (2013) find that splitting is not the main concern in relational network analysis. Instead, they and Strotmann and Zhao (2013) both find that it is the erroneous "merging" of individuals, i.e., the failure to separate the contributions of multiple individuals correctly because their names are too similar, that causes major distortions of large-scale network analysis results in relational network analysis. Especially East Asian names are prone to extreme amounts of merging. While in European cultures there are relatively few common given names but a large variety of family names, in Chinese, Korean and other East Asian cultures the opposite is the case—a small number of surnames is shared by half their populations, but given names are much more varied. The old tradition in scientific publishing to list authors by their surnames and initials works, sort-of, when science is done in European-origin cultures, but all bibliographic databases have in recent years had to move to a full-name model as research boomed in the Asian Tiger nations (e.g., PubMed/MEDLINE in 2002).

When is a scientometric network sufficiently complete and clean?

As Torvik and Smalheiser (2009) make abundantly clear, it is for all intents and purposes impossible to disambiguate the names of all the individuals in a large dataset completely and fully correctly. With absolute perfection thus out of the question, what remains is to ask when a disambiguation is "good enough", and if (and how) it is possible for a typical researcher to go about disambiguating the dataset well enough. Unfortunately, there is very little research, if indeed any, into what constitutes "good enough" for a scientometric study. The few studies that have looked into what goes wrong when individuals are not recognized correctly do give us a hint, though.

First of all, "good enough" usually means that the most important contributions of the top-ranked individuals must be absolutely correctly attributed. Whatever other good methods (e.g., name disambiguation algorithms or author registries) we may find to disambiguate our data, in the end it will therefore be necessary to manually double-check, and where necessary fix, the highest-impact individuals' data. Secondly, some statistical procedures or network measures are more vulnerable than others to name ambiguities. Local network measures (e.g., node degree) are less affected than global ones (e.g., size of connected component), and evaluative studies (e.g., ranking) are more affected than relational ones (e.g.,

correlations) (Diesner & Carley, 2013; Strotmann & Zhao, 2012).

An 80/20 scientometric data quality rule?

For ranking studies, absolute correctness is paramount, and huge efforts need to be expended to get all the top-ranked individuals just right. When the "individuals" are research institutions, this can be a daunting task. For correlative studies, on the other hand, a study by Albert, Jeong, and Barabási (2000) warns us that, while global measures of power-law distributed networks may be quite resilient to *uniformly* distributed random errors, they are also quite vulnerable to the kind of *highly skewed* error distributions that we observe for name ambiguities, for example. In the case of an extremely skewed error distribution, they observed that an error rate as low as 10%-20% completely changed the measured values for a fundamental global network metric, namely, connectivity.

We can take this as a warning that, as a rule of thumb, we generally need to aim for a roughly 90% (but definitely 80% or better) complete and correct dataset when error distributions are skewed. Note that the requirement of 80% completeness or better applies, in particular, to the underlying citation index's coverage of the field being studied: a focus on high-impact literature implies a highly skewed error distribution! On the plus side, studies on the life sciences can thus be relied upon to yield reliable results as long as their disambiguations are good. Results from *any* scientometric study on the social sciences, however, are suspect as long as they rely on these databases and these databases cover much less than 80% of the literature in those fields.

Note that an 80% data correctness requirement for a professional scientometric study would apply to the data as it is used for network statistics. When both data collection *and* cleaning are subject to highly skewed error distributions, this means that we need 90% correct data collection *and* 90% correct data cleaning to guarantee 80% correct data for analysis.

Conclusions: the bad news and the good

This, then, is the bad news for those who aim to provide a truly professional scientometric service to their community: power-law-like data *and* error distributions may mean that only nearly-complete *and* nearly-clean datasets can be trusted to serve as a reliable basis for nearly *any* type of network or statistical analysis.

The good news is that there are plenty of successful bibliometric studies that imply that this level of correctness is also usually quite sufficient for meaningful studies, as long as only "local" measures or relational statistics are required. There *are* fields that are covered to 90%+ in citation databases, e.g., the citable literature of the life sciences, and there *are* disambiguation methods (e.g., some of those reviewed in Smalheiser & Torvik, 2009 or that of Strotmann et al., 2009) that do make reliable scientometric studies possible.

However, scientometric professionalism may well require that these methods be utilized in nearly *all* future studies, and thus, that they be applied to, and adopted by, the citation databases themselves.

Acknowledgments

This Ignite Talk extends, with permission of the publisher, Section 4.4, "Disambiguation in Citation Network Analysis: Ambiguity and Power Laws," of Zhao & Strotmann (2015).

References

Albert, R., Jeong, H.W. & Barabási, A.L. (2000). Error and attack tolerance of complex networks. *Nature 406*, p.378

Barabási, A.L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, p.509

Barabási, A.L., Jeong, H., Neda, Z, Ravasz, E, Schubert, A. & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A, 311*, p. 590

Diesner, J. & Carley, K.M. (2013). Error propagation and robustness of relation extraction methods. *XXXIII International Sunbelt Social Network Conference*, Hamburg, Germany, May 2013.

Fegley, B.D. & Torvik, V.I. (2013). Has large-scale named-entity network analysis been resting on a flawed assumption? *PLoS ONE 8* (7): e70299.

Smalheiser, N. R. & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology* 43, 287.

Strotmann, A., Zhao, D. & Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology 2009 Annual Meeting*, November 6–11, 2009, Vancouver, BC, Canada

Strotmann, A. & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63* (9), p.1820

Torvik, V. I. & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 3* (3)

Zhao, D. & Strotmann, A. (2015). *Analysis and Visualization of Citation Networks.* Morgan & Claypool.

Some Features of the Citation Counts from Journals Indexed in Web of Science to Publications from Russian Translation Journals

Maria Aksenteva

ms@ufn.ru
"Uspekhi Fizicheskikh Nauk" Editorial Office, Lebedev Physical Institute, Russian Academy of Sciences, Leninskii prosp. 53, 119991 Moscow (Russia)

Introduction

As it was emphasized by Moed, H.F., Glänzel W. & Schmoch U. (2005) in their editors' introduction to the *Handbook of Quantitative Science and Technology Research*: "A most important data source for analysis of the science system is the *Science Citation Index* (SCI) and related Citation Indexes published by the *Institute for Scientific Information* (ISI–Thomson Scientific, Philadelphia, PA, USA), or, in a more recent version, ISI's *Web of Science*." Due to this very competent opinion (supported of course by major part of scientists all over the world) it is very important for proper evaluation of the science and its development in Russia to investigate how publications in Russian journals indexed in SCI and how citations to these publications were counted and recorded in SCI in previous decades and is counted and recorded now in Web of Science (WoS).

Some systematic problems with proper indexing and correct counting of citations to publications in Russian journals in SCI was revealed by brilliant founder of modern bibliometrics ("statistical bibliography") Eugene Garfield long time ago in 1974. The greatest problems (according to Garfield) occurred with so-called "translation journals": "The term Russian journals is used here as it is daily used in libraries in the United States. We are aware of its inadequacy and inaccuracy, but plead its convenience. A few of the journals are Slavic, but not Russian. The term Soviet journals might seem more appropriate, but it would not be. An important group of the journals considered is published outside the Soviet Union the so-called translation journals. Neither Russian nor Soviet, they are nevertheless the product of Russian and Soviet research. They also present, as we learned in this study, a formidable stumbling block in journal citation analysis of this type. I speak here only in terms of statistical bibliography as regards the translation journals." (Garfield, 1974).

What was (and is now) the biggest problem with indexing and counting of citations of the "translation journals"? It was (and is now) the adopted by SCI (now Web of Science) policy of the counting of citations to original publications (articles, published in Russian) and to the English version of the same article, published in "translation journals". As it was found in the present research this policy were changed several times during the period of SCI existence and this policy can significantly affect the conclusions, which could be made about Russian science in many analytical reports and investigations, based on *Web of Science* data (see, for example, Albarrán et al., 2013).

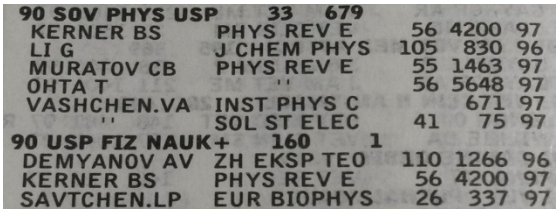
In this research we studied the style (the policy) of records for publications from Russian (and translation) journals and counting of citations to them in printed volumes of SCI in 1960-1998 years and compared these styles with the policy, adopted in the internet version of the successor of SCI (WoS) in 1990-es and now. It is possible to say after this investigation, that significant (sometimes huge) amount of citations (from the journals indexed in WoS) to Russian publications are not possible to find in WoS now without some complicated additional search. All these citations are not taken into account in many analytical reports about Russian science (especially about natural science such as physics, chemistry, biology etc.). At the same time it is not very difficult now to return back to the Garfield's old policy of records and calculations of the citations to Russian publications in translation journals, which could collect properly all citation using new possibilities of Internet linking of publications. (See, for example, UFN journal's web-site www.ufn.ru on which the citing articles are collected using CrossRef system (using Digital Objects Identifier - DOI) or www.mathnet.ru site for more precise and elegant citations collecting (Zhizhchenko & Izaak, 2009; Chebukov et al, 2013)).

Methodology and data

We compared the number of citations to an article published in "Uspekhi Fizicheskikh Nauk" (UFN) journal (or to the English translation to the same article published in "Physics-Uspekhi" (former "Soviet Physics-Uspekhi" journal until 1992 year) ─ cover-to-cover English translation of UFN journal) presented in printed volumes of SCI with the number of citations to the same article presented in Web of Science (on-line version) and with the number of citations, which could be found using CrossRef links (DOI) on www.mathnet.ru and/or www.ufn.ru web-sites (see details in Aksenteva, Kirillova & Moskaleva, 2013).

Results and discussion

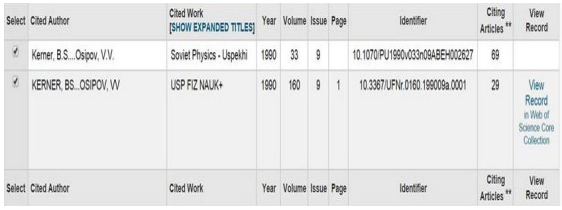
Let's consider (as a typical example) an article (Kerner & Osipov, 1990). First of all we have found that in printed volume of SCI (see Fig. 1) both Russian original article and its English translated version were indexed (citations to them were collected separately, but all citations were displayed, see Figure 1):



90 SOV PHYS USP	**33**	**679**		
KERNER BS	PHYS REV E	56	4200	97
LI G	J CHEM PHYS	105	830	96
MURATOV CB	PHYS REV E	55	1463	97
OHTA T	''	56	5648	97
VASHCHEN.VA	INST PHYS C		671	97
''	SOL ST ELEC	41	75	97
90 USP FIZ NAUK+	**160**	**1**		
DEMYANOV AV	ZH EKSP TEO	110	1266	96
KERNER BS	PHYS REV E	56	4200	97
SAVTCHEN.LP	EUR BIOPHYS	26	337	97

Figure 1. Copy from SCI (1997) for Kerner B.S.

But now in WoS (internet version) we cannot find citations to the English version of this article. It is possible to find them only by using the WoS's option "Cited References Search" (see Figure 2).



Select	Cited Author	Cited Work [SHOW EXPANDED TITLES]	Year	Volume	Issue	Page	Identifier	Citing Articles **	View Record
✓	Kerner, B.S....Osipov, V.V.	Soviet Physics - Uspekhi	1990	33	9		10.1070/PU1990v033n09ABEH002627	69	
✓	KERNER, BS...OSIPOV, VV	USP FIZ NAUK+	1990	160	9	1	10.3367/UFNr.0160.199009a.0001	29	View Record in Web of Science Core Collection
Select	Cited Author	Cited Work	Year	Volume	Issue	Page	Identifier	Citing Articles **	View Record

Figure 2. Cited references search in WoS core collection for article Kerner B.S. &Osipov, 1990.

It is possible to see on this figure, that there are 29 citations to the Russian version of this article and 69 citations to the English version of the article, but (unfortunately for the Russian journal) it is possible to view citing articles for the Russian version only (only 29 citing articles). 69 citations to the English version of this article are not taken into account in Prof. Kerner's (and of course for Prof. Osipov too) citation report, are not included into their Hirsh's indexes, are not taken into account for his laboratory and his institute bibliometrics etc. (and for Russian physics and science in general). On our web-site using CrossRef links it is possible to find 70 citing article: http://ufn.ru/ru/articles/1990/9/a/.

It is necessary to mention that for publications in UFN journal until September 2001 only citations to the Russian version are presented in WoS (but citations to the English version are not taken into account). We have checked more than one thousand articles (published in 1990-2000 years in UFN) and have found that about 67% of citations (in average) to these articles were not presented now directly in WoS (and so do not taken into account for any analytical scientometric report). According to WoS in 1990-2000 years 1190 articles were published in UFN (and indexed in WoS) and they have only 9002 citations (on April 25, 2015). Using DOI on our website we have found 14973 citations to 1167 articles, published in UFN in the same period.

Conclusions

It was found that now WoS show less than half of citations (from journals indexed in WoS) to described above article (Kerner, Osipov, 1990), but this is not an exceptional example. So all publications in Russian translated journals (indexed in WoS) lose a lot of their absolutely correct citations (about 60% in average) from journals indexed in WoS and therefore scientometrics, based on WoS direct data, underestimates the real impact of Russian scientists and science in general.

Acknowledgments

I am grateful to Prof. M.Yu. Romanovsky who has encouraged me to make this investigation.
The work was supported by the Russian Foundation for Basic Research (project No. 13-07-00672 a).

References

Aksenteva, M.S., Kirillova, O.V., & Moskaleva, O.V., (2013) On Paper Citation by Web of Science and Scopus From Translated Russian Journals (in Russian), *Nauchnaya periodika: problemy i resheniya* [Scientific periodical press: problems & solutions], No. 4(16), 4-18. Retrieved January 18, 2015 from http://ufn.ru/tribune/trib124.pdf

Albarrán, P., Perianes-Rodriguez, A., & Ruiz-Castillo, J. (2013). Differences in citation impact across countries. In *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, Vienna, Austria*. Vol. 1, p. 536. Retrieved January 18, 2015 from http://www.issi2013.org/Images/ISSI_Proceedings_Volume_I.pdf

Chebukov, D., Izaak, A., Misurina, O., Pupyrev, Yu. & Zhizhchenko, A., (2013) "Math-Net.Ru as a digital archive of the Russian mathematical knowledge from the XIX century to today", *Lecture Notes in Computer Science*, 7961 (Ed. J. Carette et al.) 344–348, arXiv: 1305.5655.

Garfield, E. (1974) "Russian Journal References and Citations in the Science Citation Index Databank". In *Journal Citation Studies*, 22. Philadelphia: ISI. Retrieved 01/18/ 2015 from: http://www.garfield.library.upenn.edu/papers/244.html

Kerner, B.S., Osipov, V.V. (1990) *Usp. Fiz. Nauk* 160 (9) 1–73 [*Sov.Phys.Usp.* 33 (9) 679–719]

Moed, H.F., Glänzel W. & Schmoch U. (Eds.) (2005) Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems, Dortrecht: Kluwer Academic Publishers.

Zhizhchenko A.B. & Izaak A.D. (2009) "The information system Math-Net.Ru. Current state and prospects. The impact factors of Russian mathematics journals", *Russian Math. Surveys*, 64**,** 4.

Semantics, a Key Concept in Interoperability of Research Information -The Flanders Research Funding Semantics Case

Sadia Vancauwenbergh

Sadia.Vancauwenbergh@uhasselt.be
ECOOM-UHasselt, Hasselt University, Research Coordination Office, Martelarenlaan 42, BE-3500 Hasselt (Belgium)

Introduction

In a knowledge-based economy, a good overview of the scientific and technological portfolio is essential for policy formation and driving knowledge transfer to the industry and the broad public. In order to enhance open innovation, the Flemish public administration has created a Flanders research information portal (FRIS, http://www.researchportal.be) that integrates information available from its data providers (research institutions, funding organizations…) using the CERIF (The Common European Research Information Format) standard. Although this standard allows for almost unlimited flexibility for modelling the research information, it has limitations when it comes down to communication to end-users, in terms of semantics. However, interoperability of research information is only meaningful when a well-defined semantics is used. This paper describes the implementation of a business semantics tool on data concepts and classifications for research funding as a means to unambiguously exchange and interpret these data.

The need of semantics

A couple of decades ago, the demands on the research community to report on research data were rather low. Results were published in preferably highly-rated journals and rather limited research reports were written. Over the years, more research data became available and the need for research databases grew. Unfortunately, these databases were predominantly developed per organization without consultation of other organizations. Moreover, because of the rather low data volume and people involved, there seemed no explicit need for defining an accompanying semantics.

However, as the research system expanded, there has been a massive increase in the amount and nature of the information stored as well as its information consumers. These changes are not only due to the advancements made in the research field itself, but are also explained by the global efforts undertaken to transfer the obtained knowledge to industry and the broad public. In Flanders, this resulted in the creation of the FRIS-portal which makes Flemish research information publicly available. This information is provided via a multitude of data providers that often use a different terminology for a similar concept or alternatively, use a similar terminology for a different concept. The correct interpretation of the information at the FRIS portal is realized by the addition of a semantic layer on top of the data by the data providers, which later on is translated to a general FRIS semantics resulting in data communication in the same language. The focus on the explicit semantic alignment with the data providers, adds further to existing initiatives like VIVO and CERIF based CMS (Guéret et al., 2013). Data unambiguity is increasingly important, in an era where many initiatives have seen light to measure and benchmark research and where public research reporting obligations are vastly growing. Obviously, the lack or incomplete definition of semantics puts large constraints on the interoperability of research information, and in extension on the policies drawn out of these data.

The Flanders research information landscape

In Flanders, research institutions receive funding from a broad range of research funding providers going from the regional to national and international level. Obviously, each funding provider has its own requirements with regards to the formats or classifications used for reporting on the resulting research output, thereby creating a multitude of largely similar research reports. Obviously, this places a large burden on the research community. Until now, the data providers tried to keep pace with this vast expansion of research reporting by improving or even creating databases, unfortunately without generally agreed upon semantics. At the same time, the data providers were feeding their information to the FRIS-portal in order to increase the visibility of the research in Flanders to third parties (i.e. companies, research institutions and individual researchers).

In line with the growing concern on the administrative burden put on the research community, a report was published by Peters et al. (2011) providing guidelines for the reduction of redundant research information reporting. Following these advices, the Flemish Department of Economy, Science and Innovation (EWI) is currently improving the FRIS-portal in order to be used as a virtual research information space, for information retrieval in a transparent and automated manner that can be used for research reporting (Figure 1) (Debruyne et al., 2011). This implicates

the use of unambiguous data concepts and research funding classifications. Until recently, funding organizations were using their own funding classification schemes which were semantically poorly defined and lacked concordance mappings to other (inter)national classifications.

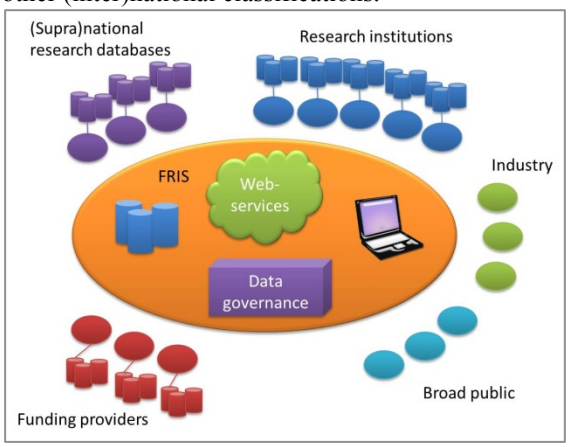


Figure 1: Representation of the FRIS design.

Funding data and classification governance

In order to add a semantic layer on top of the FRIS database layer, the Data Governance Centre® (DGC) platform of Collibra has been used. This platform allows data suppliers to manage their own data models used to describe, i.e. research funding together with the corresponding institution specific semantics (Figure 2).

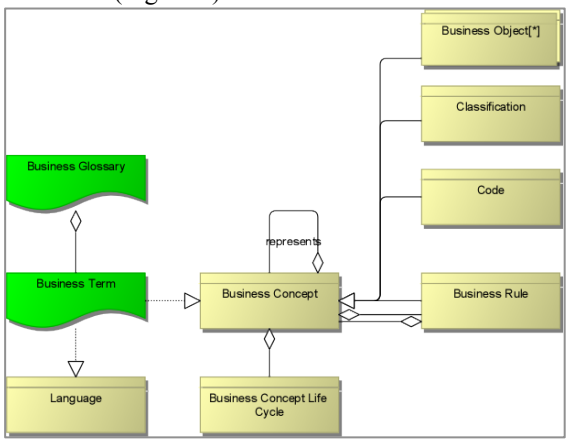


Figure 2. Incorporation of a business semantics glossary on the research funding model.

At the same time, the DGC platform has been used for the description of each individual component of the FRIS research funding model using definitions (Figure 3). By explicitly defining all concepts, the governance tool assists in the swift identification of semantic inter-organizational misalignments when mapping corresponding concepts by the stakeholders. The resulting ontologies can be exported and used to annotate data in relational databases, and hence render data meaningful. Furthermore, the DGC tool has been used for defining the semantics of classifications and code sets on research funding, which is essential when it comes down to consistent and unambiguous reporting on research funding to third parties. Obviously, the research community at large will benefit from this, as the information retrieved via FRIS will be much more reliable and accurate.

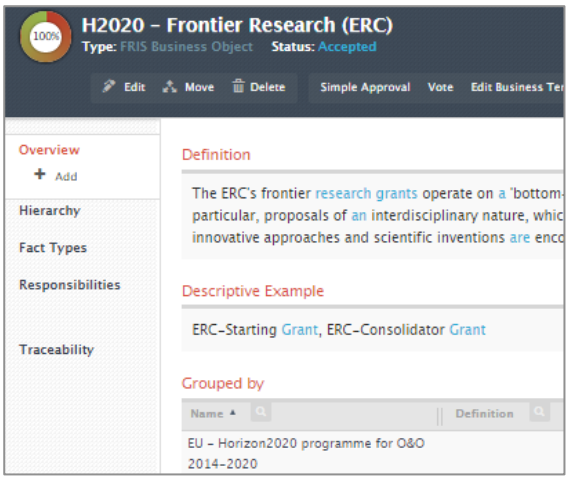


Figure 3: DGC as a governance tool for research funding classifications.

Altogether, the use of a data governance tool focused on semantics opens new avenues in terms of efficiency of the research ecosystem. Not only will governments be able to delineate better founded policies, also research administrations and researchers themselves can gain tremendously as research reporting could be automated from the FRIS-portal in a reliable manner, thereby reducing the administrative burden at the benefit of scientific discovery and innovation.

Acknowledgement

This work is part of the Classification Governance project carried out for the Expertise Centre for Research & Development Monitoring (ECOOM) in Flanders, which is supported by the Department of Economy, Science and Innovation, Flanders.

References

Debruyne, C., De Leenheer, P., Spyns, P., Van Grootel, G., & Christiaens, S. (2011). Publishing open data and services for the Flemish research information space. In *Advances in Conceptual Modeling. Recent Developments and New Directions,* Springer.

Guéret, C., Chambers, T., Reijnhoudt, L., et al. (2013). Genericity versus expressivity – an exercise in semantic interoperable research information systems for Web Science. *In: http://arxiv.org/pdf/1304.5743.pdf*

Peters, A., & Lambrechts, L. (2011). De vereenvoudiging van onderzoeksverslaggeving, een analysetraject uitgevoerd door de Vlaamse universiteiten en hogescholen en de VLIR, in opdracht van de Vlaamse Overheid (EWI).

The Information Retrieval Process of the Scientific Production at Departmental-level of Universities: A New Approach.

César David Loaiza Quintana[1] and Víctor Andrés Bucheli Guerrero[2]

[1] *cesar.loaiza@correounivalle.edu.co*
Universidad del Valle, Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería. Universidad del Valle, Sede Meléndez. Calle 13 # 100-00. Cali, Valle Del Cauca (Colombia).

[2] *victor.bucheli@correounivalle.edu.co*
Universidad del Valle, Escuela de Ingeniería de Sistemas y Computación, Facultad de Ingeniería. Universidad del Valle, Sede Meléndez. Edificio 331 oficina 2113, Calle 13 # 100-00. Cali, Valle Del Cauca (Colombia)

Introduction

Our work was focused on document retrieval from Scopus databases of the Escuela de Ingeniería de Sistemas y Computación (EISC) of the Universidad del Valle (Cali - Colombia).

The databases systems as WoS (Web of Science) or Scopus contain the knowledge produced by engineer schools. However, this information is ambiguous and the retrieving of the specific documents of one school is identity uncertainly (Pasula et al., 2003). Thus, the design of machines (search engines) to retrieve the relevant documents of engineer schools is a complex process.

After the work of Bucheli et al. (2013); Cuxac, Lamirel, & Bonvallot (2013) proposed a semi-supervised approach, mixing soft-clustering and Bayesian learning. Additionally, Huang et al. (2014) proposed a rule-based algorithm. Both approaches were for affiliation disambiguation.

We reproduced the model proposed by Bucheli et al. (2013). The results show that the model can be used to information retrieval of department-level. In addition, we proposed a new approach addressing the problem of classification using network science. The future work will be related with building a model according to the network science approach.

Methodology

Model of Bucheli et al. (2013)

We followed the methodology specified by Bucheli et al. (2013) shown in Figure 1(a).

1) The configuration of the initial search strategy proposed by Bucheli et al. (2013) was applied using the Scopus search engine to get a set **I** composed by documents that contains all the documents that belong to EISC and others that not belong to it.

2) The initial search strategy was based on a review of the research activity of the School and it proposes recovering a set of documents $\mathbf{I} = \mathbf{A} \cup \mathbf{J} \cup \mathbf{S} \cup \mathbf{O}$. The staff **S** set is made up by papers which are related to a list of school professors names explicitly. The journal set **J** is the bunch of documents published in the journals where the school has previously published. The address set **A** is related to the documents that have in their affiliation the name of the school explicitly. Finally, socio-semantic set $\mathbf{O} = \mathbf{S} \cup \mathbf{C}$, where the concepts set **C** is made up by the documents related to a bunch of research areas from a school. Every set mentioned before has an additional restriction; his documents must belong to the university that hosts the internal-level unit, in our case to the Universidad del Valle.

3) An Expert from EISC classified all the documents from the initial search and we built a relevant set **R** with **I** elements that belong to EISC.

4) We built a dataset where one paper or instance is characterized by a vector (with five positions). Each position is a binary variable, related to sets **A**, **S**, **J**, **O** and **R**, that tell us if the paper belongs or not to the corresponding set. Thus, the instance class is determined by the variable **R**.

5) Afterwards, we made the classification using the Naïve Bayes model of information retrieval illustrated in (1). It was evaluated based on all instances of the dataset. We used standard measurements over cross validation test 10 fold (Witten, 2005; Baeza-Yates, 1999). On the other hand, the publication year was taken into account as parameter of evaluation. Thus, we train the model with paper published between two specific years, for instance 1989-2010 and testing the model with papers published in the following years. This procedure was evaluated by the following years of training 1989-2011, 1989-2012 and 1989-2013.

$$p(R|J,S,O,A) = \frac{p(R)p(J,S,O,A|R)}{p(J,S,O,A)} \quad (1)$$

Proposed model based on network science

The machine learning process follows five phases: Selecting data, expert validation, co-author network building, feature extraction from network and classification, as shows the Figure 1(b).

The data selection trough the initial search strategy and the expert validation have be taken into account similarly to the review model of Bucheli et al. (2013). Here, the document corpus used is the same of evaluation model applied to the EISC, however the feature extraction changes and the features are related with network measurements.

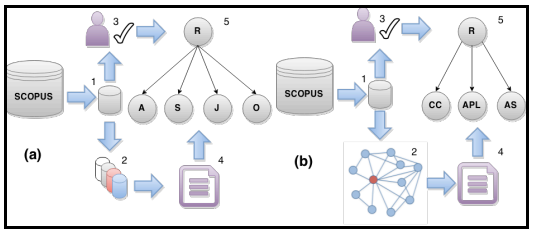


Figure 1. The five phases of the evaluated and proposed methodologies.

The document corpus contains information about co-authorship relations. Each author is identified by an ID that Scopus assigns. We build a co-authorship network, where, the network is traduced as a weighted and undirected graph in which the weight of the edges designates the number of documents where whichever two authors have participated. The new dataset is built as follows: one document or instance is a vector of values where each position is a variable related with one measurement of co-author network, in which, the specific paper was subtracted. Thus, for each instance, the authors that participated in the specific document are deleted and the measures are computed again. Additionally, the last variable **R** shows if the paper belongs or not belongs to the School. The measurements of networks are:

1. The Cluster Coefficient (**CC**): The local clustering coefficient captures the degree to which the neighbours of a given node link to each other. We use the average of all local clustering coefficients.
2. The average path length (**APL**) is the average distance between all pairs of nodes in the network.
3. The average strength (**AS**), is the average of the sum of the edge weights of each node. (Barabasi. 2012).

Finally, we develop a supervised learning environment through a Naïve Bayes Classifier and the proposed model is evaluated and compared with the model proposed by Bucheli et al (2013).

Results, discussion and future work

Table 1 shows standard evaluation measurements. Here, we introduce the cross validation fold 10 test, the measurements show in Bucheli. et al. (2013), and the evaluation for different publication years 1989-2011, 1989-2012 and 1989-2013. The results show that the model was applied to other School with similar performance measurements, in this sense the model is consistent and allows to build one search engine of department-level. Additionally, we evaluated the practical utility of the model, verifying that it is capable of doing an acceptable prediction of EISC's documents published after a specific date when it is trained with a set of documents published until that date.

In this work, we found the finger prints of department-level of universities that allow us to design search engines that retrieve relevant documents of department-level.

Table 1. Evaluation measurements of the model.

	Recall	Precision	ROC curve
EISC Univalle			
Cross Validation fold 10	0,932	1,000	0,989
Bucheli et al. (2013)			
Department of Industrial Engineering –University of Pittsburgh	0,494	0,997	0,984
Faculty of Engineering – Universidad de los Andes (Colombia)	0,954	0,992	0,965
EISC Univalle			
Training:1989-2011 Evaluation: 2012-2014	0.833	1.000	0.974
Training 1989-2012 Evaluation: 2013-2014	0.826	1,000	0.964
Training 1989-2013 Evaluation: 2014	0,786	1,000	0,939

The networks science approach is an opportunity to propose a mathematical model able to learn the structure of co-authorship network from a particular school. Then, we can design a classifier of relevant documents at department-level based on co-authorship relations. This allows making a classification with little a priori information about an organization, which turns into a more general model than Bucheli et al. (2013). We proposed a model, namely (2).

$$p(R|CC, APL, AS) = \frac{p(R)p(CC, APL, AS|R)}{p(CC, APL, AS)} \quad (2)$$

We suggest as future work to evaluate the model based on network measurements at the same school and other 3 schools of engineering from different universities.

Acknowledgments

Thanks to Convocatoria Interna, Universidad del valle 2014; Facultad de Ingeniería, Universidad del valle; and EISC.

References

Baeza-Yates, R. (1999). *Modern information retrieval*. New York: Addison-Wesley.

Barabási, A.L. (2012). *Network science book*. Center for Complex Network Research, Northeastern.

Bucheli, V., Calderón, J., Gonzales, F., Bidanda, B., Valdivia, J., & Zarama, R. (2013). Model to support the information retrieval process of the scientific production at departmental-level or faculty-level of universities. *Proc. ISSI*.

Cuxac, P., Lamirel, J. C., & Bonvallot, V. (2013). Efficient supervised and semi-supervised approaches for affiliations disambiguation. *Scientometrics*, *97*(1), 47-58.

Huang, S., Yang, B., Yan, S., & Rousseau, R. (2014). Institution name disambiguation for research assessment. *Scientometrics*, *99*(3), 823-838.

Pasula, et al. (2003). Identity Uncertainty and Citation Matching, *NIPS,* MIT Press.

Witten, I. (2005). *Data mining: practical machine learning tools and techniques*. 2nd ed., Amsterdam: Morgan Kaufman.

Efficiency, Effectiveness and Impact of Research and Innovation: a framework for the analysis

Cinzia Daraio

daraio@dis.uniroma1.it
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

Introduction, motivation and policy relevance

The main objective of this paper is to provide a framework for the assessment of the research activity and its impacts. This is a difficult task. First of all, because of the heterogeneity, partial overlapping and fragmentation of the different streams of literature. Secondly, due to the need of applying a systemic approach to account for the complexity of the research activity and its complementarities and interrelationships with teaching, third mission activities and other relevant dimensions of performance, including the inputs.

This work originated from Daraio (2015) which pointed out the unavailability of a best evidence on the "efficiency, effectiveness and impact of research and innovation" due to the lack of a suitable framework for a comprehensive analysis.

Two recent policy initiatives witness the need and call for the proposal of a general framework for assessing research and its impact. We refer to the STAR metrics in US and to the EC (2014) "Expert Group to support the development of tailor-made impact assessment methodologies for ERA" in Europe.

We discuss in the following the main dimensions of our framework which are: 1. Theory, 2. Methods, 3. Data.

Research and innovation in the theory

In theory, the following streams of literature have considered research and innovation as the main link of Science and Society interplay:

• *Economics of science and technology* as an emerging field, which draws on the fields of economics, public policy, sociology and management (Audretsch et al., 2002).

• *Growth theory* (Aghion & Howitt, 2009), within which «the residual» is considered as technology advance over time (Solow, 1957); or as our ignorance (Abramovitz, 1956). The old growth theory (Nelson & Phelps, 1966) considers as additional inputs investments in R&D and education while the new growth theory (Romer, 1986; 1994) emphasizes the influence of other factors such as technologies or efficiencies, spill-overs and incentive of agents.

• *Quantitative science and technology research*, organized as quantitative studies of science system, of technology system and of science-technology interface. The focus here is -though not exclusively- on scholarly publications and patents, it embraces bibliometrics, scientometrics (Moed, Glanzel & Schmoch, 2004) and informetrics (Egghe & Rousseau, 1990), more recently starting to consider also other non-scholarly and societal «altmetrics» dimensions (Cronin & Sugimoto, 2014).

• *Economics of innovation*, which is at the core of several different economic fields, including macroeconomics, industrial organization (strategies and interactions of innovative firms), public finance, policies for encouraging private sector innovation, and economic development (innovation systems and technology transfer) (Hall & Rosenberg, 2010).

• *Science of Science policy* (Fealing et al., 2011; National Academy of Science, 2014; Lane, 2011, 2014).

• *Science and Society interplay* (Etzkowitz & Leydesdorff, 2000; Aghion et al., 2009; Helbing & Carbone, 2012).

A neglected aspect within these streams of work is the building block of education. From the economics of education (Johnes & Johnes, 2004; Hanushek et al., 2011) we know that education is an investment in human capital analogous to an investment in physical capital. The missing link with previous streams of literature is people. People in fact carry out research and innovation activities; attend schools and higher education institutions, acquiring competences and skills. Here another link could be added with Dosi (2014).

Methods for the assessment of Research

The assessment of the performance of an activity can be carried out on its output, on its outcome (indirect output), on its productivity (partial or total factor productivity), on its efficiency, on its effectiveness, on its impact.

From a methodological point of view, a distinction between productivity and efficiency has to be done. Productivity is the ratio of the output/input. Efficiency, in the broad sense, is defined as the distance with respect to the frontier of the best performers (Daraio & Simar, 2007). The econometrics of production functions is different than that of production frontiers as the main objective of their analysis differs: production

functions look at average behaviour whilst production frontiers analyse best performers behaviour (Bonaccorsi & Daraio, 2004). Obviously, assessing the impact on the average performance is different than assessing the impact on the best performance. This distinction has been considered also recently in the theory of growth and in the managerial literature. From a methodological perspective, different approaches, both parametric and nonparametric (Badin, Daraio & Simar, 2012; Daraio & Simar, 2014) have been proposed.
On the other hand, classical methods of impact assessment (Bozeman & Melkers, 1993; Khandker et al., 2010) proved inadequate to the checklist of "sensitivity auditing" (Saltelli & Guimarães Pereira; Saltelli & Funtowicz, 2014).

Important role of data

The data dimension is characterized by a kind of "data paradox". On the one hand, we are in a "big data" world, with open data and open repositories that are exponentially increasing. On the other hand, in empirical applications «data constraints» are almost the same as those described in Griliches (1989, 1994).
We believe that a great improvement could come by the adoption of an Ontology-Based-Data-Management (OBDM) Approach (Calvanese et al. 2010; Lenzerini, 2011; Poggi et al., 2008) to integrate the heterogeneous sources of data on which the empirical analysis has to be carried out.

A framework for the analysis

A general framework to investigate and empirically assess the research activity and its impacts is derived integrating existing approaches according to three dimensions. The main building blocks of these dimensions are reported in Figure 1.

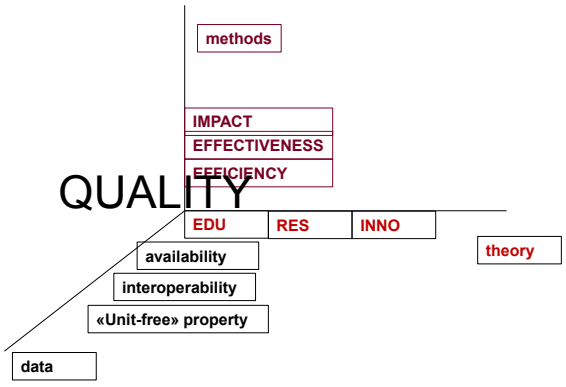


Figure 1. A framework for the analysis of research assessment and its impacts.

We propose "quality" as the overarching concept, which links together all the three dimensions. Quality should be declined along the three dimensions and by each building block. In theory, in education, a lot of progresses have been done. Much more work is needed for research and innovation. If we include quality indicators in the analysis we can move from efficiency to effectiveness. Moreover, it is the quality of education, research and innovation, which has an "impact" on the growth and development of the society. Finally, it is on the data dimension that the quality issues are of primary importance in all the three main building blocks proposed.
If we are not able to conceptualize and formalize in an unambiguous way the different meanings of «quality» for each building block proposed, we will not be able to make a real step forward in the empirical evaluation of the Efficiency, Effectiveness and Impact of Education, Research and Innovation. Third mission indicators (see Bornmann, 2013 for a survey) have a crucial role in this respect. It is indeed the role played by third mission indicators formally conceptualized as a measure of quality of higher education/research institutions, which can be used to investigate the Science-Society interplay.
For the conceptualization and formalization of the «quality» dimensions we suggest to adopt a very different approach based on: 1. Knowledge infrastructure (Edwards et al., 2013); 2. Convergence as «the coming together of insights and approaches from originally distinct fields», which «provides power to think beyond usual paradigms and to approach issues informed by many perspectives instead of few» (National Research Council, 2014).
We need to develop a knowledge infrastructure to model research and innovation and all the activities related to their (economical and societal) impacts in a systemic way. To advance towards an "open science" we have to build a common platform that has to be able to show us which data is relevant for assessing the model we selected for the analysis. In this way, the data could be analysed under different perspectives while sharing the same common conceptual characterization.

Selected References[1]

Aghion, P., David, P. A., & Foray, D. (2009). Science, technology and innovation for economic growth: linking policy research and practice in 'STIG Systems'. *Research Policy*, *38*(4), 681-693.

Bornmann L. (2013), What Is Societal Impact of Research and How Can It Be Assessed? A Literature Survey, JASIST, *64*(2), 217–233.

Daraio C. (2015), What do we know about Efficiency, Effectiveness and Impact of Research and Innovation? *Pro.of Workshop, 20th February DIAG Sapienza University of Rome*, edited by C. Daraio, Efesto Edizioni, Rome, pag. 13-25

Fealing K. H., Lane J. I., Marburger J. H. JIII, & Shipp S. S. (Eds.) (2011), *The Science of Science Policy, A Handbook.* Stanford, USA, Stanford University Press.

[1] The full list of references can be found at the author website.

Integrating Microdata on Higher Education Institutions (HEIs) with Bibliometric and Contextual Variables: A Data Quality Approach

Cinzia Daraio[1], Angelo Gentili[1] and Monica Scannapieco[2]

[1] *daraio@dis.uniroma1.it, angelo_gentili@hotmail.it*
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

[2] *scannapi@istat.it*
Italian National Institute for Statistics (Istat), Rome (Italy)

An introduction on data quality

Data quality has been addressed in different research areas, mainly including statistics, management and computer science. The statistics researchers were the first to investigate some of the problems related to data quality by proposing a mathematical theory for considering duplicates in statistical data sets, in the late 60s. The management research began at the beginning of the 80s; the focus was on how to control data manufacturing systems in order to detect and eliminate data quality problems. Only at the beginning of the 90s, computer science researchers began considering the data quality problem, specifically how to define measure and improve the quality of electronic data, stored in databases, data warehouses and legacy systems. Data quality has been defined as "fitness for use", with a specific emphasis on its subjective nature. Another definition for data quality is "the distance between the data views presented by an information system and the same data in the real world"; such a definition can be seen as an operational definition, although evaluating data quality on the basis of comparison with the real world is a very difficult task.

Data quality is well-recognized as a multidimensional concept including several distinct dimensions (Batini & Scannapieco, 2006) proposed in various contexts (Catarci & Scannapieco, 2002). A crucial dimension of data quality is data accuracy: it measures the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v is intended to represent. However, quality is more than simply data accuracy. Other significant dimensions play a role in the definition of the Data Quality concept, including completeness, consistency, and timeliness (i.e. degree of up-to-dateness), just to cite some significant ones.

Data Quality issues in data integration processes

In a data integration system, sources are typically characterized by various kinds of heterogeneities that can be generally classified into:

(i) Technological heterogeneities.
(ii) Schema-level heterogeneities.
(iii) Instance level heterogeneities.

Technological heterogeneities are due to the use of products by different providers, employed at various layers of an information and communication infrastructure.

Schema heterogeneities are principally caused by the use of (a) different data models, such as one source that adopts a relational data model and a different source that adopts a graph-based data model, and (b) different data representations, such as one source that stores addresses as one single field and another source that stores addresses with separate fields for street, civic number, and city. Schema level heterogeneities can be solved according to well-defined methods that harmonize data collected by the different sources with respect to a schema global to the whole data integration system. However, from a practical perspective, in order to make such harmonization possible it is also necessary to solve (iii) instance level heterogeneities, namely:

For overlapping data sources, same objects can be represented as different due to data quality errors. Hence, in order to resolve such conflicting representations, an object matching activity must be performed. Such activity should be as much automated as possible, especially in complex data integration systems (Zardetto, Scannapieco, Catarci, 2010).

For all sources, also those that are not overlapping, a quality control at instance-level is very useful in order to prevent the possible population of the data integration system with erroneous data. Depending on the specific types of data integration systems, such a quality control can be performed in different ways.

A Data Quality Approach to integrate HEIs microdata in a platform

For a platform supporting European Universities for Education, Research and Technology Studies, on the one hand, the lower level of disaggregation of data makes them more sensible and increases the chances of instance-level errors. On the other hand,

data collection is performed by integrating data already collected by statistical institutions by means of different statistical surveys or administrative data.

Hence, the quality control activity should have the following features:

1. It has to be applied on the overall collected data and cannot be applied to single processes producing data. Monitoring and control of processes producing data can be very useful to prevent quality problems, however, it cannot be applied to our case, due to the different nature of production processes and to the practical impossibility to revise such processes in a preventive fashion. This does not exclude of course the fact that feedbacks deriving from quality analysis could be used by organizations that produce data to revise their production processes.
2. A specific quality activity of outlier detection could be applied, by comparing data provided by "similar" sources on the same subject. Here, "similar" could mean, for instance, belonging to the same country and with analogous features such as the number of personnel. Data that are recognized as outlier by automated procedures should subsequently undergo a human analysis. This analysis can either explain the outlier on the basis of available context information, or it can recognize that the outlier is actually caused by quality problems. In this latter case, quality improvement actions must be engaged.

The following Table 1 illustrates the main sources of data which have been integrated to test the data quality approach proposed in the paper.

Figure 1 instead shows an example of outliers detection carried out through a systematic check against different distributions. The check has been done on the ratios given by number of publications divided by the number of academic staff, for all European universities in the sample.

Table 1. Main sources of data integrated

Source (link)	Description
ETER (www.eter.joanneum.at/ imdas-eter/) integrated with data from HESA for UK	Microdata on inputs outputs of higher education institutions in Europe.
Scimago Institutions Rankings (www.scimagoir.com)	Bibliometric data on scientific production and impact.
Eurostat (http://ec.europa.eu/eurostat)	Contextual factors, data at territorial level on economic and social development.

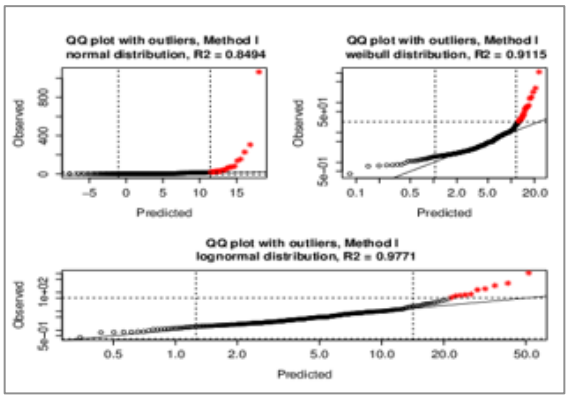


Figure 1. An example of outliers detection. Outliers are reported as stars in red: the graph top left shows outliers with respect to the normal distribution (worst fit, r-square=0.85), the one top right with respect to the Weibull distribution (r-square=0.91), the one below with respect to the lognormal distribution with the highest fit (r-square=0.98).

References

Batini, C., & Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies, and Techniques,* Springer, The Netherlands.

Catarci, T. & Scannapieco, M. (2002). Data Quality under the Computer Science Perspective. Journal of Archivi & Computer, 2.

Lepori, B., Daraio, C., Bonaccorsi, A., Daraio, A., Scannapieco, M., Gunnes, H., Hovdhaugen, E., Ploder, M. & D. Wagner-Schuster (2014), 'ETER Project, Handbook for Data Collection', Brussels, June.

Luwel, M. (2005). The use of input data in the performance analysis of R&D systems. In *Handbook of Quantitative Science and Technology Research* (pp. 315-338). Springer Netherlands.

Luwel, M. (2015), Heterogeneity of data in research assessment, in *Efficiency, Effectiveness and Impact of Research and Innovation*, Proceedings of the Workshop of the 20th February 2015, C. Daraio (ed), DIAG Sapienza University of Rome, Efesto Edizioni Rome.

Zardetto, D., Scannapieco, M., & Catarci, T. (2010). Effective Automated Object Matching. *Proceedings of the International Conference on Data Engineering (ICDE 2010)*.

Is the Humboldtian university model an engine of local development? New empirical evidence from the ETER database

Teresa Ciorciaro[1], Libero Cornacchione[1], Cinzia Daraio[1], Giulia Dionisio[1]

[1]*teresa.ciorciaro@gmail.com,* [1]*lillo-1991@libero.it,* [1]*daraio@dis.uniroma1.it,* [1]*giulia.dionisio@hotmail.it*
Department of Computer, Control and Management Engineering Antonio Ruberti, Sapienza University of Rome, via Ariosto, 25 00185 Rome (Italy)

Introduction

The higher education system, in advanced countries, has reached the point of massification (i.e. enrolment rates exceeding 50% of the relevant age cohort), while the public budget has not grown correspondingly. Universities are put under pressure to use existing resources, namely staff and funding, in the most efficient way. At the same time there is an increased pressure from the research side: the expectations of society and policy makers on the contribution of research to societal problems have grown significantly, there are new entrants in scientific arena (particularly from Asia) and the competition for funding has increased sharply. This situation creates a classical issue in public policy: we have two valuable goals (serving better mass educational needs and producing good research) between which there is tension or trade-off.

Do universities benefit from having inputs (staff and funding) that can produce jointly teaching and research, or there are efficiency-enhancing specialization effects that suggest to keep these activities under separate institutions? What is the impact of the environmental context of the universities? We focus here on the complementarity between teaching and research, which is at the core of the Humboldtian model of university (Schimank & Winnes, 2000). Is the traditional Humboldtian model of university, in which teaching and research are produced jointly by the same academic staff able to foster the economic development of the area in which the university is located? What are the main contextual factors which affect the performance of the European Humboldtian universities?

Purpose of the analysis and method

The main objective of this paper is to investigate the determinants of the efficiency scores of European universities, whose production is characterized by teaching and research outputs.

In efficiency analysis, nonparametric estimators are particularly attractive because they do not rely on restrictive parametric assumptions on the process that generates the data.

We apply a nonparametric approach, DEA (Data Envelopment Analysis, Charnes et al., 1978), which allows for multi-input - multi-output analyses, followed by a bootstrap analysis to estimate bias corrected efficiency scores and to provide confidence intervals on the efficiency scores. Given that universities in Europe face heterogeneous conditions, in a second step, we applied a semiparametric bootstrap-based approach (Simar & Wilson, 2007) to assess the statistical significance of external contextual factors on their performance.

Data and variables

Our sample is composed by 753 HEIs (Higher Education Institutions) belonging to 22 different European countries.

In the following tables we present the data analysed, the inputs, the outputs and the external factors investigated in the paper.

Table1. Data.

Data Source	Description
SCIMAGO INSTITUTION RANKING	The SIR purpose is a characterization of institutions, based on three different ranges: research, innovation and web visibility. This source uses normalized indicators, in a scale from 0 to 100, to facilitate the comparison between the institutions. The SIR database provides some bibliometric indicators for each institution, like number of publications, high quality publications, normalized impact, international collaboration and specialization index.
ETER	The European Tertiary Education Register wants to build a complete register of higher education institutions. Its database gives various information, like number of students, professors, graduates, doctorates, total incomes and expenditures. This register is developed by the Directorate General for Education and Culture of the European Commission.
EUROSTAT database	The EUROSTAT database wants to be the leading provider of high quality statistics on Europe. It contains regional data at a very disaggregated level.

Table2. Selected inputs

Input	Formula
Teaching	$\frac{\#\text{ of academic staff}}{\#\text{ of students}} * 100$
Structural	$\frac{\#\text{ of administrative staff}}{\#\text{ of students} + \#\text{ of academic staff}}$
Research	$\frac{\#\text{ of graduates at ISCED 8}}{\#\text{ of undergraduates enrolled}}$

Table 3. Selected outputs.

Output	Formula
Teaching	$\frac{\#\text{ of graduates}}{\#\text{ of students enrolled}}$
Research	$\frac{\text{output (pub)} * \text{HQP(\% high quality pub)}}{100 * (\#\text{ of academic staff} + \#\text{of graduates at ISCED 8})}$
Third mission	Percentage of third party funding.

Table 4. Selected External factors.

External factor	Description
GDP	Gross domestic product at current market prices
PAT	Patent applications
HOSP	Hospital yes/no
ER	Employment rates- age group 20-64
GERD	Total intramural R&D expenditure (GERD) at NUTS 2 level
SIZE	Size
AGE	No. of years from foundation

Modelling strategy

We estimate several partial models, i.e. models of single output production (teaching model, research model, third mission model) as well as complete models (of joint production of teaching and research, including also the third mission dimension) to analyse how the evaluation of the impact of external factors affects the production of the considered universities.

A correlation analysis is carried out to analyse the degree of association of the obtained efficiency scores with the degree of internationalization of the considered universities to account for recent results that show that is the quality of the academic staff that plays an important role to facilitate and faster third stream activities as complement of teaching and research missions.

Preliminary results and next steps

Figure 1 reports some illustrative preliminary results of the two-stage analysis conducted on the dataset.

We are going to extend the analysis in the following directions:

1. Inclusion of other third mission indicators in the input-output characterization (Geuna & Rossi, 2015), to investigate how their inclusion affects the impact of the considered external factors.
2. Apply robust nonparametric approaches (Daraio & Simar, 2007) which do not rely on the separability condition assumed by the two stage approach applied in this paper, and are more robust to outliers and extremes in the dataset as well as more flexible directional distance models (Daraio & Simar, 2014; Daraio et al., 2015a,b).

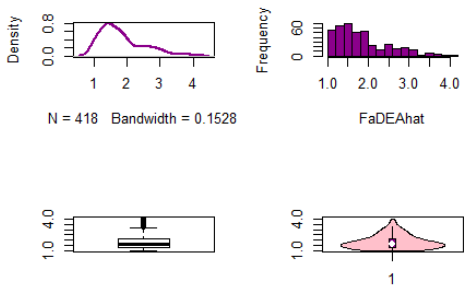


Figure1. Distribution of the European efficiency scores. Top left panel: nonparametric kernel density distribution, top right panel: histogram, bottom left panel: box plot and bottom right panel: violin plot.

Some Selected References[1]

Daraio, C., et al. (2011). The European university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy*, *40*, 148

Daraio, C., Bonaccorsi, A., & Simar, L. (2015a). Efficiency and economies of scale and specialization in European universities: A directional distance approach, *Journal of Informetrics*, *9*, 430-448.

Daraio, C., Bonaccorsi, A., & Simar, L. (2015b). Rankings and University Performance: a Conditional Multidimensional Approach, *European Journal of Operational Research, 244*, 918-930.

Daraio, C., & Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis. Methodology and applications*, Springer, New York.

Daraio, C., & Simar, L. (2014). Directional distances and their robust versions: Computational and testing issues. *European Journal of Operational Research*, 237, 358–369.

[1] See authors' webpage for a full list of references, which are removed due to space limitations.

Connecting Big Scholarly Data with Science of Science Policy: An Ontology-Based-Data-Management (OBDM) Approach

Cinzia Daraio[1], Maurizio Lenzerini[1], Claudio Leporelli[1], Henk F. Moed[1], Paolo Naggar[2], Andrea Bonaccorsi[3], Alessandro Bartolucci[2]

daraio@dis.uniroma1.it, lenzerini@dis.uniroma1.it, leporelli@dis.uniroma1.it, henk.moed@uniroma1.it
[1] DIAG, Sapienza University of Rome, via Ariosto, 25, Rome (Italy)

paolo.naggar@gmail.com,alessandro_bartolucci@fastwebnet.it
[2] Studiare Ltd.

a.bonaccorsi@gmail.com
[3] DESTEC, University of Pisa (Italy)

The OBDM approach in a nutshell

The key idea of OBDM is to resort to a three-level architecture, constituted by the ontology, the sources, and the mapping between the two. The ontology is a conceptual, formal description of the domain of interest to a given organization (or, a community of users), expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge. The data sources are the repositories accessible by the organization where data concerning the domain are stored. In the general case, such repositories are numerous, heterogeneous, each one managed and maintained independently from the others. The mapping is a precise specification of the correspondence between the data contained in the data sources and the elements of the ontology.

The main purpose of an OBDM system is to allow information consumers to query the data using the elements in the ontology as predicates. In this sense, OBDM is a form of information integration, where the conceptual model of the application domain, formulated as an ontology expressed in a logic-based language, replaces the usual global schema. The integrated view that the system provides to information consumers is not merely a data structure accommodating the various data at the sources, but becomes a semantically rich description of the relevant concepts in the domain of interest, as well as the relationships between such concepts.

Sapientia: a Platform for Developing Science of Science's Policy Models

We consider the building of descriptive, interpretative, and policy models of our domain as a distinct step with respect to the building of the domain ontology. The ontology will intermediate the use of data in the modelling step, and should be rich enough to allow the analyst the freedom to define any model she considers useful to pursue her analytic goal.

Obviously, the actual availability of relevant data will constrain both the mapping of data sources on the ontology, and the actual computation of model variables and indicators of the conceptual model. However, the analyst should not refrain from proposing the models that she considers the best suited for her purposes, and to express, using the ontology, the quality requirements, the logical, and the functional specification for her ideal model variables and indicators. This approach has many merits, and in particular:

- it permits the use of a common and stable ontology as a platform for different models;
- it addresses the efforts to enrich data sources, and verify their quality;
- it makes transparent and traceable the process of approximation of variables and models when the available data are less than ideal;
- it makes use of every source at the best level of aggregation, usually the atomic one (see examples in the following).

In this framework, exploratory data analysis, and the building of synthetic indicators, are only an intermediate step of the modelling effort that aims to the interpretation of behaviours, the explanation of differences in performance, the identification of causal chains of phenomena. That leads to the development of a policy-design model, whose inputs are policy instruments, and whose outputs are performance indicators for research activities and economic welfare.

The learning and theory building process requires feedbacks that could also concern the ontology level: the addition of new concepts and data, through the specialization of general concepts or the enlargement of the ontology commitment, could reflect the intermediate achievements of the

learning process such as the necessity of improvement of the theories submitted to test.

More often, however, a well-conceived ontology will resist to the competency test implied by new model and theories, and the most serious constraint to model development will be the impossibility of a complete mapping between the ontology and the sources, i.e. the lack of data. This is a negative result only for the short-term. In the medium and long term, the dialogue within the community of researchers that use the ontology as a workbench will result in a joint effort towards other stakeholders in order to improve detail, quality, and scope of data collection. Moreover, the shared use of logically sound definition for indicators increase the ability of the analysts to compare their studies and to test old and new theories.

Consider as an example the important issue of the assessment of the effects of scale economies on the performance of a research institution and of its affiliates. The results can widely differ if you set the analysis at different levels of aggregation: all the public research and education institutions of single countries, single universities, faculties, let's say, of Science and Technology, departments of Computer Science, research groups, or individuals within these groups.

Moreover, at different aggregation levels, the possible moderating variables or causes of different performances can widely differ. Legislation and regulation, public funding, teaching fees and duties matter at national level. Geography, characteristics of the local economic and cultural system, effectiveness of research and recruiting strategy, budgeting, infrastructures matter at the university or department level. Intellectual ability of researchers, history and stability of the group, ability to recruit doctoral students, worldwide network of contacts matter at the research groups and individuals level.

Time is a crucial dimension of research modelling. We pursue a modelling approach based on processes, i.e. collections of activities performed by agents through time. To represent the knowledge production activities, at an atomic level, we consider both stock inputs such as the cumulated results of previous research activities (those available in relevant publications, and those embodied in the authors' competences and potential), the infrastructure assets, and flow inputs as the time devoted by the group of authors to current research projects. Similarly, we can analyze the output of teaching activities, considering the joint effect of resources such as the competence of teachers, the skills and the initial education of students, and educational infrastructures and resources. Thirdly, service activities of research and teaching institutions provide infrastructural and knowledge assets that act as resources in the assessment of the impact of those institutions on the innovation of the economic system. The perimeter of our domain should allow us to consider the different channels of transmission of that impact: mobility of researchers, career of alumni, applied research contracts, joint use of infrastructures, and so on. In this context, different theories and models of the system of knowledge production could be developed and tested.

Conclusions

To bridge the gaps existing in the literature, and to integrate existing bottom-up initiatives in a coherent theoretical-based platform, we suggest an OBDM approach.

We need a change in the overall approach to the assessment of science and technology: metrics and indicators can have negative effects on the scientific community because they encourage a reductionist philosophy; on the contrary, we propose using well-defined concepts and data to build interpretative models, in order to compare and discuss theories. That can be useful both to promote a pluralistic community of analysts, and to build consensus on less superficial evaluation procedures of researchers and institutions. Moreover, indicators are often produced in closed circles, collecting ad hoc databases, with no built-in interoperability, updating and scalability features. We have to move towards an environment in which data are publicly available, collected and maintained on stable platforms, where ontologies give confidence on the precise meaning of data to people that propose models and to those that evaluate them. These repositories of knowledge can evolve following the analytical needs of the research community and the policy institutions, instead of starting from scratch each time a new research project starts. We propose our *Sapientia* ontology as a starting point to be opened, shared with the community and further developed and integrated with existing bottom-up initiatives as well as with new theories and paradigms.

Acknowledgments

Research support from the Progetto di Ateneo 2013 of the Sapienza University of Rome is gratefully acknowledged.

References

Daraio C., Lenzerini M., Leporelli C., Moed H.F., Naggar P., Bonaccorsi A. & Bartolucci A. (2015). Sapientia the Ontology of Multi-Dimensional Research Assessment*, Proc. ISSI*.

Fealing K. H., Lane J. I., Marburger J. H. JIII, & Shipp S. S. (Eds.) (2011), *The Science of Science Policy, A Handbook.* Stanford University Press.

Lenzerini M. 2011. Ontology-based data management, *CIKM 2011*: 5-6.

Poggi A., D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, & R. Rosati. (2008). Linking data to ontologies. *Journal on Data Semantics*, X: 133–173.

Incomplete Data and Technological Progress in Energy Storage Technologies

Sertaç Oruç[1], Scott W. Cunningham[1], Christopher Davis[2], Bert van Dorp[1]

[1] *s.oruc@tudelft.nl, s.cunningham@tudelft.nl, bertvandorp@gmail.com*
[1]Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5 C2.010, 2628 BX Delft (The Netherlands)

[2] *c.b.davis@rug.nl*
University of Groningen, Center for Energy and Environmental Sciences (IVEM), Nijenborgh 4, 9747AG Groningen (The Netherlands)

Abstract
Energy storage is an important topic as many countries are seeking to increase the amount of electricity generation from renewable sources. An open and accessible online database on energy storage technologies was created, incorporating a total of 18 energy storage technologies and 134 technology pages with a total of over 1,800 properties. In this database information on technical maturity, technology readiness level and forecasting is included for a number of technologies. However, since the data depends on various sources, it is far from complete and fairly unstructured. The chief challenge in managing unstructured data is understanding similarities between technologies. This in turn requires techniques for analyzing local structures in high dimensional data. This paper approaches the problem through the use and extension of t-stochastic neighborhood embedding (t-SNE). t-SNE embeds data that originally lies in a high dimensional space in a lower dimensional space, while preserving characteristic properties. In this paper, the authors extend the t-SNE technique with an expectation-maximization method to manage incompleteness in the data. Furthermore, the authors identify some technology frontiers and demonstrate and discuss design trade-offs and design voids in the progress of energy storage technologies.

Conference Topic
Mapping and visualization

Introduction
High dimensional datasets are difficult to visualize contrary to two or three dimensional data, which can be plotted comparatively easily to demonstrate the inherent structure of the data. To aid visualization of the structure of a dataset, a family of algorithms have been devised in the literature, which are collectively referred as dimensionality reduction algorithms, of which an extensive review can be found in (van der Maaten, Postma, & van den Herik, 2009).

Among these algorithms *t-stochastic neighborhood embedding* (t-SNE) is a novel machine learning technique that has burgeoning applications. t-SNE maps each data point in a given high-dimensional space to a low-dimensional space, typically to a two or three dimensional one, for visualization purposes. The algorithm does a non-linear mapping such that similar points in the high-dimensional space situated nearby each other in the low-dimensional space as well.

In its first stage, the algorithm constructs a probability distribution over pairs of high-dimensional points in such a way that similar points have a high probability of being picked. In the second stage, it constructs the same probabilities between these points in the low-dimensional space. Finally the algorithm minimizes the difference between these probabilities by minimizing Kullback-Leibler divergence between these two distributions (Van der Maaten & Hinton, 2008).

Inherently, the algorithm preserves the manifold that possibly exist in the high-dimensional data and represents this manifold in low-dimensional space. Indeed, this class of dimensionality reduction algorithms is called "manifold learning". In comparison to the more

conventional, linear dimensionality reduction techniques such as *principal component analysis* (PCA), which finds a linear mapping with an objective to find a subspace where the projection of each data point lies as close to the original point as possible, manifold learning algorithms preserve the distance between pairs of points. Because of this the manifolds are preserved as well, whereas with PCA, clusters that are far from each other in high-dimensional space might be merged in low dimensional space.

t-SNE also proves to be useful for technology analysts in monitoring target technologies. Technologies such as batteries and storage, which is the target technology in this article, have multiple characteristics that develop over time. The problem facing the analysts is that most modern data sources are unstructured in character. Unstructured data often indicates that the data is of mixed provenance and quality. Furthermore, readily available data is often a mix of actual performance results, and forecasts of potential future results. Even when performance data is available the data is rarely standardized, and therefore contains incomplete and uncertain data.

Table 1. List of technologies in the database.

Compressed Air Energy Storage (CAES)	Nickel–cadmium (NiCd) battery
Edison (NiFe) battery	Nickel-metal hydride (NiMh) battery
Flow batteries	Nickel–zinc (NiZn) battery
Flywheels	Pumped Hydro
Hydrogen storage	Saltwater (sodium-ion) batteries
Lead-acid battery	Sodium-sulfur (NaS) battery
Lithium–air (Li-air) battery	Supercapacitors
Lithium-ion (Li-ion) battery	Superconducting magnetic energy storage
Lithium–sulfur (Li-S) battery	Zinc-air battery

Table Table 1 shows typical sources used in appraising technological development. The data varies by provenance – it is provided through a mix of academic, commercial, government, non-profit and media organizations. Furthermore, the data itself pertains to technologies at different stages of development, and in different modes of deployment or development. An exemplary data source, discussed in the next section, compiles research and development data concerning storage and battery technologies.

Despite the mixed quality of the data sources, such data is useful and should be incorporated into quantitative analyses. In this paper we are primarily concerned with technometric approaches to modelling technology (Coccia, 2005). In particular we are concerned with utilizing such data to produce technological frontiers. Such frontiers are useful for anticipating the future rate of growth, and can be used for developing coordination mechanisms such as technology roadmaps (Phaal, Farrukh, & Probert, 2004).

Evidence and belief need not be mutually incompatible. Bayesian statistical techniques acknowledge that data is often collected in an open, rather than controlled, experimental framework (Gill, 2004). As a result the necessity for belief prevails in the collection of data. There are beliefs concerning the quality of data, the underlying system relationships, and the nature and number of underlying cases to be measured. What is significant then is that prior beliefs concerning the data are acknowledged, that these beliefs actually encompass the true state of the world, and that these beliefs are consistently updated in light of new data. These are requirements which are achievable given the appropriate collection, treatment, and handling of mixed data.

What is required therefore is a technique for handling complexly structured data, for judging cases and similarities, and for managing incomplete data. This paper approaches the problem through the use and extension of *t-stochastic neighborhood embedding* (t-SNE). The technique is used to develop a non-linear manifold of technological performance, and to use this manifold to manage incompleteness in the data. This builds on a long-established technique for handing missing data known as the expectation-maximization procedure (Dempster, Laird, & Rubin, 1977). In the next section, the paper details a database of storage and battery technologies. In the subsequent section, a method is proposed for dealing with this semi-structured data, and in specific, for dealing with uncertain and incomplete technological information.

Data Sources

This work builds upon data collected from Enipedia,[1] a website that collects, organizes and visualizes open data related to energy systems. One of the initiatives on the website has focused on gathering information related to energy storage technologies.

Energy storage is an important topic as many countries are seeking to increase the amount of electricity generation from renewable sources. An issue with renewable energy is that the amount of generation is often variable and can exceed or fall short of the amount that is demanded. If there is an excess of production, then not all of the electricity can be fed into the grid. If there is an undersupply, then power plants relying on fossil-fuels must often be relied on in order to help meet demand. To address this variability, large-scale energy storage could be used to store energy during periods of excess renewable electricity production, and then supply this energy during periods of increased demand.

A key problem is that large-scale energy storage does not currently exist, aside from pumped-storage hydroelectricity plants which can only be built in locations with suitable geography. The development of economically feasible large-scale energy storage technologies will be a major game changer in the energy sector as it can support a larger integration of renewables and decrease reliability on electricity generation from fossil sources.

The research indicated that a number of energy scenarios and simulations fail to include models on energy storage, and lack accurate data on technologies. Also, forecasting is often not included, while battery technologies and costs are rapidly evolving. By these needs, an accessible and open technology database was created, incorporating a total of 18 energy storage technologies and 134 facilities or technology pages with a total of over 1,800 properties. In this database,[2] information on technical maturity, technology readiness level and forecasting is included for a number of technologies.

An overview of sources of technology information on the potential and future demand for energy storage indicates that a number of technologies and solutions focus on applications with small time-scales, such as frequency and voltage control, load shifting, diurnal storage, output smoothing, mobility and reserve grid capacity. Far few technologies and facilities focus on providing seasonal and large-scale grid storage. For a number of these technologies, installations with a lower technology readiness level have been included to provide some numbers on feasibility.

Developing metrics on comparing these technologies was done through an iterative design scheme, incorporating metrics relevant to a range of applications. It was observed that a number of technologies cannot be described fully, as information is missing or the ranges in which information sources report the information are exceptionally wide. Also, the definitions found for some technologies, such as Li-ion, are weaker than those found for other

[1] http://enipedia.tudelft.nl

[2] http://enipedia.tudelft.nl/wiki/Electricity_Storage

technologies. Furthermore, metrics are often made available on a systems level, and information on other levels needs to be translated to this system level.

Table 2. Variable number, name and description

No.	*Variable Name*	*Description*
1	Case	Case number
2	Product	Product name
3	Technology	Technology type
4	Year	Reference year
5	Institutional Data	Indicator whether observation is institutional
6	Technology Readiness Level[3]	Technology maturity level
7	Investment per Unit Power	Investment unit power (EUR/KW)
8	Investment per Unit Energy	Investment cost per unit energy (EUR/KWh)
9	Efficiency	Energy efficiency
10	Cycles	Life span in cycle times
11	Energy Density	Energy density (WH/L)
12	Power Density	Power density (WH/Kg)
13	LCoE[4]	Levelized cost of energy

Method

The chief challenge in managing unstructured data is understanding similarities between technologies. This in turn requires techniques for analysing local structures in high dimensional data. The technique of choice for this is t-stochastic neighborhood embedding (van der Maaten & Hinton, 2008). Finding a manifold which represents the data is useful for developing lower dimensional representations of the data. Such a manifold is inherently non-linear, and by necessity it preserves the local structures in the data at the expense of finding any global structures which might be present. For this analysis we adopt an implementation of the algorithm created in Matlab (van der Maaten, 2007).

The t-SNE technique has previously been used in technometrics. Cunningham and Kwakkel (2014) investigate a case of electric vehicle and hybrid electric vehicle designs and technologies. The case benefitted from the use of a non-linear fitting technique since the designs considered differ substantially in fundaments. As a result different designs highlight fundamentally distinct kinds of engineering trade-offs. The case also demonstrated a potential convergence across multiple technologies. Other patterns of technological evolution on a manifold, in addition to convergence, are identified in the paper.

Other technometric approaches utilize a linear, or quasi-linear technological frontier. Many of these approaches also assume a constant rate of technological change as the frontier advances over time. These alternative approaches are useful for single technologies with well-understood morphologies. Such techniques are also suitable for technologies where there are suitable indicators of performance, outcome, or merit. The techniques are less useful for analyzing broader fields with a heterogeneous base of technology. In such fields different technological trade-offs may be at work, and the pace of technological change may be discontinuous or punctuated. Indeed, the technologies themselves each may be valued for different purposes and outcomes.

[3] http://en.wikipedia.org/wiki/Technology_readiness_level
[4] http://en.wikipedia.org/wiki/Cost_of_electricity_by_source

A desirable method must be suitable for use with diverse data types. Before applying t-SNE to the data set of Table 2, the data is first transformed and normalized. Transforming the data eases a search for locally similar data points. Furthermore, the normalization of the data helps address difficulties associated with variables being measured in different units, potentially highly discrepant in scale. The choice is made to take the logarithm of the data whenever the data is right skewed. Logistic transformation is used to create more normal-like distributions than the actual.

As previously noted, a major challenge in addressing such data sets is the presence of missing data. The principle technique for handling missing data in the statistical literature is known as the expectation-maximization procedure. This powerful technique has been extended to address the estimation of missing model parameters, as well as missing data, and later become a mainstay of machine learning techniques. Modern machine learning procedures are now availed of much faster algorithms than expectation-maximization procedures; nonetheless the technique has had a powerful effect on the field.

The expectation-maximization procedure consists of two steps. In the first, or expectation step, the missing data is replaced with an expected value. Initially the expected value can be set by the mean of the data, or even by replacing the missing data with random values. Then in the maximization step, a model of the data is selected and applied. After an initial modeling step, further estimates of expected values derived from the model can be derived. These expected values become new expected values for additional rounds of the modelling procedure. After repeated cycles of expectation and maximization the estimated values converge, and the full model of the data is derived. The technique has the benefit of replacing missing values with neutral values consistent with an assumed model of the data. The technique therefore makes the best use of available data that is possible, rather than excluding whole variables or cases because they are incomplete.

Unstructured data in this domain is not just incomplete, but also uncertain. This is expressed with reported ranges of expected performance data. In order to treat this data, an upper bound and a lower bound on the data is reported, using two distinct model variables. When the data is certain, the upper and lower bound of the variable is identical. In subsequent model runs a constraint is imposed on the expectation maximization procedure – the maximum estimated upper bound on missing data must be greater than the lower bound. When estimated variables do not satisfy this criteria they are either not updated, or both the upper and lower bounds are replaced with averages.

Every point on the manifold estimated by t-SNE is associated with a potential technological design. Thus the t-SNE model is generative – it reports the expected best fit to the data, and also anticipates new cases or designs which have not yet been reported. Nonetheless, technological constraints or other factors may mean that parts of the manifold are not populated with new designs. Interpolation using the manifold can proceed following two directions. A locally linear direction of change can be interpolated from the data given specific examples or cases. Or, a weighted average of surrounding points can be used given their relative proximity on the technological manifold.

Analysis

The following section details a complete procedure for analysis, as depicted in Figure 1Figure. The procedure begins with preprocessing the data. The raw data includes lower and upper bounds for various attributes. Thus, we made a choice to create two different features for each of such variables, e.g., both “Energy density lower bound” and “Energy density upper bound” features.

The next step identifies and masks out the missing data. The process is facilitated by the use of data structures (for instance in Python or Matlab) where the missing data is identified using

indicator values. A data matrix therefore contains two layers – the first layer stores the data itself, and the second layer contains a bit matrix for masking. The bit matrix indicates where the data is complete or non-missing, or incomplete and missing.

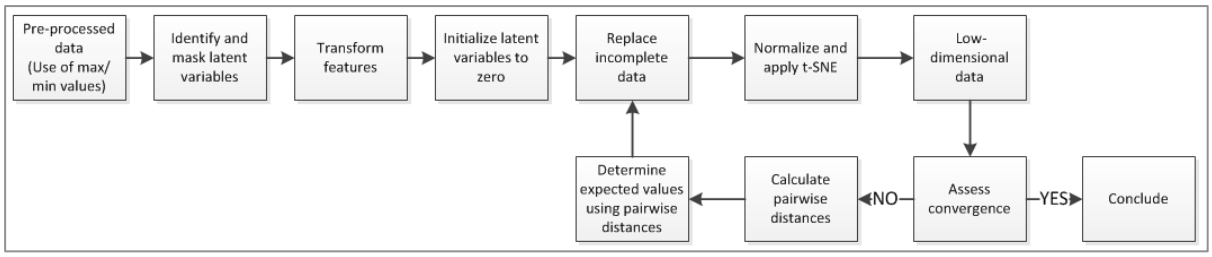


Figure 1. A Flow Chart of the Analysis Procedure.

Then the features are transformed and normalized to normal-like distributions. The following state initializes the missing variables to zero, which is in effect the mean of the normalized features. In subsequent iterations of the algorithm more refined estimates of the missing data are made. This brings us through the initialization and the first maximization step of the algorithm.

The data is complete, and can now be fitted using the t-SNE algorithm. The major output of the algorithm is a set of coordinates for all the cases – in this example there were 118 points. Intermediate outputs, such as data coordinates and scatter plots are produced.

Next, convergence of the algorithm is tested by comparing the current imputed high dimensional representation to the high dimensional representation of the previous iteration. Obviously this step is skipped for the first iteration.

If the algorithm has not converged, then pair-wise similarities between the points are evaluated as the next procedure. The purpose of this comparison is to determine the closest peers of any given technology. The basis for this comparison is the Euclidean distance between two points in the three-dimensional space as output from the t-SNE algorithm. The distance is then scaled according to the negative exponential of the squared distance between the two points. The total distance is then re-scaled to sum to 100% percent to create weightings for updating the originally missing variables in the data. The idea here is to calculate the new values for the missing data such that these values are closer to the related data points implied by the low dimensional data. Using pair-wise distances, a new expected set of values is established and finally the high dimensional representation is updated. The model converges when there is negligible differences between the consecutive imputed high dimensional representations.

Results and Visualization

This section discusses some results of the t-SNE analysis, visualizes and interprets some of the results, instead of all, due to space limitations, and displays the technologies according to their respective dates of introduction or their forecasted date of introduction. These colors suggest that the frontier of technological performance is gradually moving outward (to the upper right) over time. This is further illustrated in Figure 3.

Technological development, at least as measured by year of introduction is a somewhat noisy variable. Nonetheless, in Figure 3, we can qualitatively place three frontier lines. The first is dated 10 1985, the second to 2010, and the third to 2035. It seems plausible given the figure that the rate of technological change is higher among battery technologies than it is among storage technologies. This is demonstrated by the comparative “fanning out” of the battery technologies over time.

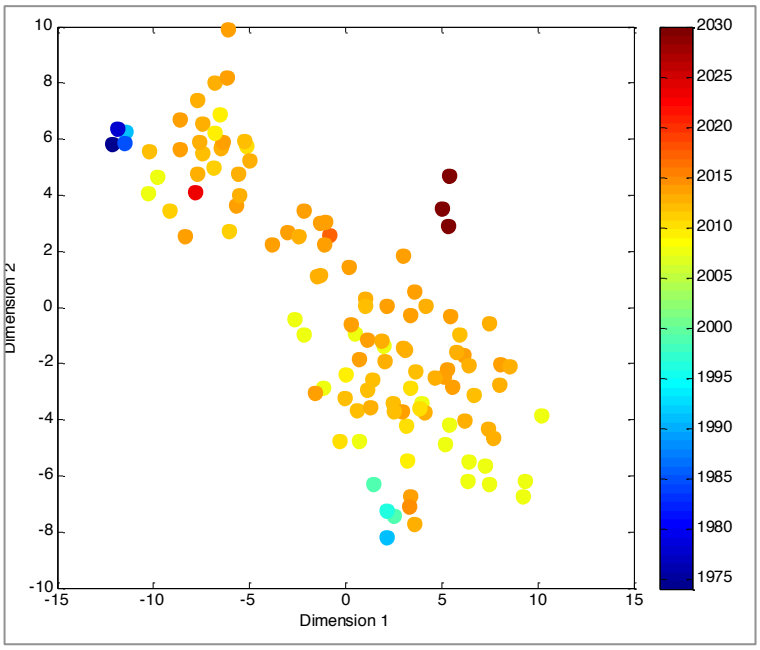


Figure 2. Technologies Positioned by t-SNE and Colored by Date of Introduction

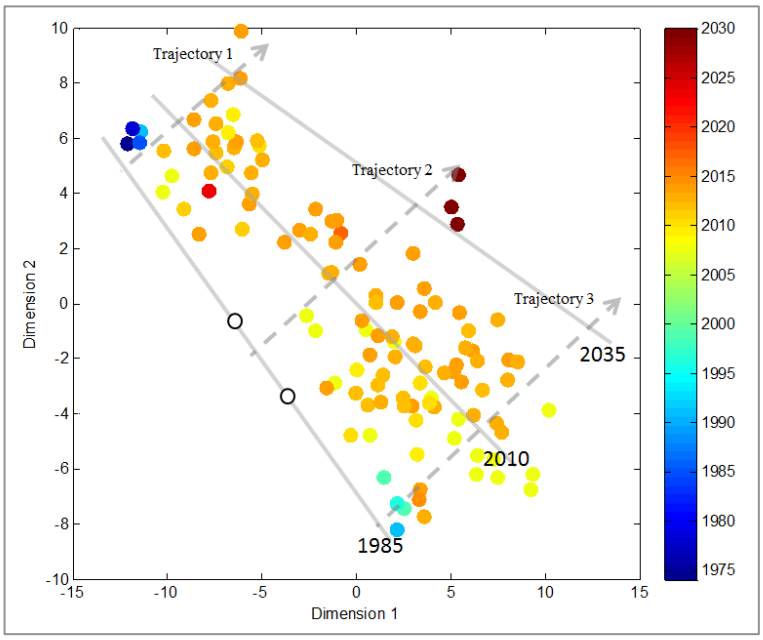


Figure 1. Technological Trajectories

In Figure 3 three technological trajectories are displayed. Changes in technological performance, based on benchmark technologies on or near the trajectory are calibrated. Then the three trajectories are compared with one another to determine whether there are common elements of change across the trajectories.

Figure 4 describes a potential trade-off in the design and selection of battery and storage technologies. In general the trade-off is between the respective cost and advantages of storage technologies versus batteries. Storage technologies are more robust, providing more cycles of operation at a lower levelized cost of energy. This comes at the cost of having a lower energy density, a lower technology readiness level, and a lower efficiency. In contrast battery technologies offer more energy density, are more readily available on the market, and operate

at a higher level of efficiency. In consequence, batteries are less robust, operating for fewer cycles, and requires a higher levelized cost of energy to be paid out.

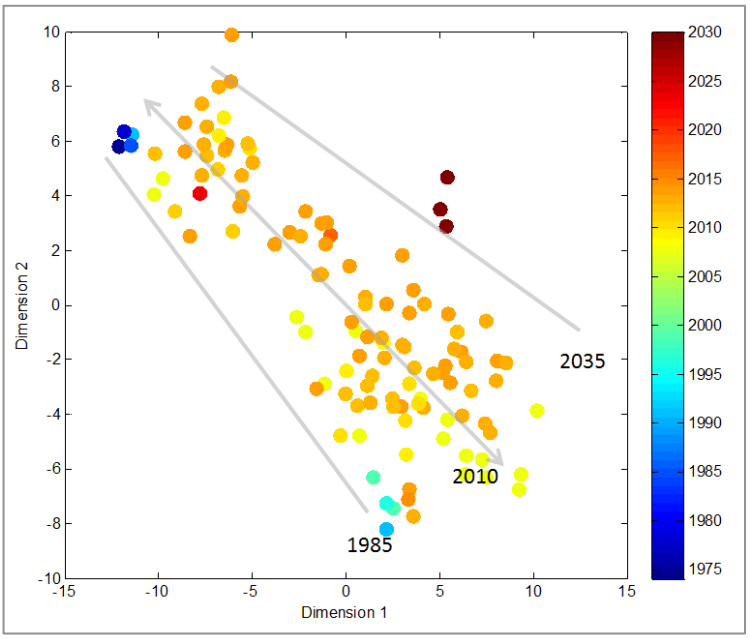


Figure 4. Design Trade-Offs.

There are three design voids on the manifold as shown in Figure 5. These are areas in the space of potential design which have not been explored. One space, design void 1, occurs along the 1985 technological frontier. The space is sparsely explored, although by 2010 a flywheel technology has emerged to occupy the space. The next two voids lie along the 2035 frontier. Because we are not yet on the 2035 frontier, these voids may be unanticipated breakthroughs. Design void 2 is in the space of high performing storage systems, and design void 3 is in the space of high performing batteries. One organization, EASE, anticipates a number of 2030 battery technologies on or beyond this frontier.

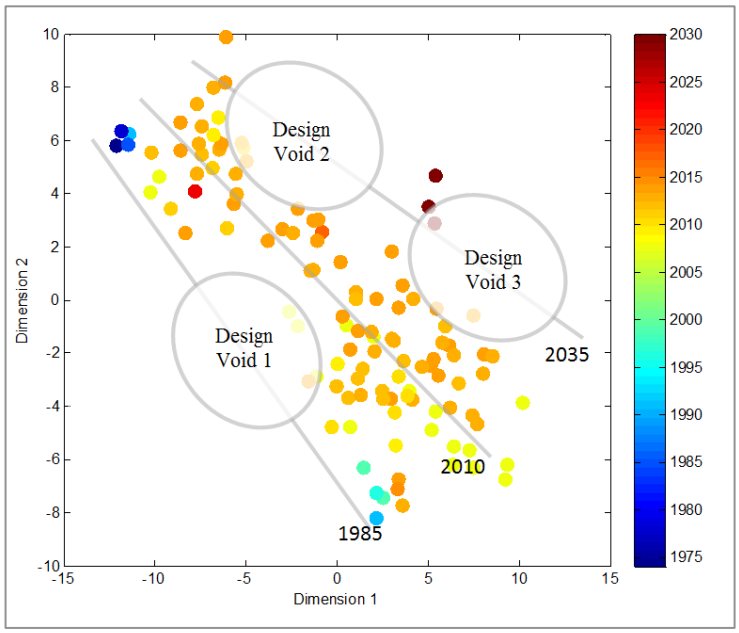


Figure 5. Design Voids.

Table 1. Historical and Emerging Designs.

	Void 1	*Void 2*	*Void 3*
Year	2013	2012	2030
InstitutionalData	0.01	0.79	0.99
TRL	8	6	9
Investment lowerbound	1,093	69	103
Investment upperbound	1,149	131	147
InvestmentEURperKW lowerbound	1,244	729	574
InvestmentEURperKW upperbound	1,262	1549	898
Efficiency lowerbound	0.767	0.709	0.785
Efficiency upperbound	0.849	0.809	0.847
Cycles lowerbound	4,265	11,306	3456
Cycles upperound	4,554	70,551	9804
EnergyDensity lowerbound	40	5	105
EnergyDensity upperbound	60	11	186
Power Density lowerbound	131	82	158
PowerDensity upperbound	220	210	295
LCoE lowerbound	0.149	0.074	0.056
LCoE upperbound	0.506	0.224	0.123

Table 3 provides, by interpolation, the performance characteristics of the technologies in the three voids mentioned previously. The exemplary void 1 technology is most likely a battery. The year of introduction suggests that there have been too few lower technology exemplars, so that the performance here is likely highly overstated. There should likely be a lower power and energy densities, and a lower levelized cost of energy. The closest existing technology is the "Wemag AG Li-Mn storage plant."

The void 2 technology, likely a storage device, should afford dramatically reduced investment and investment per kilowatt hour over previous technologies. The cycle times should be up to an order of magnitude higher than the void 1 exempla. While the power density may not be affected much from its 1985 peer, the energy density is likely to be reduced. The levelized cost of energy may be half of the previous levels of the void 1 technology. The year of introduction is too early, suggesting still higher energy and power densities over those listed. The closest existing technology is an advanced compressed air energy storage device.

The exemplary void 3 technology is most likely a battery. It will require an order of magnitude less unit investment, although the investment in terms of euros per kilowatt may be up to one half of previous levels. Cycle times will be improved, and energy densities may be doubled or even tripled over previous technologies. Power densities will also be somewhat improved. The levelized cost of energy will be three or four times lower than the equivalent technologies from 1985. The technology as anticipated is closest to some of the forecasted lead-acid battery advances for the year 2030.

Conclusions

In this paper, a database of energy storage technologies with various corresponding attributes is examined. The authors described a method to manage incompleteness of the data. The described method synthesizes t-SNE technique, which is a novel dimensionality reduction technique, with long-established expectation-maximization technique. The completed database later used for building a technology frontier that shows the progress of technology in

time, discussing the design trade-offs in the technology and finally identifying some design voids in the progress of the technology.

The technique described in this paper can be complementary to wide variety of technometrics or evolutionary technology dynamics approaches which make use of high dimensional technology data.

The technique performs better especially in visualization than other dimensionality reduction applications such as feature selection or feature extraction for two reasons. Firstly, it uses expectation maximization to impute the missing variables, which manages the incomplete data in such a way that the imputed variables have minimal weighting in producing the low dimensional map. Hence, it has least effect on the derivation of the lower dimensional map. Secondly, the t-SNE technique itself is a more suitable approach compared to other dimensionality reduction algorithms such as incumbent Principal Component Analysis (PCA). PCA aims to keep variation in the data and does not care about the pairwise relationships between data points, whereas manifold learning techniques such as t-SNE performs better in keeping similarities.

As a follow up to this work, more applications of this techniques next to the technology trajectories and design voids, as showcased in this paper, are yet to be explored. The promise of this technique is its complementary position in various technometrics analysis, which is yet to be fulfilled.

Furthermore, a methodological study regarding the validation of the technique using controlled experiments on a complete data set is on the research agenda of the authors.

Acknowledgement

Authors thank *Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities (BYTE)* project for funding this research.

References

Coccia, M. (2005). Technometrics: Origins, historical evolution and new directions. *Technological Forecasting and Social Change, 72*(8), 944-979.

Cunningham, S. W. & Kwakkel, J. H. (2014). *Technological frontiers and embeddings: A visualization approach.* Paper presented at the International Conference on Management of Engineering & Technology (PICMET), 2014 Portland, Oregon.

Delft University of Technology. (2014). Enipedia. Retrieved June 20, 2015 from: http://enipedia.tudelft.nl.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1-38.

Gill, J. (2004). *Bayesian Methods: A Social and Behavioral Sciences Approach* (Third ed.). Boca Raton, FL, USA: Chapman & Hall / CRC Press.

Phaal, R., Farrukh, C. J. P., & Probert, D. R. (2004). Technology roadmapping -- A planning framework for evolution and revolution. *Technological Forecasting and Social Change, 71*, 5-26.

van der Maaten, L. (2007). An introduction to dimensionality reduction using MatLab. *Report, 1201*(07-07), 62.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

van der Maaten, L. J., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research, 10*(1-41), 66-71.

Bibliometric Characteristics of a "Paradigm Shift": the 2012 Nobel Prize in Medicine

Andreas Strotmann[1] and Dangzhi Zhao[2]

[1] *andreas.strotmann@gmail.com*
ScienceXplore, F.-G.-Keller-Str. 10, D-01814 Bad Schandau (Germany)

[2] *dzhao@ualberta.ca*
University of Alberta, 3-20 Rutherford South, Edmonton, Alberta (Canada)

Abstract
This research-in-progress paper reports bibliometric characteristics that illustrate and give credence to the claim of the Nobel Prize committee that its 2012 Nobel Prize in Physiology or Medicine was awarded for a "paradigm shift". An all-author co-citation analysis (ACA) of stem cells research 2004-2009 provides an interesting characterization of this paradigm shift, which was triggered by a mid-2006 publication by the younger of the two 2012 laureates. In particular, while ACAs of 2-year time slices for the period consistently indicate the presence of a single cohesive subfield in which the "paradigm shift" occurred, with some fluctuation in membership throughout the period, an ACA of the entire six year period shows instead a closely interlinked pair of subfields, which on closer inspection turn out to represent the pre- and post-paradigm shift states of the same subfield. This bibliometric characterization also correctly identifies the name of the researcher primarily responsible for the paradigm shift, namely, Shinya Yamanaka, as that of the dominant post-shift cited author in that subfield. The relative lack of dominant figures in the subfield in the pre-shift period also underlines the area's pre-paradigmatic state of multiple conflicting and relatively unsuccessful research directions attempting to address a fundamental crisis in that field at that point.

Conference Topics
Mapping and Visualization; Citation and Co-citation Analysis; Methods and Techniques

Introduction
The 2012 Nobel Prize in physiology or medicine was awarded to John B. Gurdon and Shinya Yamanaka for having triggered, the latter with a discovery first reported in his mid-2006 publication (Takahashi & Yamanaka, 2006), "a paradigm shift in our understanding of cellular differentiation" (Nobel.org, 2012).

In the present paper, we report bibliometric evidence and characteristics for this paradigm shift. Results from this study may contribute to research that combines relational and evaluative citation analysis methods to extend the research problems that are addressed by citation analysis.

Methodology
We examined the evolution of the stem cell research during 2004-2009 through an author co-citation analysis (ACA) of three 2-year time slices using the same dataset as in Zhao and Strotmann (2011), which reported results from a study of the full 6-year time period. We adapted methods from that study.

The data set was constructed by retrieving about 60,000 full PubMed records of stem cell research articles published during 2004-2009 with MeSH heading "stem cells", enriched by their cited references from Scopus records corresponding to these PubMed records (Strotmann & Zhao, 2009). Automatic author name disambiguation was performed on this dataset (Strotmann, Zhao, & Bubela, 2009).

For each of the three 2-year time slices, the 200 most highly cited authors were identified by fractional author citation counting, and their exclusive all-author co-citation counts were

calculated (Zhao & Strotmann, 2008). An exploratory factor analysis with oblique rotation was performed on each of these co-citation matrices (SPSS Direct OBLIMIN) with the number of factors to extract determined by Kaiser's rule of eigenvalue greater than one. Only factor loadings greater than 0.3 were retained in the factor analysis results in order to focus on the most important relationships.

The visualization used here is similar to that in Strotmann and Zhao (2012), improving on the one introduced in Zhao and Strotmann (2008). It visualizes directly the results of a factor analysis, with authors as square, and factors (research specialties) as circular nodes. An author node is colored according to the factor that it loads most highly on in the pattern matrix result of the factor analysis. Node sizes are proportional to citations received (author nodes) or to the sum of member author citations weighted by each author's loading (factor nodes). The visualization merges information on both the pattern and the structure matrix results of the obliquely rotated factor model, using the latter for automatic layouting (Kamada-Kawai algorithm in Pajek) and the former for gray-scale values of lines that link authors to the factors that they load on. Interpretation of the factor nodes (i.e., research specialties identified) proceeded exactly as in earlier papers, by manually examining highly co-cited papers of authors that load highly on a factor.

Results

Figures 1-3 show the intellectual structure of the stem cell research field for three consecutive 2-year periods.

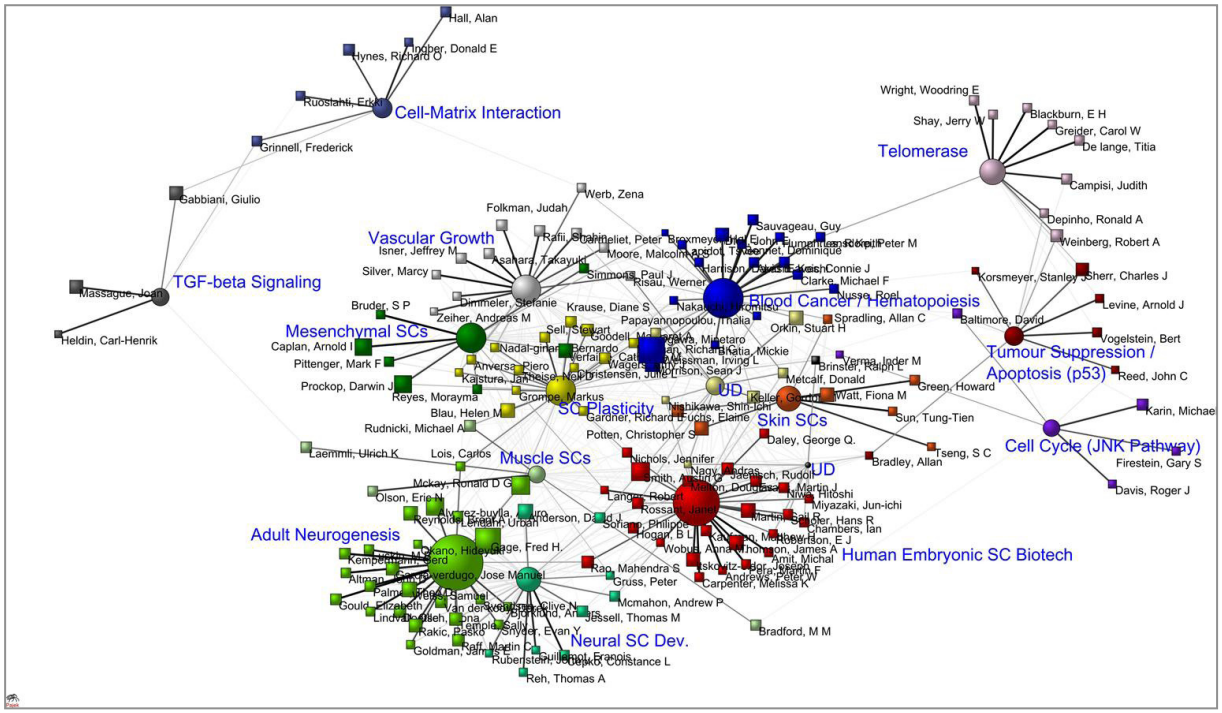


Figure 1. ACA of stem cell research 2004-05.

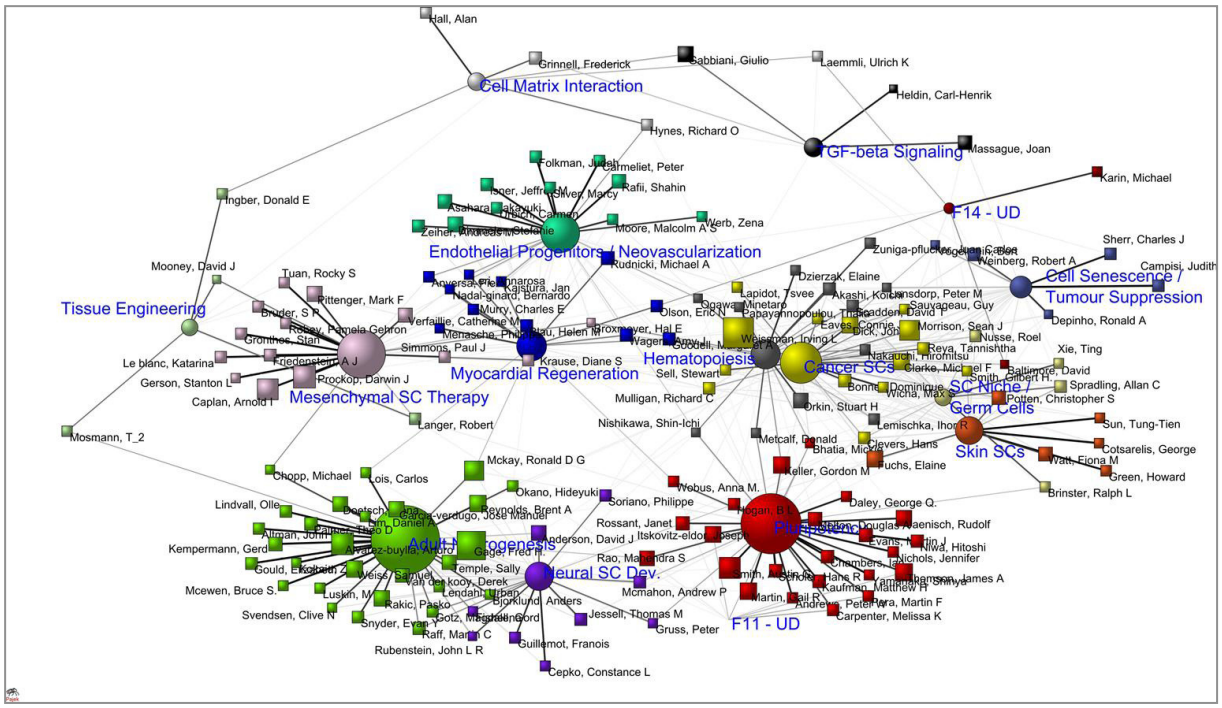


Figure 2. ACA of stem cell research 2006-2007.

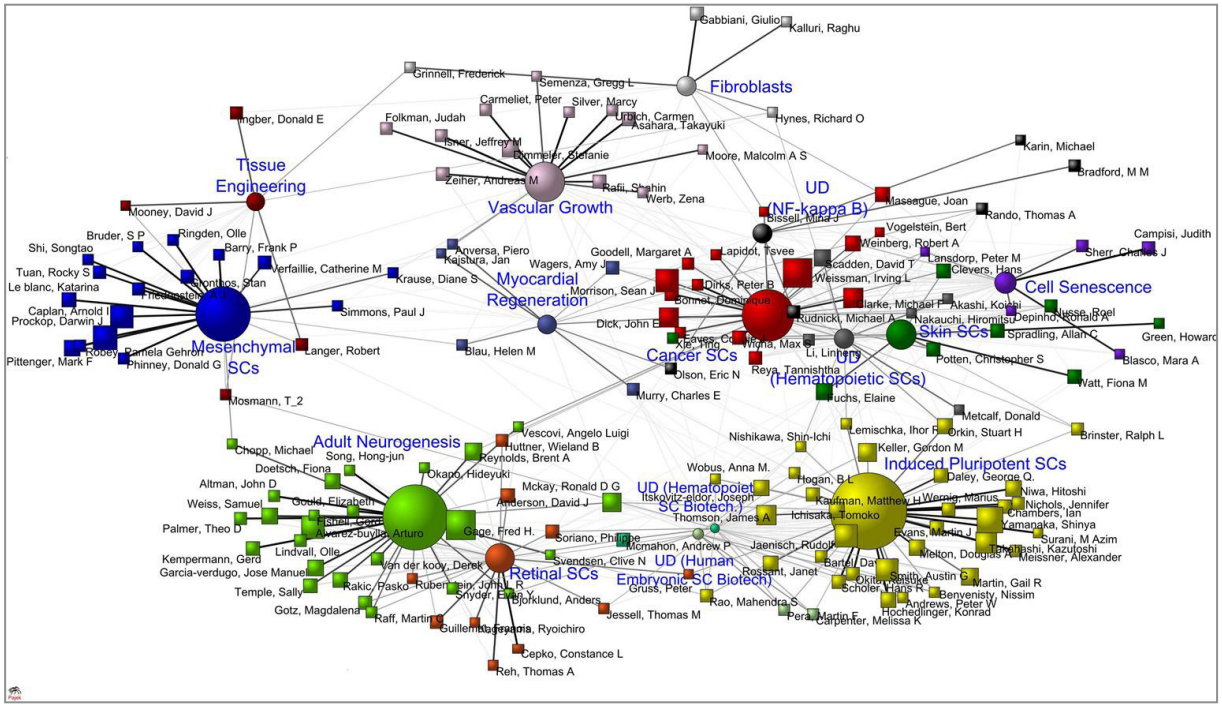


Figure 3. ACA of stem cell research 2008-2009.

While many interesting features of the international stem cell research field may be observed by examining these maps closely, we focus here on one particular major development in this field during the 2004-2009 time period as seen from changes over time. During the entire 2004-2009 time period, a subfield is shown prominently in the bottom right area of these maps as one of the two dominating specialties in stem cell research (the other being neural stem cells, bottom left). However, the entire focus appears to be shifting from (human) embryonic stem cell research in 2004-2005 (Fig. 1) through the study of pluripotency in

2006-2007 (Fig. 2) to the study of (human) induced pluripotent stem cells in 2008-2009 (Fig. 3). With this renewed focus on induced pluripotent stem cells, this subfield overtook the Neural stem cells specialty to become the most prominent specialty in the entire stem cell field in 2008-2009.

The transformation of this subfield is linked to the phenomenal rise of Shinya Yamanaka in these maps. Yamanaka was awarded the 2012 Nobel Prize in physiology or medicine for his discovery of induced pluripotent stem cells in mid-2006. He was not a highly influential researcher yet in 2004-05 as measured by citation impact (his name does not appear in Fig. 1); his name emerges in 2006-2007 (a small square in Fig. 2) and dominates this subfield by 2008-09 (the largest square in Fig. 3) with a citation impact reaching that of the two long-time most highly influential authors in the entire stem cell research field: Irving Weissman in the cancer stem cells specialty (red) and Fred Gage in the Neural stem cells area (green).

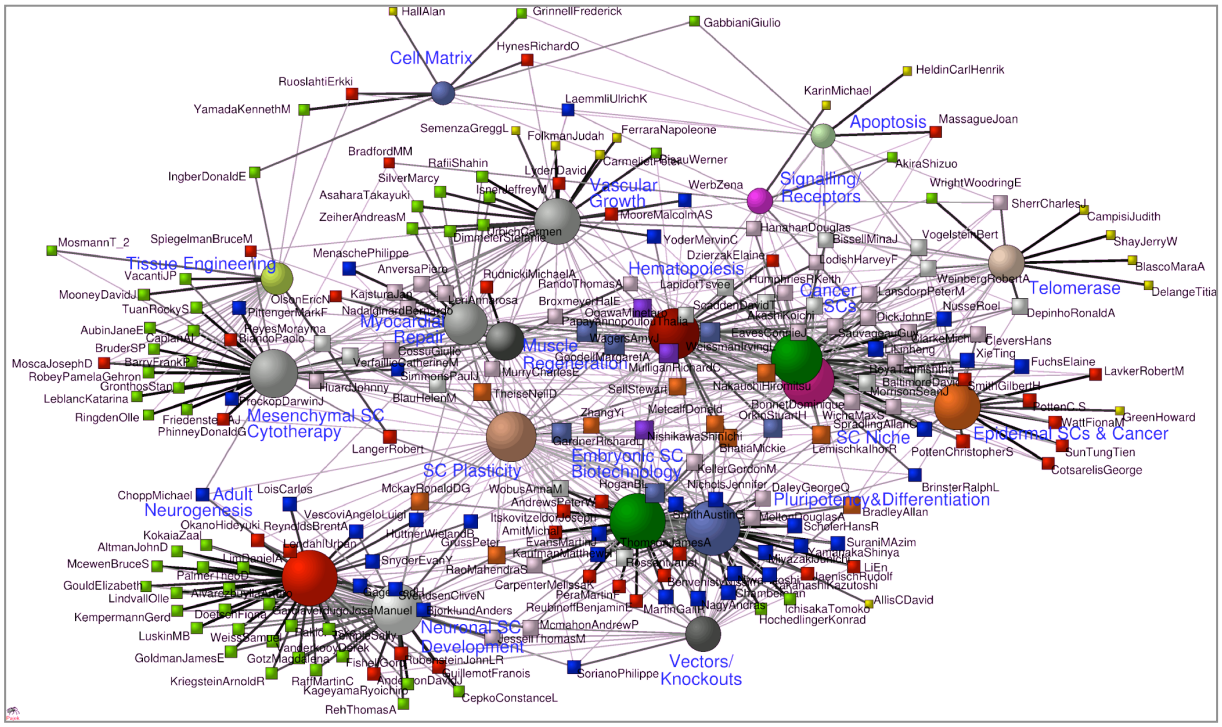


Figure 4. ACA of stem cell research 2004-2009.

By contrast, Figure 4, reproduced from (Zhao & Strotmann, 2011), which covered the entire 2004-2009 period in a single visualization, shows this subfield as consisting of two heavily interlinked research areas (bottom center), namely embryonic stem cell research (left, green) and (induced) pluripotent stem cell research (right, blue). This clarifies that what at first blush looks like it might have been a gradual change within this subfield when considering only Figures. 1-3 in fact constitutes a major in-place shift of research focus. Taken together with Figures 1-3, this confirms that the entire knowledge base for this subfield of stem cell research shifted from the former to the latter within just a couple of years of the publication of the key transformative paper – a true paradigm *shift* indeed. Most authors in this subfield co-loaded strongly on both these areas in the 6-year visualization, indicating a widespread realignment of researchers. A major paradigm shift becomes apparent.

Discussion

Kuhn's main criterion for a scientific revolution, or paradigm shift, is that something previously unthinkable becomes standard knowledge in a scientific field and a major crisis within the field is resolved as a result (Kuhn, 1970). In the case of stem cell research,

Yamanaka found that differentiated cells can be "reset" (induced) to undifferentiated (pluripotent) state, which essentially reverses the arrow of time in cell development biology, something previously unthinkable indeed.

It had been known in principle since Gurdon's 1960s paper (Yamanaka's co-laureate) that adult cells could be turned into even totipotent cells. For decades, stem cell research had been attempting to make this process feasible and controllable for therapeutic use, hoping someday to be able to regrow any type of damaged tissue (hence, the term regenerative medicine). The insurmountable research problem was a practical one: all methods for manipulating cells to this end produced stem cells that carried an unacceptably high risk of growing into malignant cancers rather than viable organs. Yamanaka's methods appear to have been the first (among uncountable failed attempts by others) to promise a fully viable resetting of cell development to the pluripotent or even totipotent state.

At the same time, Yamanaka's methods promised "safe", "natural", and abundant sources of pluripotent stem cells for research on early stages of cell development, which provided an immediate solution to a major social crisis that faced stem cell research in this subfield. This crisis came from the huge ethical and legal problems of obtaining and handling the embryonic stem cells that it required. By triggering a "natural" reset switch of much less problematic adult cells to the pluripotent state, as it were, the resulting stem cells not only side-stepped the ethically problematic use of embryos as a source, but did so without the kinds of major intervention such as genetic manipulation that had severely limited the usefulness of earlier versions of such cells for studying the "natural" biology of cell development.

As the Committee points out, Yamanaka's solution was also quite simple, so that human embryonic stem cell research was able to rapidly shift its entire focus to the study of induced pluripotent stem cells, in the remarkably short time of just a couple of years. Yamanaka's methods became standard knowledge very quickly – "textbooks were rewritten".

In the visualizations produced from an ACA of the type we performed here, this paradigm shift is characterized, somewhat paradoxically, by a stable visual appearance of the affected research subfield, accompanied by a shift in topic focus (factor labels). That a major topic shift took place can be confirmed through an analysis of a larger time slice spanning the triggering event, as we saw above. The initiator of the paradigm shift, Yamanaka, stands out as the author whose node shows explosive growth in citations received within the area as the shift occurs. The success of the paradigm shift is also seen from a rapid growth spurt of the shifting subfield relative to other subfields.

Interestingly, our visualization appears to also capture the "pre-paradigmatic" stage of this subfield, during which no single proposed solution managed to dominate the field (or subfield) that is undergoing a crisis (Kuhn, 1970). Unlike e.g. Gage in Neural stem cell biology or Weissman in bone marrow stem cell medicine research, whose citation impacts (indicated by relative node sizes) clearly dominated their respective subfields, no individual stood out in the embryonic stem cell research to that degree in Figure 1 (2004-2005). By 2008-2009, however, with the paradigm shift from embryonic to (induced) pluripotent stem cells as primary research tools completed, Yamanaka clearly plays that role in this area.

This ACA was actually performed, and Figures 1-4 were created, well before the 2012 Nobel Prize was announced (Strotmann & Zhao, 2011; Zhao & Strotmann, 2011). It appears that this paradigm shift could in principle have been identified and the 2012 Nobel Prize predicted through bibliometric studies of this kind (we did identify it as a "major development" of the field). Now that we have an idea what to look for, we could perhaps proactively look for patterns of this kind in bibliometric research in order to identify scientific breakthroughs and to make interesting predictions for major research awards. Research of this kind could enhance previous attempts to predict who among millions of scientists might qualify for the

honor of a Nobel Prize (Garfield & Malin, 1968) by combining relational and evaluative citation analysis methods to provide more convincing evidence.

Conclusions

This paper provides bibliometric evidence that the 2012 Nobel Prize in Physiology or Medicine was indeed awarded for a paradigm shift, through ACA of three consecutive 2-year time periods of stem cells research 2004-2009 compared to a single 6-year ACA for the same data. The success of this paradigm shift is seen on the ACA maps from the explosive growth in node size (citations received) of the researcher whose research initiated the shift, along with a complete shift of research focus in a subfield of stem cells research and a rapid growth spurt of this shifting subfield relative to other subfields. An ACA of the full period confirms that a major shift in the knowledge base of the subfield took place over this short time period; indeed, it shows signs of moving from a Kuhnian "pre-paradigmatic" to a "normal science" stage.

We hope that results from this study will contribute to research that combines relational and evaluative citation analysis methods to extend the research problems that are addressed by citation analysis.

Acknowledgments

This project was funded in part by the Social Sciences and Humanities Research Council of Canada.

References

Garfield, E., & Malin, M. (1968). Can Nobel Prize winners be predicted? *135th Annual Meeting, AAAS, Houston, Texas*.

Kuhn, T. (1970). *The structure of scientific revolutions*. Enlarged (2nd ed.). University of Chicago Press.

Nobelprize.org (2012). The 2012 Nobel Prize in Physiology or Medicine - Advanced Information. Retrieved June 2, 2015 from: http://www.nobelprize.org/nobel_prizes/medicine/laureates/2012/advanced.html

Strotmann, A. & Zhao, D. (2011). Evolution of stem cell research 2004-2009. A citation analysis perspective. *Stem Cells Europe. Edinburgh, 20.-21. July 2011*.

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820-1833.

Takahashi, K. & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663-676.

Zhao, D. & Strotmann, A. (2008). Information Science during the first decade of the Web: An enriched author co-citation analysis. *Journal of the American Society for Information Science and Technology*, *59*(6), 916-937.

Zhao, D. & Strotmann, A. (2011). Intellectual structure of the stem cell research field. *Scientometrics*, *87*(1), 115-131.

Bibliometric Mapping: Eight Decades of Analytical Chemistry, with Special Focus on the Use of Mass Spectrometry

Cathelijn J. F. Waaijer[1] and Magnus Palmblad[2]

[1] *c.j.f.waaijer@cwts.leidenuniv.nl*
Centre for Science and Technology Studies, Faculty of Social and Behavioural Sciences, Leiden University, P.O. box 905, 2300 AX, Leiden (the Netherlands)

[2] *n.m.palmblad@lumc.nl*
Center for Proteomics and Metabolomics, Leiden University Medical Center, Postzone L4-Q, P.O. Box 9600, 2300 RC Leiden (the Netherlands)

Introduction[a]

Bibliometric mapping tools and other scientometrics analyses may be used to study the historical development of a research field. In our paper, we use automatic bibliometric mapping tools to visualize the history of analytical chemistry from the 1920s until the present, with special focus on the application of mass spectrometry (MS).

Data and methods

Co-word maps were based on noun phrases (nouns and preceding adjectives) parsed from titles and abstracts of all papers published between 1929 and 2012 by *Analytical Chemistry*, a key journal in the field of MS. Maps were constructed by determining the co-occurrence of noun phrases and visualized using VOSviewer software (Waltman & van Eck, 2010).

Results

Evolution of topics in analytical chemistry 1929-2012

Co-word maps were based on all texts published in *Analytical Chemistry* except for advertisements (1929-1995) or on all articles, letters and reviews published in *Analytical Chemistry* (1996-2012). Table 1 shows a summary of the different clusters in the co-word maps (due to space constraints, the maps themselves could not be included).

The maps show that inorganic chemistry has been an important topic within analytical chemistry for a long time; from 1929 until 1990 there were one or more clusters on inorganic chemistry. In the 1991-2000 period it was merged with the topics of electrochemistry and sensors. Much attention was given to (the development of) different apparatuses between 1929 and 198. A cluster on general and editorial issues can be found in almost every period. Topics that have developed over time include electrochemistry, chromatography and mass spectrometry. Electrochemistry shows up as its own cluster in the 1951-1960 period, but terms relating to the subject can also be found in the inorganic chemistry and metals cluster from 1941. This suggests the topic of electrochemistry has developed from inorganic chemistry and metals to form its own subfield. Chromatography is apparent in the maps from the 1951-1960 period onwards; mass spectrometry from the 1971-1980 period. The maps suggest the widespread use of mass spectrometry in analytical chemistry primarily developed through its coupling to chromatography; for the 1971-1980 period terms relating to mass spectrometry can be discerned in the maps, but the cluster is still dominated by chromatographic techniques and applications. However, from the 1981-1990 period, mass spectrometry broke off and formed its own subfield. Finally, from 2001 a cluster on separations and microfluidics emerged. This cluster also contains terms relating to theory and simulations (of such microfluidic systems).

Use of different techniques in analytical chemistry

Next, we analyzed the development and use of a number of techniques *within* analytical chemistry. As a proxy, we determined how many articles mentioned the technique in their titles during the 1929-2012 period. This shows that titration techniques reached their publication peak in the 1950s, gas chromatography in the 1960s, and liquid chromatography in the 1980s (Fig. 1). Of these techniques, only the latter was still mentioned in the titles of over 5% of papers published in the 2001-2012 period. On the other hand, microfluidics is an example of a technology not mentioned before 1990 that has really taken off in this 2001-2012 period. A technique not mentioned to a great extent in the titles of *Analytical Chemistry* papers is nuclear magnetic resonance (NMR). As the co-word maps already suggested, the mention of mass spectrometry increased throughout the entire period. Whereas in the 1929-1940 period none of the *Analytical Chemistry* papers mentioned mass spectrometry in their title, the percentage of papers that did increased to eighteen in the 2001-2012 period (Fig. 1). This indicates *Analytical Chemistry* has made a shift towards the publication of research using mass spectrometry instead of other techniques.

Table 1. Main topics in mass spectrometry within the field of analytical chemistry.

Clusters per period
1929-1940
Apparatuses
Inorganic chemistry
Gases
Industrial applications, hydrocarbons and food
1941-1950
Apparatuses
Inorganic chemistry: gases/halogens
Inorganic chemistry: metals
Industrial applications and hydrocarbons
Organic and food chemistry
General/editorial
1951-1960
Apparatuses
Inorganic chemistry: metals
Electrochemistry
Chromatography
General/editorial
1961-1970
Inorganic chemistry
Electrochemistry
Chromatography
General/editorial and "informatics"
1971-1980
Apparatuses
Inorganic chemistry
Gases
Electrochemistry
Chromatography
General/editorial
1981-1990
Inorganic chemistry
Electrochemistry
Chromatography
Mass spectrometry
General/editorial
1991-2000
Inorganic chemistry, electrochemistry and (bio)sensors
Chromatography
Mass spectrometry and proteomics
Electrophoresis
General/editorial
2001-2012
Mass spectrometry
Detection, electrochemistry and (bio)sensors
Small molecules and quantitation
Separations, microfluidics, and theory and simulations

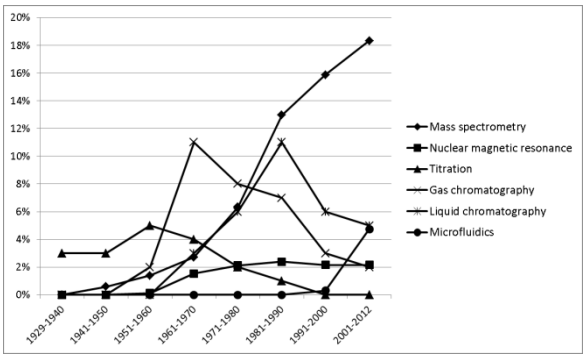


Figure 1. Use of different techniques in *Analytical Chemistry*. Search terms used were "mass spectro*", "nuclear magnetic resonance" or "NMR", "titration", "gas chromato*", "liquid chromato*", and "microfluid*", searched against the titles of *Analytical Chemistry* papers.

Additional work

Additional results, such as the trends in research topics in analytical chemistry research using MS, an assessment of which research fields use MS, and a citation network of research using MS, will be included on our poster.

Endnote

[a]A manuscript with the same title has been published in *Analytical Chemistry* as a Feature.

References

Waltman, L. & van Eck, N. J. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*, 523-538.

Introduction of "Kriging" to Scientometrics for Representing Quality Indicators in Maps of Science

Masashi Shirabe[1]

[1]*shirabe.m.aa@m.titech.ac.jp*
Tokyo Institute of Technology, Oookayama 2-12-1 S6-5, Meguro
152-8550, Tokyo (Japan)

Introduction

Maps of science are an effective technique, especially for non-experts, to facilitate intuitive understanding of science activities, even though they could be cut both ways. Among such maps, science overlay maps have received adequate attention from scientometrics researchers (Perianes-Rodríguez et al., 2011; Grauwin & Jensen, 2011; Klaine et al., 2012; Leydesdorff, Rotlo, & Rafols, 2012; Boyack & Klavans, 2013; Gorjiara & Baldock, 2014). Actually they are an attractive approach "to visually locate bodies of research within the sciences, both at each moment of time and dynamically." (Rafols, Porter, & Leydesdorff, 2010)

To produce science overlay maps, (1) we draw a basemap, which contains positional information of nodes from bibliographical data, then (2) we overlay other information on the basemap by assigning the information (i.e., indicators like publications and citations) to the nodes with such factors as colors and/or size of circles representing the nodes.

To think more abstractly, an essence of science overlay maps is "sharing" of positional information of nodes by different science maps, which are similar in concept to thematic maps in geography. What makes such "sharing" possible is the stability of global maps (Rafols, Porter, & Leydesdorff, 2010). This perspective could broaden choices of expressions in science overlay maps to improve our understandings. For example, VOSviewer (Van Eck & Waltman 2010) provides five different views, i.e., label view, density view, scatter view, cluster view, and cluster density view, for a fixed set of positional information of nodes. By switching these views, we can understand phenomena behind the maps deeply and multidimensionally. Therefore, introducing a new way to project bibliographical information on given maps is expected to expand availability of science overlay maps, just as a new method to produce thematic maps does in geography.

From this perspective, the author first pays attention to density view provided by VOSviewer. By mapping journals in the fields of Business, Business-Finance, Economics, Management, and Operations Research & Management Science, Van Eck and Waltman (2010, p. 529) explain functionality of the density view as follows: "The density view immediately reveals the general structure of the map. Especially the economics and management areas turn out to be important. These areas are very dense, which indicates that overall the journals in these areas receive a lot of citations." As they pointed out, this view is helpful to outline the macro structures of maps and to show which areas in the maps are important. Basically, however, density view can be used only for representing quantitative indicators, because "the item density of a point in a map depends both on the number of neighboring items and on the weights of these items." (p. 533) If citations were used as weights of items, the density map might be seen to show "quality" of areas. Actually, citation densities are only a representation of quantities. That is particularly evident in assuming to represent quality (impact) indicators like proportion of top 10% publications in the density view.

Judging from many scientometrics studies rely on density or heat maps (e.g., Pinto, Pulgarin, & Escalona, 2014), it would be reasonable to assume that graphical representations like the density view to represent quality indicators on science maps is very helpful to outline the structures of bibliographical data and to show which areas in maps of science are efficient, superior, or highly shared. Then, this paper introduces "kriging" to scientometrics for representing quality indicators.

Data

The author uses a data platform that consists of datasets from SCI Expanded, PubMed, and USPTO patent databases. By adopting matching methods developed in Shirabe (2014), records in PubMed are linked to those in SCI expanded, and non-patent references in the face sheets of US utility patents are also matched to records in SCI Expanded. As a result, three databases can be analyzed in an integrated fashion by using this platform.

This platform contains the product set (number of items is 8.5 millions) of SCI expanded (articles, reviews, letters, notes, and articles & proceedings papers; their database years are between 1992 and 2011) and PubMed (their publication years are between 1991 and 2012) as well as science citations of US utility patents registered between 1991 and 2012.

Method

First "macro and micro" basemaps are constructed by co-occurrence analysis of MeSH terms (Leydesdorff & Opthof 2013), where VOSviewer is used for mapping and clustering. For making the macro map, all the items of the product set are included in the analysis, and only third layer descriptors are treated as subjects of co-occurrence analysis. For that, lower layers' MeSH terms are replaced by their higher taxon. For making the micro map, only items containing mesenchymal stromal cells, mesenchymal stromal cell transplantation, totipotent stem cells, multipotent stem cell, induced pluripotent stem cells, pluripotent stem cells, and embryonic stem cells as their MeSH terms are included in analysis. Top 150 MeSH terms (except highly shared terms) are used in co-keyword analysis. Thus, this micro map is a map of pluripotent stem cell research.

Secondly, sets of data overlaying on the basemaps are produced. For that, positional data (i.e., two-dimensional position coordinate) of nodes produced by VOSviewer are transmitted to SAGA (Böhner, McCloy, & Strobl, 2006). Then, overlaying data for density maps (by Gaussian kernel function) or those for isograms (by kriging) are calculated from bibliographic indicators and overlaid on the basemaps.

Results

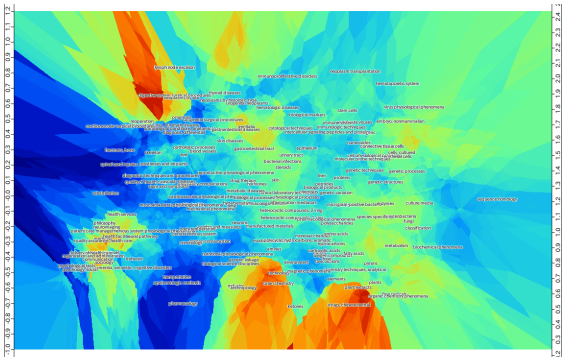


Figure 1. Japanese Share of Life-Science Papers cited by US Patents Registered between 2001-11.

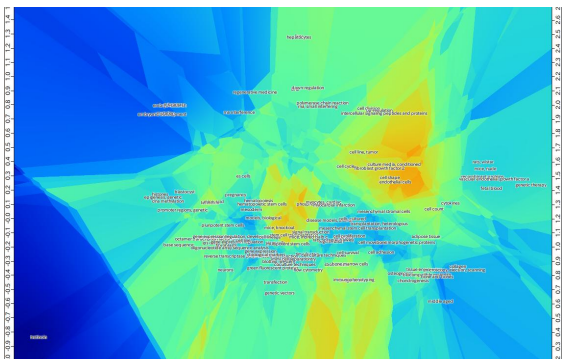


Figure 2. Japan's Relative Frequencies of Top 10% Cited Papers in Stem Cell Research.

The above figures show examples of overlay maps to represent quality indicators. They make it easier to understand the quality of Japanese research outputs intuitively and multidimensionally either at macro or micro level.

Acknowledgments

This research is partly supported by JST/RISTEX research funding program "Science of Science, Technology and Innovation Policy".

References

Böhner, J., McCloy, K.R., Strobl, J. (Eds.) (2006). *SAGA – Analysis and Modelling Applications*. Göttinger Geographische Abhandlungen, Vol. 115.

Boyack, K.W., & Klavans, R. (2013). Creation of a Highly Detailed, Dynamic, Global Model and Map of Science. *JASIS&T*, *65*(4), 670–685.

Gorjiara, T., & Baldock, C. (2014). Nanoscience and nanotechnology research publications: a comparison between Australia and the rest of the world. *Scientometrics*, *100*(1), 121-148.

Grauwin, S. & Jensen, P. (2011). Mapping scientific institutions. *Scientometrics*, *89*(3), 943-954.

Klaine, S. J., Koelmans, A. A., Horne, N., Carley, S., Handy, R. D., Kapustka, L., Nowack, B., & von der Kammer, F. (2012). Paradigms to Assess the Environmental Impact of Manufactured Nanomaterials. *Environmental Toxicology and Chemistry*, *31*(1), 3-14.

Leydesdorff, L., & Opthof, T. (2013). Citation analysis with Medical Subject Headings (MeSH) using the Web of Knowledge: A new routine. *JASIS&T*, *64*(5), 1076–1080.

Leydesdorff, L., Rotlo, D., & Rafols, I. (2012). Bibliometric Perspectives on Medical Innovation Using the Medical Subject Headings of PubMed. *JASIS&T*, *63*(11), 2239–2253.

Perianes-Rodríguez, A., O'Hare, A., Hopkins, M. M., Nightingale, P., & Rafols, I. (2011). Benchmarking and visualising the knowledge base of pharmaceutical firms (1995-2009), *Proceedings of ISSI 2011*, Vol. II, 656-661.

Pinto, M., Pulgarin, A., & Escalona, M. I. (2014). Viewing information literacy concepts: a comparison of two branches of knowledge. *Scientometrics*, *98*(3), 2311-2329.

Rafols, I., Porter, A. L. & Leydesdorff, L. (2010). Science Overlay Maps: A New Tool for Research Policy and Library Management. *JASIS&T*, *61*(9), 1871–1887.

Shirabe, M. (2014). Identifying SCI covered publications within non-patent references in U.S. utility patents. *Scientometrics*, *101*(2), 999-1014.

Van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for

bibliometric mapping. *Scientometrics*, *84*(2), 523-538.

The *Technology Roots spectrum*: a New Visualization Tool for Identifying the Roots of a Technology

Eduardo Perez-Molina[1]

[1]*eperezmolina@icloud.com*
CTB, Universidad Politecnica de Madrid (Spain)

Introduction

The purpose of this work is to present a new tool for identifying the technological foundations, or roots, of a specific technology in the whole range of existing technologies. The idea is to go back to the date before a specific technology existed as such—its origin date—and to evaluate the influence of every existing technology in relation with it. Our tool is based on the role played by prior art patent citations as a historical footprint. The documents cited in the prior art search reports by patent examiners against patent applications in a particular —new—technology link the new emerging techniques to the conventional existing ones. The nature of this particular set of references, namely *who* produced the citations—the patent examiner in place of the author—and *why* they are cited—the evaluation of the novelty and non-obviousness—, is unique within the body of bibliographic references (Meyer, 2000), and explicitly points to temporal and conceptual proximity. These two factors seem fundamental to the study of history and technology.
The *Technology Roots spectrum* (*TR spectrum*) is a tool for visualizing the components at the origin of the specific technology under study, showing their relative weight as bars in a graph containing the whole range—the spectrum—of technologies. It uses the computer to exploit the network formed by prior-art citations in patent publications and the classification codes assigned to them. This tool can be used to study the history of technology and, as a technology indicator of technological origins, can also be used for defining technology metrics.

Data Collection Methodology

The data collection methodology is shown in Figure 1. First, we select the whole collection of patents published in a specific technology using classification codes. For example, if this technology is graphical user interfaces (GUI), we must use the IPC code G06F3/048, literally "Interaction techniques based on graphical user interfaces" (IPC codes and titles can be consulted at http://www.wipo.int/). In this way we get the specific "technology" collection. From this set we extract all the citations from its search reports building the "citations" collection. Then, we keep the patents filed before the specific technology has emerged, in this case 1975 (Reimer, 2005) and we obtain the "Roots" collection.

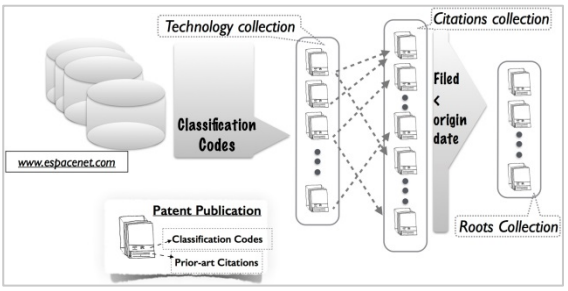


Figure 1. Data collection path

The *TR spectrum*

The set of selected patents—the "Roots" collection—is formed by patent publications disclosing technology methods, concepts, devices or systems intertwined with different aspects of the specific technology under study and filed (and therefore developed) before this technology existed —the origin's date. Analysing in turn the codes assigned to them provide us with indications of the technological foundations of the technology under study. This is why we use the expression: *Technology Roots*. Furthermore, every patent publication in the "Roots" collection is classified with a code representing a technology chosen between all possible existing technologies, this is why we use the term: *spectrum*.
The *TR spectrum* is built by aggregating the classification codes allocated to each document within the "roots" collection, and ordering this dataset in a sequence in accordance with the IPC scheme at a certain level of granularity—*section, class, sub-class, group or sub-group*—(WIPO, 2014). Changing the level of granularity we zoom out or zoom in on the techniques to have different conceptual resolutions and in consequence we can identify more technical details or we can have global views of technical fields. Figure 2 (top graph) shows the *TR spectrum* for computer graphics (CG) at the IPC *class* level. This spectrum was built using the IPC codes G06T11 (2D image generation), G06T13 (Animation), G06T15 (Image rendering), G06T17 (3D image modelling for computer graphics) and G06T19 (Manipulation of 3D models) for the "technology" collection, and the origin date was set at 1960 (Perez-Molina, 2014).
Following our methodology the "technology" collection contained 32,034 documents. Then, all

the patent publications cited in their search reports made a "citations" collection with 83,719 documents. Finally, the "roots" collection is formed by 344 patents.

A tool for studying the history of technology

The direct analysis of the main components of the spectrum provides us with an indication about the technological foundations of a specific technology. Looking, for example, at the computer graphics *TR spectrum* at IPC-*class* level (see Figure 2 top graph), it is straightforward to note that the foundations of CG are mainly in computers, electrical devices and electronics, and photography (the right-hand side of the spectrum), and to a lesser extent in medicine (left) and mechanics (left-center). The main components are G06 (computation), G01 (measuring), G09 (Education, cryptography, displays and seals), H04 (electric communications) and G03 (photography and cinematography).

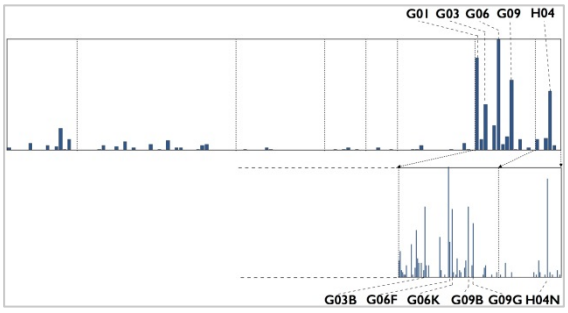


Figure 2. C.G. *TR spectrum* at IPC-*class* level (top) and partial view of the CG *TR spectrum* at IPC-*subclass* level (bottom)

At finer granularity, in other words, aggregating the dataset at the level of *sub-classes*, we have more precision in these technologies already identified. Then, it is clear from the partial view of the TR-spectrum at IPC *sub-class* level (see Figure 2 bottom graph) the importance of digital processing (G06F), television (H04N), photography (G03B), pattern recognition (G06K), educational appliances (G09B) and display control circuits (G09G). If, for instance, we are interested to know which specific technology is behind educational appliances, we zoom in on this spectral component, discovering that the most populated group is simulators (G09B9), and zooming in again we find in particular flight simulators (G09B9/08).

A tool for technology metrics

The *TR spectrum* contains information about the technological influences at the origin of a specific technology. It forms a sort of technology affiliation fingerprint of its origins, thereby it can be used as a technology identifier in technology metrics.

We have used it to get an indication of the relative distances between technologies. The different spectral bin values of the *TR spectrum* are considered as coordinates in a *technology-roots* space, thereby every particular *TR spectrum* is a point in this space. Then, applying *multi-dimensional scaling* (Wickelmaier, 2000) we have reduced the dimensionality for visualizing the relative positions of technologies. Figure 3 shows the results for four technologies—computer graphics (CG), graphical user interface (GUI), computerized tomography (CT) and Airbags—using Euclidean distance.

At present we are experimenting with other distance metrics more suitable for classification spaces.

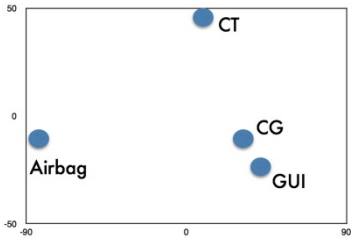


Figure 3. Relative position of CG, GUI, CT and Airbags after applying *multidimensional scaling* to its respective TR spectrums

Conclusions

We have introduced a new visualization tool—the *TR spectrum*—for identifying the technological foundations of a specific technology. We also have briefly disclosed the application of this tool for studying the history of technology and its use as a technology indicator.

References

Meyer, Martin (2000). What is special about patent citations? Differences between patent and scientific citations. *Scientometrics*, *49*(1), 93-123.

Perez-Molina, E. (2014). The Technological Roots of Computer Graphics. *IEEE Annals of the History of Computing*, *36*(3), 30-41.

Reimer, J. (2005). A History of the GUI. Retrieved, December 2014, from http://arstechnica.com/.

WIPO (2014). International Patent Classification - Guide. Retrieved, February 05, 2015, from http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf.

Wickelmaier, F. (2003, May 4). An introduction to MDS. Retrieved, May 16, 2015, from https://homepages.uni-tuebingen.de/florian.wickelmaier/pubs/Wickelmaier2003SQRU.pdf.

Modelling of Scientific Collaboration based on Graphical Analysis

Veslava Osinska[1], Grzegorz Osinski[2] and Wojciech Tomaszewski[2]

[1]*wieo@umk.pl*
Nicolaus Copernicus University, Institute of Information Science and Book Studies
ul. Bojarskiego 1, 87-100 Torun (Poland)

[2]*grzegorz.osinski@wsksim.edu.pl; wojciech.tomaszewski@wsksim.edu.pl*
Institute of Computer Science, College of Social and Media Culture
sw. Jozefa 23/35, 87-100 Torun (Poland)

Introduction

An analysis of the interrelationships between elements within dynamic structure typically involves perturbation methods based on the minimum energy. In result, the researchers use minimum distance-based algorithms and therefore the shortest path between the various components of the system. However, the history of science development shows that collaboration between the researchers in different disciplines becomes effective and fruitful when scientific explorations do not follow the "shortest possible" roads.
In current work authors present a novel approach, how to analyse and evaluate the possible collaborations ways in a small team of researchers (number of nodes is less than 100) participating in the project network KnowEscape COST Action.[1]

Data, metrics and assumption

Analysed dataset consists of 83 records characterized each member of COST network. Input data organized in 83x83 matrix, describe two years collaboration within such activities as: mobility, events organization, publishing (also for former years) and project management. The dataset was gathered using KnowEscape website (knowescape.org), ResearchGate and Mendeley services.
To describe the mutual relationships between members the graph based on Mycielski concept was constructed (Larsen, Propp & Ullman, 1995). The authors identified graphically four attractors of maximum energy. The clique represents each researcher's pair, and arbitrarily large chromatic number means any combination of disciplines. Presented visualisation (Fig. 1) was generated by using the Poincare section (PS) of the 3D space which is defined by all ties between team's members (Tamassia, 2000).
The main problem concerns identification subgroups categories with regard to scientific activity. The matrix was generated using selected nodes and links through Poincare projection (Clifford, Azuaje, & McSharry, 2006).

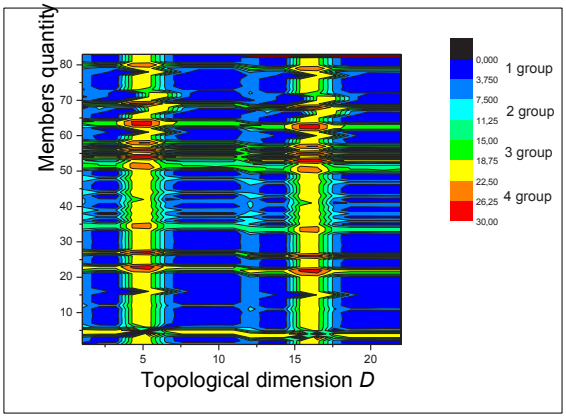


Figure 1. An iterated visualization of discrete distance routes.

Obtained iterated visualization of discrete distance routes is shown on Figure 1. As a final result we observe four clear clusters. All participants were divided on four groups by describing appropriate roles in social network: leaders, connectors, performers and outliers.
This approach was tested using algorithms adopted from medical data analysis for time series (Swierkocka-Miastkowska & Osinski, 2007, Mazur, Osinski, Swierkocka, 2009).
The authors evaluate also the dynamics of total activity by using fractal dimension (FD) of each PS image. FD is the measure of nonclassical geometry shapes and can be used as a pattern's complexity parameter (Osinska 2012).
Fractal dimension was obtained by Higuchi algorithm, so the resulting maps help to discover possible opportunities for further development of cooperation between the scientists.

Visual results

All members' activities represented by matrixes are summarized and full collaboration is weighted by appropriate real numbers. Popular application *Gephi* allows finding collaboration groups and revealing the scientists with basic roles: leader,

[1] This research is sponsored by National Science Center (NCN) under grant 2013/11/B/HS2/03048/ Information Visualization methods in digital knowledge structure and dynamics study.

subgroup leader, connector, outsider and so on. By using force directed layout (*force atlas 2*) the authors have obtained clarify configuration presented on Figure 2. As expected, the central point is occupied by the real team's leader. The closer node to central one represents the scientist who is more active in collaboration with the team's leader.

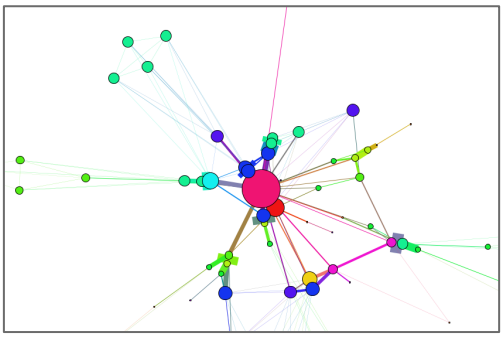
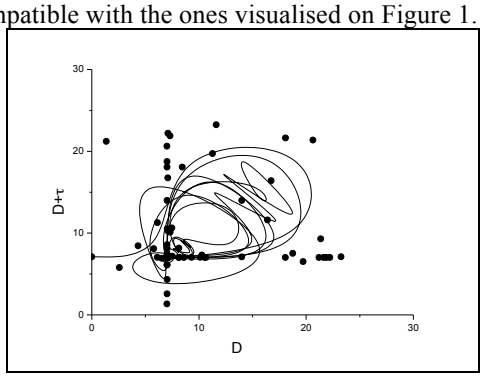


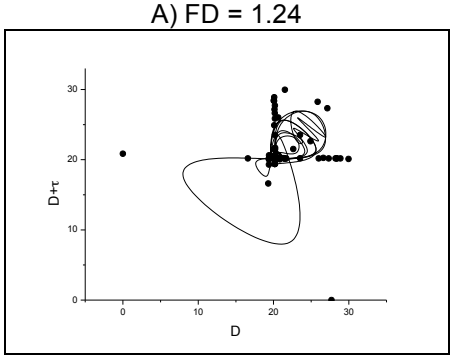
Figure 2. The graph of full activity of team's members.

Network visualisation exposes also some subgroups where intrinsic collaboration (mainly in publishing) is significant. The scientists within these groups share a common feature: geographic localisation. They work in the same country.

Simple quantitative proportional correlations between identified groups on a graph are compatible with the ones visualised on Figure 1.



A) FD = 1.24



B) FD =1.39

Figure 3. Two variations of collaboration between scientists with different social roles: A) Leader-performer; B) performer-performer.

Next step, calculation of fractal dimension, was accomplished for combinations of representatives of different groups, for example: leader-performer, subleader-leader, connector-performer and so on.

Two variations of collaboration with appropriate FD are shown on Figure 3. Fractal dimension is always lower for every pairs composed from the leader or subleader compared to the performers and connectors.

Conclusions

The authors propose new parameters for the prediction of a stable way of scientific collaboration. First is the shape of Poincare section (Return Map Poincare). For inhomogeneous academic groups where there is no self-consistency (like in this work), the level of nonlinearity can also reflect collaboration potential. It is proportional to the quantity of curves on Figure 3. The second indicator – FD shows the possibility to cooperate as well as its dynamics.

Higher fractal dimension in the case of performers can be explained by larger dynamics of predictive collaboration. This indicates the pattern is more complex. It means the pair covers significant collaboration potential.

Visualisation can help discover possible opportunities for further development of scientific cooperation. Therefore, we can observe common career landscapes of the various members and groups.

References

Clifford, G.D., Azuaje, F. & McSharry, P. (2006). *Advanced Methods and Tools for ECG Data Analysis.* Artech House Publishers.

Larsen, M., Propp, J., & Ullman, D. (1995). The Fractional Chromatic Number of Mycielski's Graphs. *Journal of Graph Theory*, *19*, 411-416.

Mazur, R., Osinski, G., Swierkocka, M. & Mikolaiczik, G. (2009). Evaluation of the dynamics of energetic changes in the brain stem respiratory centre in the course of increasing disorders of consciousness. *Activitas Nervosa Superior. 51*(5112), 69-72.

Osinska, V. (2012). Fractal Analysis of Knowledge Organization in Digital Library. In A. Katsirikou & Ch. Skiadas (Eds.) *New Trends in Qualitative and Quantitative Methods in Libraries* (pp. 17-23). World Scientific Publishing Company.

Swierkocka-Miastkowska, M. & Osinski, G. (2007). Nonlinear analysis of dynamic changes in brain spirography. Results in patients with ischemic stroke. *Clinical Neurophysiology*, *118*(12), 2822.

Tamassia, R. (2000). Graph Drawing. Ch. 21 In J.-R. Sack & J. Urrutia (Eds.) *Handbook of Computational Geometry* (pp. 937-971). Amsterdam, Netherlands: North-Holland.

Monitoring of Technological Development - Detection of Events in Technology Landscapes through Scientometric Network Analysis

Geraldine Joanny[1], Adam Agocs[2], Sotiri Fragkiskos[2], Nikolaos Kasfikis[2], Jean-Marie Le Goff [2] and Olivier Eulaerts[1]

[1] *geraldine.joanny@ec.europa.eu*
Joint Research Centre, European Commission, Brussels (Belgium)

[2] CERN, Geneva, (Switzerland)

Introduction

Monitoring technological development is an important challenge for research organisations and regulators. For decision-makers, the detection of early signals of technology maturation is key to designing proper standards and regulations. Anticipating the arrival of new technologies also allows policy-makers to develop and implement fit-for-purpose research or industrial policies. Scientometric analysis (in this case using both publications and patents) is a powerful tool to monitor technological fields and can be used to detect events in the lifecycle of a technology (Rotolo et al., 2014).

Objectives

- to analyse different cases (historical) of technological change by monitoring the evolution of patterns of collaboration between research organisations, the apparition of new keywords and/or subject categories in articles as well as changes in quantitative data such as patent or publication counts;
- to investigate whether network analysis can be used for the detection of events related to technological change;
- to identify potential indicators of technological maturation useful in the context of early warning to regulators.

Methods

Results relating to 4 technologies are presented here. Publications for each technology were retrieved from the Web of Science Core Collection database and patents from Thomson Innovation. To select the technologies, a semantic search was used in the abstract, title and author keywords of the publications.

Different network landscapes were then created using the retrieved patents and publications: sociograms showing how organisations collaborate together (through co-publishing and co-patenting); keywordgrams based on co-occurrence of author keywords in articles; and subject-category-grams based on subject categories given by Thomson Reuters. These three types of network landscapes were created and analysed for each technology.

Results

Shale Gas and horizontal drilling

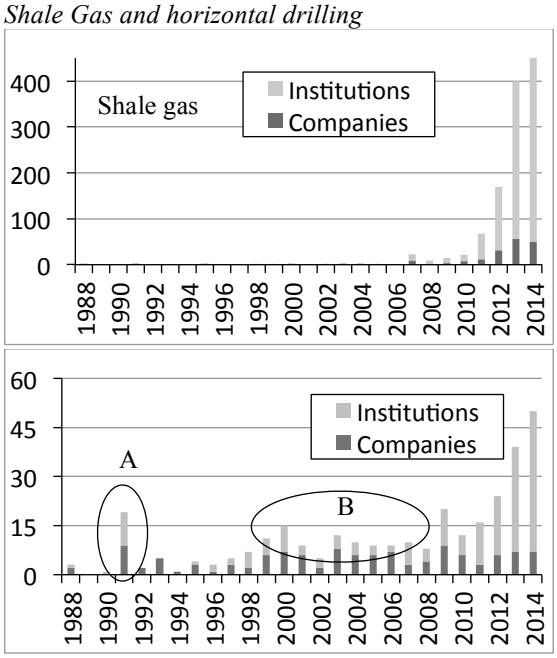


Figure 1. Number of patents and publications for horizontal drilling and shale gas from 1988.

Figure 1 shows that the number of patents and publications mentioning "shale gas" in the abstract, title or keywords started to increase noticeably in 2007 and boomed from 2011 onwards. By contrast, articles mentioning horizontal drilling, one of the key enabling technologies for "shale gas" appeared earlier (A) and rose from the year 2000 onwards (B). In addition, comparison with press content analysis shows that the rise in articles mentioning "shale gas" correlates with an increase of occurrences of press articles about shale gas (data not shown), which leads to think that this rise does not correspond to a technological trend. This shows that for the prediction of technological change the subjacent technologies - not the broad concepts - are more meaningful for the early detection of technological change.

The 2nd graph of Figure 1 shows the need to build composite indicators to avoid false positive signals. The peak of publication activity in 1991 is indeed not correlated to increased activity in other

indicators such as volume of patents or variation of number of players, for example (data not shown).

3D-printing - Detection of new uses of a technology
The number of patents and publications on fused-deposition modeling (a key enabling technology of 3D-printing) is growing steadily from 1995 to nowadays (data not shown). The subject categories of the journals in which the selected publications were published are manifold and evolve in time. As shown in Figure 2, from 1998 to 2014 a few clusters of new subject categories appear. In 1998 the articles relating to fused deposition modeling were belonging to engineering, material science and automation, which are categories describing the core of this technology. Categories describing applications of 3D-printing appear as of 2001, i. e., earlier than the entry of the first 3D printer on the market (2009).

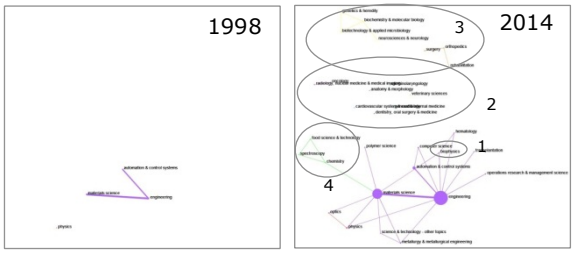


Figure 2. Subject categories for publications on fused-deposition modeling in 1998 and 2014. The circles show appearance of new non-core subject categories. 1. Biophysics (2001), 2. Radiology (2004), dentistry (2005), oncology (2006) 3. Genetics, Biochemistry (2007), Neurosciences (2008) 4. Food science and chemistry (2011).

CRT - Detecting substituting technology
The study of the author keywords for publications related to cathode ray tube (CRT) allowed to observe the emergence of the replacing technology, Liquid Crystal Display, in the CRT space. Figure 3 shows various synonyms of LCD in the keywordgram for CRT. The LCD nodes are quite big, showing their relative importance. The keyword LCD or its synonyms appear in 35 out of 649 publications or 5% of the publications.

Silicon wafer for microelectronic and for solar cell
Two application lifecycles can be observed for silicon wafers by analysing the number of related publications and patents (data not shown). These two lifecycles culminate respectively around the years 2000 and 2010. Analysing the keywordgram for the selected publications we see the keyword "silicon solar cells" appearing in 1999, and being increasingly used until 2011. Figure 4 shows its co-occurrence with other keywords in 2014. The emergence of this keyword reflects the apparition of a new use of silicon wafers for solar applications.

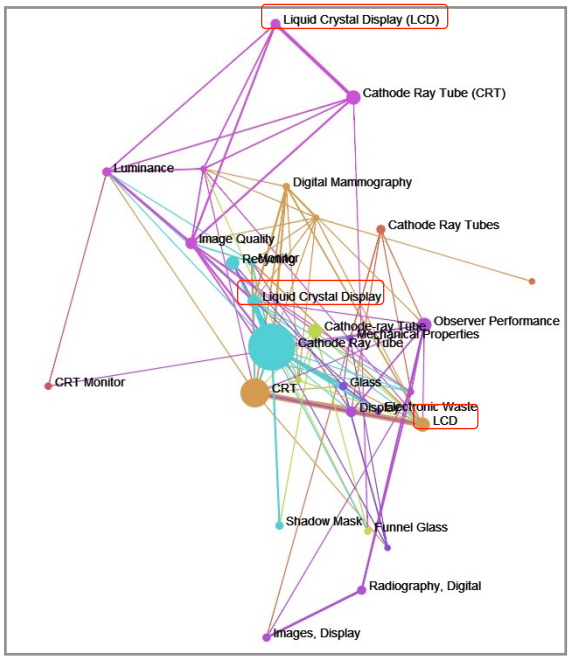


Figure 3. Author keywords view for Cathode Ray Tubes in 2014.

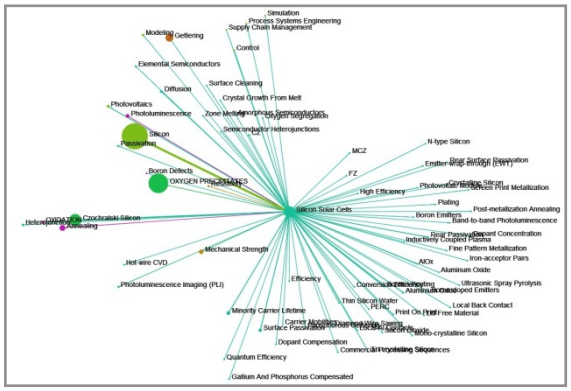


Figure 4. Centric view of keyword "Silicon Solar Cells" and its co-occurrence with other author keywords in the publications space relating to Silicon wafers.

Conclusions

Our study suggests that network analysis can be used for the detection of events relating to technological change.
We have identified several types of indicators that could be combined in order to design an early warning system to alert decision-makers of changes in technology landscapes.

References

Rotolo, D., Rafols, I., Hopkins, M. & Leydesdorff, L. (2014). Scientometric mappings as strategic intelligence for tentative governance of emerging science and technologies. *SPRU Working Paper Series,* 10, 1-40.

Analysis of R&D Trend for the Treatment of Autoimmune Diseases by Scientometric Method

Eunsoo Sohn[1], Oh-Jin Kwon[1], Eun-Hwa Sohn[2] and Kyung-Ran Noh[1]

[1] *essohn@kisti.re.kr*, [1] *dbajin@kisti.re.kr*, [1] *infor@kisti.re.kr*
Korea Institute of Science and Technology Information, Seoul 130-741 (Korea)

[2] *ehson@kangwon.ac.kr*
Kangwon National University, Gangwon-do 245-905 (Korea)

Introduction

Autoimmune diseases (AD), referred to as abnormal immune responses of body against self-antigen, are caused by the loss of immunologic self-tolerance resulting in damage to the cells, tissues and organs. The National Institute of Health (NIH) lists more than 80 autoimmune diseases that affect varied organs of the body including rheumatoid arthritis, multiple sclerosis, systemic lupus erythematosus and so on.

Significant advances of AD have been made in the understanding of clinical and pathological mechanisms involved but, to date, a few elements have been identified as being responsible for the autoimmune process. With a better understanding of the causes and treatments of AD, many potential novel therapies have recently been developed and evaluated, focusing on cellular or molecular targets.

Although there have been several research activities carried out with scientometric tools to evaluate scientific output for individual autoimmune diseases such as rheumatoid arthritis, Crohn's and Behchet's disease (Shahram et al., 2013), there was no scientometric studies on the entire autoimmune disease to date. Density-equalizing algorithms, scientometric methods and large scale data analysis were applied to evaluate quality and quantity of scientific researches in rheumatoid arthritis (Schöffel et al., 2010). Various scientometric analysis including literature-related discovery (LRD), text-mining was more broadly performed to produce knowledge discovery such as gene expression and proteomic studies. Data mining and bioinformatics approaches for autoimmune biomarker discovery studies were also attempted (Kostoff, 2014).

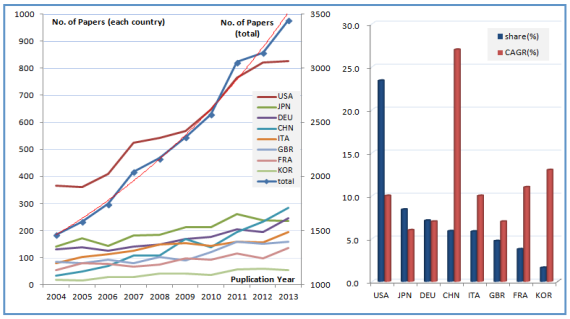
The purpose of this study is to analyze the status and trends of treatments for AD using scientometric methods, and intend to give researchers and policy-makers valuable information in the field of AD.

Data and Methods

Publications associated with the treatment of AD were retrieved from Elsevier's SCOPUS database. The query to collect data for scientometric analysis was as follows: "TS=(autoimmun*) AND TS=(therap* OR treatment*)" Total 23,587 articles published during recent 10 years (2004-2013) were collected and analyzed. Microsoft Excel, KITAS, NetMiner and VOSviewer software were combined to analyze bibliometric data. KITAS software from KISTI (Korea Institute of Science and Technology Information) was used for data extracting and cleaning. NetMiner and VOSviewer software were also used for clustering and mapping.

Results and Discussion

Figure 1 shows R&D trends over time in major countries, and the share and CAGR (compound annual growth rate) of each country based on scientific papers regarding treatments of AD. Over the last 10 years, there has been a significant growth in performance of papers with CAGR 10% in this field. Although the US quantitatively represents the largest share (23.4%), China shows the most rapid CAGR 26.6% followed by Korea (13.2%). Especially in the field of AD, Japan and Germany show a strong tendency compared with other general aspects of pharmaceuticals.



(a) (b)

Figure 1. The changes of number of papers (a) and the share and CAGR by major countries (b).

2-mode network in Figure 2 shows the co-occurrence between main countries and keywords extracted from papers, which can help identifying; which country related to; which kind of autoimmune diseases or therapeutics or treatment technologies. Circle nodes represent countries and the size of each node indicates the number of publications. The degree of relationships is

indicated by the thickness of the link and the distance between two nodes.

Keywords are divided into 2 groups, different types of AD at the bottom of Figure 2 and its technical terms at the top. In terms of the disease, high prevalence of AD including rheumatoid arthritis, multiple sclerosis, type I diabetes have shown a high correlation with US. Japan is estimated to be active in the field of autoimmune pancreatitis, autoimmune hepatitis, and Germany seems active in multiple sclerosis and type I diabetes. In particular, autoimmune thyroiditis shows a high correlation with Japan, Germany and Italy rather than US. As shown in the top of Figure 2, US is very active across all areas of the field. Advanced immunotherapies with cell-based technologies using dendritic cell, regulatory T cell (T-reg) are particularly revealed to be active in Japan and Germany as in the US.

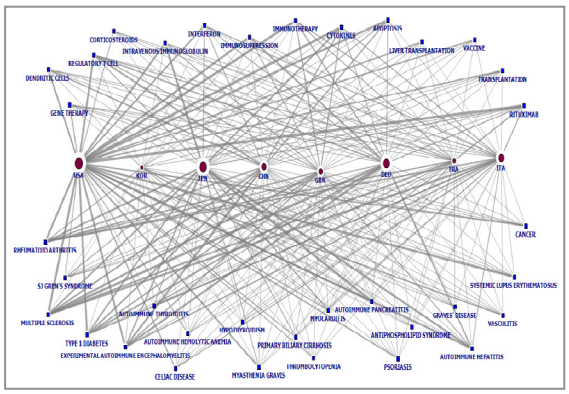


Figure 2. 2-mode network of the major countries and keywords related to autoimmune diseases.

Figure 3 provides the knowledge mapping for AD treatment drawn by co-word analysis, which shows the hot topic field or an increasing R&D productivity trend for AD treatment. To find out changes in R&D trends for treatment of AD, the dataset was divided in two time periods: 2004 to 2006 and 2011 to 2013. Several changes are found in the map of the past 3 years (2004-2006) compared with the last 3 years (2011-2013).

Figure 3 shows an experimental study using experimental autoimmune encephalomyelitis (EAE) animal model of multiple sclerosis has been disappeared in the last map (2011-2013). As time passed, clinical studies on many diseases considered to be autoimmune have been conducted with various organs and systems including endocrine, hepatobiliary, vascular systems. In addition, cell-based immune therapies with regulatory T cell (T-reg) or Th17 cells gradually have emerged in the last map (2011-2013). Immunomodulatory effects of mesenchymal stem cell (MSC) are also shown in the second figure of Figure 3. This might imply that a targeted immune therapy had been developed and successfully utilized in treating AD patients.

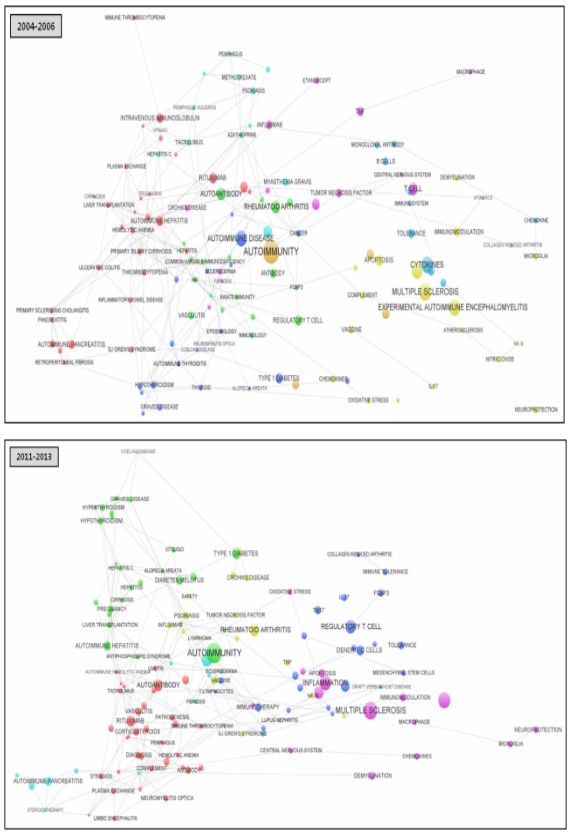


Figure 3. Co-word knowledge mapping product for the treatment of autoimmune disease.

In this study, we investigated present R&D status and trend for the treatment of AD using scientometric analysis methods. The trend in advanced R&D for the treatment of AD was identified through knowledge mapping techniques such as co-word analysis of articles and visualization technology. The results show that each country has progressive development of AD therapeutics with any other aspect. Additionally, the approach to identify the molecular and cellular mechanisms of AD underlying the immune tolerance has been increased.

References

Kostoff, R.N. (2014). Literature-related discovery: common factors for Parkinson's Disease and Crohn's Disease. *Scientometrics*, *100*, 623-657.

Schöffel, N., Mache, S., Quarcoo, D., Scutaru, C., Vitzthum, K., Groneberg, D.A., & Spallek, M. (2010). Rheumatoid arthritis: Scientific development from a critical point of view. *Rheumatoloty International*, *30*(4), 505-513.

Shahram, F., Jamshidi, A.R., Hirbod-Mobarakeh, A., Habibi, G, Mardani, A., & Ghaemi, M. (2013). Scientometric analysis and mapping of scientific articles on Behcet's disease. *International Journal of Rheumatic Diseases, 16*(2), 185-192.

Analysis of Convergence Trends in Secondary Batteries

Young-Duk Koo[1] and Dae-Hyun Jeong[2]

[1] *ydkoo@kisti.re.kr*
Korea Institute of Science & Technology Information, Gyeonggi-Inchon Branch, 145 Gwanggyo-ro, Yeongtong-gu, Suwon0si, Gyeonggi-do (Korea)

[2] *gregori79@kisti.re.krt*
Korea Institute of Science & Technology Information, Dept. of Creativity Implementation, 66, Hoegi-ro, Dongdaemun-gu, Seoul (Korea)

Introduction

Convergence refers to the creation of new technologies (or industries, markets) through the combination of two or more technologies (or industries, markets), which is promoted by technical changes, innovations, and technology diffusion, and plays a key role in changing gradual innovations to destructive innovations. Furthermore, convergence is a key factor in accelerating changes in the growth curve of technologies and the life cycle of products (Pennings & Puranam, 2001). This study was conducted to analyze convergence trends in secondary batteries and find their implications. For this purpose, useful papers and patent data for analysis were selected, collected, and processed to calculate the convergence index. This attempt is expected to provide the foundation for predicting convergence by identifying major causes that accelerate convergence. To effectively measure convergence status in this study, the diversity index suggested by Yegros Yegros et al. (2003) was used. The diversity index, which is used to measure interdisciplinary studies, considers three aspects: variety, balance, and disparity. An interdisciplinary study means the integration of different disciplines, thereby creating new academic disciplines. In this study, the convergence index was derived by the integration of different technologies into one technology.

Method of Analysis

For this purpose, the diversity index suggested by Yegros Yegros et al. (2013) was used for analysis, and IPC International Patent Classification) was used for the analysis of patents. IPC codes are assigned to individual patents and multiple codes can be specified depending on the case. In this study, IPC codes were used to analyze the convergence phenomena in secondary batteries (Stirling, 1998, Purvis et al., 2000, Stirling, 2007). The equation for each variable is given below.

$$\text{Variety} = n$$

$$\text{Balance} = -\frac{1}{\ln(n)}\sum_i p_i \ln p_i \quad (1)$$

$$\text{Disparity} = \frac{1}{n(n-1)}\sum_{ij} d_{ij} \quad (2)$$

(d_{ij} = 1-cosine coefficient)

In this equations, n means that number of IPC codes and p_i means that ratio of i IPC code.
In this study, U.S. patents about secondary batteries that had been opened or registered between January 1, 1998 and December 31, 2011 were analyzed with the IPC code for secondary batteries H010-010 using the USPTO database. In this study, we use patent data until 2011 because patent data is valid until 2011.

Table 1. Search formula for secondary batteries

Data	Search formula	Number of patents
USPTO	IPC=H01M-010*, PY=19880101~20111231	8,181

Result and Discussion

The measurement of variety through the number of IPC subclasses about patents in secondary batteries by year showed that the variety value was increasing sharply over time. In particular, the variety value greatly increased after 2009 when the number of applicants in medium- and large-sized secondary batteries increased rapidly, indicating that the variety value of secondary batteries increased with the active research related to medium- and large-sized secondary batteries. The measurement of balance by year showed that the balance value decreased between 1988 and 2000, and steadily increased again after 2003. This suggests that with the beginning of the development of the medium- to large-sized secondary batteries, research and development of various technologies have been carried out to develop the required technologies. The measurement of disparity values by year showed that the disparity value has been decreasing over time. This suggests the decreasing distance between technologies and the progress of convergence.

Figure 1. (left) Trend of variety by year; (middle) Trend of balance by year; (right) Trend of disparity by year.

In particular, the distance between technologies has become very low after 2001. As analyzed above, with the emergence of medium- to large-sized secondary batteries, convergence with other technology fields such as eco-friendly cars and solar cells has been going on.

Figures 2 and 3 show the network structure of IP codes for secondary batteries by period (1988-2000, 2001-2011). The node size indicates the number of IPCs and the length of link indicates the distance between different IPCs. The network structure of IPC codes shows that IPCs have gathered together since 2001, indicating that the relationships among different technologies have been strengthened and the distances shortened since 2001. Furthermore, IPCs related to new application fields for medium- and large-sized secondary batteries such as solar cells and wind power energy have appeared, and the distance between them and the representative IPC for secondary batteries has become closer since 2001. In other words, with the research and development of medium- and large-sized secondary batteries since 2001, the convergence in secondary batteries has become conspicuous.

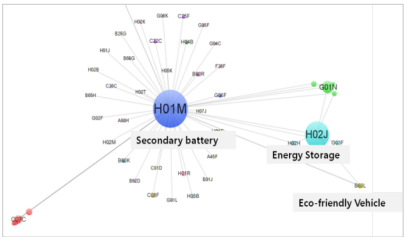


Figure 2. IPC network structure (1988-2000)

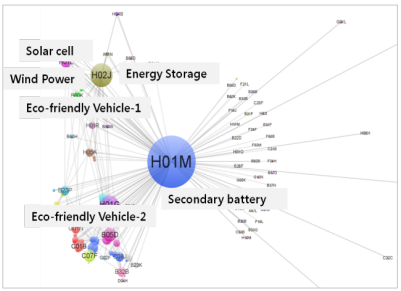


Figure 3. IPC network structure (2001-2011)

Conclusion

In this study, we analysis of convergence trend using patent data of secondary battery. As a result, it can be summarized as follows: First, as passing by year, convergence of secondary battery has increased, especially, in terms of variety and balance. This means that as increasing convergence, various field has merged and increased similarity between fields. Second, as the comparing result of IPC mapping between 1998-2000 and 2001-2011, convergence in secondary batteries is greatly increasing around the medium- and large-sized secondary batteries with the progress of convergence with eco-friendly vehicles, wind power energy, and solar energy and the decreasing distance between technologies. Predicting the convergence trends in secondary batteries has great implications to countries and companies in that they allow us to predict future industries and search for new markets and strategic partners. Furthermore, considering that existing studies used patents in a limited way due to limitations of patent analysis and limited use of time-series patent data so far, the analysis in this study was useful.

Acknowledgments

This work was supported by Construction of SMB Support System based on Industry-University-Institute Knowledge Ecosystem (K-15-L04-C01-S01).

References

Pennings, J.M. & Puranam, P. (2001). Market convergence & firm strategy: new directions for theory and research, *ECIS Conference*. *The Future of Innovation Studies*.

Purvis, A. & Hector, A. (2000). Getting the measure of biodiversity. *Nature*, *405*, 212-219.

Stirling, A. (1998). On the economics and analysis of diversity. *SPRU Electronic Working Paper.*

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, *4*(15), 707-719.

Yegros Yegros, A., D'Este Cukierman, P., & Rafols, I. (2013). Does interdisciplinary research lead to higher citation impact? *35th DRUID Celebration Conference.*

Can Scholarly Literature and Patents be Represented in a Hierarchy of Topics Structured to Contain 20 Topics per Level? Balancing Technical Feasibility with Human Usability

Michael Edwards[1], Mahadev Dovre Wudali[2], James Callahan[3], Paul Worner[4], Jeffrey Maudal[5], Patricia Brennan[6], Julia Laurin[7] and Joshua Schnell[8]

[1]*michael.edwards@thomsonreuters.com*
[2]*mahadev.wudali@thomsonreuters.com*
[3]*jim.callahan@thomsonreuters.com*
[4]*paul.worner@thomsonreuters.com*
[5]*jeff.maudal@thomsonreuters.com*
Data Center Operations, Thomson Reuters, 610 Opperman Drive, Eagan, Minnesota 55123

[6]*patricia.brennan@thomsonreuters.com*
[7]*julia.laurin@thomsonreuters.com*
[8]*joshua.schnell@thomsonreuters.com*
Intelectual Property & Science, Thomson Reuters,
1500 Spring Garden St, Philadelphia, Pennsylvania 19130

Introduction

The Intellectual Property & Science division of Thomson Reuters curates millions of records a year covering scholarly literature (Web of Science®), patents and intellectual property (Derwent World Patent Index®) and life sciences discovery (Cortellis®). These millions of records could be connected through billions of potential relationships, such as that represented by a citing relationship between literature and patents, or by different documents that pertain to similar topics. By building these relationships using machine learning techniques we hope to unite information from different data sources to enable extraction of knowledge such that the whole is greater than the sum of the parts, with minimal human effort required.

However, connecting these documents in a meaningful way is challenging from both a technological perspective as well as a usability perspective. As shown in Figure 1, studying citation patterns among approximately 250,000 articles from the Web of Science, or 1/200 of the full data set, generates a citation graph that, while rich with information, is extremely difficult to use to understand knowledge flows.

This challenge is the focus of our presentation. For this research project, we have created a graph of the topics represented in a subset of the scholarly literature and granted patents, in order to explore ways to constrain the visualization of this topic graph to emphasize usability. While many additional research areas remain, our initial findings suggest that such constraint enables users to easily explore the knowledge graph in way that maximizes understanding while minimizing user effort.

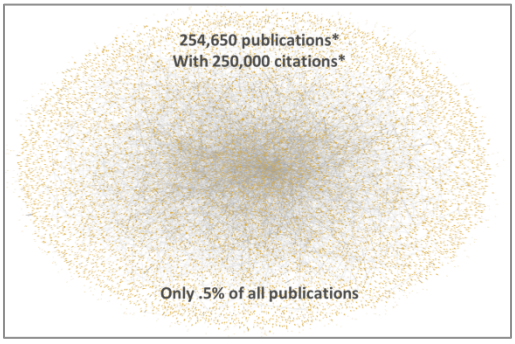


Figure 1. Ball and stick diagram of the citing relationships among a select set of publications from Web of Science®.

Generation of the Topic Graph

We chose to use topic modelling based on the latent dirichlet allocation (LDA) algorithm (Blei, Ng & Jordan, 2003) to generate connections between documents that reflect the shared knowledge among scholarly articles and granted patents. From Web of Science, we selected 27 million publications published since 1990 that had abstracts in English. Our past experience with LDA topic modelling led us to take a hierarchical approach to clustering the documents based on topics. We created a tree of over 1 million topics for the corpus, parceling out the topics into manageable chunks (20 at a glance) which were a better fit for human perception. We also created our own algorithm for applying these topics to patents, demonstrating a flexible, unsupervised technique for combining two distinct content sets. We found that the hierarchy we produced generally exhibited 4 to 5 levels of depth to the terminal nodes or documents.

Understanding the Knowledge Graph

We created the Epiphany tool to more effectively navigate the corpus of scholarly articles, using both browse and search interactions. As shown in Figure 2, the tool supports drill-down (e.g. 2.6 million articles assigned to an algorithm-focused topic; left side green), as well as search, (e.g. 8 topics strongly related to "genetic programming"; right side orange). This allows users to interact with topics and the relevant documents to understand the underlying data.

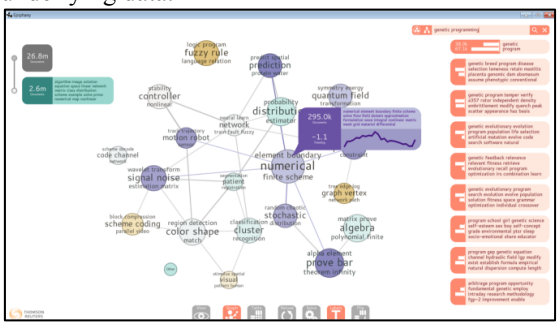


Figure 2. Screenshot of Epiphany tool showing topic clusters matching "genetic programming" search criteria.

Drilling down into the topic details is show in Figure 3. At the top in purple are statistics on the topic itself including the number of documents closely associated with the topic, the most frequent terms and the Trending metric score for the topic.

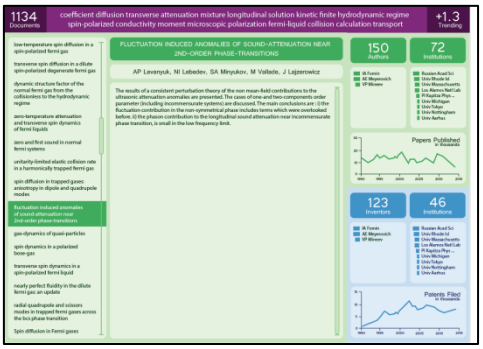


Figure 3. Screenshot of Epiphany tool Topic Details screen.

The right side of the panel contains two statistics sections, one in green for scientific papers and one in blue for patents. The header for each of the sections includes counts of the unique number of authors (or inventors) and unique number of institutions (or assignees) responsible for creation of the documents associated with the topic. Below these counts are a breakdown of the most commonly mentioned authors (inventors) and institutions (assignees). Finally, the bottom part of the statistics section is a graph of the proportion of documents assigned to this topic out of all documents published for each year.

Project Outcomes

The purpose of this research project is to test the application of scalable machine learning techniques to generate a knowledge graph that is accessible to the analyst. Now that we have developed the Epiphany tool, we have begun using it to gather feedback on this approach from a cross section of potential users. We expect to present that feedback at the ISSI2015 conference specifically to answer the question of whether a topic graph of millions of records of scholarly literature and granted patents can indeed be represented in hierarchical structure with a maximum of 20 topics at each level.

Acknowledgments

We acknowledge the support of the Intellectual Property & Science staff and the Data Center Operations staff for improvements made to this research project.

References

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993-1022.

A Sciento-Text Framework for Fine-grained Characterization of the Leading World Institutions in Computer Science Research

Ashraf Uddin[1], Sumit Kumar Banshal[2], Khushboo Singhal[3] and Vivek Kumar Singh[4]

[1]*mdaakib18@gmail.com,* [2]*sumitbanshal06@gmail.com,* [3]*khushbusinghal18@gmail.com,* [4]*vivek@cs.sau.ac.in*
Department of Computer Science, South Asian University, New Delhi (India)

Introduction

This paper describes our experimental framework for a text analysis based fine-grained characterization of leading world institutions in Computer Science (CS) research. Though the present paper uses CS research output data from Web of Science, it can be extended and applied to any discipline and data source. The existing well-known ranking systems, such as ARWU[1],Times Higher Education World University rankings[2], QS World University Rankings[3], SIR[4], Leiden Ranking[5] and Webometrics[6], only present an overall (or for a whole discipline) rank of institutions. These rankings may not be helpful if one is interested in knowing centers of excellence in research in a particular area (say Artificial Intelligence or Software Engineering in CS). Such fine-grained characterization could be very useful for different purposes. Prospective students looking to work in a particular specialized area may look at the fine-grained characterization and select institutions accordingly. Academicians or industry professionals looking for collaboration in a particular area can use the information for selecting potential institutions for collaboration. Similarly, funding agencies and policy making bodies in a country may identify institutions strong in different specialized areas of research. The other advantage of this kind of sciento-text characterization is that it is completely automated, verifiable and does not use any perceptual scores for ranking (such as reputation survey and perceptual scores of QS). Our system thus proposes a framework that uses scientometric data to produce a fine-grained research strength characterization of institutions and to rank them in order of their research excellence in a particular area.

Data Collection

We have demonstrated the working and suitability of our approach for CS domain. We obtained research output data for CS domain for the period 1999 to 2013 indexed in Web of Science (WoS). The data has been collected through an institution-wise search and we collected data for top 100 most productive institutions. A total of 261,154 records were obtained. This data constitutes about 34% of the total worldwide CS domain research output (784,920 records in total) for the period 1999-2013.

Sciento-Text Based Analytical Framework

Since our main objective is to produce a fine-grained characterization and consequential rankings, we had to first assign every research output to one or more particular research specialization. We identified a total of 11 major thematic areas (specializations) in CS domain research output. The 11-classes are based on perusal of data, some recent work (Gupta et al., 2011; Uddin et al., 2015) and recent research trends in the discipline. We processed each record in the data, extracted its 'title', 'author keywords' and 'abstract' fields and obtained the text contents of these fields. For classifying a record (research paper) to belong to one or more of the 11 thematic areas (specializations), a simple Naïve Bayes (NB) text classifier is used. The names of the 11 classes are embedded in table 1. For obtaining training data for the NB classifier, we used a keyword-match strategy for a part of the data. First of all, we created a term-profile for each thematic area (through a manual annotation by three independent annotators). Then, each record is checked for occurrence of any term from the term-profile of the 11 thematic classes, in its 'author keyword', 'title' and 'abstract' fields, in a sequential manner. Those records which get an exact match of keywords with one or more of the 11 thematic classes are assigned that class label. The assigned records then serve as training set for NB classifier, which is then used to classify the remaining unclassified records. In this manner, we classify each record to belong to one or more of the 11 thematic classes. After assigning thematic class to each record, we partitioned the data into 11 groups. Now, we have research output data for each of the major thematic areas (specializations) from the 100 most productive institutions of the world. This information is now used to first produce a plot of the research output landscape of the 100 most productive institutions and then to identify top ranking institutions in all the thematic areas. For ranking we use a simple average of scientometric indicator values for these

[1] http://www.shanghairanking.com/
[2] http://www.timeshighereducation.co.uk/world-university-rankings/
[3] http://www.topuniversities.com/university-rankings
[4] http://www.scimagoir.com/
[5] http://www.leidenranking.com/
[6] http://www.webometrics.info/

Table 1. Thematic Area Wise Top Ranking Institutions.

AI	CT	CHA	CN	CSA	CG	DBMS	IM	OS	SIP	SE
NTU	NTU	INRIA	INRIA	UCB	INRIA	NTU	TU	INRIA	NTU	INRIA
UCB	MIT	IBM	NTU	INRIA	SJTU	HU	INRIA	TU	UL	UCB
TU	INRIA	TU	UCB	KL	NTU	INRIA	MS	KL	UCB	HU
MS	UL	NTU	TU	NTU	UT	MIT	NUS	HKPU	NUS	UL
UGR	UM	GIT	CUHK	UL	UL	UL	HU	IBM	UIUC	MIT
CUHK	UTA	UCB	HIT	CMU	UW	NUS	NTU	UM	MS	NTU
INRIA	PSU	INTEL	UNC	TU	KL	MS	SU	UW	INRIA	UNC
HKPU	CMU	MS	UL	GIT	TU	MPG	CUHK	UCSD	TAU	UMCP
HU	UCL	PUC	SU	MIT	CUHK	CU	UL	NTU	TU	TU
UL	SU	CMU	GIT	MPG	IBM	IBM	MIT	UCB	KL	IBM

AI : Artificial Intelligence, CT: Computation Theory, CHA: Computer Hardware & Architecture, CN: Computer Networks ,CSA: Computer Software & Applications, CG: Cryptography, DBMS: Database Management System, IM: Internet & Multimedia, OS: Operating System, SIP: Signal & Image Processing, SE: Software Engineering

institutions, namely TP (Total Papers), TC (Total Citations), ACPP (Average Citations Per Paper), and HiCP (Highly Cited Papers). The absolute scores are first normalized to 0-100 range and then a simple arithmetic average is computed. One such similar ranking work (without thematic areas) is presented in a past literature (Ma et al., 2008).

Results and Conclusion

Our framework produces a detailed characterization of research output along the major research themes by the 100 most productive institutions of the world. The Figure 1 presents a plot of TP and TC values along the 11 research themes for the whole set of 100 institutions. Top ranking institutions identified in all 11 thematic research areas for the given period are listed in table 1. It can be seen that many of the institutions are almost available in each list but with different rank positions. Thus the presented results verify the importance of ranking institutions in different thematic areas rather than doing it for a broader research field. The paper thus presents an interesting framework for fine-grained characterization of leading world institutions and to identify the top ranking institutions in different thematic areas of CS domain. The work is extendable to other disciplines and data sources. The work may benefit more if we would have incorporated the number of researchers and graduate students for better insightful result but unfortunately obtaining those data for each institution is cumbersome and time consuming. See http://www.viveksingh.in/publications/issi2015/appendix.pdf for the full names of institutions.

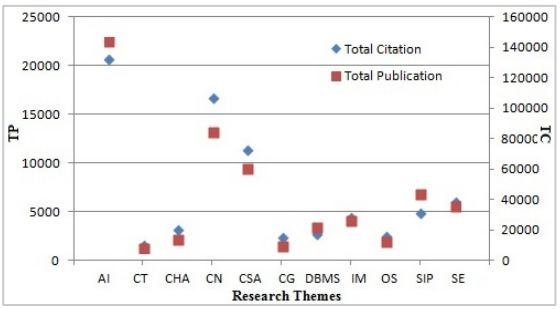


Figure 1. Thematic Area Wise Research Output and Citations.

Acknowledgments

This work is supported by research grants from Department of Science and Technology, Government of India (Grant: INT/MEXICO/P-13/2012) and University Grants Commission of India (Grant: F. No. 41-624/ 2012(SR)).

References

Gupta, B.M., Kshitij, A., & Verma, C. (2011). Mapping of Indian computer science research output, 1999–2008. *Scientometrics*, *86*(2), 261–283.

Ma, R.., Ni, C., & Qiu, J. (2008). Scientific research competitiveness of world universities in computer science. *Scientometrics*, *76*(2), 245–260.

Singh, V.K., Uddin, A., & Pinto, D. (2015). Computer Science Research: The Top 100 Institutions in India and in the World. Submitted in *Scientometrics*.

Influence of *Human Behaviour and the Principle of Least Effort* on library and information science research

Yu-Wei Chang

yuweichang2013@ntu.edu.tw
National Taiwan University, Department of Library and Information Science, No. 1, Roosevelt Rd., Taipei (Taiwan)

Introduction

The principle of least effort (PLE), a concept advanced by the American linguist George Kingsley Zipf, indicates that people complete tasks by choosing the way of least effort among various options (Zipf, 1949). To prove that the PLE is an indication of human nature, Zipf analyzed numerous empirical data collected from various human activities and used mathematical formulae to explain his findings. Zipf explained the PLE in detail in his classic 1949 entitled *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (HBPLE).

The PLE represents a common human behavior; it may thus be expected that the HBPLE has become visible in various fields and applied to various human activities. HBPLE was also compared with similar theories and was reconceptualized in the field of library and information science (LIS) (Austin, 2001; Gratch, 1990). The LIS publications on PLE have indicated that the concept of the PLE is connected to various topics (Bronstein, 2008; Chrzastowski, 1995, 1999; Kim, 1982; Wang, 2001).

This paper presents partial results of a research project for exploring the interdisciplinary influences of HBPLE. The focuses is this paper are on which concepts and citation functions of HBPLE were cited by authors of LIS articles that were published between 1949 and 2013. We analyzed citation frequency trends and the research topics of citing articles to identify emerging trends in the influence of HBPLE on LIS research and to determine which topics in LIS research have involved applying the concepts in HBPLE. In addition, citation context analysis was used to identify the cited concepts and the citation functions of HBPLE; thus, whether the PLE was the most frequently cited concept in HBPLE and the reasons HBPLE was cited were identified. The results may contribute to the understanding how a classic book on linguistics has influenced LIS research.

Methodology

The bibliographic records of LIS articles citing HBPLE published between 1949 and 2013 were searched and collected from the database Web of Science. The LIS journal candidates had to be included in the subject category of "Information Science and Library Science" in the 2012 Journal Citation Reports and the subject category of "Library and Information Science" in the database provided by Ulrichsweb.com. The publication language of articles had to be English and only research articles were collected. Regarding the search strategy used for collecting the citing articles, search terms were combined in two designated fields: the cited author field and publication year of the cited work.

A citing article could have two or more citation contexts referring to HBPLE. Each in-text citation was defined as an independent citation context. Of the 274 citing articles, three were excluded from the dataset because of citation errors existed between the in-text references and reference lists (two articles), or because full-text articles could not be obtained (one article). Finally, we analyzed 260 citing articles including 310 citation contexts. The records of cited concepts were analyzed and divided into several categories. The classification scheme of citation functions was developed based on a temporary classification scheme devised after reviewing previous studies and was modified during the analysis process. The main topic of each citing article was also coded.

Results

Topics of citing articles

Table 1 shows that HBPLE is more associated with bibliometrics and information retrieval research than are other research topics.

Table 1. Distribution of citing article topics.

Topics	No. of articles	Percentage
Bibliometrics	121	46.5
Information retrieval	64	24.6
Information behavior	24	9.2
Information system	12	4.6
Information service	7	2.7
Collection development	7	2.7
Information science	7	2.7
Knowledge organization	7	2.7
Management	5	1.9
Scholarly communication	3	1.2
Resource allocation	2	0.8
Information literary	1	0.4
Total	260	100.0

Cited concepts and citation functions

Table 2 shows the distribution of 17 cited concepts in 11 citation functions. The most frequently cited concept was "Zipf's law" and was mainly used for comparison with other bibliometric laws, whereas the second-most cited concept, the "PLE," was mainly used as evidence.

Among 201 citation contexts referring to the concept of "Zipf's law," 52.2% used the term "Zipf's law," 28.4% used other terms, such as "Zipfian distribution," "power law," "hypobolic distribution," and "rank-size law," and 19.4% contained a statement to describe or imply the concept of "Zipf's law." Although Zipf's law is a well-known informetrics law, not all authors have used the formal term "Zipf's law" to refer to the law emphasizing the relationship between word rank and word frequency.

Although the concept of the PLE, which is derived from Zipf's law, is the focus of HBPLE, the number of citation contexts referring to the PLE was lower than that referring to "Zipf's law." This result ran counter to our assumption that the number of citation contexts referring to the concept of the PLE would be highest. This implies that citing behavior is complicated and that various motivations for citing publications also affect the visibility of cited publications.

Table 2. Distribution of cited concepts according to citation functions.

Cited concepts	Citation functions											
	E	C	RS	H	R	D	E	F	Exp	T	M	Total
Zipf's law	29	38	30	27	21	22	17	7	4	5	1	201
Principle of least effort	15	13	8	6	11	7	1	4	8	3		76
HBPLE	2		2	2	2		1					9
Word distribution	3	1	1				1					6
Human behavior			2									2
Information cycle	2											2
Publication productivity			1	1								2
Rank				1								1
Sample size							1		1		1	3
Information nonuse			1									1
Language analysis	1											1
Lotka's law						1						1
Richer effect										1		1
R.Y. Chao	1											1
Signal information theory								1				1
Social physics								1				1
Optimization problem								1				1
Total	53	52	45	37	34	30	21	14	13	9	2	310

Note: (1)E: Evidence. (2)C: Comparison. (3)RS: Related studies. (4)H: History. (5) R: Relationship (6)D: Definitions. (7)E: Examples. (8)F: Further reading. (9)Exp: Explanations. (10)T: Terms. (11) M: Methods.

The 17 cited concepts were examined by year. Figure 1 shows large fluctuations for the two concepts of "Zipf's law" and the PLE; opposing trends appear. A "falling after rising" trend was observed in the concept of "Zipf's law" whereas a "rising after falling" trend was evident for the concept of the PLE. These opposing trends have resulted in a decreased difference in the annual percentage between the top two cited concepts. Although a close relationship exists between the PLE and Zipf's law, they exert an evidently different influence.

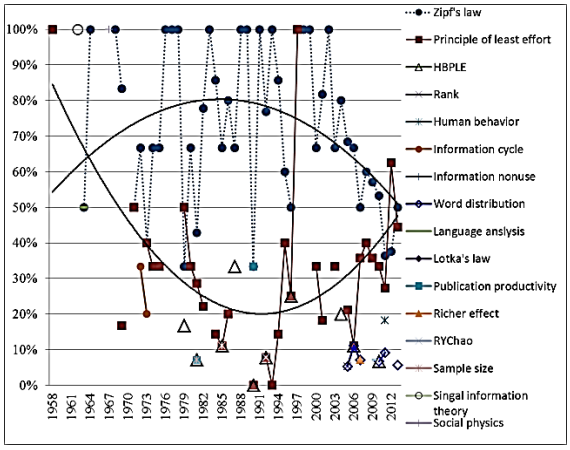


Figure 1. Changes in the percentage of cited concepts by year.

Acknowledgments

This research was supported by a grant from the Ministry of Science and Technology of Taiwan (MOST 103-2410-H-002-172).

References

Austin, B. (2001). Mooers' law: In and out of context. *Journal of the American Society for Information Science and Technology*, 52(8), 607-609.

Chrzastowski, T. E. (1995). Do workstations work too well? An investigation into library workstation popularity and the `principle of least effort.' *Journal of the American Society for Information Science*, 46(8), 638-641.

Chrzastowski, T. E. (1999). E-journal access: the online catalog (856 field), Web lists, and 'The principle of least effort.' *Library Computing*, 18(4), 317-322.

Gratch, B. G. (1990). Exploring the Principle of Least Effort and its value to research. *College and Research Libraries News,* 51(8), 727-728.

Kim, K. S., Sin, S. C. J. (2011). Selecting quality sources: Bridging the gap between the perception and use of information sources. *Journal of Information Science*, 37(2), 178-188.

White, H. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology*, 52(2), 87-108.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

Document type assignment accuracy in citation index data sources

Paul Donner

donner@forschungsinfo.de
Institute for Research Information and Quality Assurance (iFQ), Schützenstr. 6a, 10117 Berlin (Germany)

Introduction

The observed citation counts of publications can be divided by the average of a reference set of similar publications in order to get a relative impact measure. It is customary to define the reference set by publication date, scientific discipline and document type. Different document types (DT) have very different citation distributions, leading to very different results in calculations of indicators when separating reference sets by DT and disregarding this kind of normalization (Sirtes, 2012). Thus, when computing relative impact, the correctness of the assignment of document types to publications is crucial. The correctness of DT assignment in citation indexes has been called into question by studies of van Leeuwen et al. (2007), drawing attention to the treatment of letters and 'research letters' from medical journals as the same type in Web of Science and by Harzing (2003), illustrating how WoS is using some highly questionable assignment criteria. In this contribution DT assignments in WoS (Thomson Reuters, 2013) and Scopus (Elsevier, 2014) by their respective staff are compared to those of the publishers.

Methods and data

For this study data licenced from Thomson Reuters Web of Science and Elsevier Scopus and loaded into SQL databases was used. The databases are part of the infrastructure of the German Competence Centre for Bibliometrics project. Random samples of document identifiers were drawn from the WoS records, stratified by DT as assigned in WoS, restricted to items published in journals. Subsamples of the document types 'article', 'review' and 'letter', as well as of records not assigned to any of those three types (here called 'other') were taken. This follows the convention of distinguishing between 'citable items' and others. They were linked to the Scopus records detailing the same documents using DOIs. It follows that only documents with a DOI are used. In the resulting sample table, only the WoS and Scopus document identifiers and the DOI are saved in a row. The rows were randomized.

To each sample record, bibliographic description data comprised of article title, first author family name and initials, publication year, journal name, volume and issue were queried from the WoS data and saved along with record IDs into a separate table. Student assistants were tasked to search for the article abstract web pages online using the bibliographic information to query Google Scholar and web search. On the individual article web page of the journal, they were instructed to find the officially assigned document type, if specified, and code it as article, letter, review, other or not found. If no type was stated but it was clearly deducible from the abstract or title, this was also accepted.

A sample of 528 publications was analyzed so far, on which the following provisional results are based. For a further 90 publications, no certain DT assignment was possible. Found (true) DT and Scopus/WoS DT were tabulated and classified as true/false positive/negative. From those counts precision and recall were computed for each DT and combined precision and recall as weighted by DT occurrence frequency in the databases. The effect of false DT assignment on publication normalized citation score is measured in percent deviation.

Results

The results depicted in Fig. 1 show that in both citation indexes the accuracy of correct DT assignment is quite poor. WoS gives the correct DT in about 72%, Scopus in about 80% of cases (as weighted by shares of DT in the databases). On average WoS finds about 81% of publications of a given DT while Scopus will return about 73%. Error bars for the DT specific results are 95% posterior probability Bayesian credible intervals for the binomial proportion, using a flat beta prior with both shape parameters set to 1.

These findings necessarily have an adverse effect on the mean field/DT/year specific expected citation rates used as reference standards in obtaining normalized publication level citation scores. To give an idea of the magnitude of this effect, the normalized article citation score (3-year citation window) for publications that were assigned an incorrect DT in WoS was calculated following Waltman et al. (2011).

The differences between incorrect and correct score in percent of the correct score are plotted as a histogram in Fig. 2. Publications with zero citations are not used (N_0=34), since no difference could manifest.

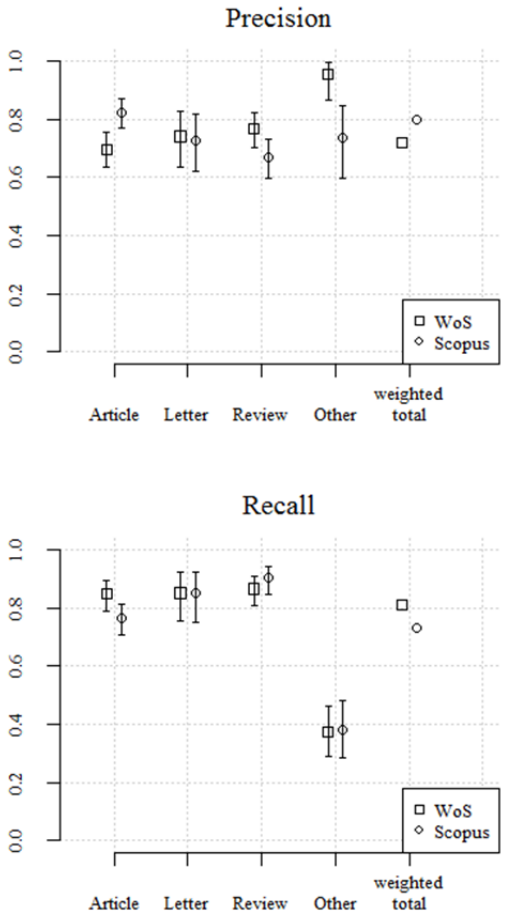


Figure 1. Precision and recall per document type in WoS and Scopus (N=528).

Conclusion

Document type assignment is unreliable in both Web of Science and Scopus and will cause large errors in publications' normalized citation scores and consequently derived indicators such as field-normalized mean citation rate.

References

Elsevier B.V. (2014). Scopus Content Coverage Guide 07.14. [accessed 2015/02/06] http://www.elsevier.com/__data/assets/pdf_file/0011/242489/Content-Coverage-Guide.pdf

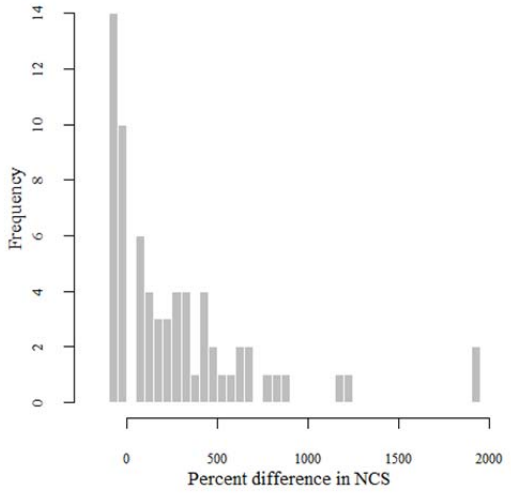


Figure 2. Percent difference in normalized citation score per document for those with wrong DT assignment in WoS (N=68).

Harzing, A.W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics*, 93(1), 23-34.

Sirtes, D. (2012) How (dis-) similar are Different Citation Normalizations and the Fractional Citation Indicator? (And How it can be Improved). É. Archambault, Y. Gingras & V. Larivière (Eds.), *Proceedings of 17th International Conference on Science and Technology Indicators (STI)*, Montréal: Science-Metrix and OST, 894-896.

Thomson Reuters (2013). Web of Science® Help. Searching the Document Type Field. [accessed 2015/02/06], http://images.webofknowledge.com/WOKRS59B4/help/WOS/hs_document_type.html

Van Leeuwen, T. N., Van Der Wurff, L. J., & De Craen, A. J. M. (2007). Classification of "research letters" in general medical journals and its consequences in bibliometric research evaluation processes. *Research Evaluation*, 16(1), 59-63.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics,* 5(1), 37-47.

Measuring the Impact of Arabic Scientific Publication: Challenges and Proposed Solution

Raad Alturki [1]

[1]*ralturki@ccis.imamu.edu.sa*
Department of Computer Science, Al-Imam Mohammad Ibn Saud Islamic University,
P.O.Box: 5701, Riyadh:11432 (Saudi Arabia)

Introduction

Citation Indices are very useful tools that were firstly used to help finding articles easily and then, used to provide information about research output. They can be used as indicator to measure research performance, provide information about trends in research and compare and rank the research output of countries, institutes and authors. It is well known that English is the universal language for science and technology and that have resulted in having many citation indices like Web of Science (Formerly ISI) and SCOPUS. It has been reported in the literature that such Indices overlook and hide publications in other languages (van Leeuwen et al., 2001) and that -with other reasons- have resulted in having indices for other languages like Chinese, Portuguese and Korean. Arabic publications is one of the least represented in the scientific community despite its been spoken by more than 200 million which makes it the fifth spoken language in the world (Gordon Jr., 2005). This work investigates the possibility of making a Citation Index for Arabic literature and addresses the challenges associated with that. This is supported by initial implementation of web based Arabic Citation Index (ACI).

Challenges

This section discusses challenges associated with non-English citation indices with special focus on the one dealing with Arabic literature. In order to have citation index for any language, it is very important to make it integrate with other English-based indices. Non-English citation indices should be able to read citations from other indices in order to see how any article or language is impacting the scientific community. This raises some issues of how to make cross languages referencing; if an article written in Chinese has cited other article in Korean, how the Chinese/Korean indices will identify this citation. This problem is not easy to be solved unless if there is a well established standardization for citations which allows identifying any article in any language. Such identifier should be unique across the globe and can be used in every citation. Luckily, Digital Object Identifier (DOI) can be used to serve this purpose while the adoption of using DOI in referencing is not yet being very popular as citation styles are still not considering that as part of the cited article. Having DOI as a compulsory in each citation style makes it easier for articles to be identified, then cited and discovered in citation indices across languages.

Unfortunately, there is no enough information about the scientific contribution written in Arabic. One of the most accurate information we found is the number of periodicals that have ISSN. According to a report by ISSN foundation, in 2012 there were 4489 new periodical record in Arabic which makes it the 26th most registered language in the world. The ISSN records do not represent only scientific journals but it registers any types of periodical. Also, there is a report by Thomson Reuters about the contribution of Arab countries recorded in their databases. The report shows that the number of scientific documents produced in those countries is around 13,574 in 2008 (Adams et al., 2011) where most of the written articles are in English. In fact, there are many journals written in Arabic that are not well recognized in the internet and digital libraries. We have noticed that Arabic scientific journals are still focusing on publishing printed format with no much focus on the electronic version.

In reality, there are some digital libraries that aggregate articles of major Arabic journals and provide electronic versions of such articles. However, having seen some of the main digital libraries and aggregators in Arabic, we still believe such aggregators have some issues as they provide the articles as scanned documents that cannot be indexed automatically. Also, such digital libraries do not have the full bibliographic information like title, abstract, authors, year of publishing, publisher name, volume, ISSN and list of references. Having bibliographic information is vital for building any citation index as they are the raw data to draw the relationship between article and scientific work in term of citations. If bibliographic information is not available for any reason, the PDF electronic version of the article could be used to extract the bibliographic information. Extracting such information from any electronic file can be done with some challenges if the article is saved as text rather than picture. The process becomes very

sophisticated if article is saved as picture where scanning should be done properly. Then Arabic text recognition algorithm should be used to recognize text used when current algorithms in Arabic are not reliable and accuracy rate is low.

Additional challenge in working with Arabic literature is the lack of standardization of the structure and the location of different section in articles. Any software that scan or parse the paper will make some assumptions of the location of the title, authors and abstract. Google scholar software that extract bibliographic information from files directly without having bibliographic information assumes that first line is the title which is written in large font. It has been stated in a study of Arabic journals that "instructions to authors" are generative and are not precise enough (Alkholaifi, 2001). That results in having different interpretations of instructions specially in using referencing style. Variations in formatting could happen at different places of the article, including authors' names, authors' salutation (Dr, professor), availability of abstract and list of references. List of references can be written in mixture of two languages at the same time (Arabic and English) which makes extraction harder. The extraction program should be able to work with different languages at the same time and be able to differentiate between different citing styles.

Extracted Information from article could include errors that can be stored in the index. The program should be aware of such errors and correct them before storing. Detecting errors is not an easy task as it should understand the context of the information. Names sometimes could be recognized as error or misspelled words as some names could have different variations or do not have a direct meaning especially if the name is not Arabic. After the information about any specific word is stored in the index, a query can be done to find a specific article or articles in certain subject. For this reason, search query should be able to consider all possible errors that user might have done when entering the keywords beside the stemming and lemmatization process that happens at indexing phase. In fact, there are several Arabic spelling correction techniques (Manning et al., 2006; Attia et al., 2012; Larkey et al., 2002; Rytting et al., 2011; Shaalan et al., 2012). Using such techniques will be of great important in implementing any Arabic based citation index. These techniques in Arabic are similar to other languages with few differences include the morphological analysis and context understanding of the language where Arabic language is complex in comparison to English.

The proposed system

The overall architecture of the system is shown in Figure 1 where it shows the five main components: Crawler, Parser, Matcher, Database and User Interface. This architecture is inspired by the typical design of search engines as they share similar concepts. One major difference between the two systems is that citation indices use citations as way to rank and measure the impact of an article whereas search engines normally uses the links and other metrics as a way to rank sites and documents.

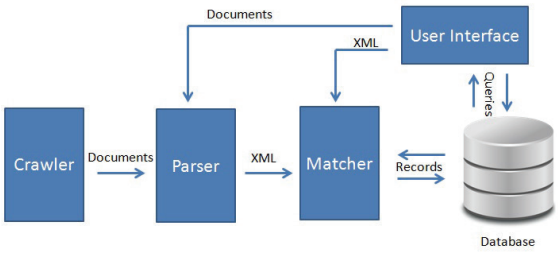


Figure 1. The proposed Architecture of ACI.

References:

Adams, J., King, C., Pendlebury, D., Hook, D. & Wilsdon, J. (2011). Global research report. Middle East, *Evidence*, Thompson Reuters.

Alkholaifi, M. (2001). Documenting citations: an analytical study of publishing policy in some journals. *Journal of King Fahd National Library* vol. 6.

Attia, M., Pecina, P., Samih, Y., Shaalan, K. F., & van Genabith, J. (2012). Improved Spelling Error Detection and Correction for Arabic. *Proc. COLING,* 103-112.

Gordon Jr, R. G. (2005). Ethnologue: Languages of the World, Dallas, Tex.: SIL International. *Online version*: http://www. ethnologue. com.

Larkey, L. S., Ballesteros, L., & Connell, M. E. (2002). Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis. *Proc. 25th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 275-282.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge: Cambridge University Press.

Rytting, C. A., Zajic, D. M., Rodrigues, P., Wayland, S. C., Hettick, C., Buckwalter, T., & Blake, C. C. (2011). Spelling correction for dialectal Arabic dictionary lookup. *ACM Transactions on Asian Language Information Processing* (TALIP), *10*(1), 3.

Shaalan, K. F., Attia, M., Pecina, P., Samih, Y., & van Genabith, J. (2012). Arabic Word Generation and Modelling for Spell Checking. *Proc. LREC*, 719-725.

Van Leeuwen, T., Moed, H., Tijssen, R., Visser, M., & Van Raan, A. (2001). Language biases in the coverage of the Science Citation Index and its consequencesfor international comparisons of national research performance. *Scientometrics*, *51*(1), 335-346.