Eigenfactor: Detailed methods

Here we describe the methods used to compute the eigenfactor score and other journal statistics featured at www.eigenfactor.org. The purpose of the eigenfactor algorithm is to estimate the relative influence of each of approximately 111,000 reference items in our data set, based on the frequencies with which they are cited by the approximately 7000 core journals listed by Thompson Scientific in their Journal Citation Reports. The algorithm works by computing eigenvector centrality weights for the value of citations from the ~ 7000 source journals, and then calculating weighted citation rates for each of the $110,000$ reference items.

1 Citation Data

We draw our data from the 2004 Journal Citation Reports (JCR) published by Thompson Scientific. From these data, we extract a set of annual *cross-citation matrices* **Z**, which indicate how often each of approximately 7000 source journals listed in the JCR have cited a much larger set of reference items (the ~ 7000 source journals, plus many additional journals, newspapers, books, and other materials). We then compute $\mathbf{M}(y,d)$, a d-year cross citation matrix for year y as follows:

$$\begin{aligned}\mathbf{M}(y,d) &= \sum_{k=1}^{d}\mathbf{Z}(y-k,y), \quad \text{where}\\ \mathbf{Z_{ij}}(y_1,y_2) &= \text{Citations from journal } j \text{ in year } y_2 \text{ to articles published in journal } i \text{ in year } y_1.\end{aligned}$$

When constructing **M**, we omit all self-citations[1], setting the diagonal entries to 0. In this paper, we work with $\mathbf{M}(2004,5)$; this a 5-year cross-citation

[1] We ignore self-citations for several reasons. First, we want to avoid over-inflating journals that engage in the practice of opportunistic self-citation and to minimize the incentive that our measure provides for such practice. Second, we do not have self-citation information for the journals not listed in the JCR. Considering self-citations for JCR-listed but not non-listed journals would systematically over-value the journals in the former set relative to the latter. Third, if self-citations are included, some small journals with unusual citation patterns appear as nearly-dangling nodes, and thus receive dramatically over-inflated scores. The tendency of the JCR data set to list some outgoing citations under a single composite item "others" — which we cannot use our calculations because we do not know where they are directed — exacerbates this problem.

matrix for the year 2004. Hereafter we suppress the arguments of $\mathbf{M}$ unless necessary to avoid confusion.

2 Calculating Eigenfactor

To compute the weights indicating the influence of a citation from each of our $\sim$ 7000 source journals, we extract $\mathbf{M'}$, the square 7000 by 7000 submatrix of $\mathbf{M}$ indicating how often each of these source journals cites each of the other source journals[2].We normalize $\mathbf{M'}$ by the column sums (i.e., by the total number of outgoing citations from each journal) to create a column-stochastic matrix $\mathbf{N}$:

$$\mathbf{N_{ij}} = \frac{\mathbf{M_{ij}}}{\sum_k \mathbf{M_{kj}}}$$

Following Google's PageRank approach, we define a new stochastic matrix P as follows[3]:

[2] We omit from $\mathbf{M'}$ those journals which publish few than 12 articles per year, and those "dangling node" journals which do not cite any other journals in the set of source journals. This later procedure is conducted iteratively; after removing the first round of dangling nodes, some new dangling nodes are created, and additional rounds of removal may be needed until all journals remaining in $\mathbf{M'}$ cite at least one other journal in $\mathbf{M'}$. After this process, we are left with a square matrix of cross citations among approximately 6000 source journals.

[3] Under our stochastic process interpretation, the matrix $\mathbf{M'}$ corresponds to a random walk on the citation network, and the matrix $\mathbf{P}$ corresponds to the Markov process which with probability α follows a random walk on the journal citation network, and which with probability $(1-\alpha)$ "teleports" to a random journal in proportion to the number of articles published by each journal. Rather than using the leading eigenvector of $\mathbf{M'}$ for our journal weights, we compute the leading eigenvector of the matrix $\mathbf{P}$. We do so for a number of reasons.

1. The stochastic matrix $\mathbf{M'}$ may be non-irreducible or periodic. Adding the teleport probability $1-\alpha$ ensures that P is both irreducible and aperiodic, and therefore has a unique leading eigenvector by the Perron-Frobenius theorem.
2. Even if the network is irreducible, without teleporting, rankings can be unreliable and highly volatile when some components are extremely sparsely connected. Suppose, for example, that a citation network comprises two fields are connected only by the citations of two journals, one in each field. The relative weight of each field would then be set solely by the relative frequencies with which these two journals cited the other field. Similarly, teleporting keeps the system from getting trapped in small nearly-dangling clusters. If a small clique of journals are occasionally cited from outside but rarely cite out of clique itself, the Markov process characteried by $\mathbf{M'}$ can become trapped in this portion of the citation network for a very long period in time, effectively overvaluing the journals in this clique. Teleporting corrects this problem by reducing the expected duration of a stay in these small cliques.

$$\mathbf{P} = \alpha\mathbf{N} + (1-\alpha)\mathbf{A},$$

where, with $\mathbf{e}^T$ as a row vector of 1's, $\mathbf{A} = \mathbf{a}.\mathbf{e}^T$ is a matrix with identical columns a, such that a_i = (articles in journal i) / (total articles).

We define the journal influence vector $\mathbf{f}$ as the leading eigenvector of $\mathbf{P}$; corresponds to steady-state fraction of time spent at each journal represented in $\mathbf{P}$. The journal influence vector $\mathbf{f}$ thus gives us our weights for the 6000 source items.

The weighted number of citations received by each of the 110,000 reference items is given by $\mathbf{M}.\mathbf{f}$. We define the *eigenfactor* w_i of journal i as the percentage of the total weighted citations that journal i receives from our 6000 source items. We can write the vector of eigenfactors as

$$\mathbf{w} = \frac{100\,\mathbf{M}\,\mathbf{f}}{\mathbf{e}^T\mathbf{M}\,\mathbf{f}}.$$

3 Calculating Article Influence

Eigenfactor provides a measure of the total influence that a journal provides, rather than a measure of influence per article. Impact factor, by contrast, measures the per-article influence of a given journal. To make our results comparable to impact factor, we need to divide the journal influence by the number of articles published. Let $\mathbf{a_i}$ be the total number of articles in journal i over the census period (5 years, in our case). We normalize $\mathbf{a}$ to give us a vector of the fractions of articles that each journal i contributes to the total literature:

$$b_i = \frac{a_i}{\sum_j a_j}$$

Finally, we compute the *article influence* as the ratio of the fractional contribution of journal i to the total eigenfactor ($w_i/100$) to the fractional contribution of journal i to the total articles published (b_i):

$$g_i = \frac{w_i}{100\,b_i}.$$

We teleport to a journal with probability proportional to the number of articles published by that journal in order to avoid over-inflating the influence of small journals and under-inflating the influence of large ones. This is important because the journals in the social sciences are much smaller, on average than the journals in the sciences. As a result, an unweighted teleportation process, in which one teleports to each journal with equal probability, overestimates influence of articles in social science journals relative to science journals because the teleportation process.

4 Assigning field classifications

Thompson Scientific provides field classifications for the 7000 journals listed in the Journal Citation Reports; each of these journals is assigned to one or more of 205 categories. We do not have field classifications for the 100,000 additional journals that we list, but we can use their citation patterns to assign them to primary categories. For each journal i, the distribution of incoming citations from the 7000 source journals is given by the $i-th$ column of the cross citation matrix M.

$$\mathbf{v}(i) = \mathbf{M}_{\cdot i}$$

Similarly, for each of the 205 JCR categories, we can construct a citation vector $\mathbf{v}(\mathrm{cat})$ representing the incoming citations to journals in this category:

$$v(\mathrm{cat}) = \sum_{j \in \mathrm{cat}} M_{\cdot j}$$

We assign each journal i to the field cat such that the angle between their respective citation vectors is the smallest. In other words, journal i is assigned to the category which maximizes

$$\frac{\mathbf{v}(i).\mathbf{v}(\mathrm{cat})}{\|\mathbf{v}(i)\| \, \|\mathbf{v}(\mathrm{cat})\|}$$

where $\| \cdot \|$ is the Euclidean norm.

Eigenfactor and Eigenfactor.org were developed by Jevin West, Ben Althouse, Martin Rosvall, Ted Bergstrom, and Carl Bergstrom at the University of Washington and the University of California Santa Barbara.