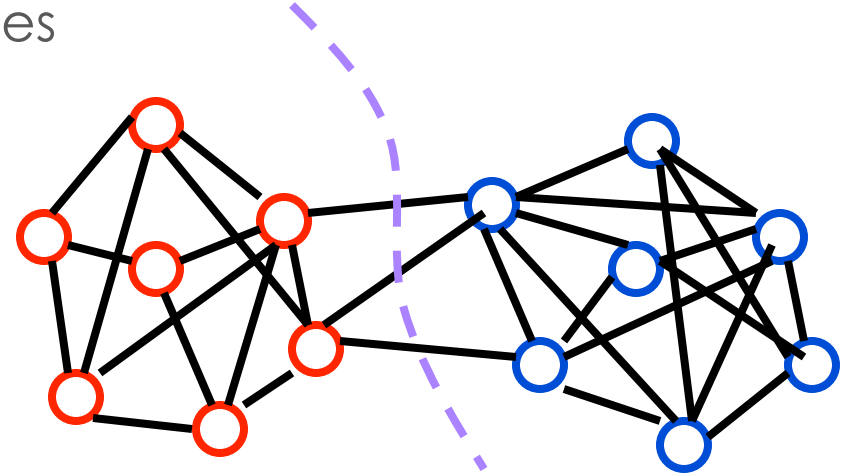
Community finding vs. other approaches

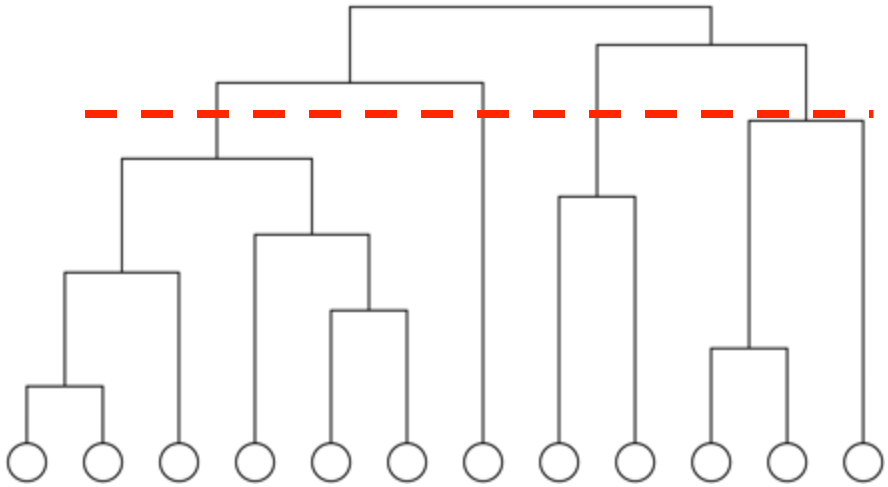
- Social and other networks have a natural community structure
- We want to discover this structure rather than impose a certain size of community or fix the number of communities



- Without "looking", can we discover community structure in an automated way?

Hierarchical clustering

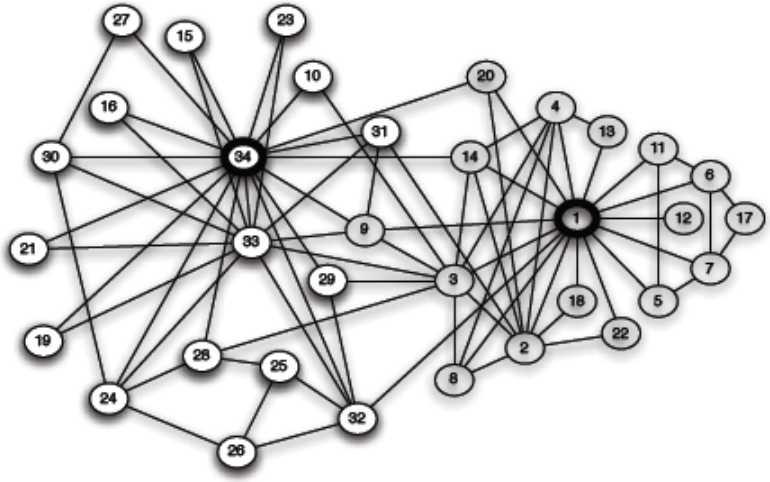
- Process:
  - after calculating the "distances" for all pairs of vertices
  - start with all n vertices disconnected
  - add edges between pairs one by one in order of decreasing weight
  - result: nested components, where one can take a 'slice' at any level of the tree



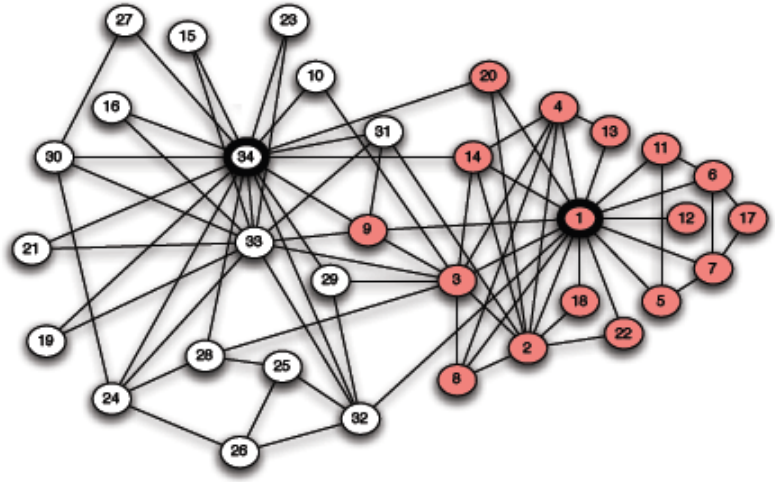
Hierarchical clustering

- Process:
  - after calculating the weights W for all pairs of vertices
  - start with all n vertices disconnected
  - add edges between pairs one by one in order of decreasing weight
  - Efficient and successful implementation in Pajek…

Zachary Karate Club

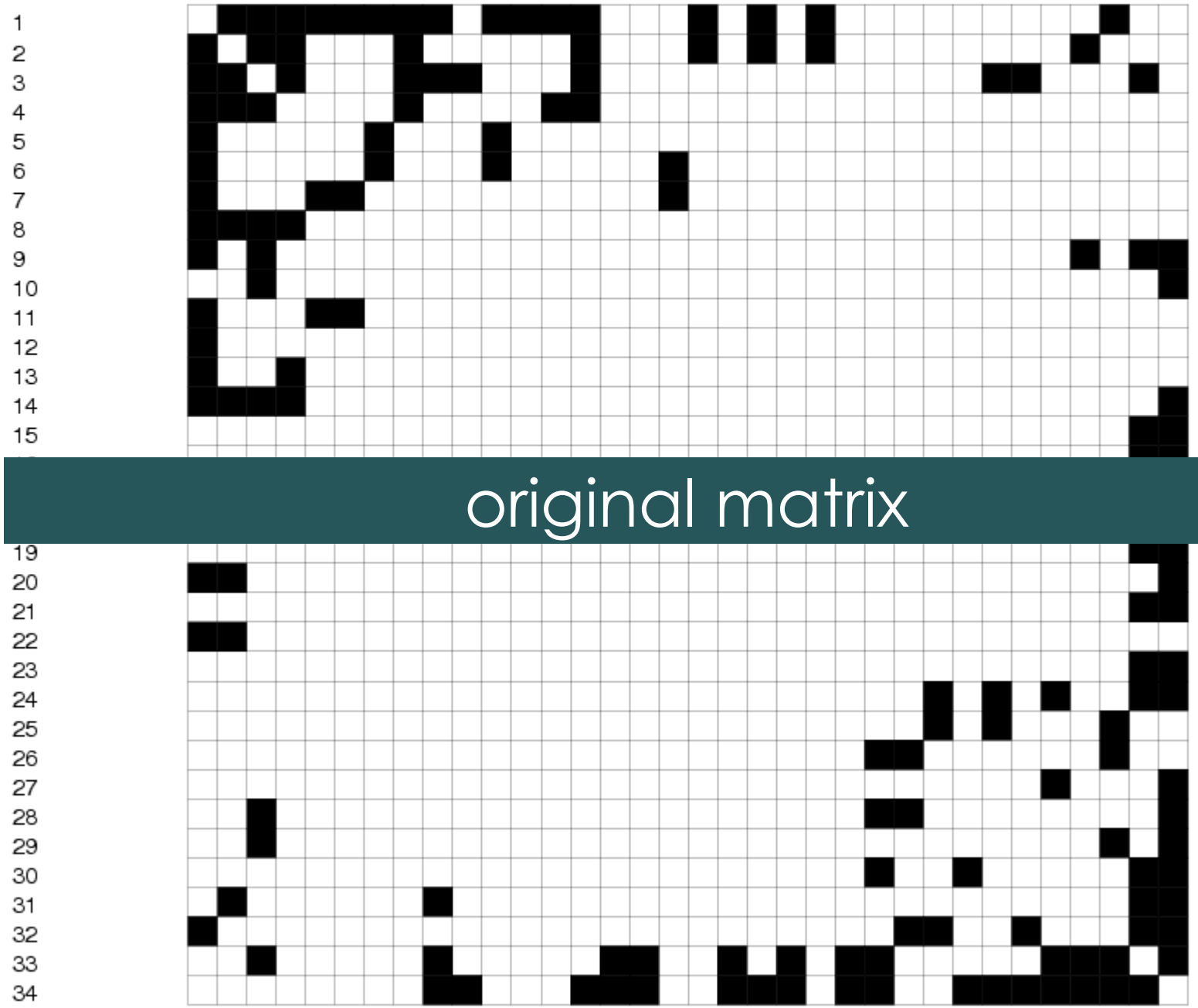


(a) *Karate club network*



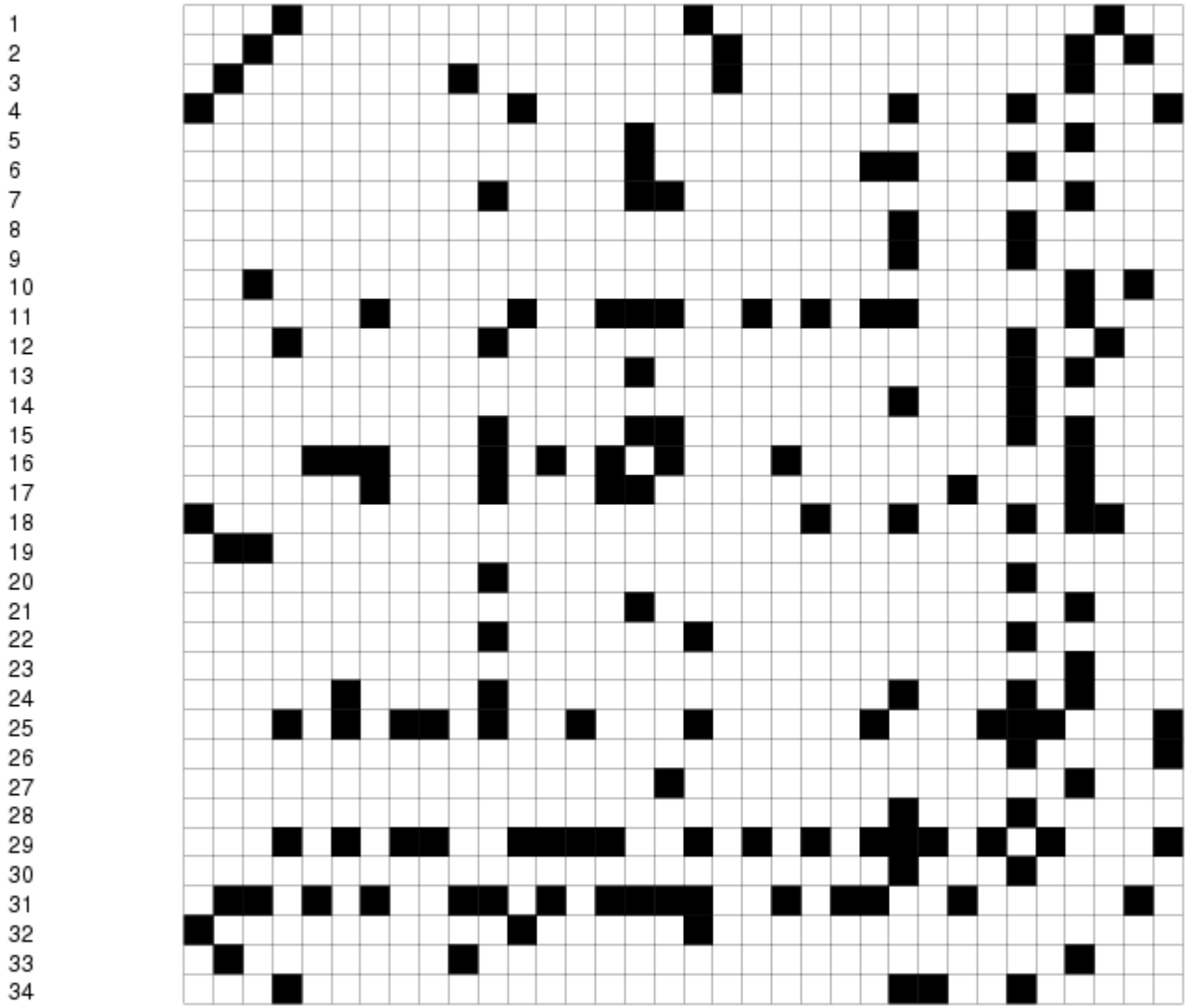
(b) *After a split into two clubs*

source:Easley/Kleinberg

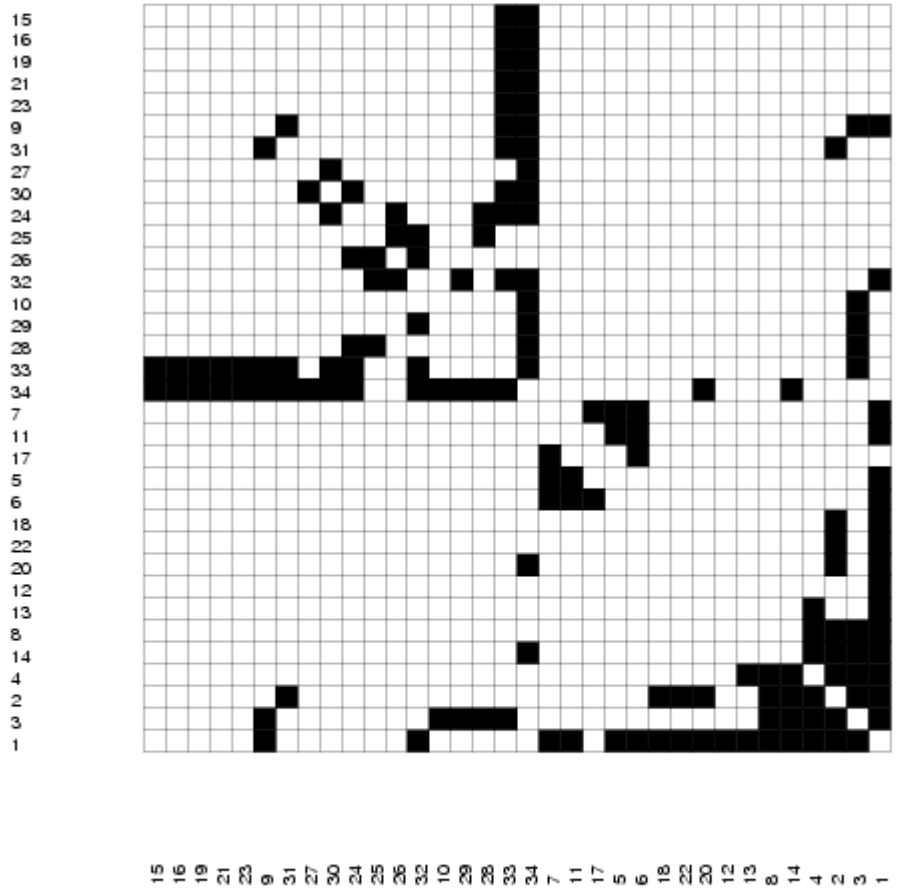


original matrix

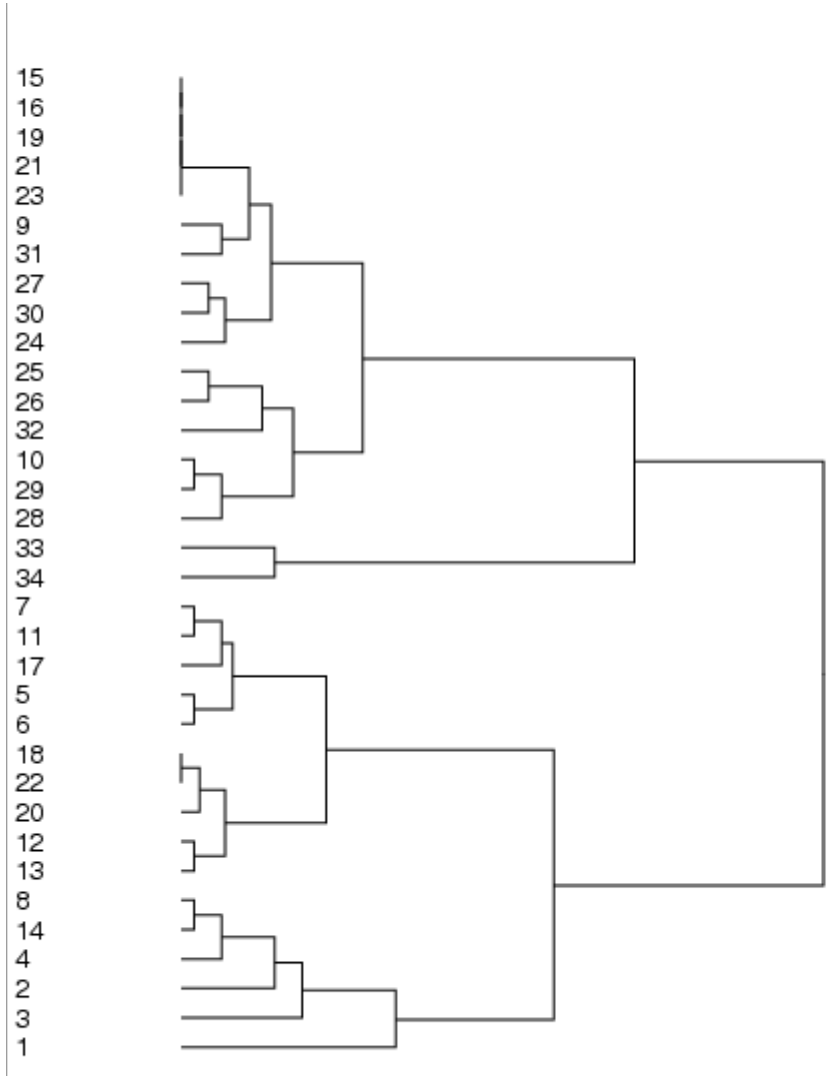
randomized karate club matrix



permuted matrix



dendrogram

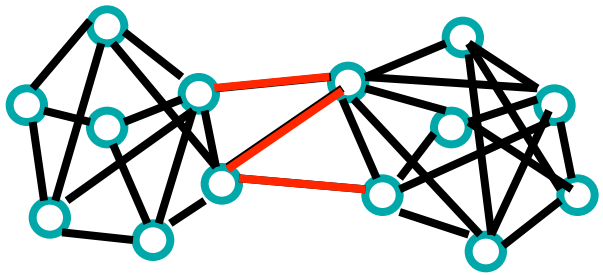

15
16
19
21
23
9
31
27
30
24
25
26
32
10
29
28
33
34
7
11
17
5
6
18
22
20
12
13
8
14
4
2
3
1

betweenness clustering

- Algorithm
  - compute the betweenness of all edges
  - while (betweenness of any edge > threshold):
    - remove edge with highest betweenness
    - recalculate betweenness

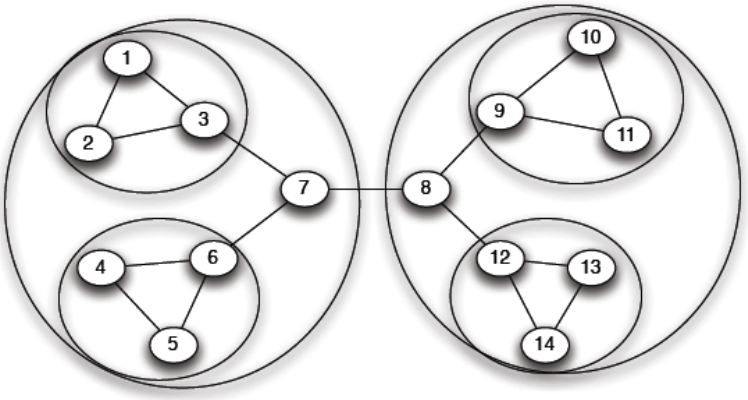
- Betweenness needs to be recalculated at each step
  - removal of an edge can impact the betweenness of another edge
  - very expensive: all pairs shortest path – $O(N^3)$
  - may need to repeat up to N times
  - does not scale to more than a few hundred nodes, even with the fastest algorithms

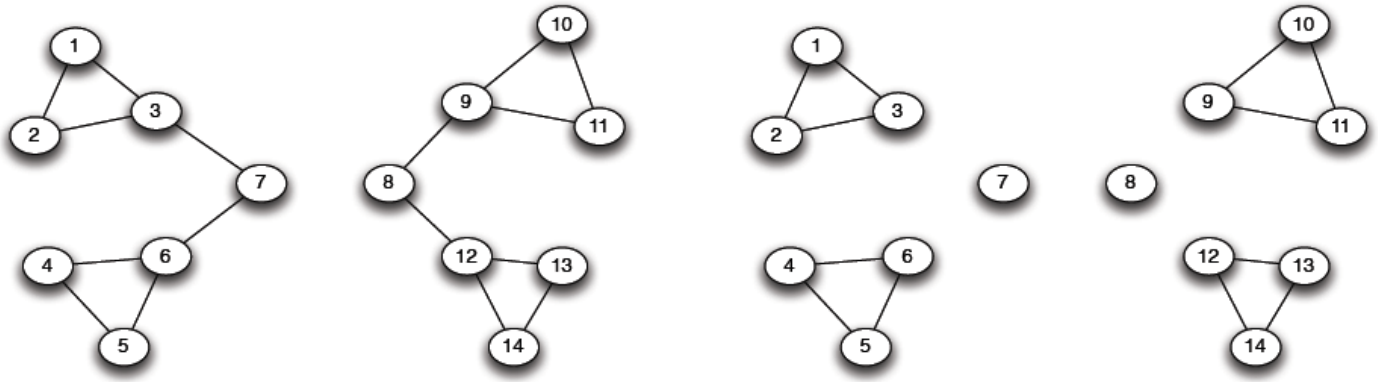
betweenness clustering algorithm



betweenness clustering:

- successively remove edges of highest betweenness (the bridges, or local bridges), breaking up the network into separate components

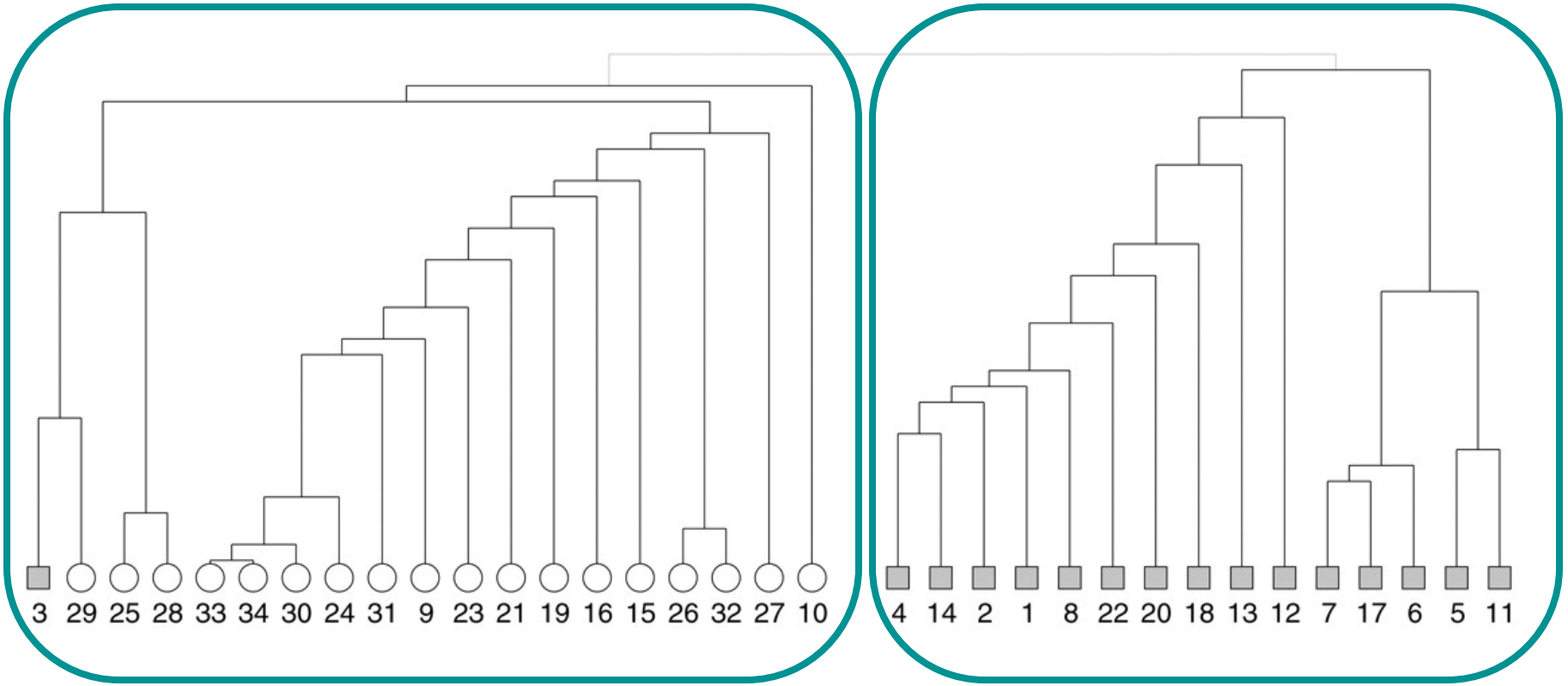




(a) *Step 1*

(b) *Step 2*

betweenness clustering algorithm & the karate club data set



source: Girvan and Newman, PNAS June 11, 2002 99(12):7821-7826

Modularity

- Consider edges that fall within a community or between a community and the rest of the network
- Define modularity:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w)$$

if vertices are in the same community

adjacency matrix

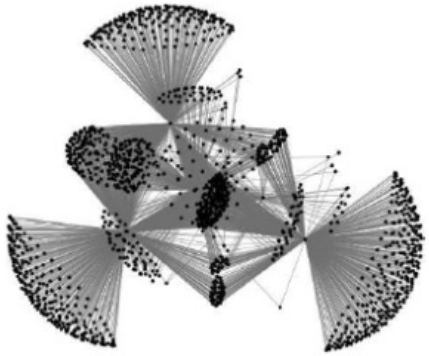
probability of an edge between two vertices is proportional to their degrees

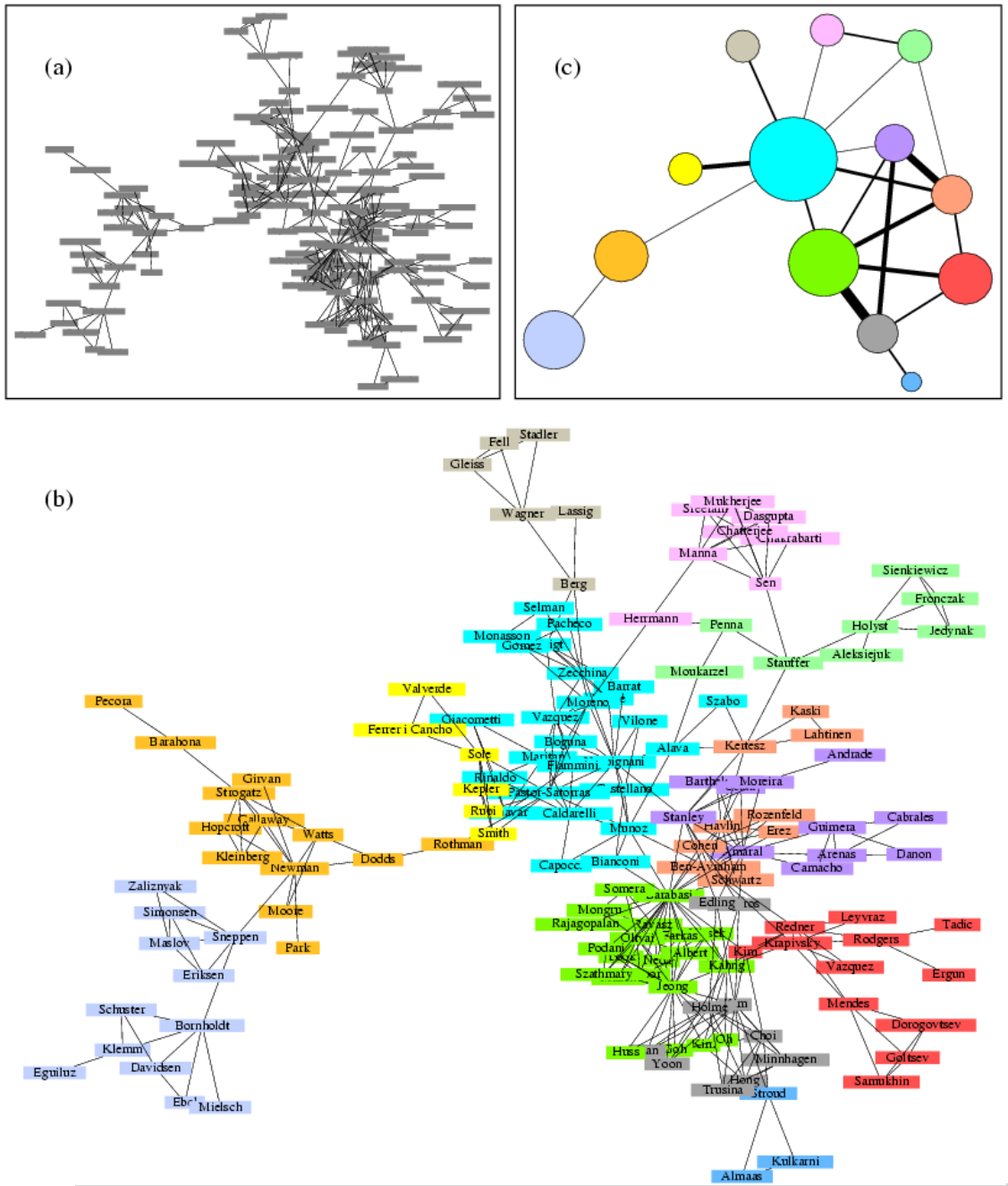
- For a random network, Q = 0
  - the number of edges within a community is no different from what you would expect

Finding community structure in very large networks
Authors: Aaron Clauset, M. E. J. Newman, Cristopher Moore 2004

Modularity

- Algorithm
  - start with all vertices as isolates
  - follow a greedy strategy:
    - successively join clusters with the greatest increase ΔQ in modularity
    - stop when the maximum possible $\Delta Q <= 0$ from joining any two
  - successfully used to find community structure in a graph with > 400,000 nodes with > 2 million edges
    - Amazon's people who bought this also bought that…
  - alternatives to achieving optimum ΔQ:
    - simulated annealing rather than gre





Reminder of how modularity can help us visualize large networks

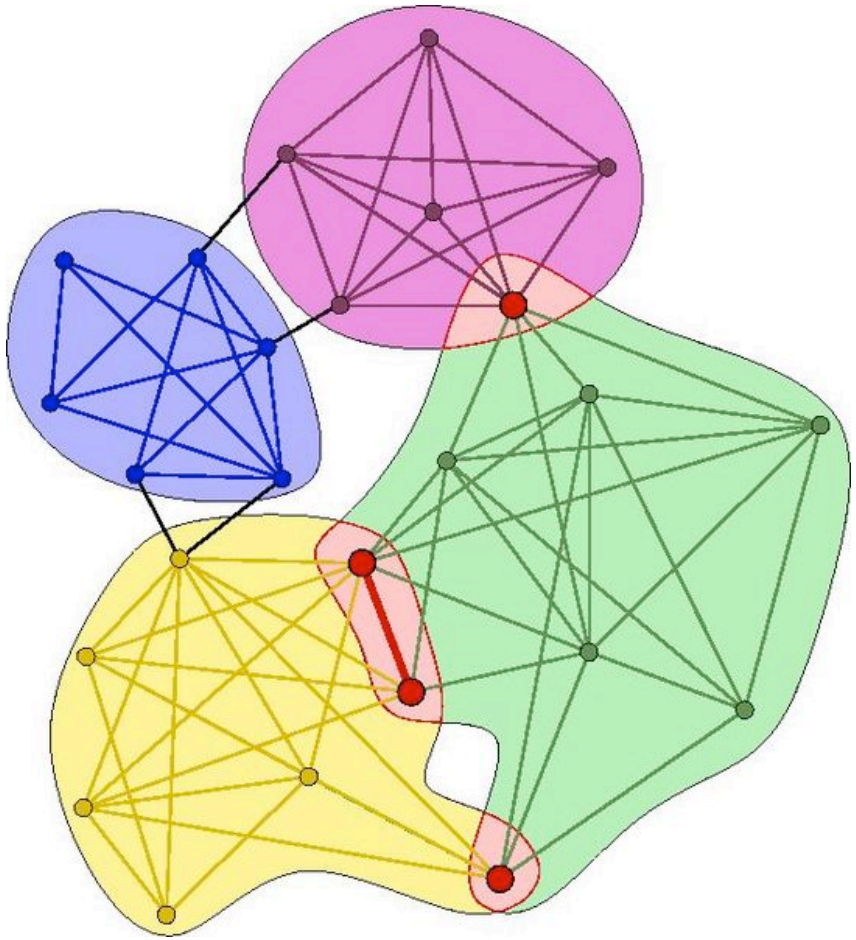
What if communities overlap?

- Recent research has found that for communities such as Orkut and FlickR, community finding algorithms cannot identify communities of more than ~100 nodes
- Statistical Properties of Community Structure in Large Social and Information Networks by J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. *International World Wide Web Conference (WWW)*, 2008. [Video]

Clique finder

- http://cfinder.org

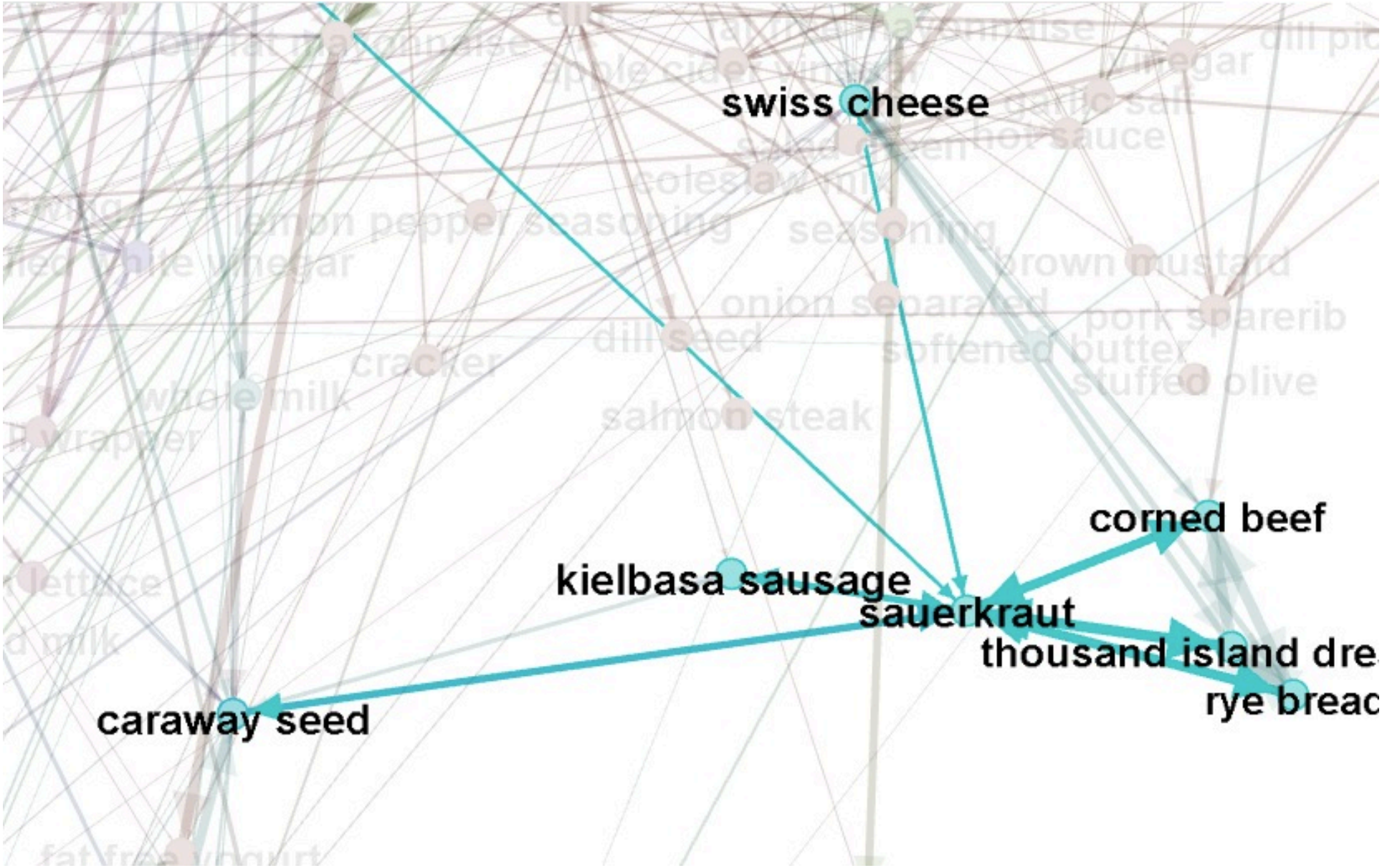
Uncovering the overlapping community structure of complex networks in nature and society G. Palla, I. Derényi, I. Farkas, and T. Vicsek: Nature 435, 814–818 (2005)



wrap up

- community structure is a way of 'x-raying' the network, finding out what it's made of
- you can look for specific structures
  - k-cliques, k-cores, etc.
- but most popular is to discover the "natural" community boundaries

For your assignment: community finding & ingredients ☺



An information theoretic approach

- How to most concisely describe a random walk on the network using huffman codes? Prefixes become communities…

- http://www.***pnas***.org/content/105/4/1118
  http://www.mapequation.org/mapgenerator/index.html