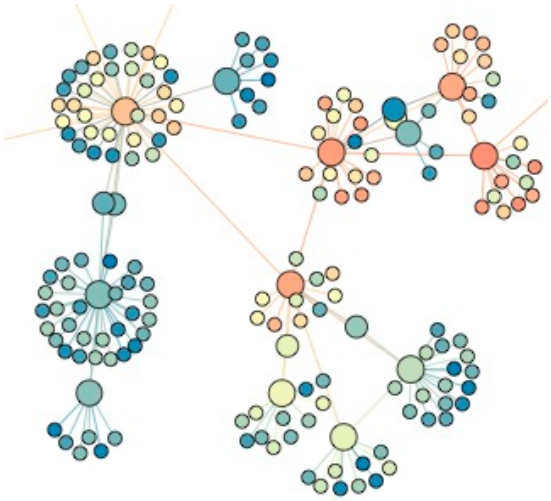


community structure

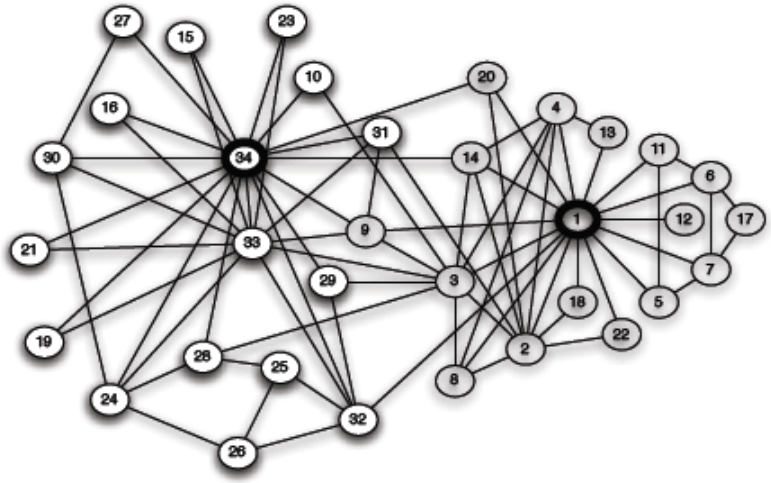
http://barabasi.com/networksciencebook/content/book_chapter_9.pdf

Outline

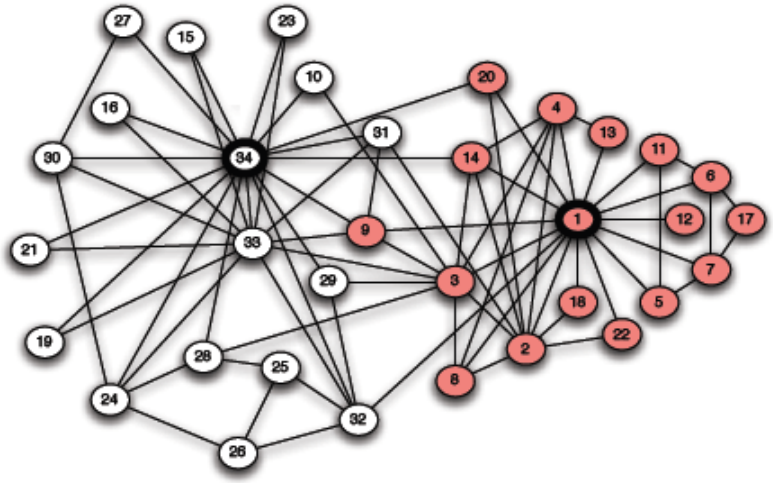
- why do we look for community structure?
- we need to define it in order to find it
- approaches to finding it

In network science we call a community a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities.

Why look for community structure?Zachary Karate Club



(a) *Karate club network*



(b) *After a split into two clubs*

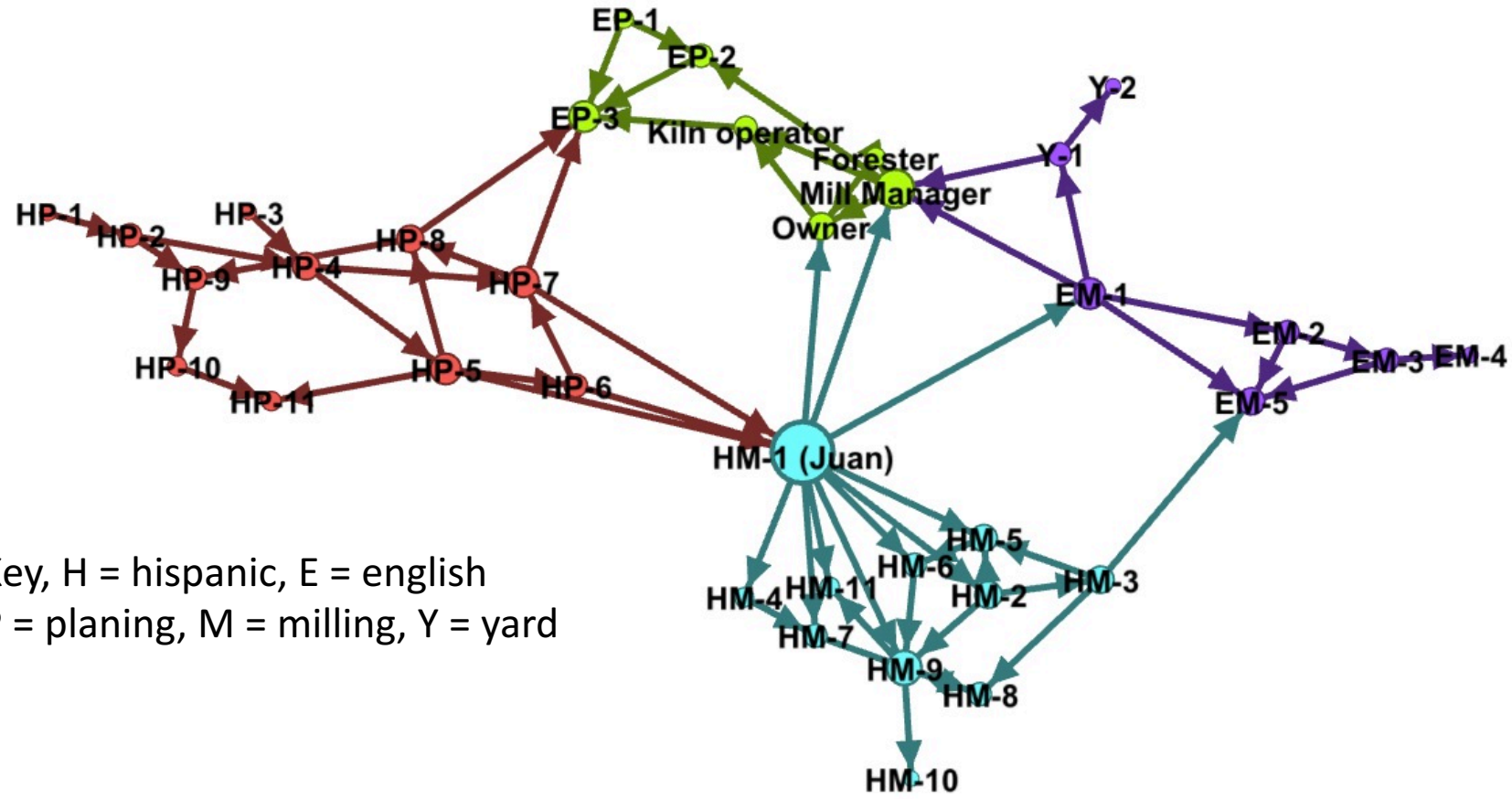
source:Easley/Kleinberg

Why look for community structure?

The existence of communities is rooted in who connects to whom, hence they cannot be explained based on the degree distribution alone. To extract communities we must therefore inspect a network's detailed wiring diagram.

Fundamental Hypotesis 1: A network's community structure is uniquely encoded in its wiring diagram

Why look for community structure?

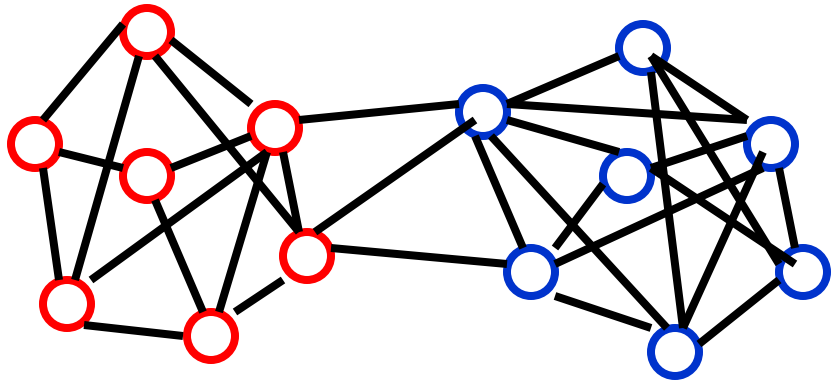


Key, H = hispanic, E = english
P = planing, M = milling, Y = yard

Sawmill network: source Exploratory Social Network Analysis with Pajek

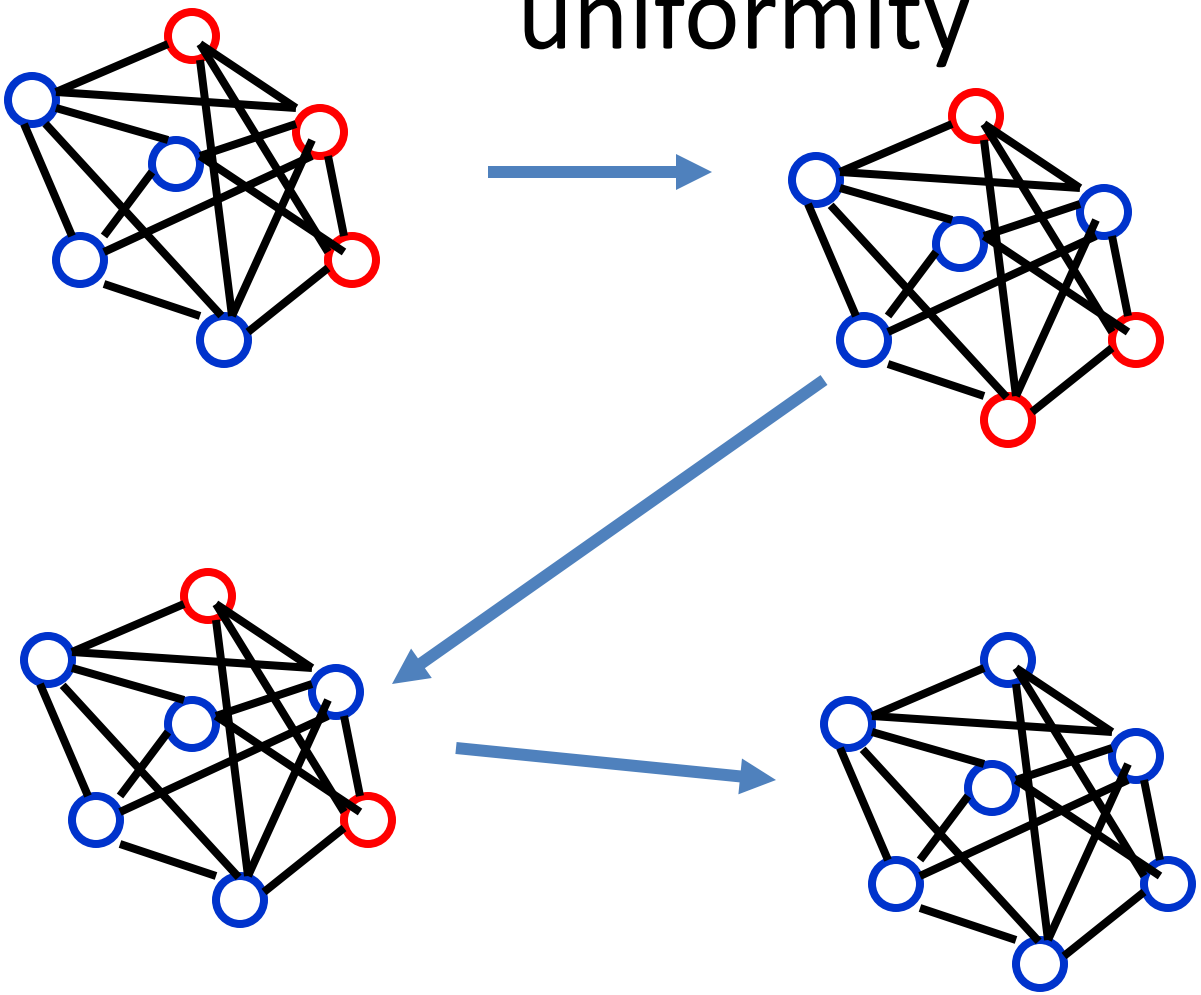
Why care about group cohesion?

- opinion formation and uniformity



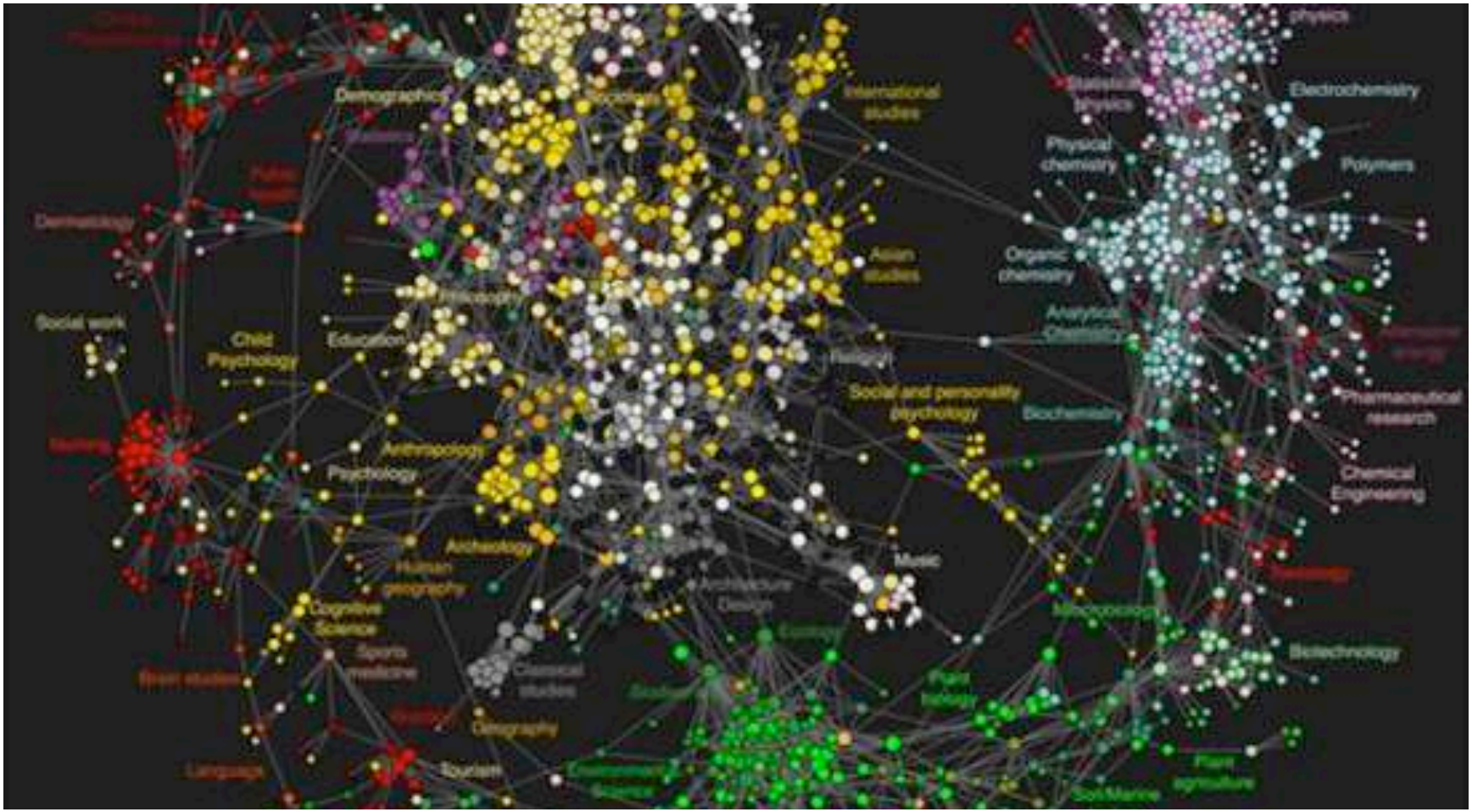
- if each node adopts the opinion of the majority of its neighbors, it is possible to have different opinions in different cohesive subgroups

within a cohesive subgroup – greater uniformity



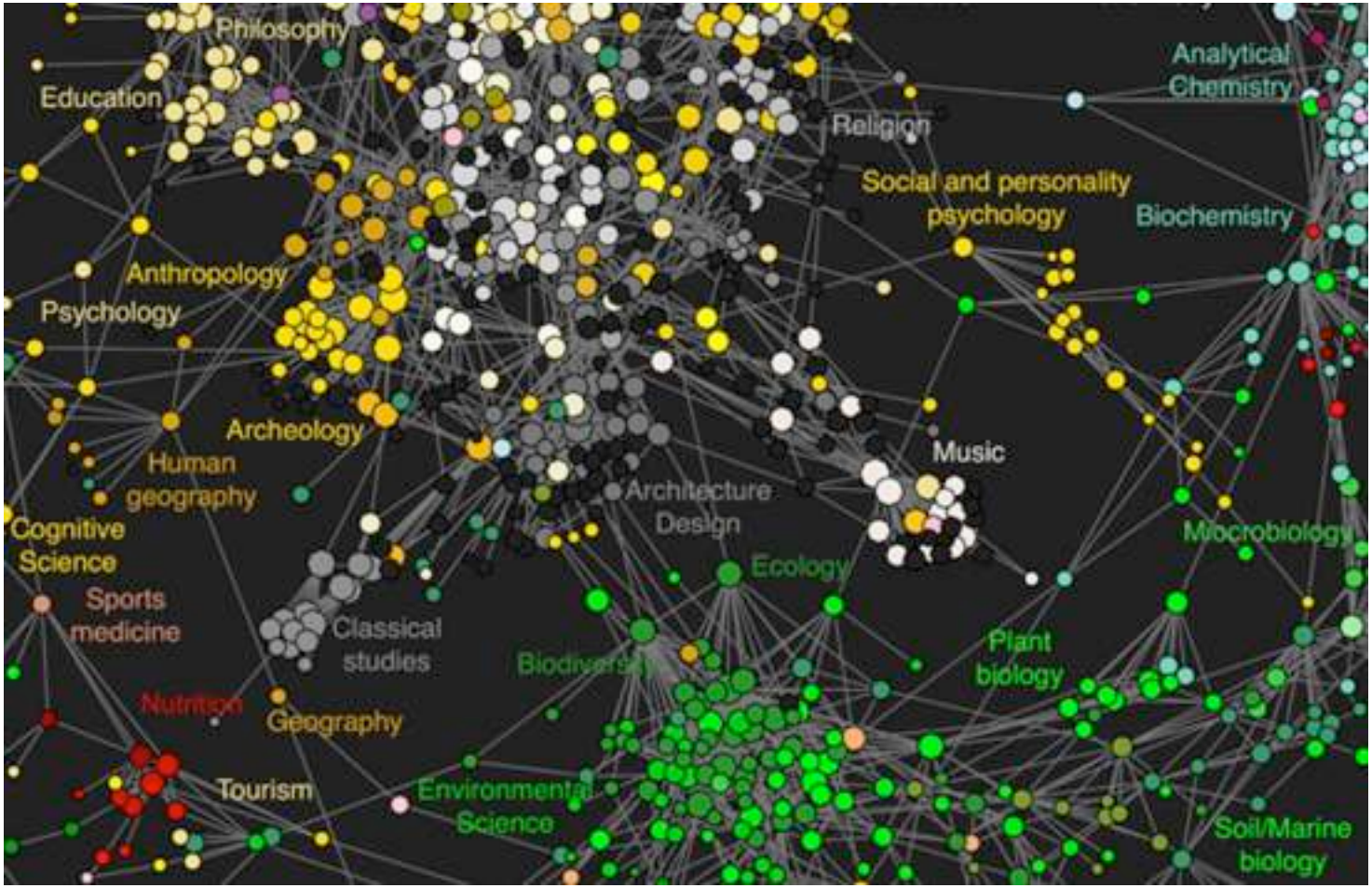
high-res maps of science

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004803



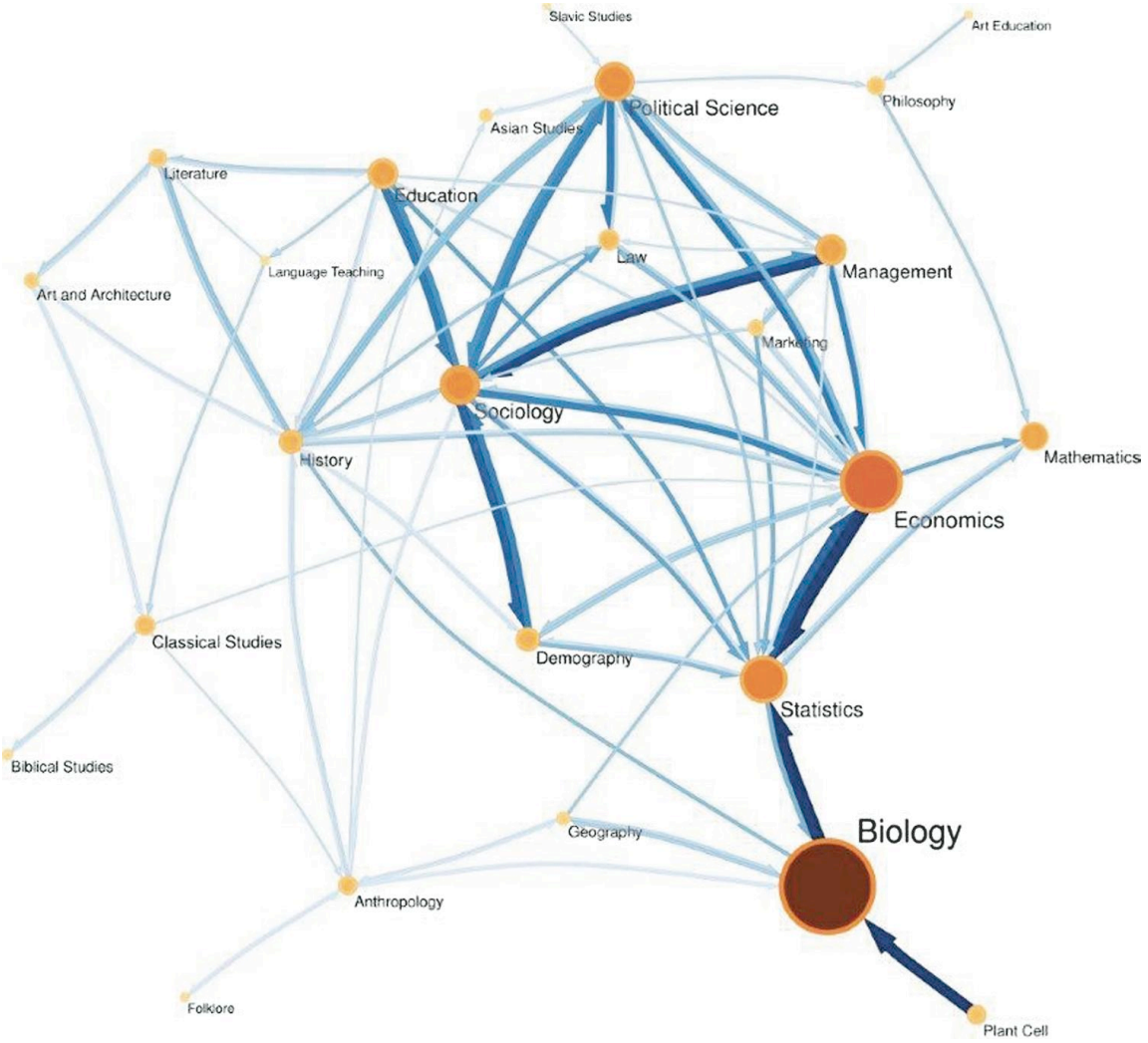
high-res maps of science

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004803



high-res maps of science

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004803



What makes a community?

- mutuality of ties
  - everybody in the group knows everybody else
- frequency of ties among members
  - everybody in the group has links to at least k others in the group
- closeness or reachability of subgroup members
  - individuals are separated by at most n hops
- relative frequency of ties among subgroup members compared to nonmembers

What makes a community?

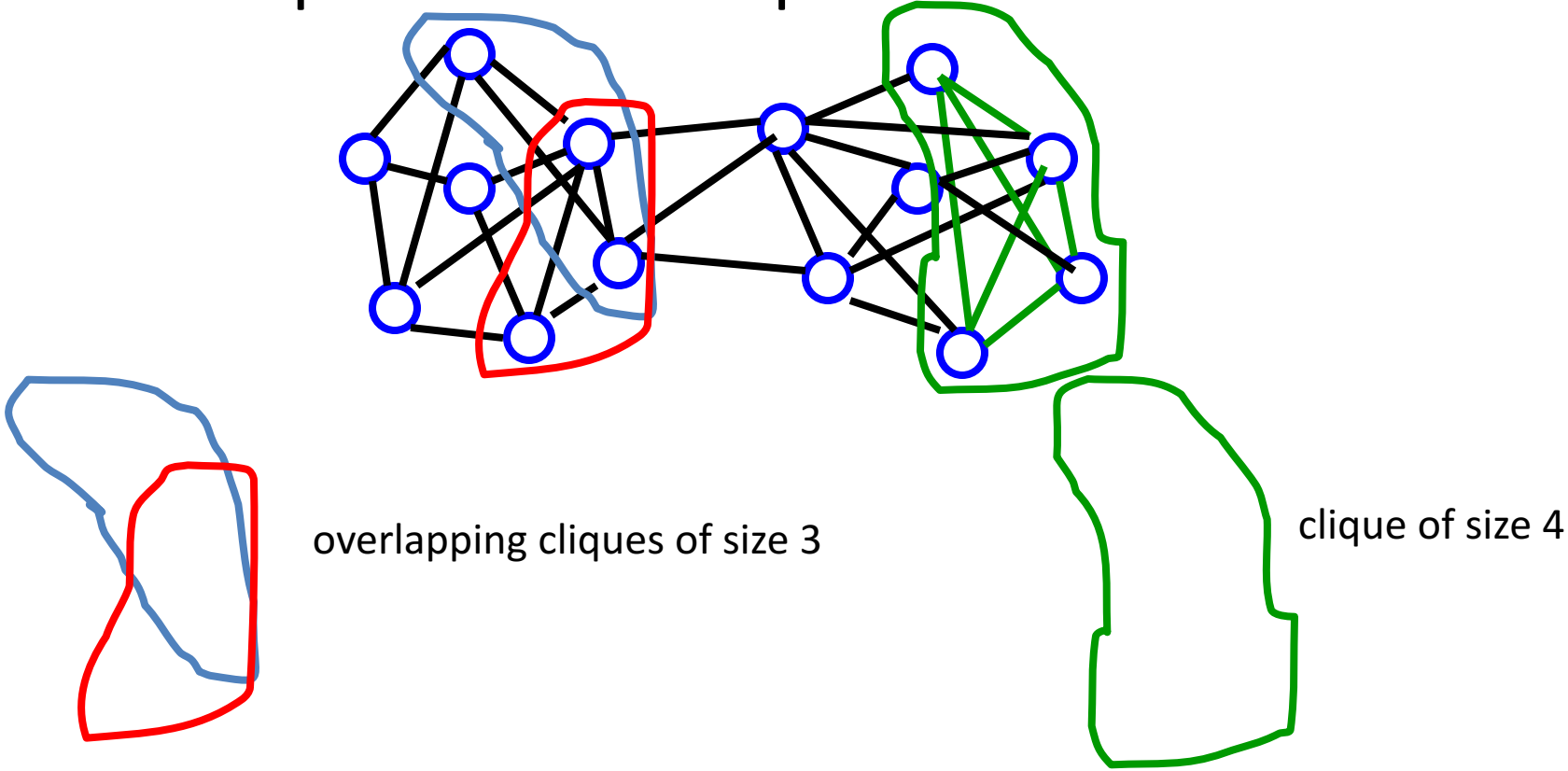
Fundamental hypotesis 2: A community is a locally dense & connected subgraph in a network

While this hypothesis considerably narrows what would be considered a community, it does not uniquely define it. Indeed, several community definitions are consistent with it:

- Maximum cliques
- Strong and Weak communities

Maximum Cliques

- Every member of the group has links to every other member
- Cliques can overlap



Drawbacks of cliques

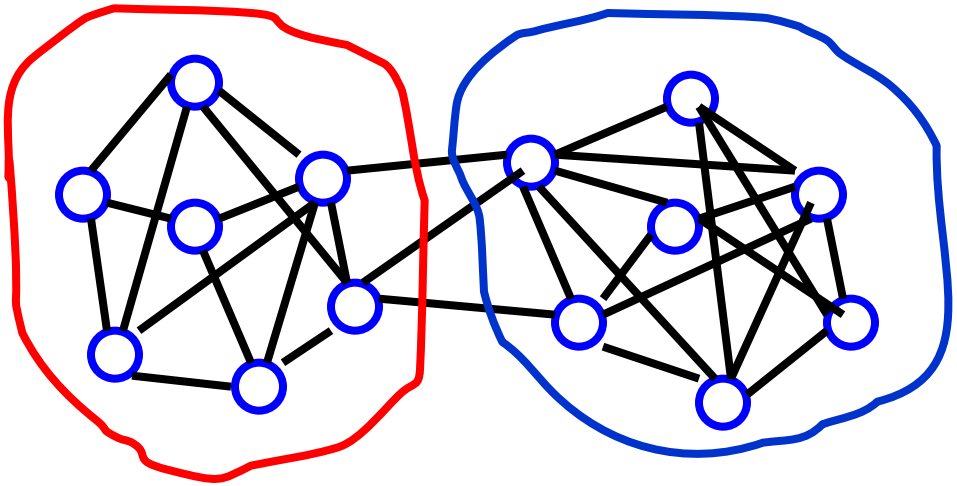
- Not robust
  - one missing link can disqualify a clique
- Not interesting
  - everybody is connected to everybody else
  - no core-periphery structure
  - no centrality measures apply
- How cliques overlap can be more interesting than that they exist

Strong and Weak Communities

- C is a strong community if each node within C has more links within the community than with the rest of the graph
- C is a weak community if the total internal degree of a subgraph exceeds its total external degree

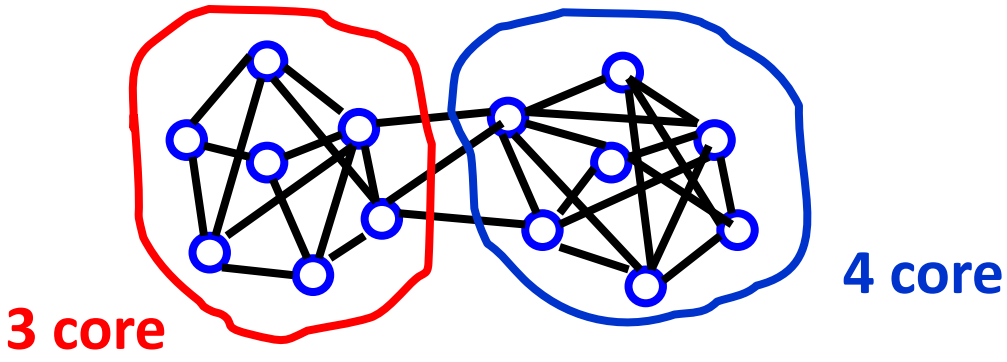
k-cores: similar idea, less stringent

- Each node within a group is connected to k other nodes in the group

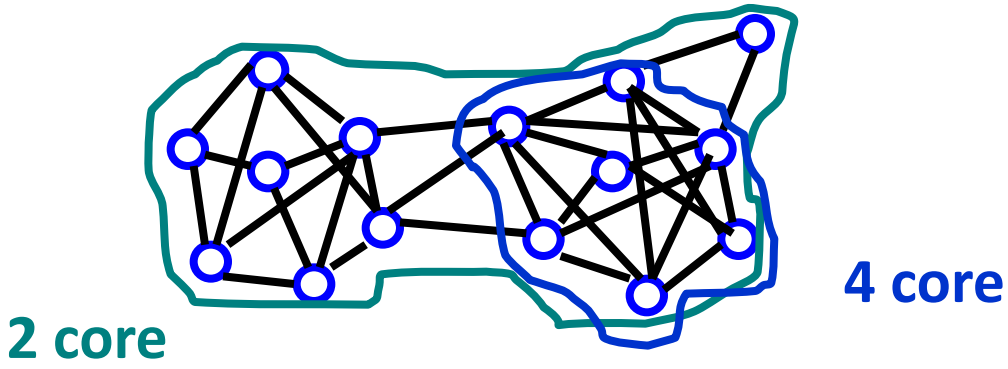


k-cores

- Each node within a group is connected to k other nodes in the group

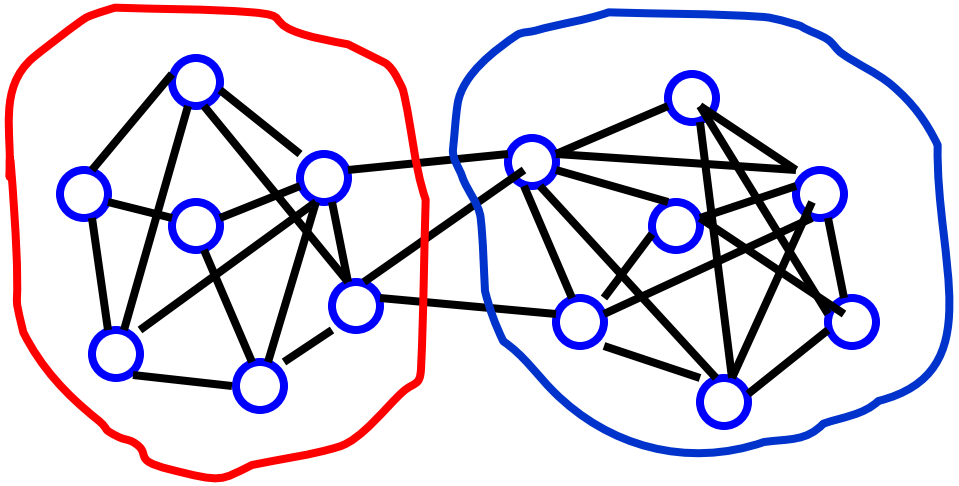


- but even this is too stringent of a requirement for identifying natural communities



subgroups based on reachability and diameter

- n – cliques
  - maximal distance between any two nodes in subgroup is n

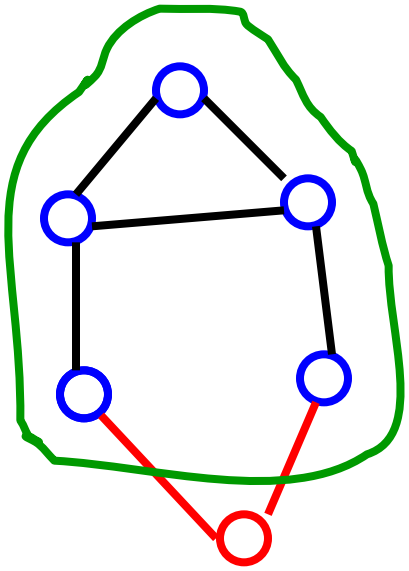


2-cliques

- theoretical justification
  - information flow through intermediaries

considerations with n-cliques

- problem
  - diameter may be greater than n
  - n-clique may be disconnected (paths go through nodes not in subgroup)



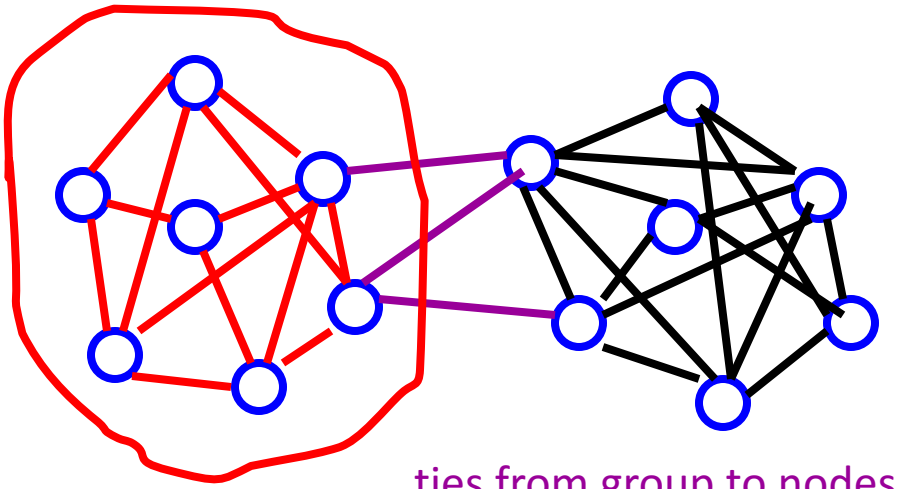
2 – clique
diameter = 3

path outside the 2-clique

- fix
  - n-club: maximal subgraph of diameter 2

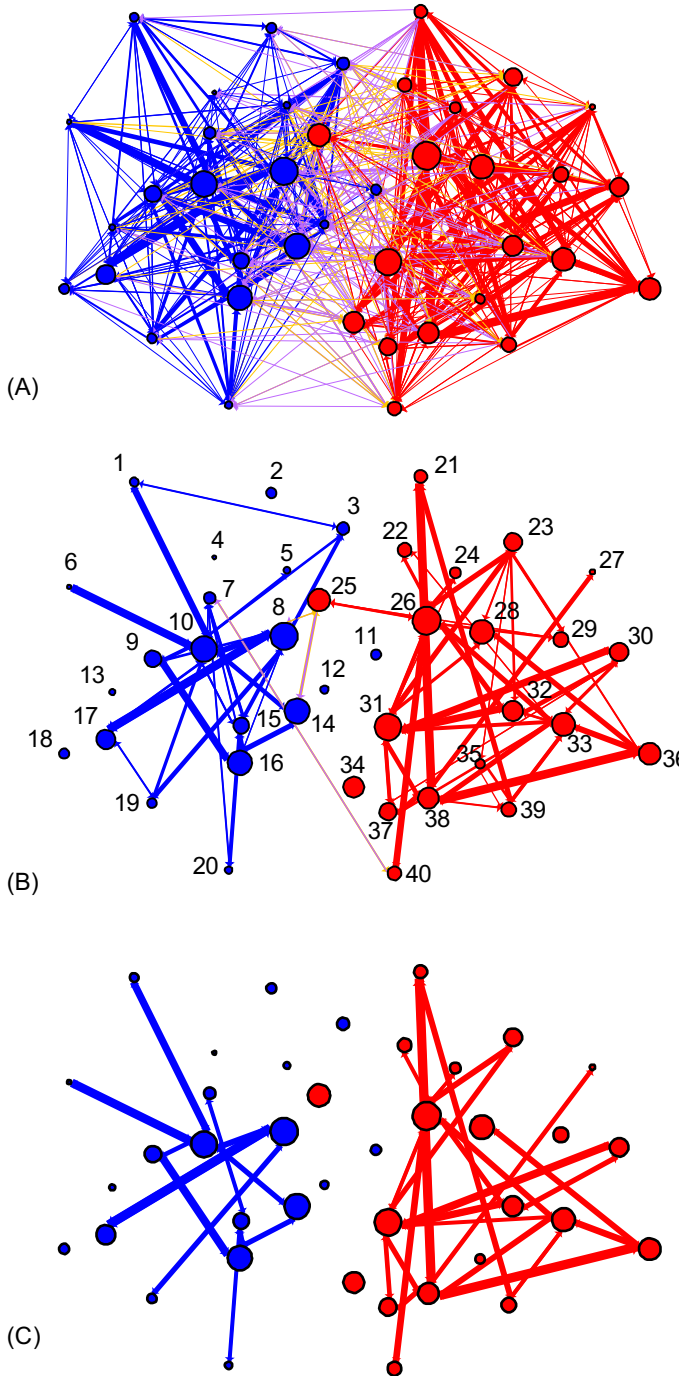
p-cliques: frequency of in group ties

- partition the network into clusters where vertices have at least a proportion p (number between 0 and 1) of neighbors inside the cluster.



within-group ties

ties from group to nodes external to the group



1 Digbys Blog
2 James Walcott
3 Pandagon
4 blog.johnkerry.com
5 Oliver Willis
6 America Blog
7 Crooked Timber
8 Daily Kos
9 American Prospect
10 Eschaton
11 Wonkette
12 Talk Left
13 Political Wire
14 Talking Points Memo
15 Matthew Yglesias
16 Washington Monthly
17 MyDD
18 Juan Cole
19 Left Coaster
20 Bradford DeLong

21 Jawa Report
22 Voka Pundit
23 Roger L Simon
24 Tim Blair
25 Andrew Sullivan
26 Instapundit
27 Blogs for Bush
28 Little Green Footballs
29 Belmont Club
30 Captain's Quarters
31 Powerline
32 Hugh Hewitt
33 IND Journal
34 Real Clear Politics
35 Winds of Change
36 Allahpundit
37 Michelle Malkin
38 WizBang
39 Dean's World
40 Volokh

Example: political blogs

(Aug 29th – Nov 15th, 2004)

A) all citations between A-list blogs in 2 months preceding the 2004 election

B) citations between A-list blogs with at least 5 citations in both directions

C) edges further limited to those exceeding 25 combined citations

only 15% of the citations bridge communities

source: Adamic & Glance, LinkKDD2005

What makes a community?

These community definitions refine our notions of communities. At the same time they indicate that we do have some freedom in defining communities.

And how many communities?

How many ways can we group the nodes of a network into communities?

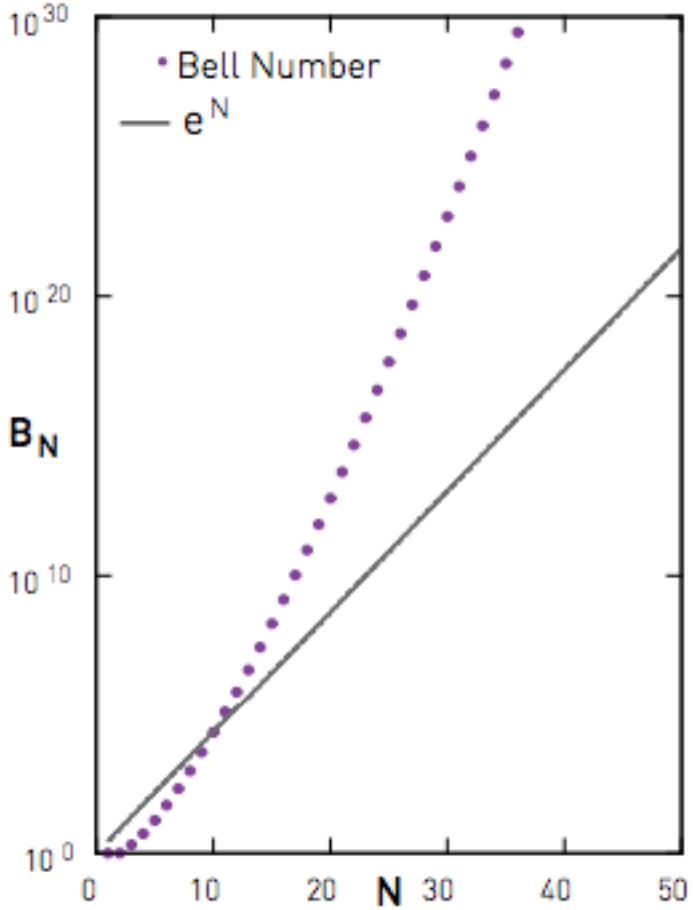
To answer this question consider the simplest community finding problem, called **graph bisection**: We aim to divide a network into two non-overlapping subgraphs, such that the number of links between the nodes in the two groups, called the cut size, is minimized.

We call a partition a division of a network into an arbitrary number of groups, such that each node belongs to one and only one group

Brute force approach for graph partitioning.

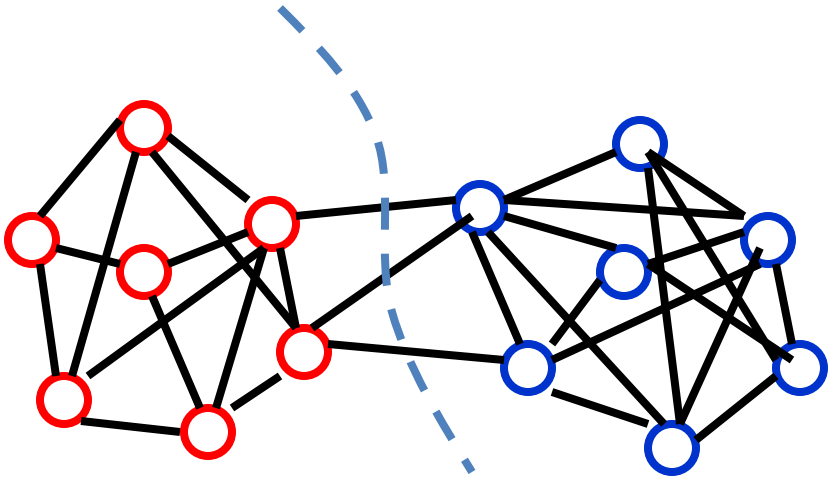
We can compute all possible divisions into two groups and choose the one with the smallest cut size.

Problem: computationally infeasible, the number of partitions of a network of size N is provided by the Bell number



The number of possible ways we can partition a network into communities grows exponentially or faster with the network size N

- 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?

Community detection

While in graph partitioning the number and the size of communities is predefined, in community detection both parameters are unknown.

We need algorithms that can identify communities without inspecting all partitions

Hierarchical clustering

To uncover the community structure of large real networks we need algorithms whose running time grows polynomially with N. Hierarchical clustering helps us achieve this goal.

After calculating the similarity matrix (whose elements indicate the distance of node i from node j), two approaches:

- agglomerative algorithms merge nodes with high similarity into the same community

- divisive algorithms isolate communities by removing low similarity link that tend to connect communities

Both procedures generate a hierarchical tree, called a dendrogram, that predicts the possible community partitions

Agglomerative procedures: the Ravasz algorithm

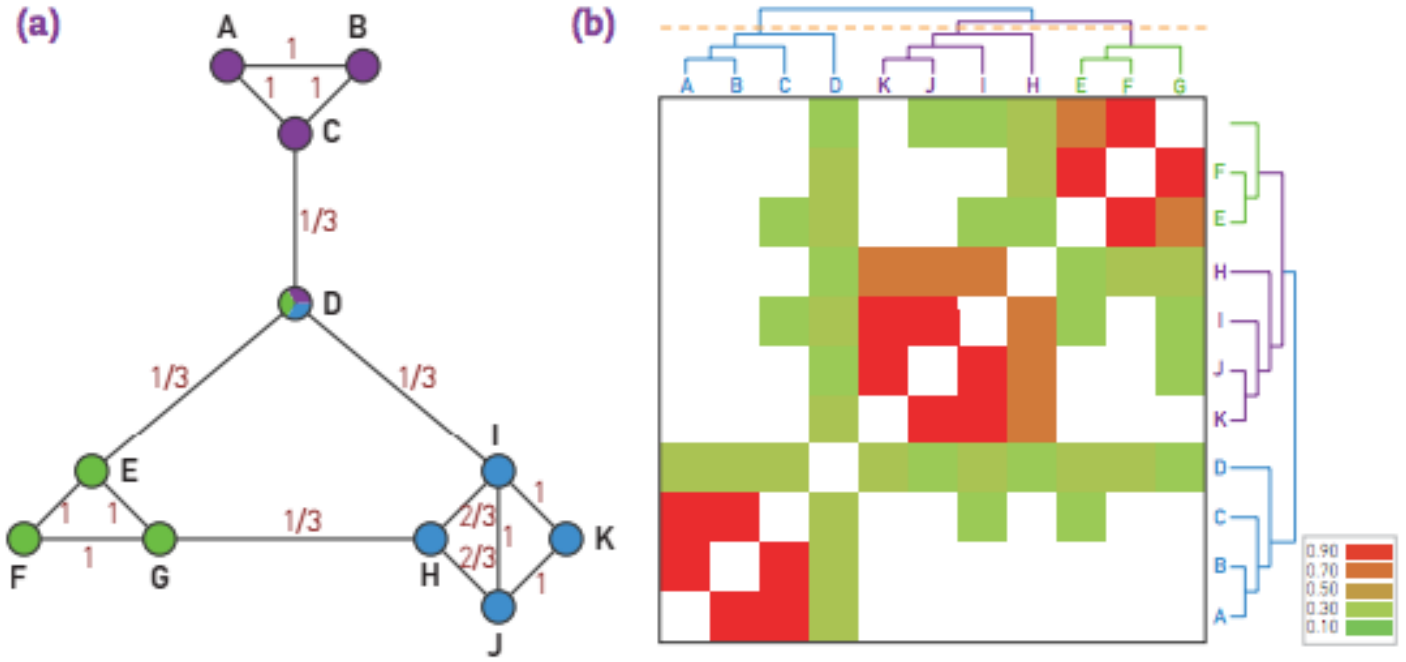
- Step 1 Define the similarity matrix

$$x_{ij}^{0} = \frac{J(i,j)}{\min(k_i, k_j) + 1 - \Theta(A_{ij})}$$

The topological overlap matrix is such that nodes that connect to each other and share neighbors likely to belong to the same community have a large similarity

Heaviside step function: zero for x<=0, 1 for x> 0

Agglomerative procedures: the Ravasz algorithm



- X_{ij}^0=1 if nodes i and j have a link to each other ahd have the same neighbors (A-B)
- X_{ij}^0=0 if i and j do not have a common neighbors, nor do they link to each other (A-E)
- Members of the same dense local neighborhood have high topological overlap (H-I-J-K)

Agglomerative procedures: the Ravasz algorithm

- Step 2: Decide group similarity

The Ravasz algorithm uses the average cluster similarity method, defining the similarity of two communities as the average of xij over all node pairs i and j that belong to distinct communities

Agglomerative procedures: the Ravasz algorithm

- Step 3 Apply Hierarchical Clustering

1. Assign each node to a community of its own and evaluate xij for all node pairs.
2. Find the community pair or the node pair with the highest similarity and merge them into a single community.
3. Calculate the similarity between the new community and all other communities.
4. Repeat Steps 2 and 3 until all nodes form a single community.

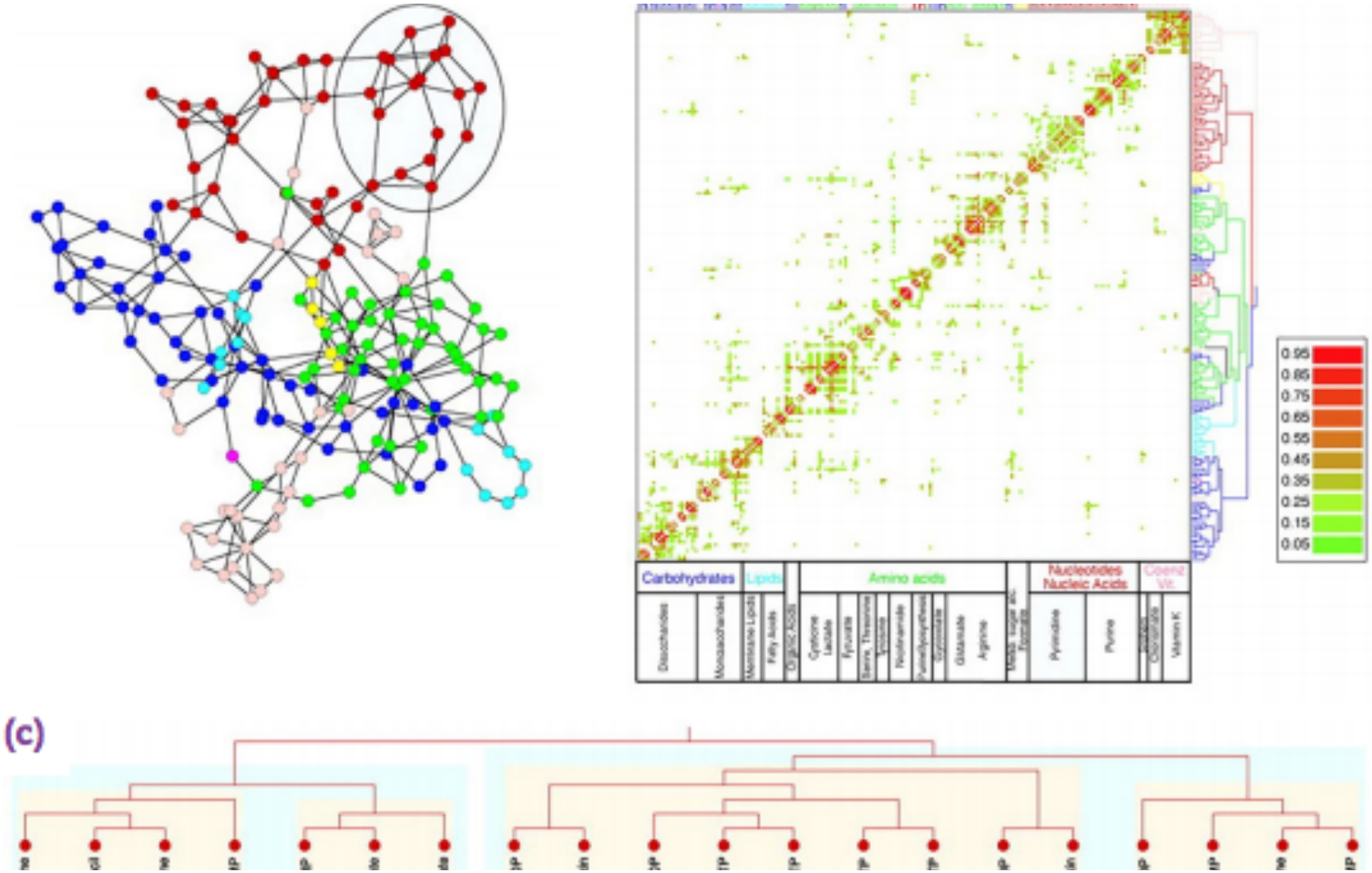
Agglomerative procedures: the Ravasz algorithm

The pairwise mergers of Step 3 will eventually pull all nodes into a single community. We can use a dendrogram to extract the underlying community organization.

The dendrogram visualizes the order in which the nodes are assigned to specific communities.

To identify the communities we must cut the dendrogram. Hierarchical clustering does not tell us where that cut should be.

E-coli metabolic network



Agglomerative procedures: the Ravasz algorithm

Computational complexity:

Combining Steps 1-4, we find that the number of required computations scales as $0(N^2) + 0(N^2) + 0(NlogN)$. As the slowest step scales as $0(N^2)$,

the algorithm's computational complexity is $0(N^2)$. Hence hierarchal clustering is much faster than the brute force approach, which generally scales

as $0(e^N)$.

Divisive algorithm: betweenness clustering or the Girvan-Newman algorithm

- Step 1: Define Centrality

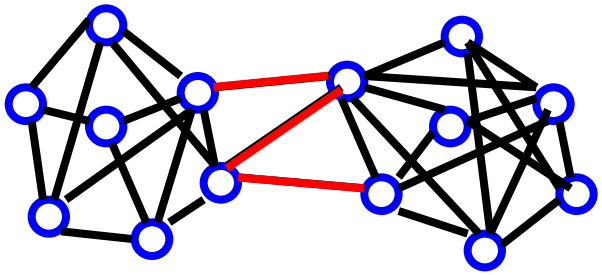
In divisive algorithms x_{ij}, called *centrality*, selects node pairs that are in different communities. Hence we want x_{ij} to be high (or low) if nodes *i* and *j* belong to different communities and small if they are in the same community The fastest of the three centrality measures is ***link betweenness***, defining x_{ij} as the number of shortest paths that go through the link (*i, j*). Links connecting different communities are expected to have large x_{ij} while links within a community have small x_{ij}.

Divisive algorithm: betweenness clustering or the Girvan-Newman algorithm

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

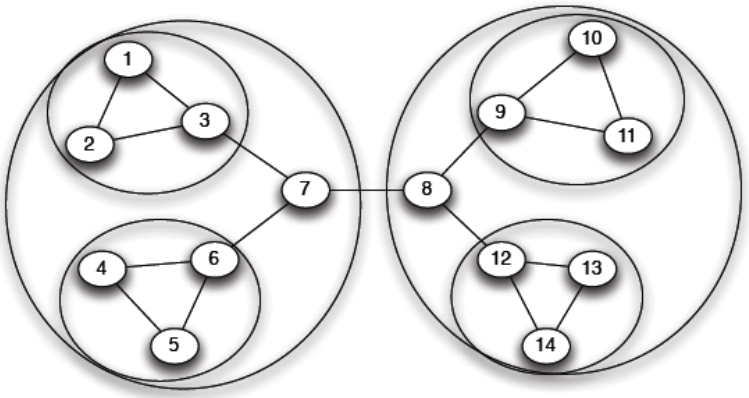
- Betweenness needs to be recalculated at each step
  - removal of an edge can impact the betweenness of another edge
  - very expensive: all pairs shortest path – $O(N^3)$
  - may need to repeat up to N times
  - does not scale to more than a few hundred nodes, even with the fastest algorithms

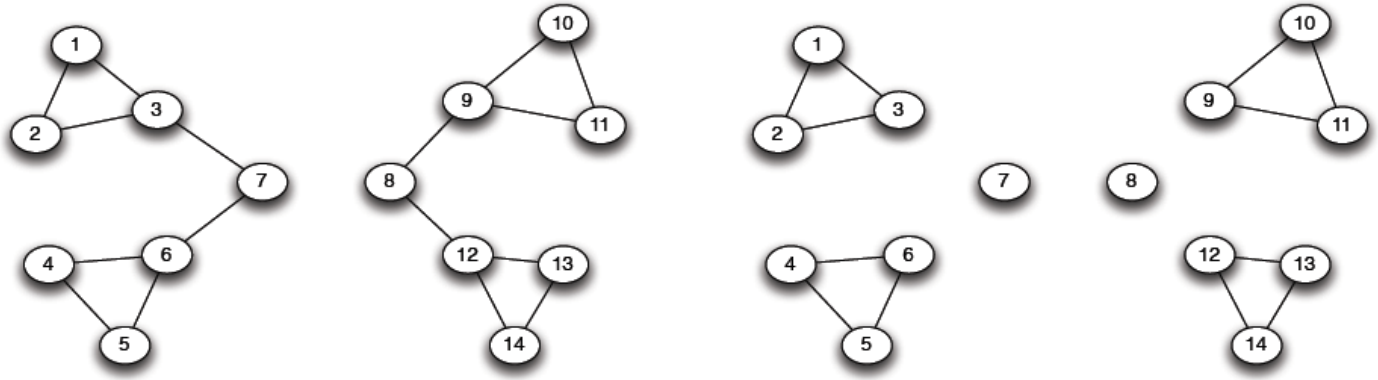
betweenness clustering algorithm



betweenness clustering:

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

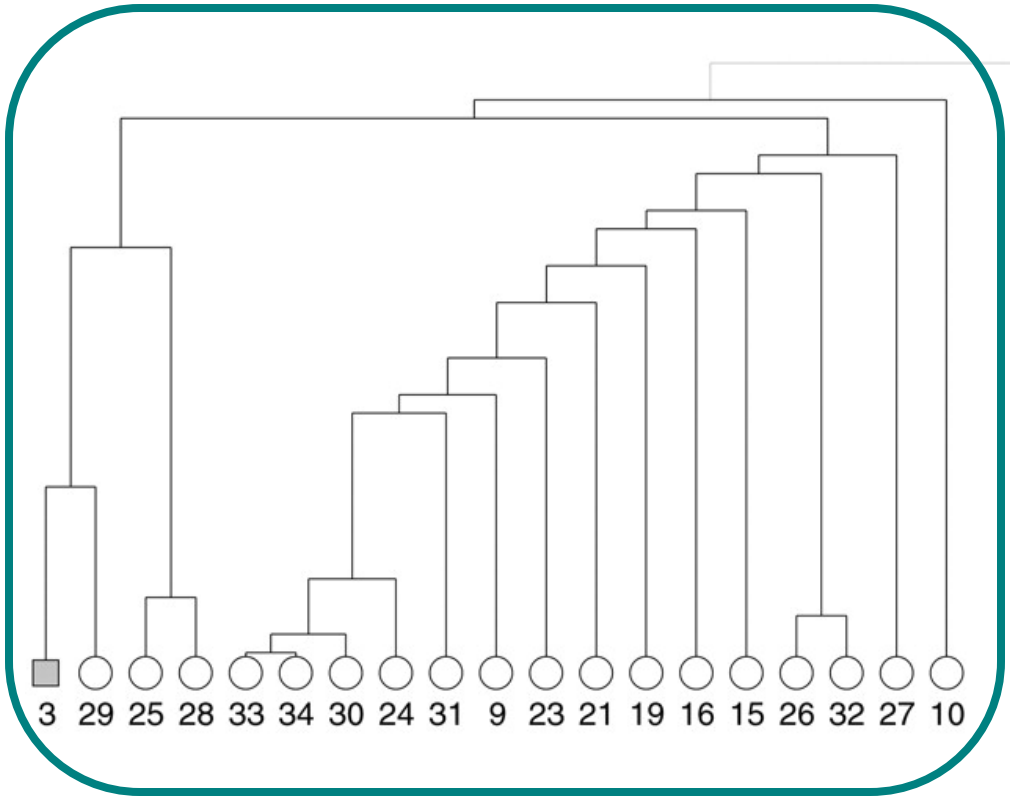


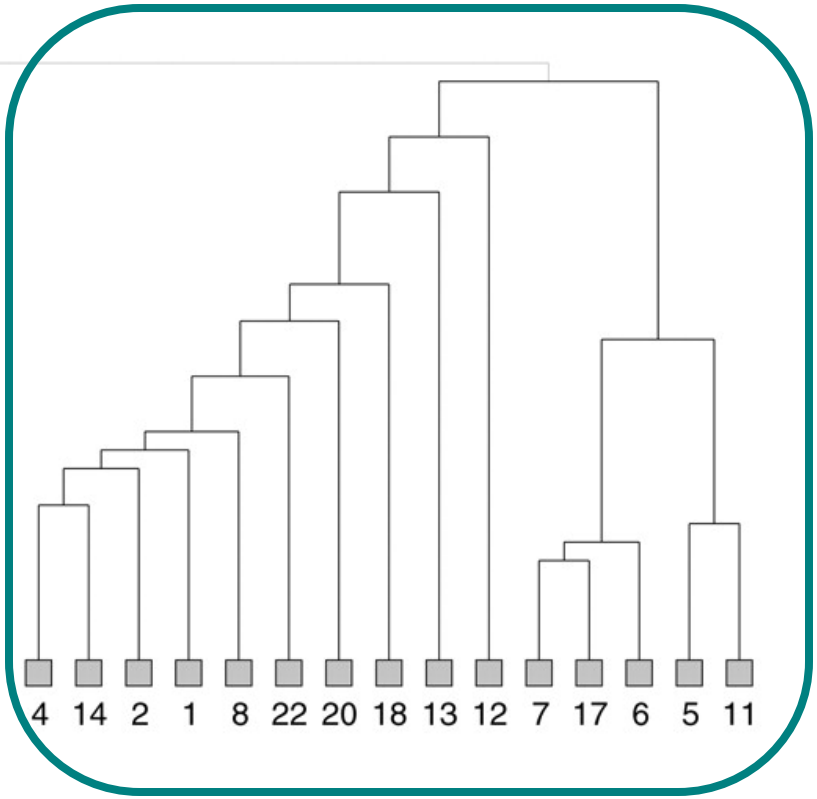


(a) *Step 1*

(b) *Step 2*

betweenness clustering algorithm & the karate club data set

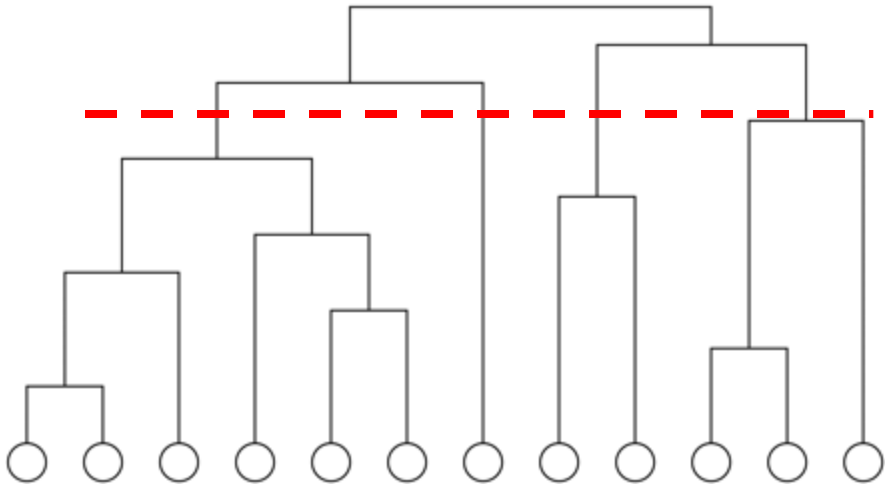




source: Girvan and Newman, PNAS June 11, 2002 99(12):7821-7826

In principle hierarchical clustering does not require preliminary knowledge about the number and the size of communities. In practice it generates a dendrogram that offers a family of community partitions characterizing the studied network. This dendrogram does not tell us which partition captures best the underlying community structure.

Indeed, any cut of the hierarchical tree offers a potentially valid partition. This is at odds with our expectation that in each network there is a ground truth, corresponding to a unique community structure



Modularity

Fundamental Hypothesis 3: Randomly wired networks lack an inherent community structure.

Systematic deviations from a random configuration allow us to define a quantity called *modularity* wich we can use to decide if a particular community partition is better than some other one

Novel approach to community detection

Modularity

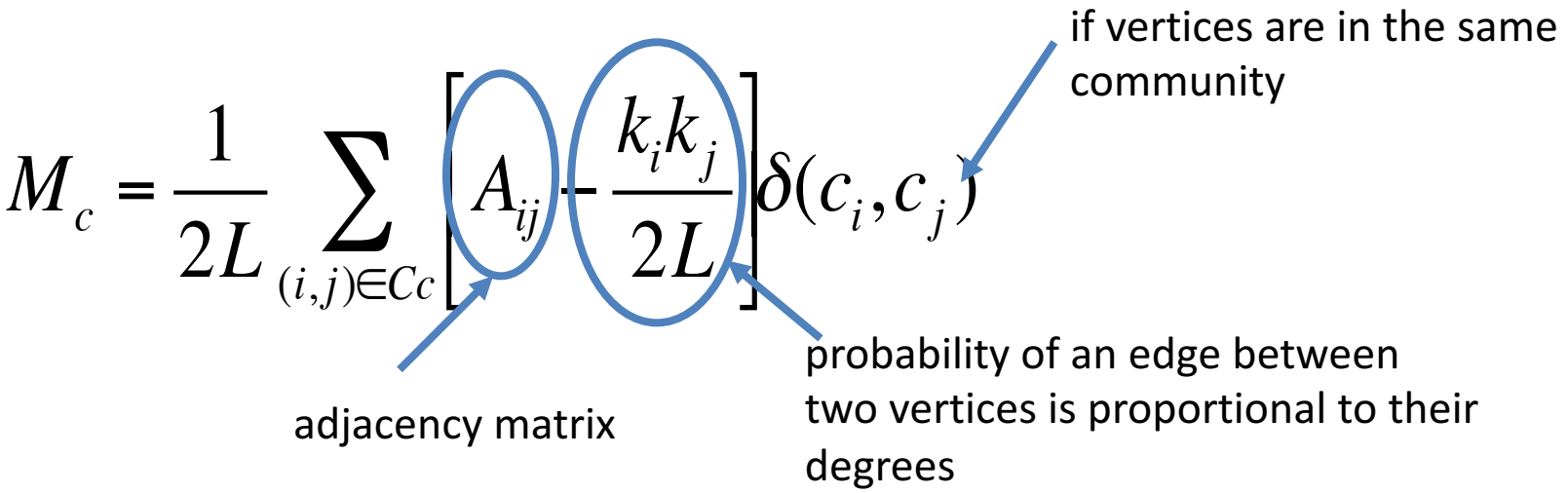
Consider a network with N nodes and L links and a partition into n_c communities, each community having N_c nodes connected to each other by L_c links, where c=1,...,n_c.

If L_c is larger than the expected number of links between the N_c nodes given the network's degree sequence, then the nodes of the subgraph C_c could indeed be part of a true community, as expected based on the Density Hypothesis.

We therefore measure the difference between the network's real wiring diagram (Aij) and the expected number of links between i and j if the network is randomly wired (pij)

Modularity

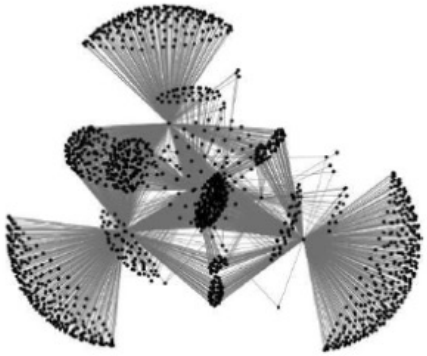
Define modularity:



- For a random network, M = 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

1. start with all vertices as isolates
2. follow a greedy strategy:
   - successively join clusters with the greatest increase ΔM in modularity
   - stop when the maximum possible $\Delta M <= 0$ from joining any two
3. Repeat step 2 until all nodes merge into a single community, recording M for each step
4. Select the partition for which M is maximal

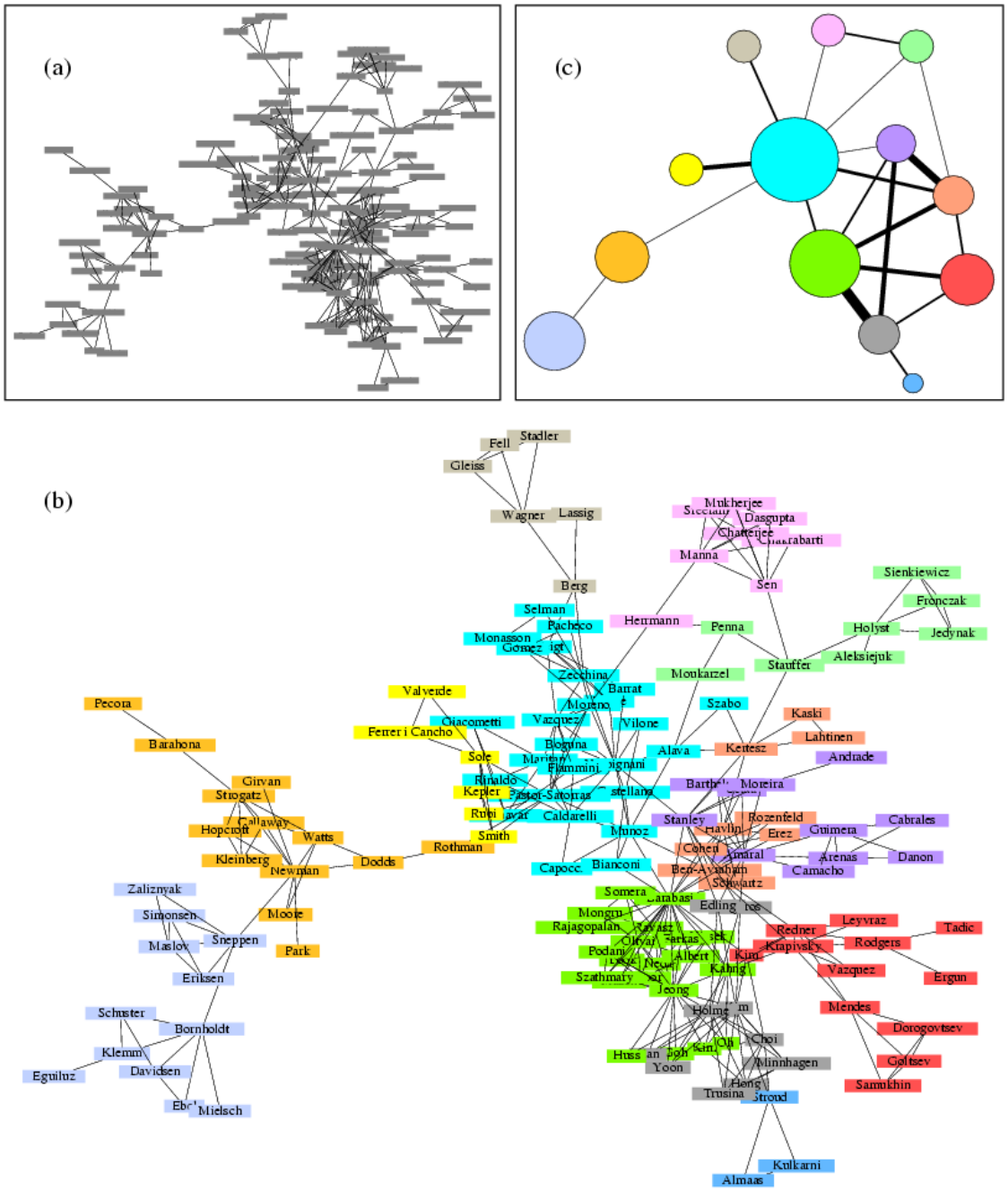
Successfully used to find community structure in a graph with > 400,000 nodes with > 2 million edges

- Amazon's people who bought this also bought that…

Modularity

Computational complexity:

Since the calculation of each ΔM can be done in constant time, Step 2 of the greedy algorithm requires O(L) computations. After deciding which communities to merge, the update of the matrix can be done in a worstcase time O(N). Since the algorithm requires N-1 community mergers, its complexity is $O[(L + N)N]$, or $O(N^2)$ on a sparse graph. Optimized implementations reduce the algorithm's complexity to $O(N\log^2 N)$



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]

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