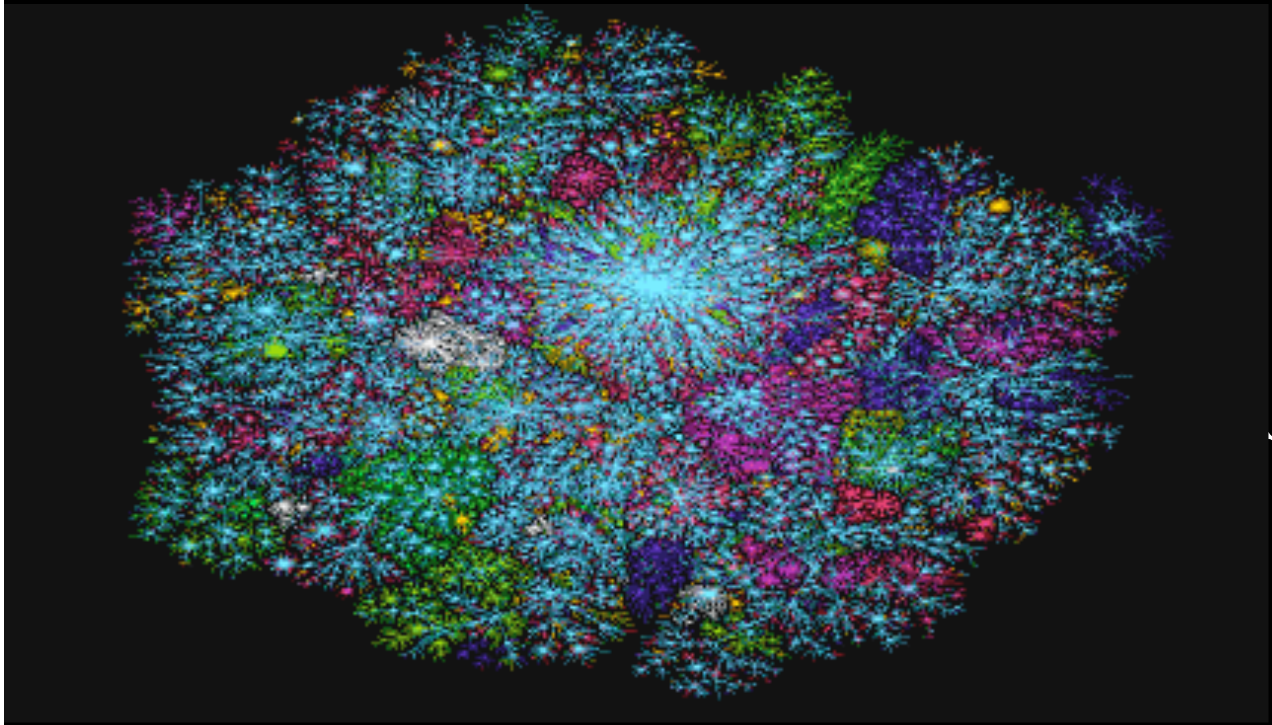
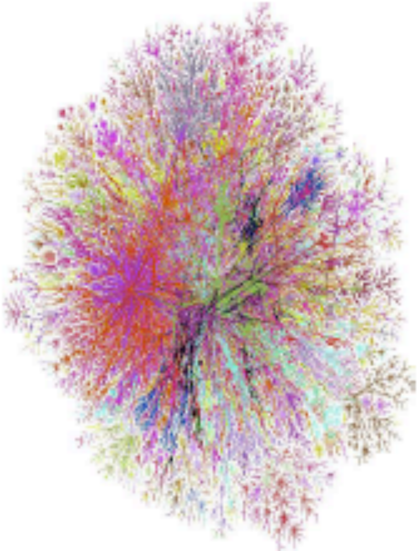
Mixing Patterns and Community Structure in Networks

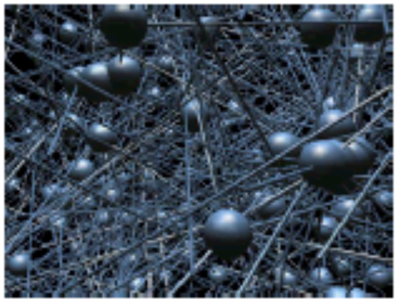


Michelle Girvan

Santa Fe Institute

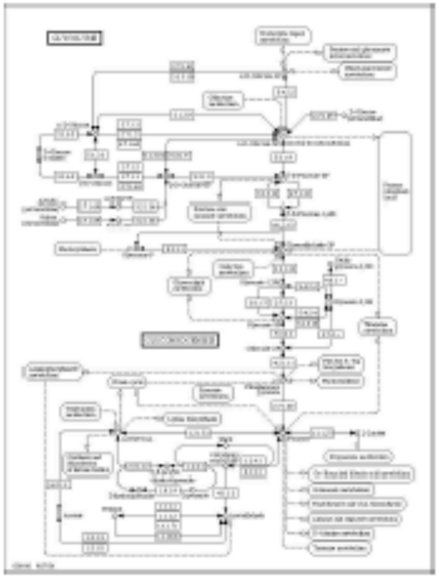


The Internet



Visualization of a
social network

A neural network



A metabolic network

The Physics of Society

“Whatever concerns the human species, considered en masse, belongs to the domain of physical facts; the greater the number of individuals, the more the individual will is submerged beneath the series of general facts which depend on the general causes according to which the society exists and is conserved.” --Quetelet

Laplace

Quetelet

Buckle

Maxwell

Boltzmann

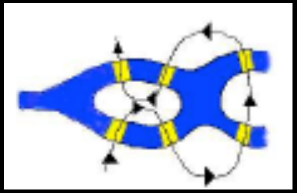
“The molecules are like so many individuals, having the most various states of motion, and the properties of gases only remain unaltered because the number of these molecules which on the average have a given state of motion is constant.” --Boltzmann

Traditional vs. Complex Systems Approaches to Networks

Traditional Questions:

Social Networks:
Who is the most important person in the network?

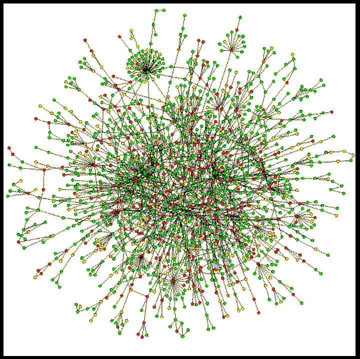
Graph Theory:
Does there exist a cycle through the network that uses each edge exactly once?



Complex Systems Questions:

What fraction of edges have to be removed to disconnect the graph?

What kinds of structures emerge from simple growth rules?



Areas of Network Research

Structural Complexity

- The wiring diagram could be an intricate tangle, far from perfectly regular or perfectly random. *Example: community structure*
- The network could include different classes of nodes
- The edges could be heterogeneous with different weights, directions and signs.

Dynamical Complexity

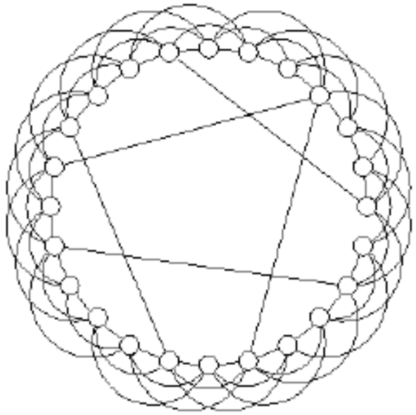
- Dynamics on the network: processes could be taking place on the fixed network.
- Dynamics of the network: the network itself could be evolving in time.

Structural properties of real-world networks

The small world effect

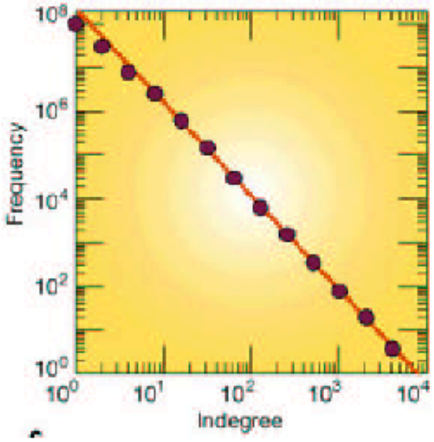
Diameter small compared to system size.

Clustering



Two of your friends are likely to be friends with one another.

Degree distributions

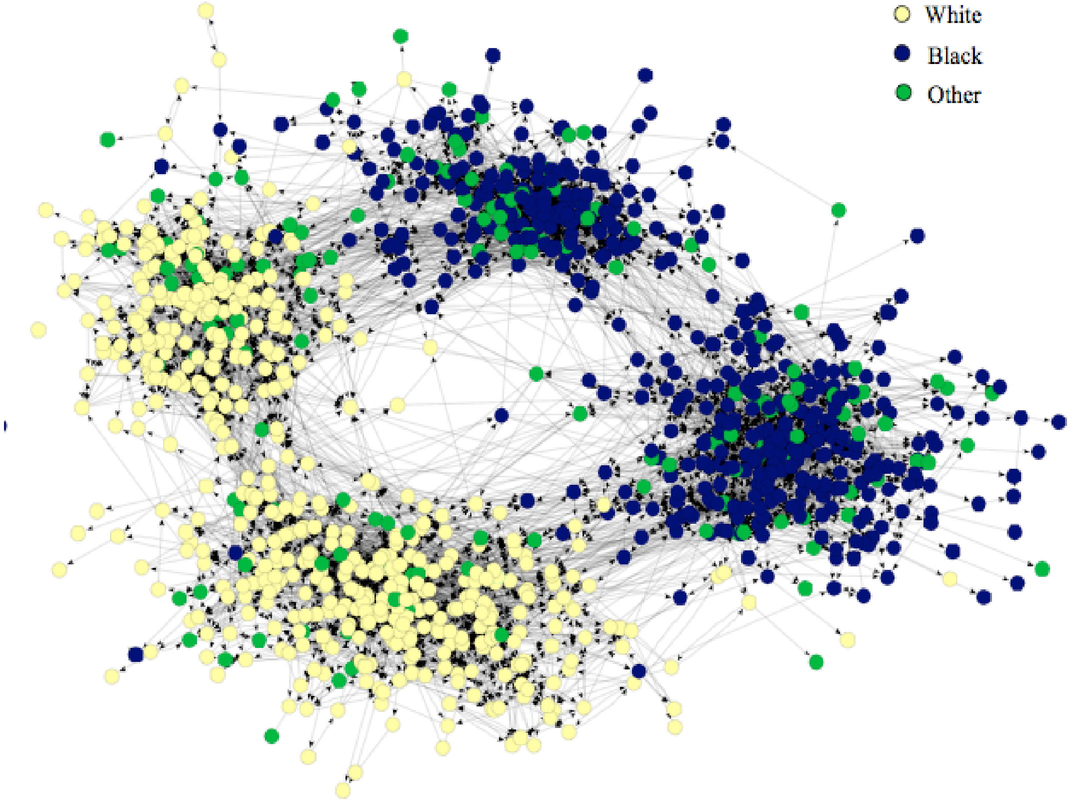


World-Wide Web

Degree distributions are often heavy tailed.

Assortative Mixing

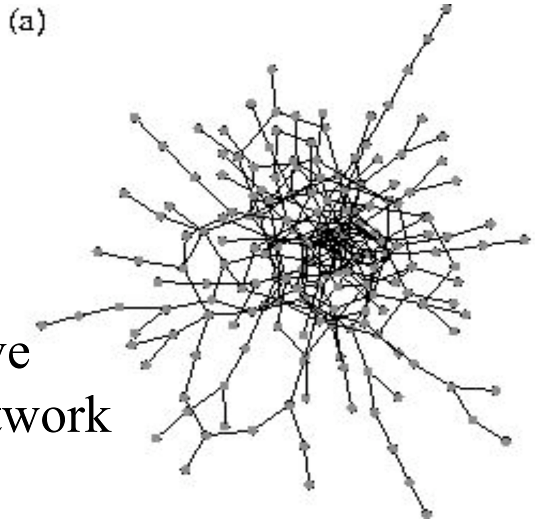
In assortatively mixed networks, like vertices tend to connect preferentially to one another.



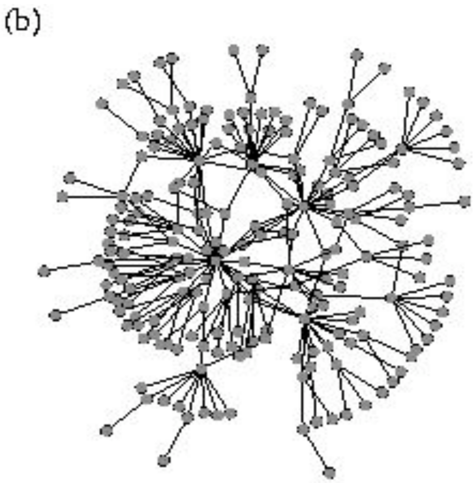
Friendship network of students in a U.S. school. Friendships are determined by asking the participants, and hence are directed, since A may say that B is their friend but not vice versa. Vertices are color coded according to race, as marked, and the split from left to right in the figure is clearly primarily along lines of race. The split from top to bottom reflects a division between middle school and high school students.

Assortative Mixing by Degree

- A network is said to be assortatively mixed by degree if high degree vertices tend to connect to other high degree vertices
- A network is disassortatively mixed by degree if high degree vertices tend to connect to low degree vertices.

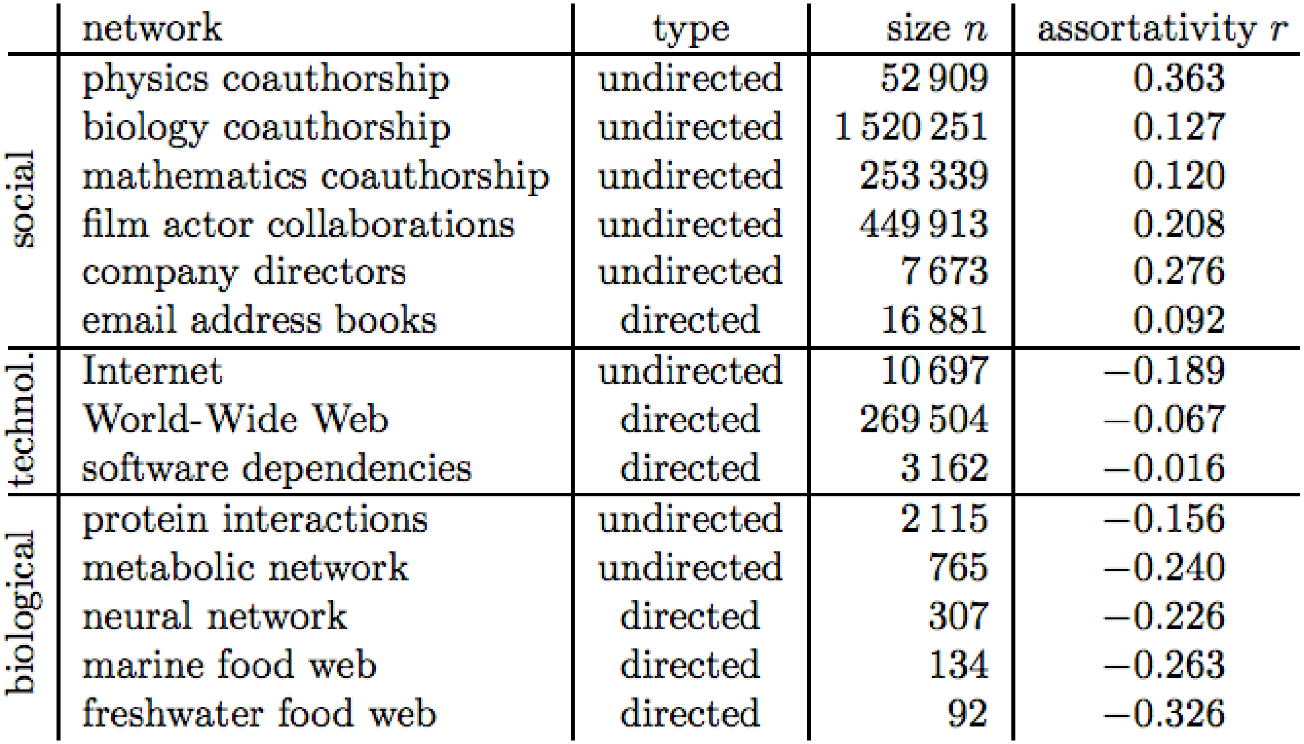


Assortative
Scale-free network



Disassortative
Scale-free
network

Measured assortativity for various networks

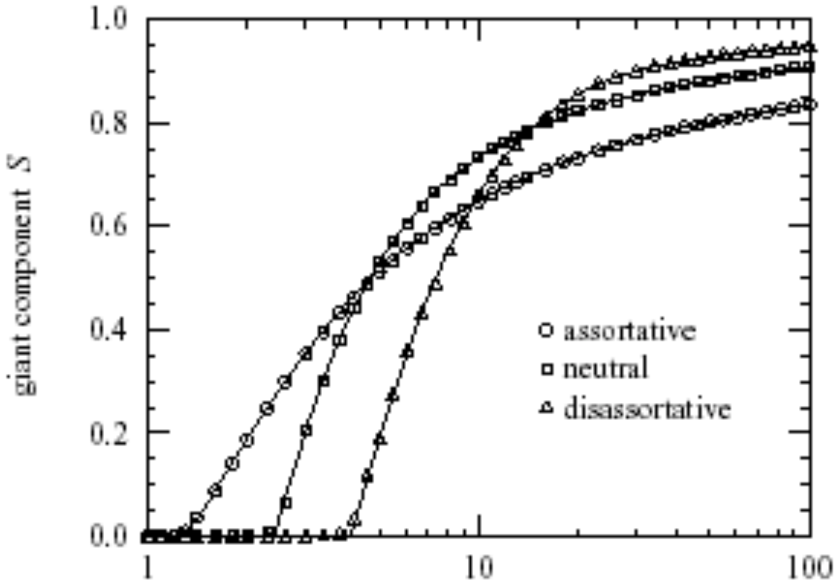


	network	type	size n	assortativity r
social	physics coauthorship	undirected	52 909	0.363
	biology coauthorship	undirected	1 520 251	0.127
	mathematics coauthorship	undirected	253 339	0.120
	film actor collaborations	undirected	449 913	0.208
	company directors	undirected	7 673	0.276
	email address books	directed	16 881	0.092
technol.	Internet	undirected	10 697	−0.189
	World-Wide Web	directed	269 504	−0.067
	software dependencies	directed	3 162	−0.016
biological	protein interactions	undirected	2 115	−0.156
	metabolic network	undirected	765	−0.240
	neural network	directed	307	−0.226
	marine food web	directed	134	−0.263
	freshwater food web	directed	92	−0.326

M.E.J Newman and M. Girvan, *Mixing Patterns and Community Structure in Networks* (2002).

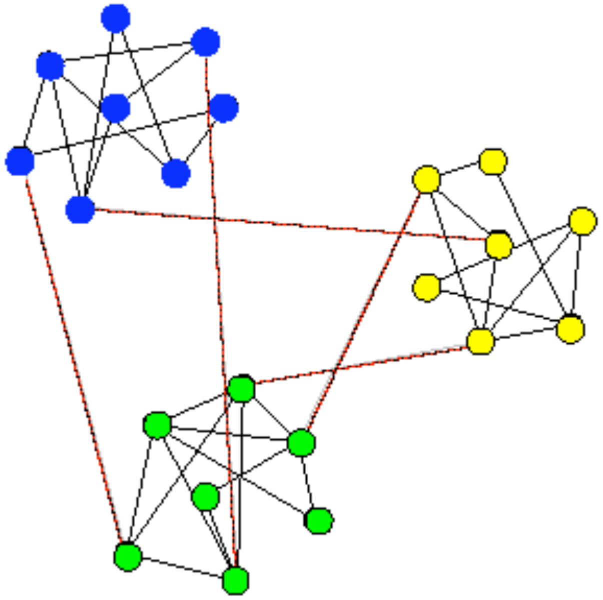
How does assortative impact information flow?

- In assortatively mixed networks, the epidemic transition occurs sooner as the transmissability is increased. In other words, epidemics occur at lowe transmissability than in neutral or disassortative networks.
- In parameter regimes where epidemics readily occur, the number of infected individuals is generally lower for assortatively mixed networks.

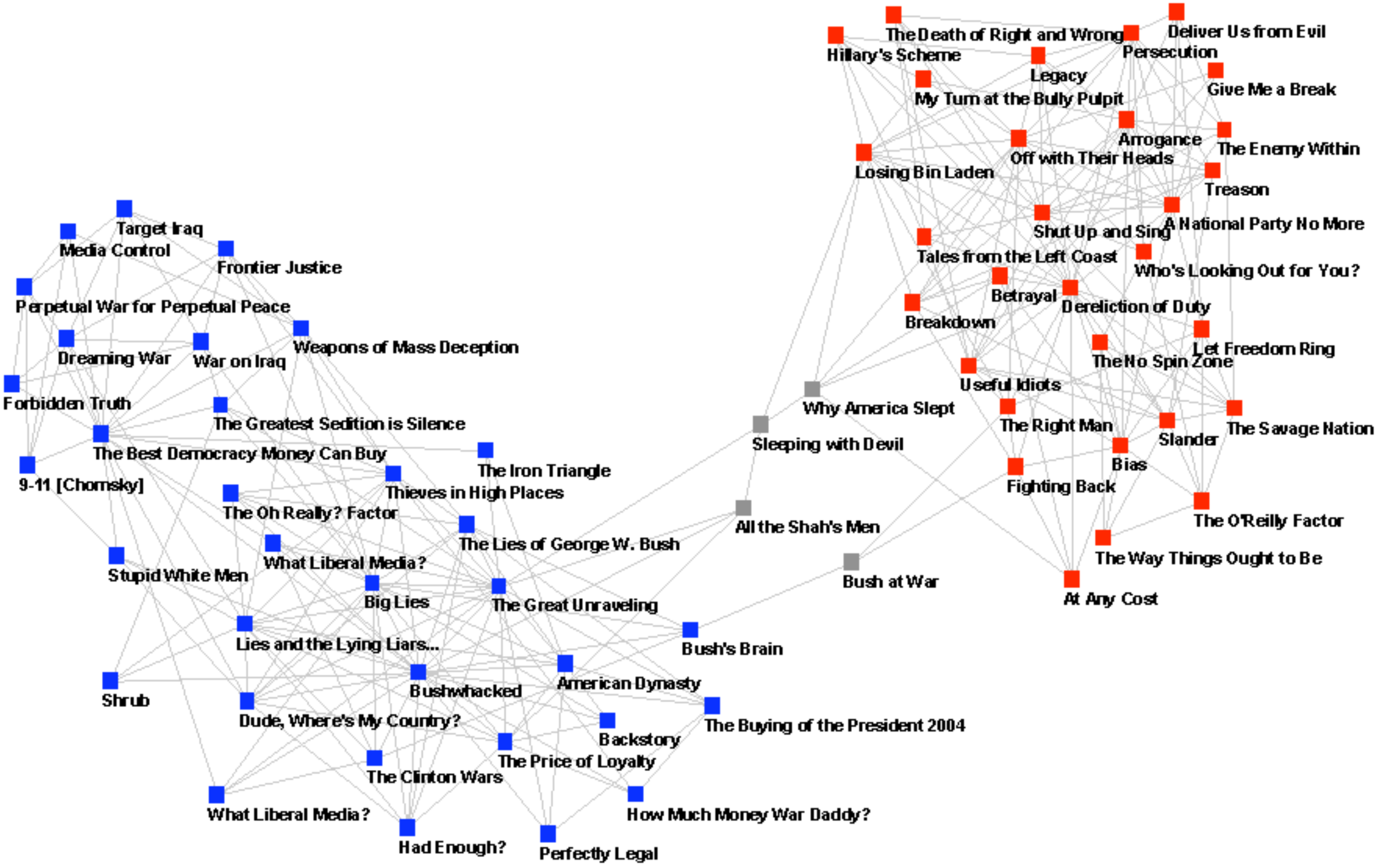


Community Structure in Complex Networks

We define communities in a network as groups of nodes between which connections are sparse but within which connections are relatively dense. More on community structure to come in lecture 3.

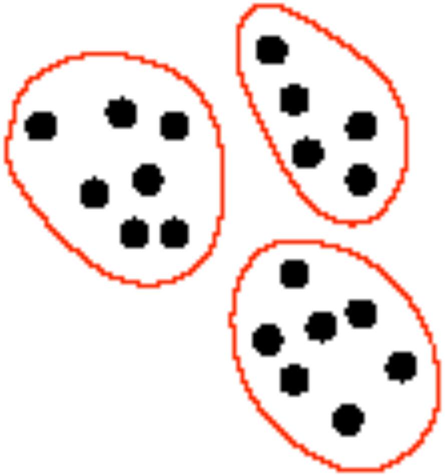


Community Structure in Political Books

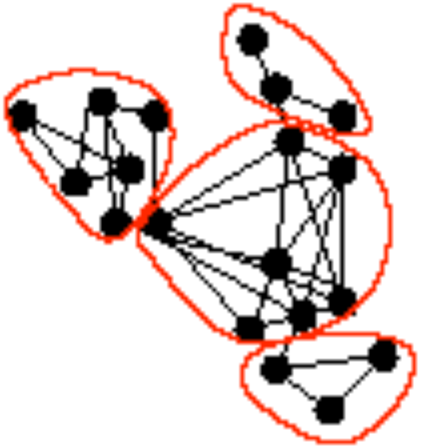


Detecting Modularity

- We are interested in network clustering, which differs from ordinary data clustering.
- In network clustering, relationships between vertices are determined by flows through other vertices.
- In data clustering, relationships between vertices can be determined independently of other vertices
- Traditional methods for network clustering have involved transformation of the network into a data clustering problem.



Data Clustering

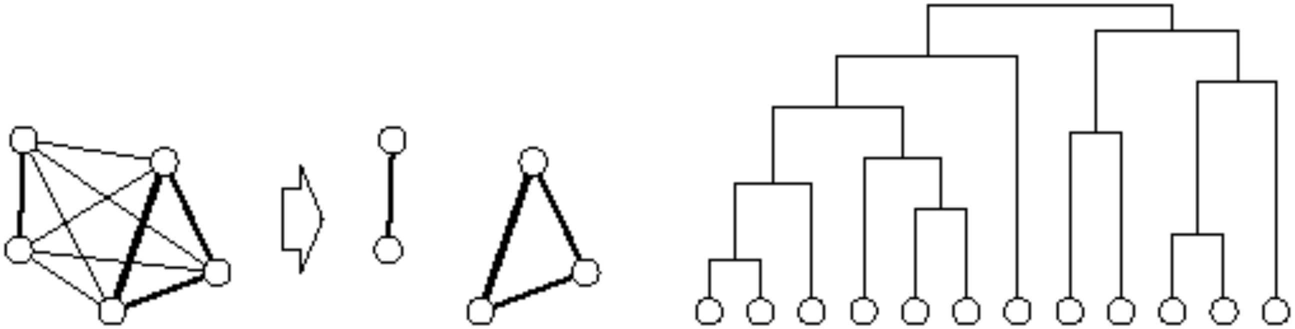


Network Clustering

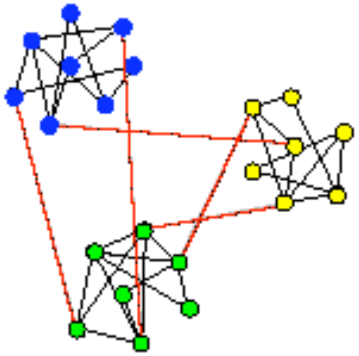
Traditional Method for Detecting Community Structure

- Calculate a weight W_{ij} for every pair *i,j* of vertices in the network.
- Start with the set of vertices with no edges between them and add edges one by one in order of decreasing weight.
- As edges are added, the resulting graph shows a set of increasingly large components, which are taken to be the communities.
- Example weight matrix W for adjacency matrix A:

$$\mathbf{W} = \sum_{\ell=0}^{\infty} (\alpha \mathbf{A})^{\ell} = [\mathbf{I} - \alpha \mathbf{A}]^{-1}.$$



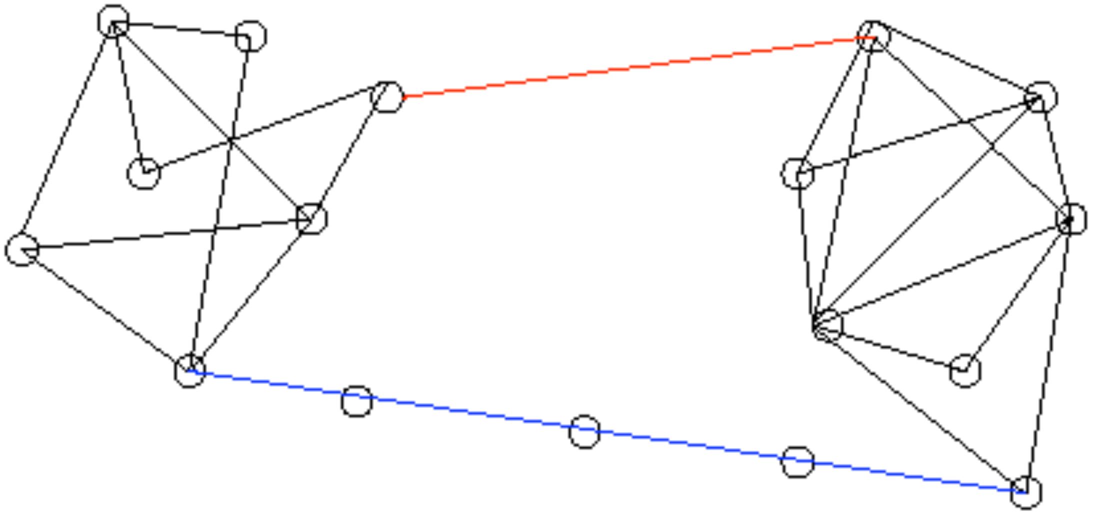
New Method for Detecting Communities



- Instead of looking for the cores of communities, we try to detect community structure by identifying boundaries.
- Boundary detection based on centrality indices:
  - Node betweenness -- The betweenness centrality of a vertex *i* is the number of shortest paths between pairs of other vertices which run through *i*.
  - Edge betweenness -- Similarly, the betweenness of an edge *j* is the number of shortest paths between pairs of nodes which run along *j*.

Algorithm for Detecting Communities

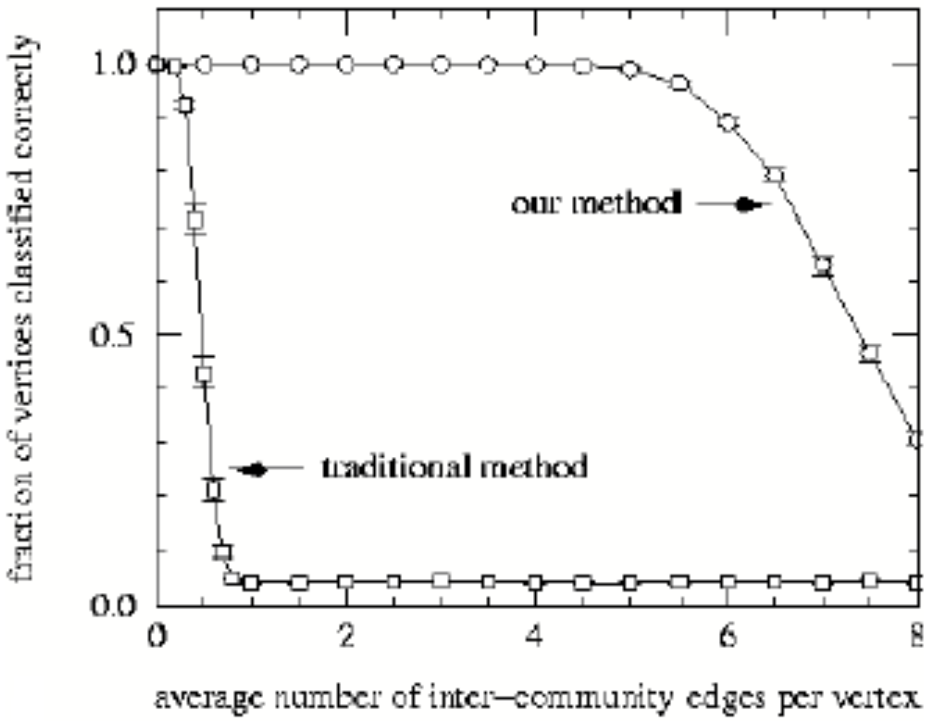
1. Calculate the betweenness for all edges in the network.
2. Remove the edge with the highest betweenness.
3. Recalculate betweennesses for all edges affected by the removal.
4. Repeat from step 2 until no edges remain.



Algorithm Run Time

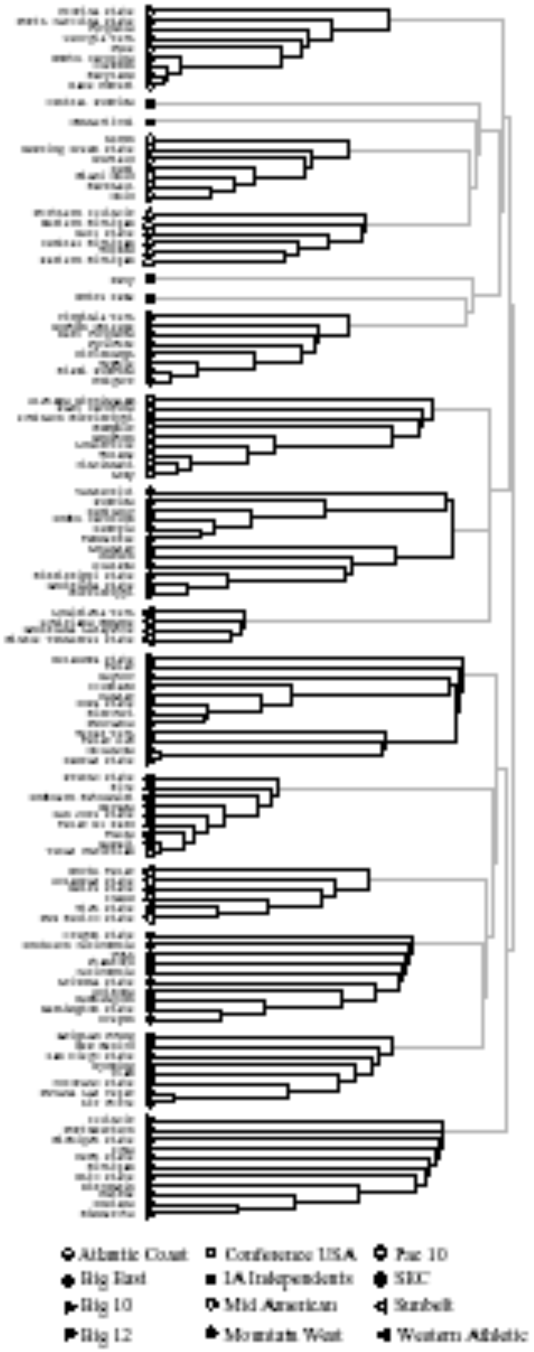
- The betweenness values for all edges can be calculated in $O(mn)$ time, where m is the number of edges and n is the number of nodes.
- The entire algorithm runs in worst case $O(m^2n)$ time because the betweenness values must be recalculated after each edge removal.
- In graphs with strong community structure, the run time may be significantly reduced.

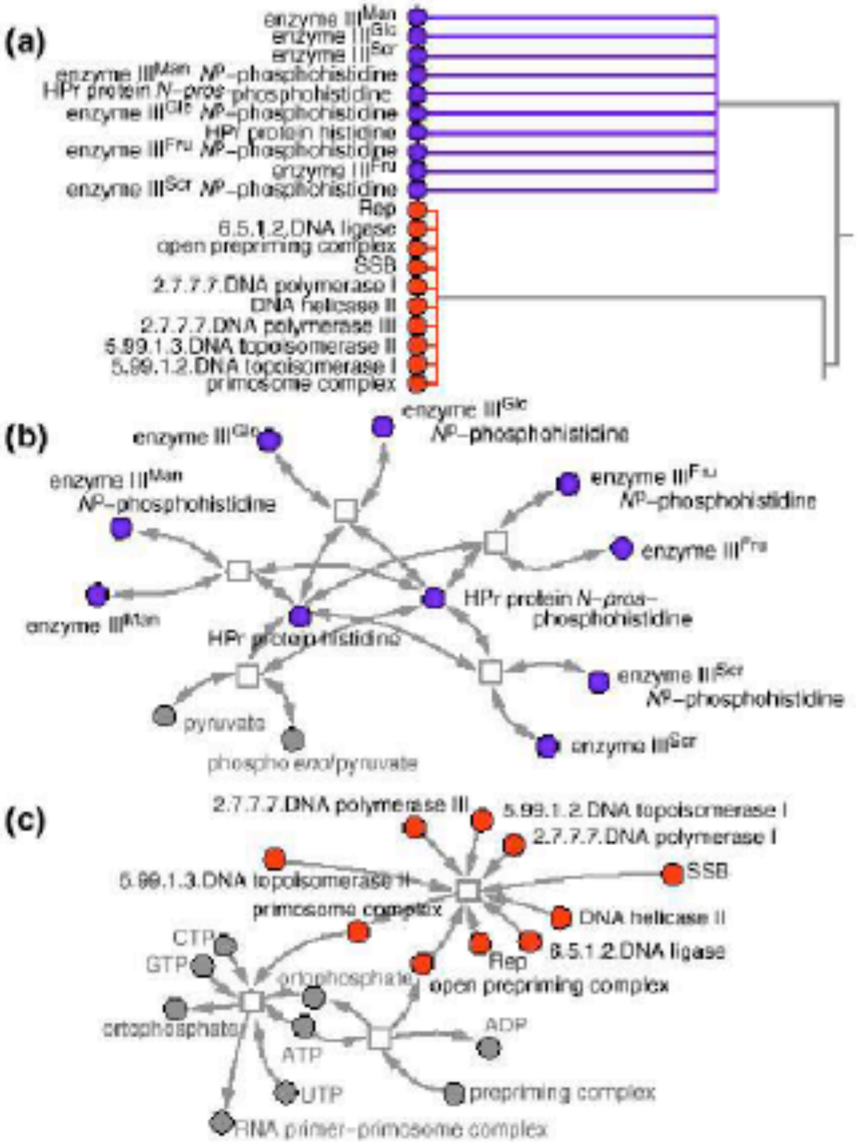
Testing the Algorithm



The fraction of vertices correctly classified as the number z_{out} of inter-community edges is varied. The circles are results for our method; the squares for a standard hierarchical clustering method based on max-flow. The graphs were generated with the following parameters: N=128, z=16. Each point is the average over 100 realizations of the graphs.

College Football




(a)
enzyme III^{Man}
enzyme III^{Glc}
enzyme III^{Scr}
enzyme III^{Man} N^{P}–phosphohistidine
HPr protein N–pros–phosphohistidine
enzyme III^{Glc} N^{P}–phosphohistidine
HPr protein histidine
enzyme III^{Fru} N^{P}–phosphohistidine
enzyme III^{Fru}
enzyme III^{Scr} N^{P}–phosphohistidine
Rep
6.5.1.2.DNA ligase
open prepriming complex
SSB
2.7.7.7.DNA polymerase I
DNA helicase II
2.7.7.7.DNA polymerase III
5.99.1.3.DNA topoisomerase II
5.99.1.2.DNA topoisomerase I
primosome complex
(b)
enzyme III^{Glc}
enzyme III^{Glc} N^{P}–phosphohistidine
enzyme III^{Man} N^{P}–phosphohistidine
enzyme III^{Fru} N^{P}–phosphohistidine
enzyme III^{Fru}
enzyme III^{Man}
HPr protein N–pros–phosphohistidine
HPr protein histidine
enzyme III^{Scr} N^{P}–phosphohistidine
pyruvate
phospho enol pyruvate
enzyme III^{Scr}
(c)
2.7.7.7.DNA polymerase III
5.99.1.2.DNA topoisomerase I
2.7.7.7.DNA polymerase I
5.99.1.3.DNA topoisomerase II
SSB
primosome complex
DNA helicase II
CTP
GTP
6.5.1.2.DNA ligase
Rep
ortophosphate
open prepriming complex
ortophosphate
ADP
ATP
UTP
prepriming complex
RNA primer–primosome complex

Biochemical Network of *M. pneumoniae*

An exerpt of the hierarchical tree for the biochemical pathways of *M. pneumoniae* and the corresponding subnetworks associated with sugar import (b) and DNA replication (c). Taken from P. Holme, M. Huss, and H. Jeong, Subnetwork hierarchies of biochemical pathways. *Bioinformatics* **19**, 532-538 (2003).

Quantifying the Community Structure

- Each level of the hierarchical tree represents a division of the network into a certain number of communities.
- Let e_{ij} be the fraction of edges that connect a node in community i to a node in community j. For a system of S communities, e is an S x S matrix.
- Define $a_i=\Sigma_j e_{ij}$, *i.e.* a_i is the fraction of edges attached to vertices in community i.
- The level of community structure can then be quantified with a modularity coefficient Q:

$$Q = \Sigma_i e_{ii} - \Sigma_i a_i^2$$

which has the following properties:

- $Q = 0$ when there is no community structure.

- Q approaches 1 as S gets large and the network displays perfect modularity. $Q = 1 - 1/S$ for a perfectly modular network of S equally sized communities.

- When the network is divided into a certain number of communities, Q is largest when the communities are equal in size.

- The minimum value of Q likes in the range $-1 < Q < 0$

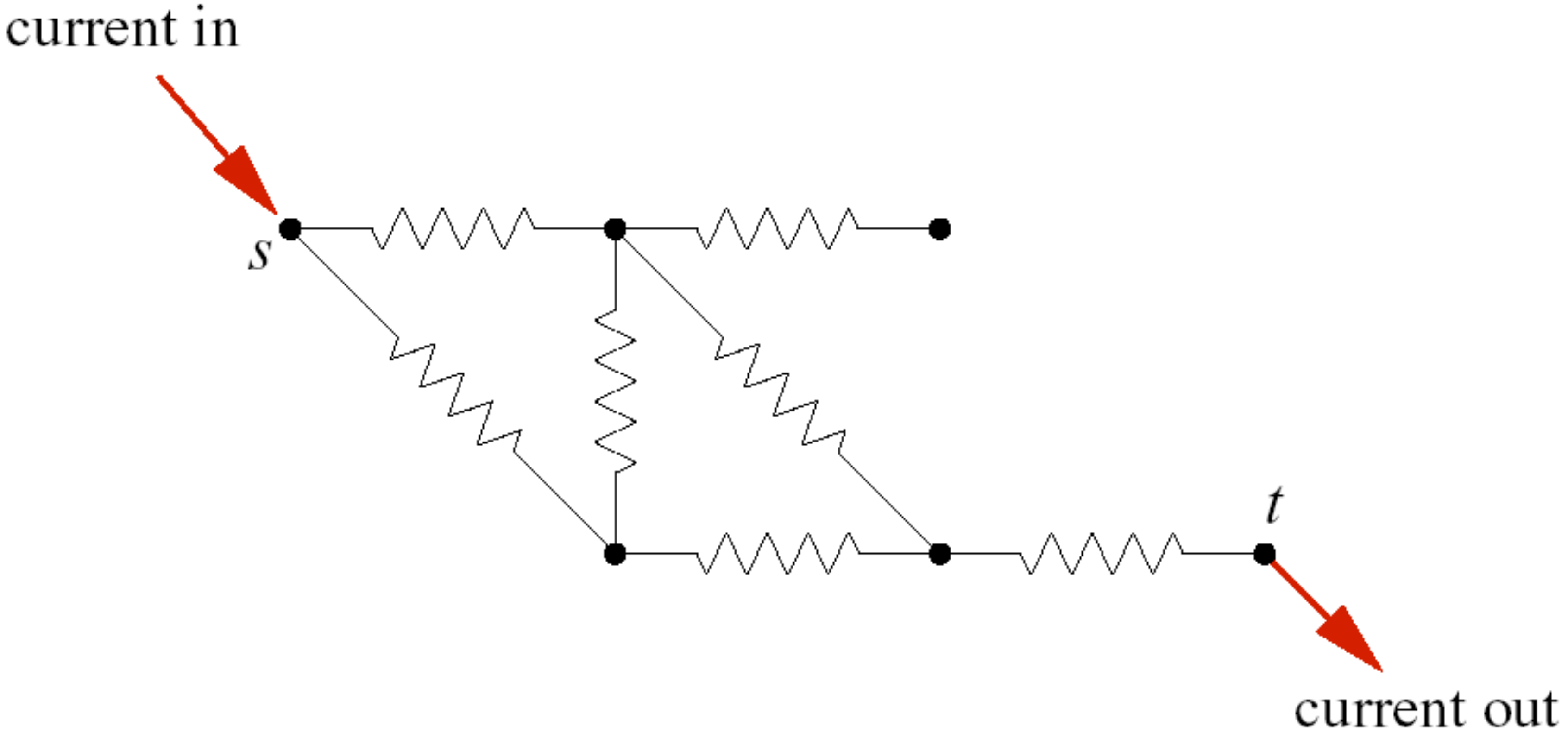
Another approach: random walks

Instead of counting the number of shortest paths between pairs of nodes which run through an edge, we can count the number of random walks which go through the edge as a measure of betweenness.

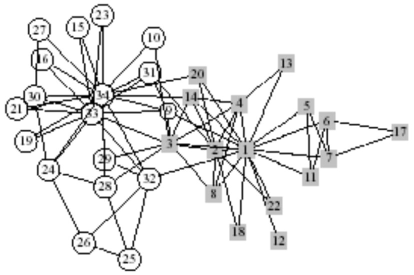
- Let N be the number of nodes.
- Let A be the adjacency matrix.
- Let D be a diagonal matrix where d_{ii} is degree of node i, $d_{ii} = k_i$.
- Let R be the random walk matrix. r_{ij} represents the probability of moving to i given that we start at j. $R = AD^{-1}$. $[R^n]_{ij}$ gives the probability of being at i after exactly n steps given that we started at j.
- Let M_t be a random walk matrix with absorbing site t (the target vertex). $M_t = A_t D_t^{-1}$
- Let $\mathbf{s}$ be the source vector whose components are all zero except for a single 1 in the position corresponding the the source vertex s.
- Accounting for paths of all lengths, the mean number of times that a walk traverses the edge from v to w is $k_v^{-1}[(I - M_t)^{-1}]_{vs}$.
- The contribution from node pair (s, t) to the random walk betweenness of edge (v, w) is defined as the net number of times the walk passes along the edge.

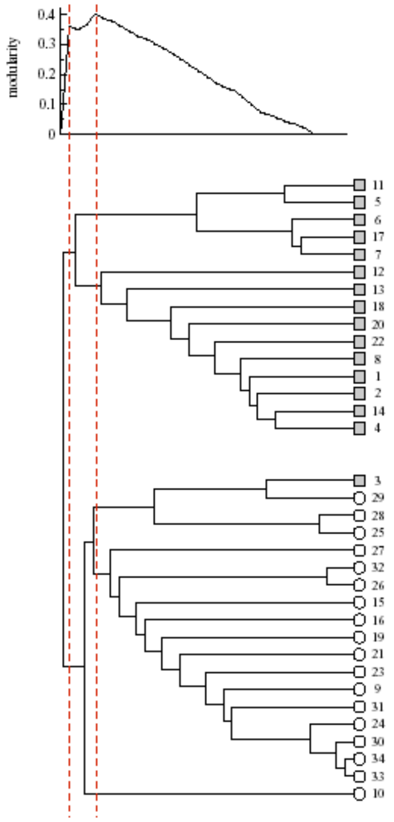
We find that the random walk method gives almost identical results to the original betweenness method.

Current-flow betweenness

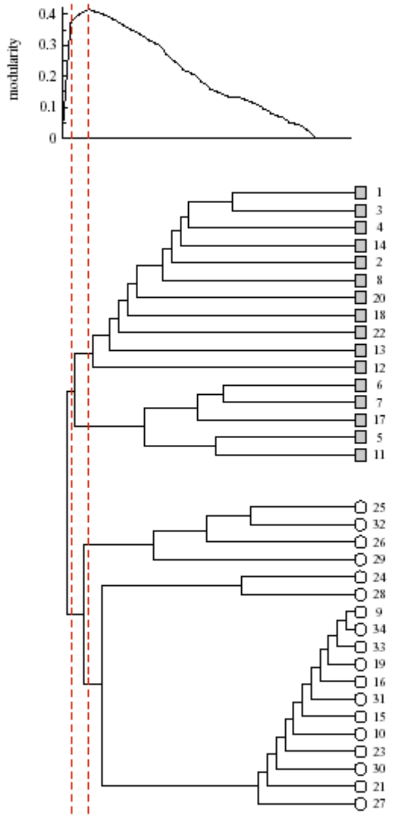


Zachary's Karate Club

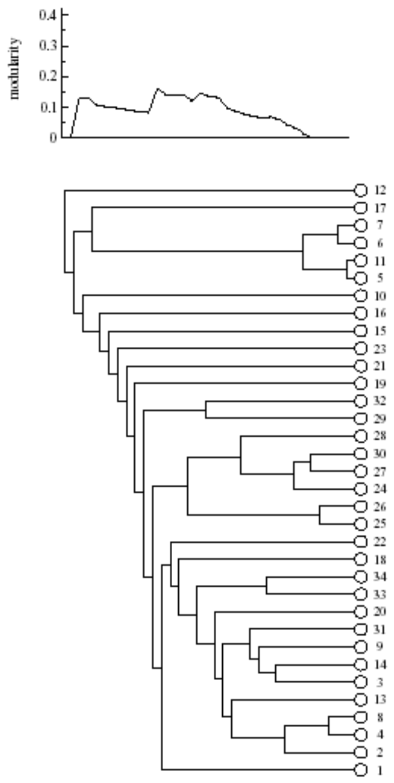




shortest path



random walk



shortest path
without recalculation