

Networks

Carlos Carvalho, Mladen Kolar and Robert McCulloch

11/12/2015

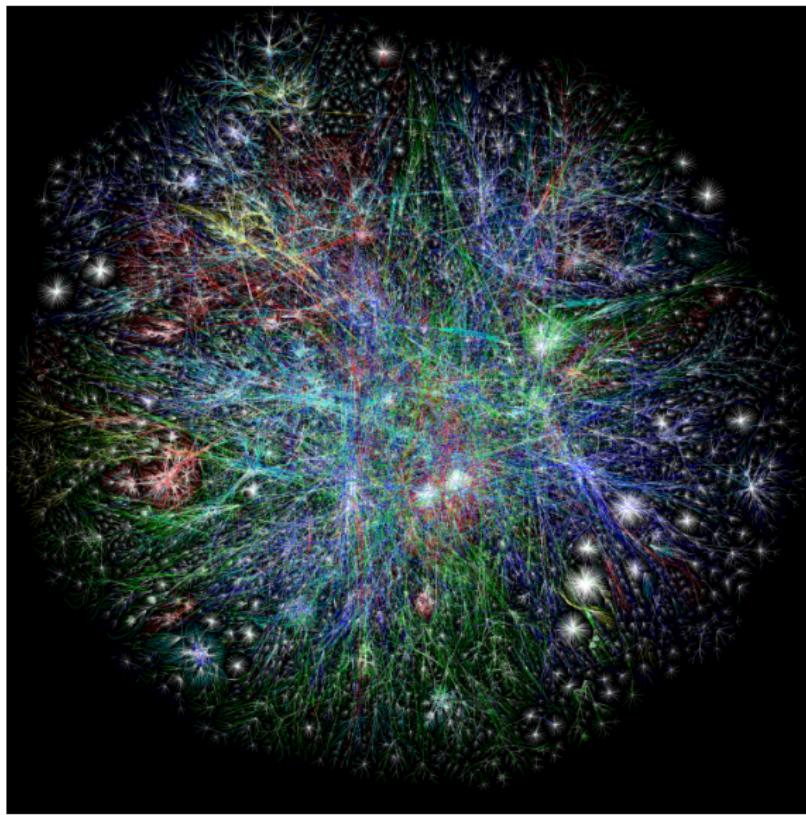
Networks are everywhere

Our world is complex

- ▶ Societies are collections of individuals who interact
- ▶ Communication systems link electronic devices
- ▶ Information and knowledge is organized and linked
- ▶ Our genes interact to regulate processes in our body
- ▶ Our brain processes thoughts using billions of interconnected neurons

How to make sense of these complex systems?

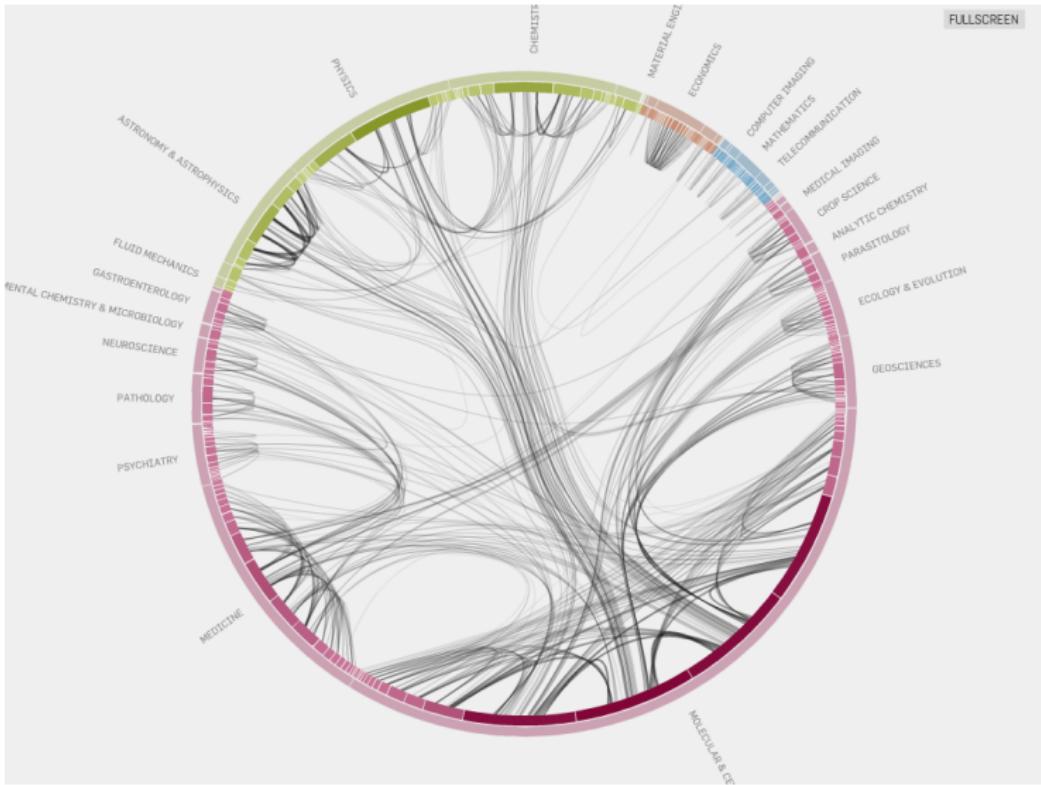
Internet — 50 billion Webpages



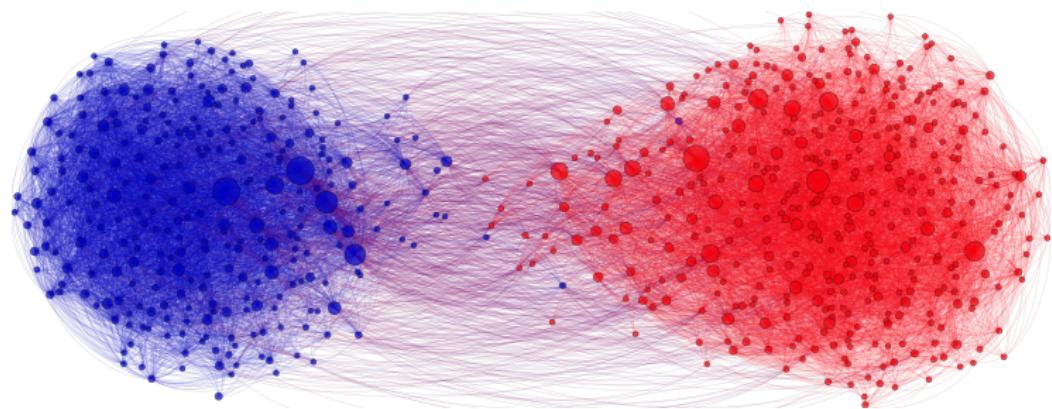
Facebook — 1.2 Billion Users



Citation Network — 250 Million Articles

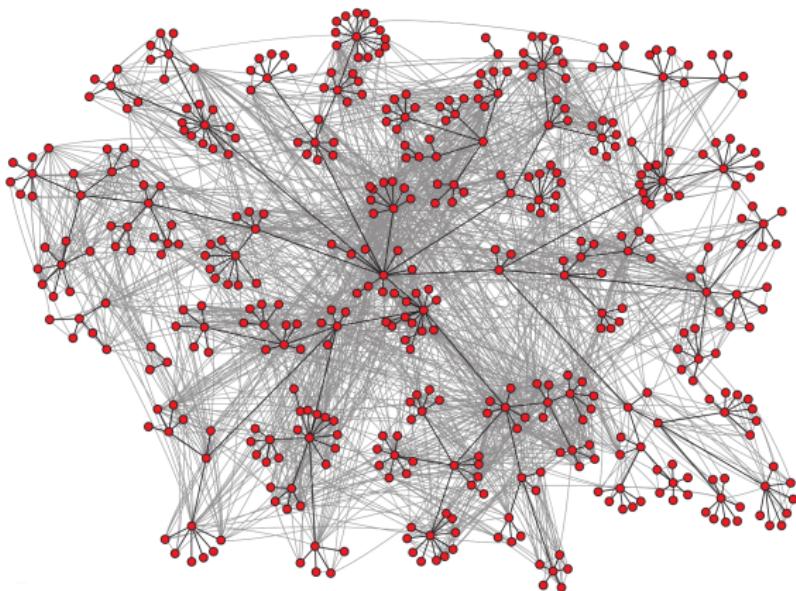


Media networks



Connections between political blogs (Adamic, Glance, 2005)

Organizational networks



Email exchange network

Organizational networks

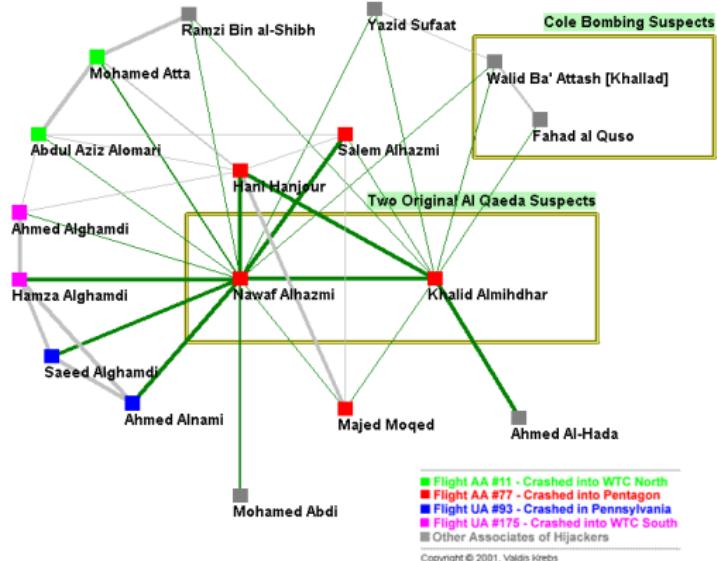
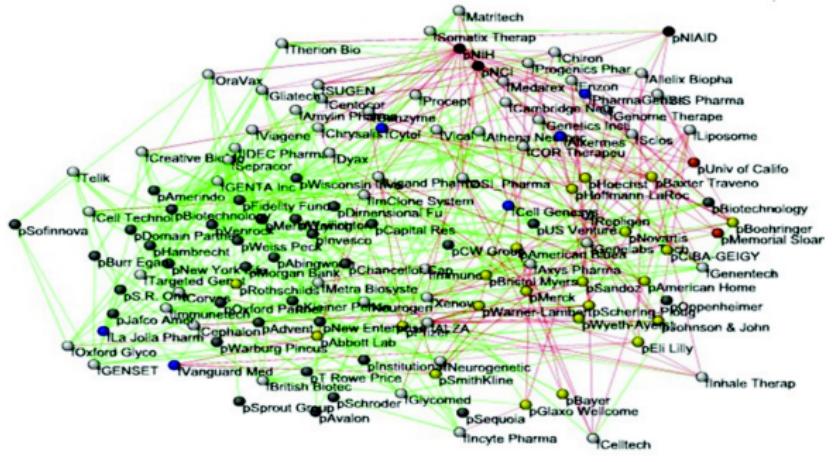


Figure 2 - All nodes within 1 step [direct link] of original suspects

9/11 terrorist network (Krebs, 2002)

Economic networks



Nodes:

Companies

Investment

Pharma

Research Labs

Public

Biotechnology

Links:

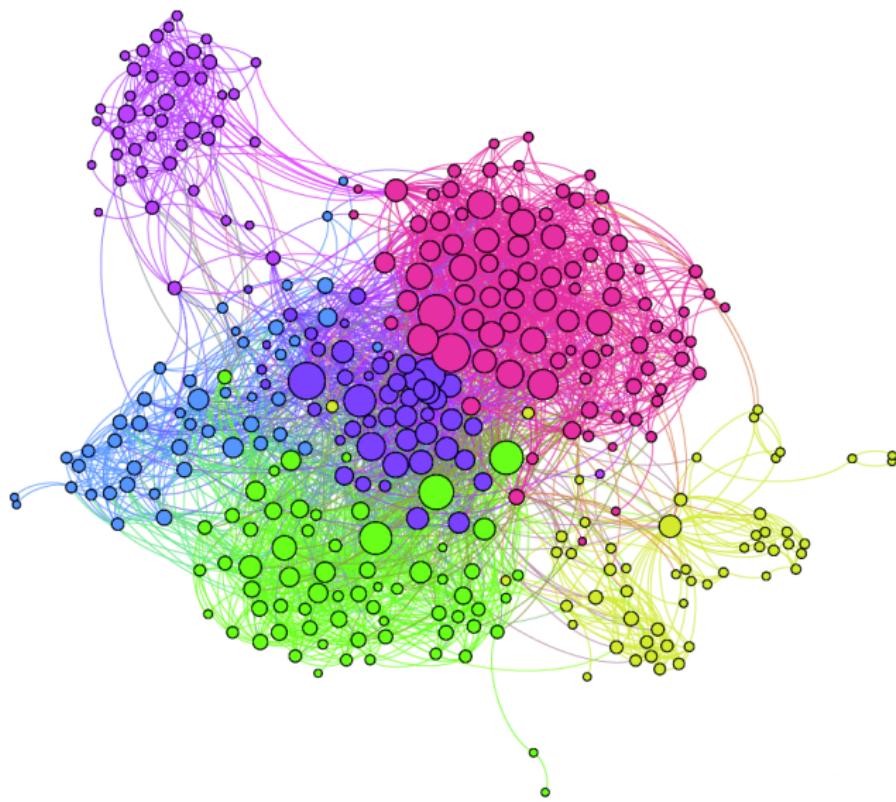
Collaborations

Financial

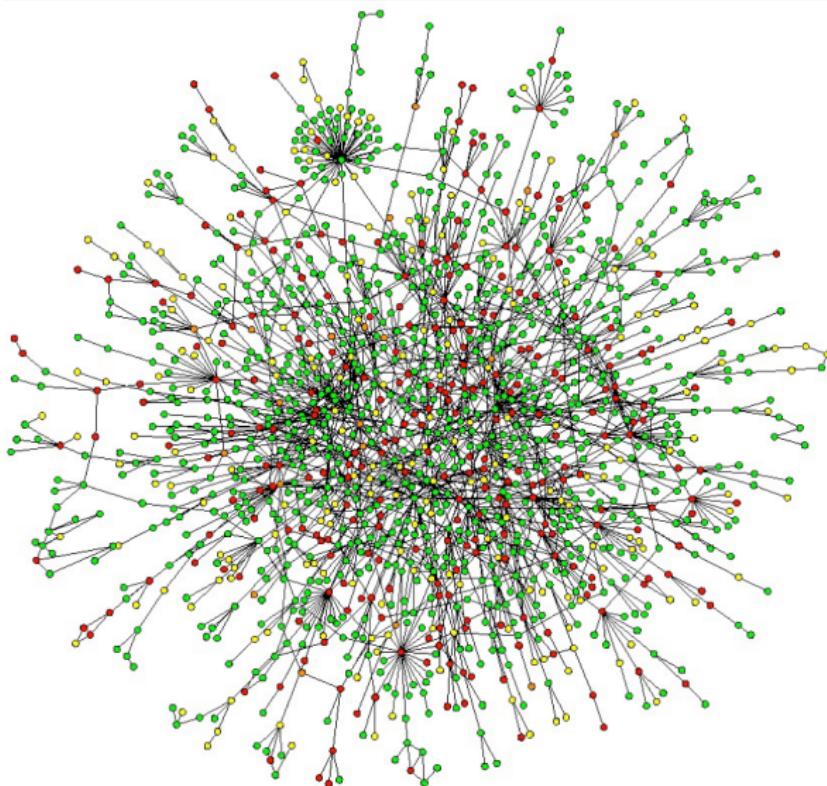
R&D

Biotech companies (Powell-White-Koput, 2002)

Ego networks



Biological networks



Protein-protein interactions

Many more examples



who follows whom?



who calls whom?



who buys what?

Networks as a unifying tool

One set of tools to help us understand problems arising in diverse fields.

As with all data analysis, we start by exploratory analysis. Often a lot of insights can be gained through visualization.

However, how to effectively visualize large networks is an open problem.

Why should you care?

Rich data easily accessible on millions of users producing content, exchanging ideas

- ▶ dataset, developers APIs, crawl the web

Learn about behaviors, preferences, trends

Applications: Reputation management

- ▶ consumer brand analytics
- ▶ marketing communication
- ▶ product reviews

Why should you care?

Applications: Data driven policy making

- ▶ who supports which political candidate
- ▶ law enforcement: gang members boast on social media about their activities
- ▶ citizen unrest: protests being organized through twitter

Application: Social media marketing

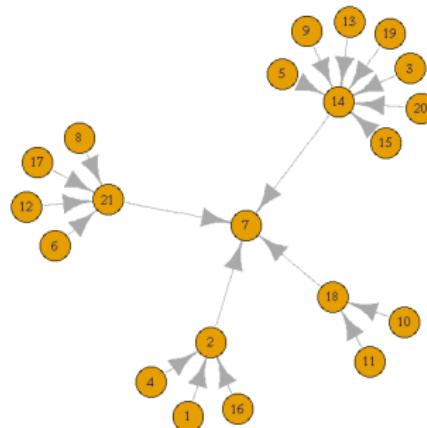
- ▶ viral marketing and personalized recommendation
- ▶ online users are brand advocates

www.socialmediaexaminer.com/new-studies-show-value-of-social-media

Why should you care?

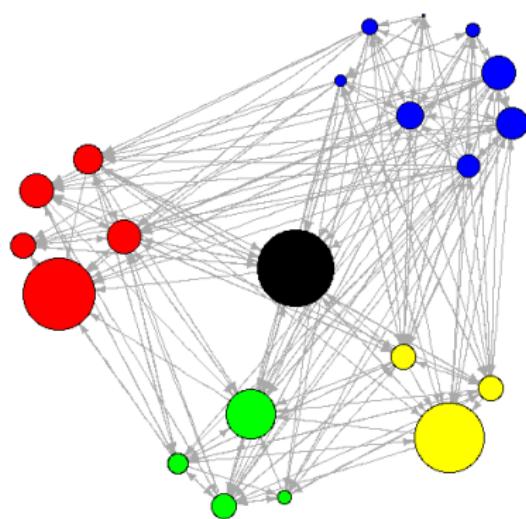
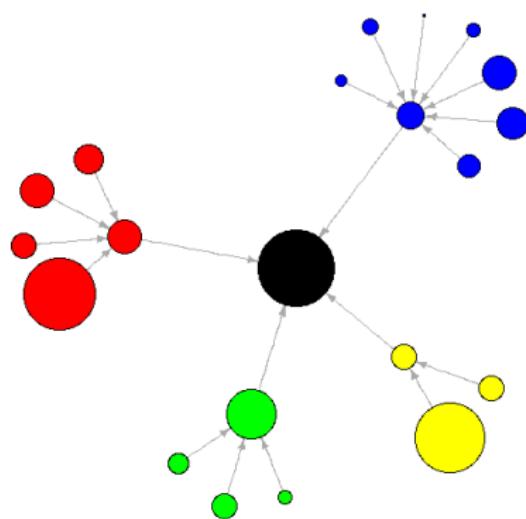
Application: Human behavior analysis

- ▶ identify members of different social groups
- ▶ identify topics of group conversations



Reports to

Advice from



Today's topics

Representation of networks

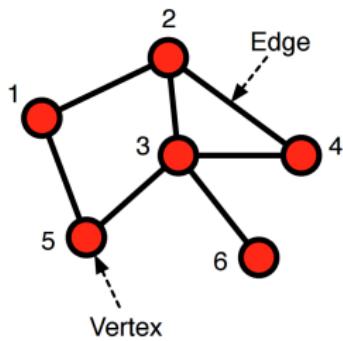
Simple networks statistics

Community detection

Basics: How to represent a network?

We will use graphs, which consist of vertices and edges.

Visually



Adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Adjacency list

1	2	5		
2	3	1	4	
3	2	5	4	6
4	2	3		
5	1	3		
6	3			

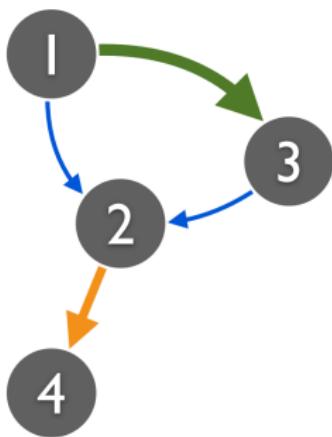
Edge list

$$\{(1, 2), (1, 5), (2, 3), (2, 4), (3, 5), (3, 6)\}$$

Basics: How to represent a network?

More exotic networks

Visually



Adjacency matrix

		Target node			
		1	2	3	4
Source node	1	0	1	3	0
	2	0	0	0	2
	3	0	1	0	0
	4	0	0	0	0

Adjacency list

1	2, 3
2	4
3	2

Edge list

- 1, 2, 1
- 1, 3, 3
- 2, 4, 2
- 3, 2, 1

Detour to R

See *plottingScript.R*

Descriptive statistics of networks

Degree of a node

- ▶ number of edges connected to a node $k_i = \sum_j A_{ij}$

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Degree distribution

k	$\Pr(k)$
1	1/6
2	3/6
3	1/6
4	1/6

Centrality: How important is a node?

Normalize degree of each node with the maximum degree in the network.



Does this capture what we consider important?

Closeness centrality

Distance d_{ij} between nodes i and j is the number of edges between on the shortest path between i and j .

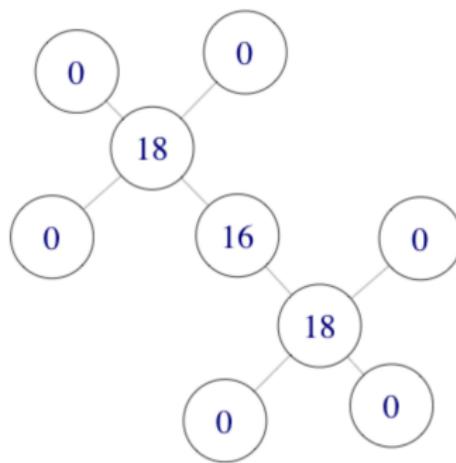
$$\text{closeness_centrality}(i) = \frac{n - 1}{\sum_{j \neq i} d_{ij}}$$



What matters is how close to everybody else a node is, that is, to be easily reachable or have the power to quickly reach others.

Betweenness centrality

A node is important if it lies on many shortest-paths, so it is essential in passing information through the network.



Betweenness centrality

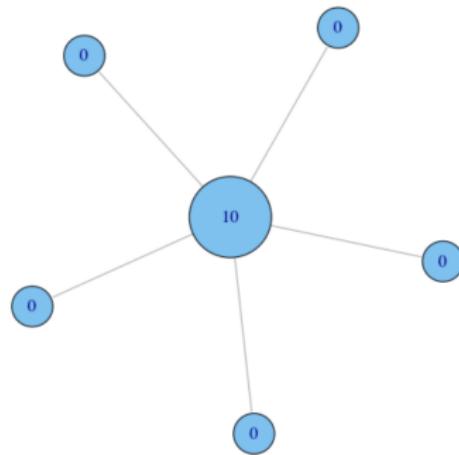
How often a node serves as the “bridge” that connects two other nodes.

$$\text{betweenness_centrality}(i) = \sum_{jk} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

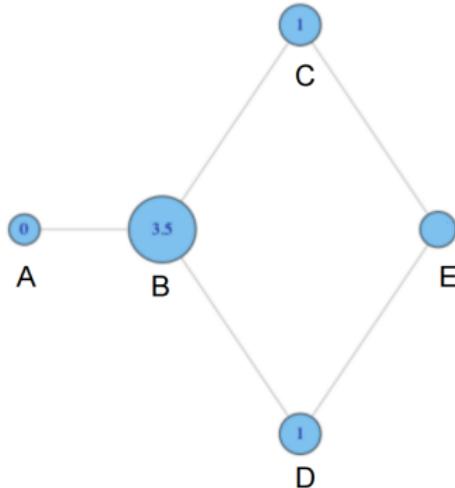
- ▶ $\sigma_{jk}(i)$ number of shortest paths from j to k that go through i
- ▶ σ_{jk} number of shortest paths from j to k

Strength of weak ties

Betweenness centrality

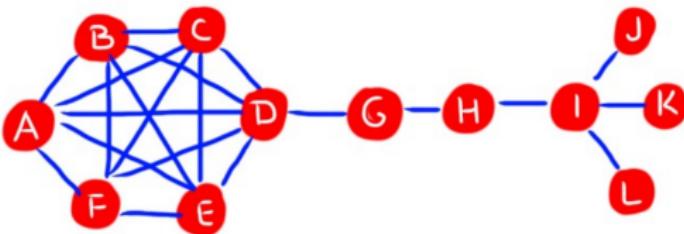


Betweenness centrality



- ▶ Why is betweenness of C and D equal to 1?
- ▶ What is betweenness of E?

Betweenness centrality



- ▶ Which node has high betweenness but low degree?
- ▶ Which node has high degree but low betweenness?

Eigenvalue centrality

A node is central if it is connected to other central nodes.

$$\text{eigenvector_centrality}(i) \sim \sum_j A_{ij} \cdot \text{eigenvector_centrality}(j)$$

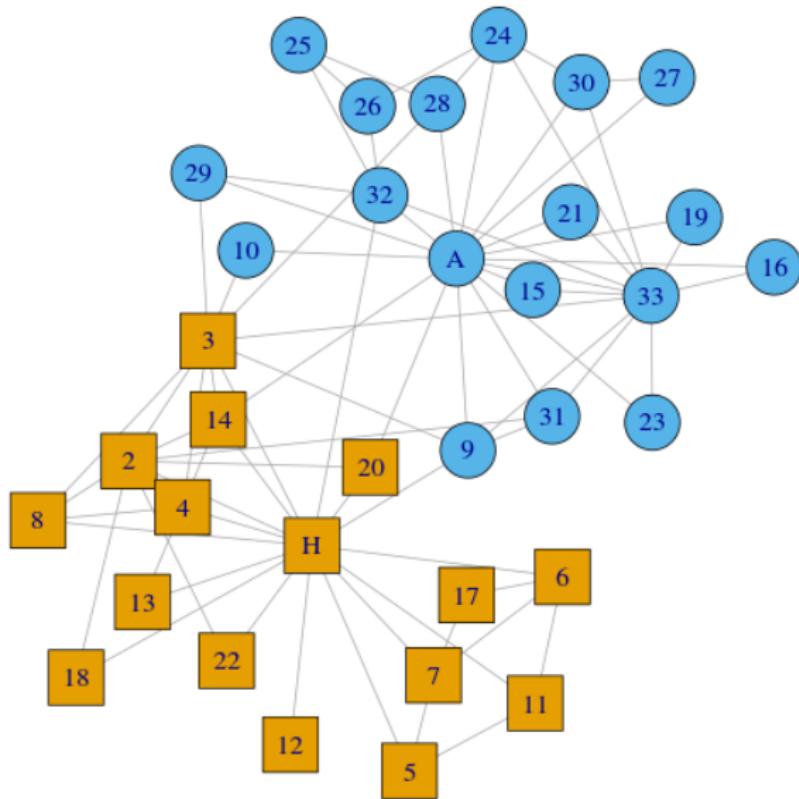
Page rank — extension to directed networks

A random walker following edges in a network for a very long time will spend a proportion of time at each node which can be used as a measure of importance.



Karate club example

Zachary's karate club network (H: Instructor, A: Club president)



Caveats about centrality

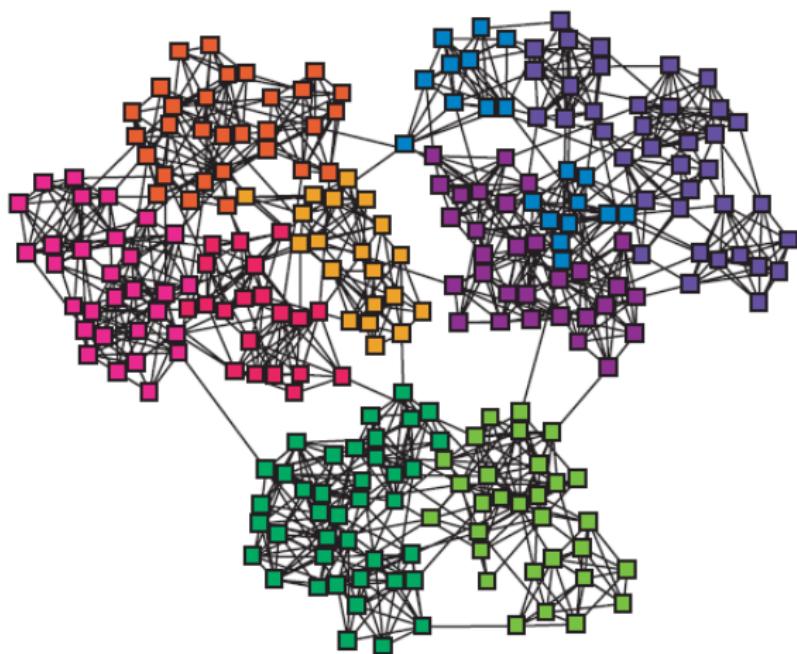
Each measure of centrality is fundamentally a proxy of some underlying network process.

If the particular network process is irrelevant or unrealistic for a given network, then any measure of centrality based on that process will produce nonsense.

Should be used mainly in an exploratory manner, to gain some insight into the general structure and pattern of a network.

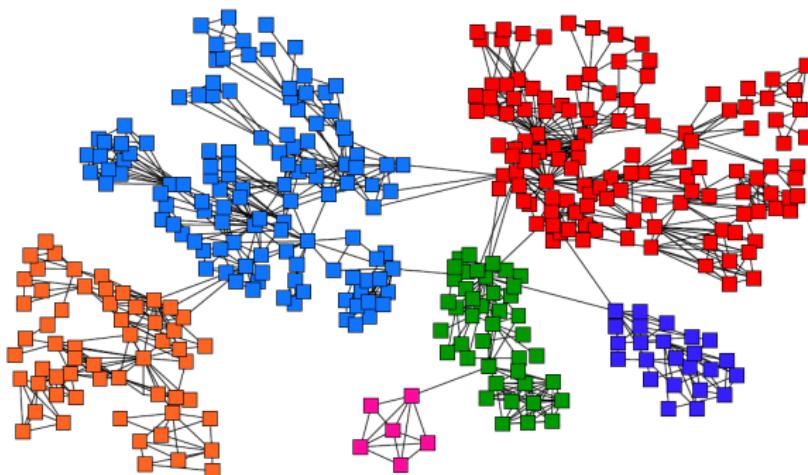
Community detection

Networks are often organized in modules, clusters, and communities

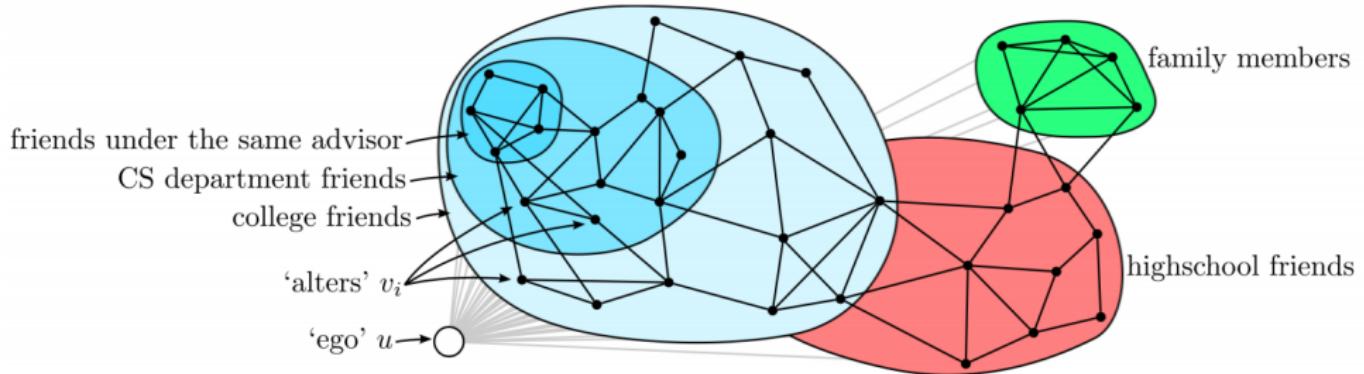


Community detection

Goal is to identify meaningful communities. One reasonable definition is to find groups of nodes that are densely connected, but have few edges with nodes from other communities.



Friend groups within ego-nets



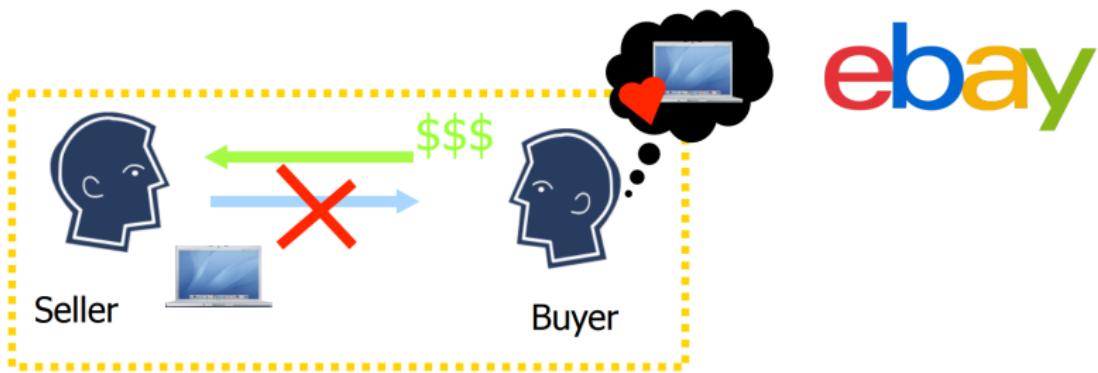
Fraud in Online Auctions

Auction sites: Attractive target for fraud

- ▶ 63% of complaints to Federal Internet Crime Complaint Center in U.S. in 2006

Average loss per incident: = \$385

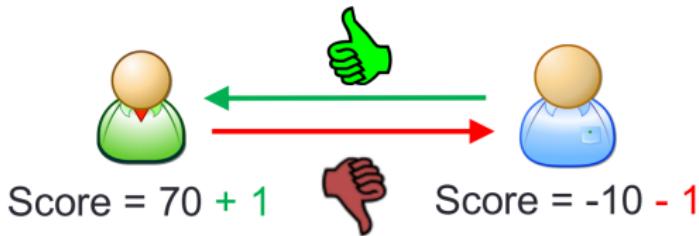
Often non-delivery fraud



Individual features (for example, geography), are too easy to fake.

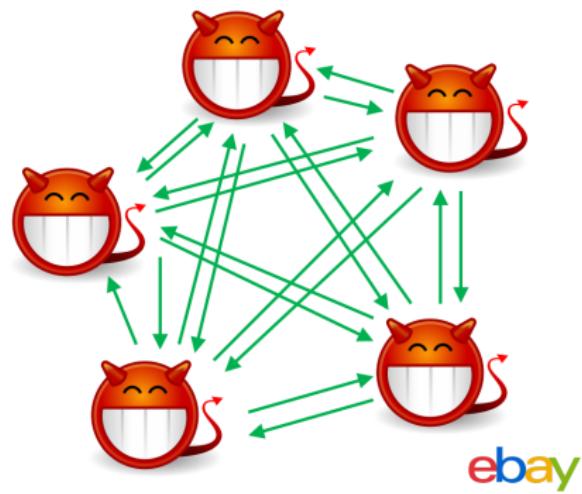
Given a graph of user interactions, what does fraud look like and how can we catch it?

Each user gets a reputation score based on peer feedback

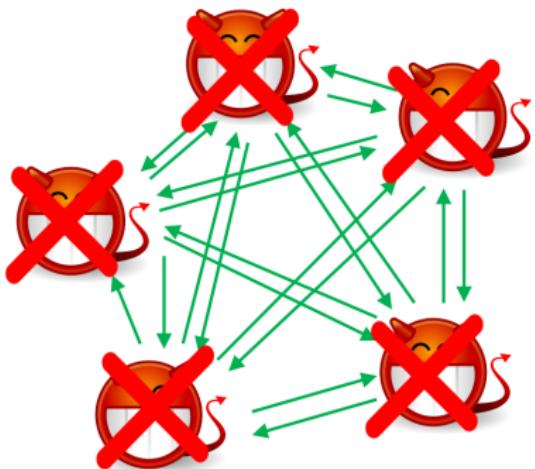
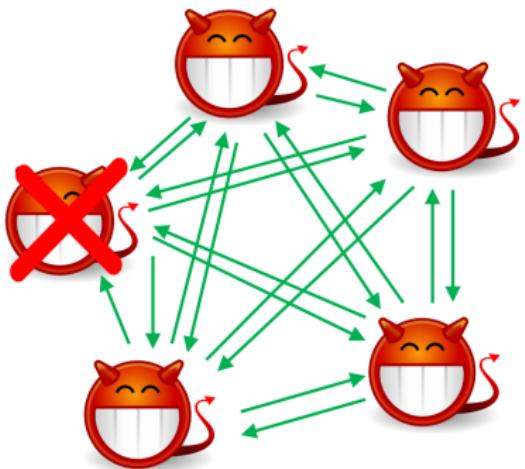


Fraudsters need to keep a high reputation score. How do they game the system?

Do they all just give each other positive reviews?

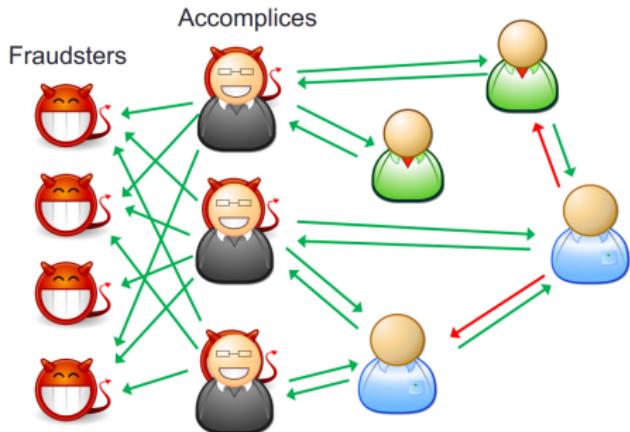


No, because if one is caught they are all revealed.



Fraudsters form near-bipartite core of 2 roles:

1. Accomplices: Trade with honest, looks normal
2. Fraudsters: Trade with accomplices; Fraud with honest



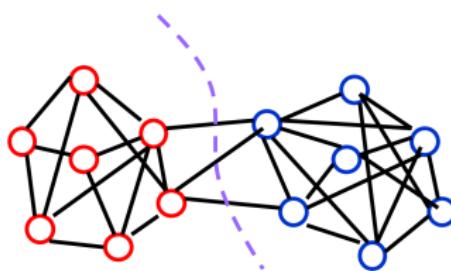
	Fraud	Accomplice	Honest
Fraud	ϵ	$1 - 2\epsilon$	ϵ
Accomplice	0.5	2ϵ	$0.5 - 2\epsilon$
Honest	ϵ	$(1 - \epsilon)/2$	$(1 - \epsilon)/2$

Community detection

Networks have a natural community structure.

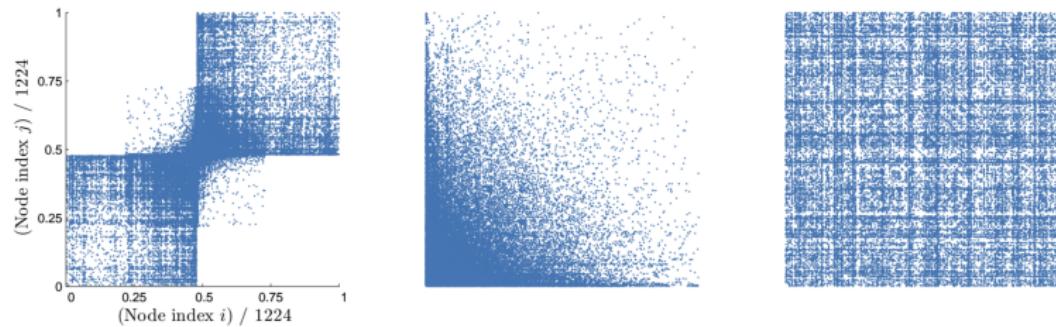
We want to discover this structure automatically.

Without “looking”, can we discover community structure in an automated way?



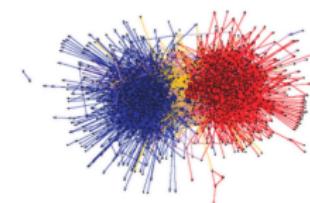
What we have access to is a graph (adjacency matrix).

Example: Political weblog data



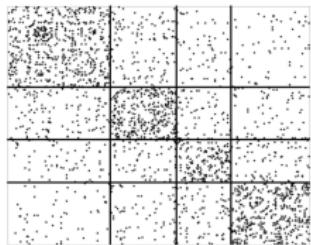
- Left: 586 liberal, 638 conservative
- Middle: Sorted by degree
- Right: Randomly permuted

⇒ Party “labels” reveal block structure

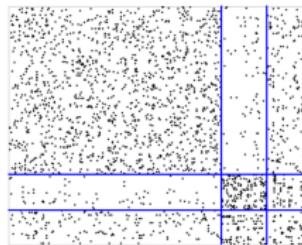


Example: Add Health (1994)

Survey data on high-school friendships



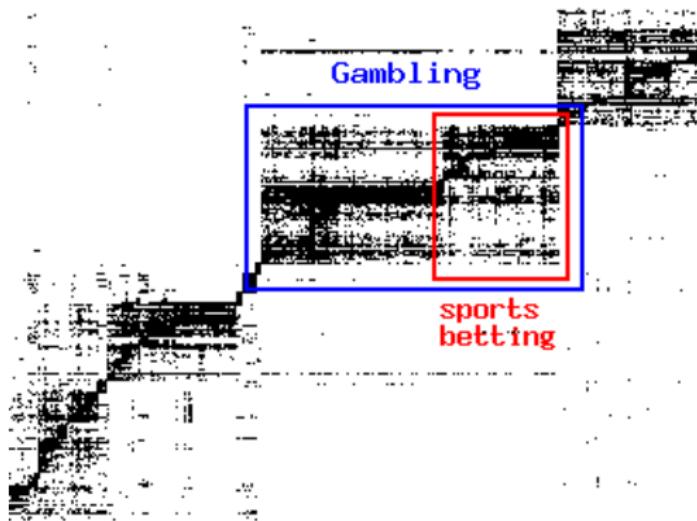
students grouped by
year (black lines)



students grouped by
race (blue lines)

Example: Micro-Markets in Sponsored Search

Find micro-markets by partitioning the query-to-advertiser graph.



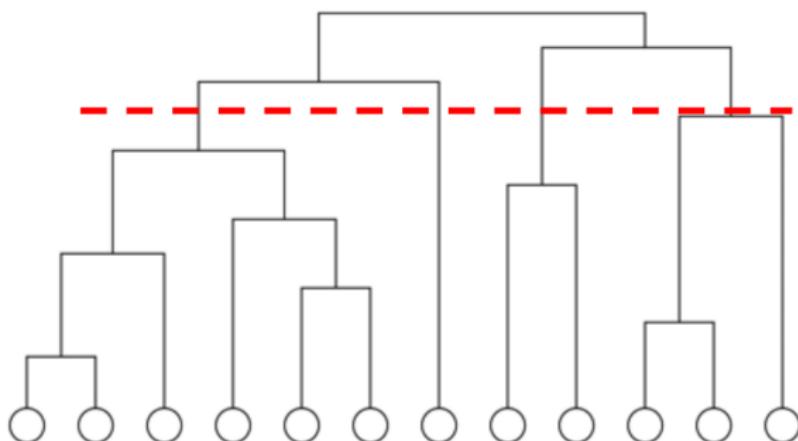
Hierarchical clustering

Compute the “distances” for all pairs of vertices

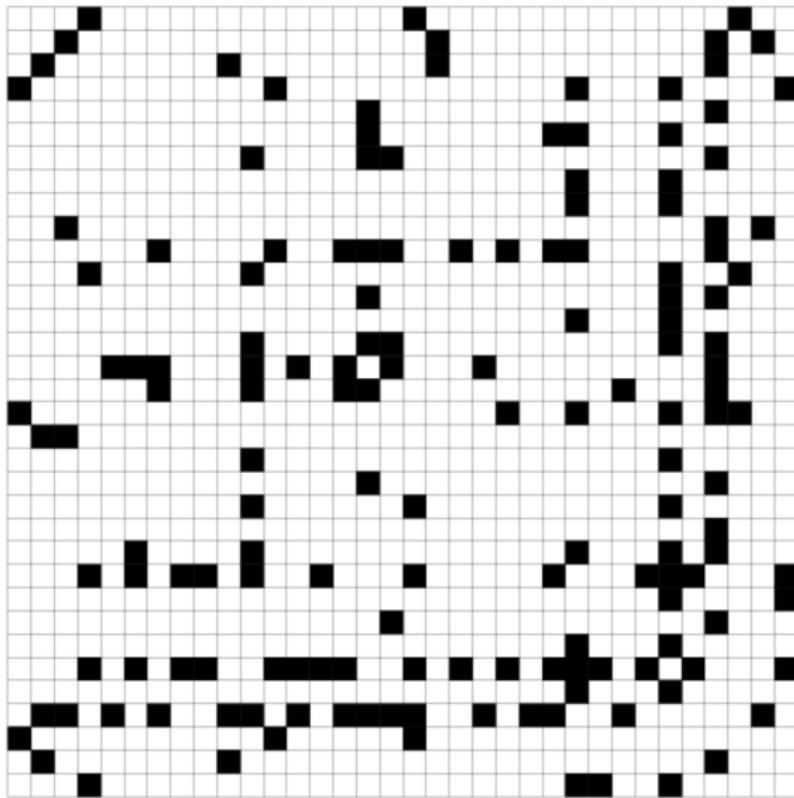
Start with all n vertices disconnected

Add edges between pairs one by one in order of decreasing weight

Output: nested components, where one can take a “slice” at any level of the tree

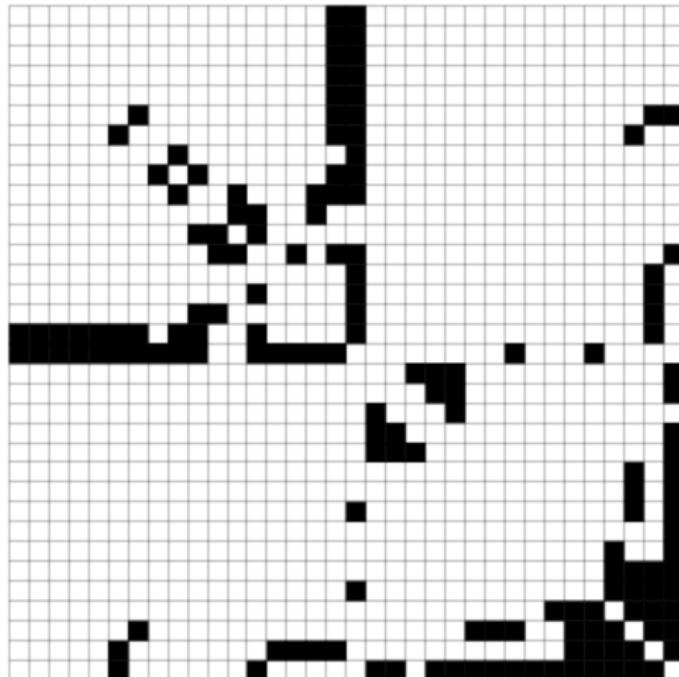


Karate club: Permuted adjacency matrix



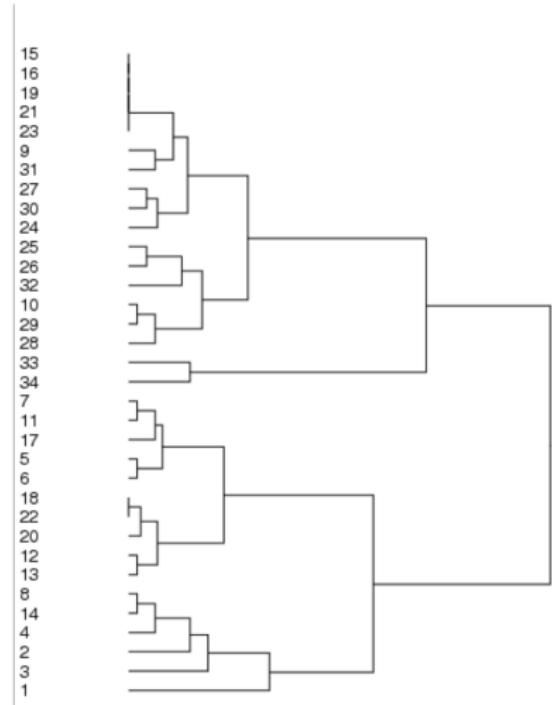
Karate club: Reordered adjacency matrix

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



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

Karate club: Dendrogram



Betweenness clustering: Girvan-Newman

Compute the betweenness of all edges

While (betweenness of any edge > threshold):

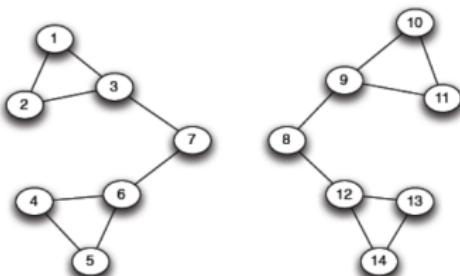
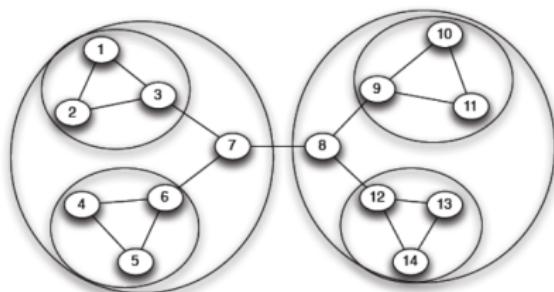
- ▶ remove edge with highest betweenness
- ▶ communities are
- ▶ recalculate betweenness

Betweenness needs to be recalculated at each step as removal of an edge can impact the betweenness of another edge

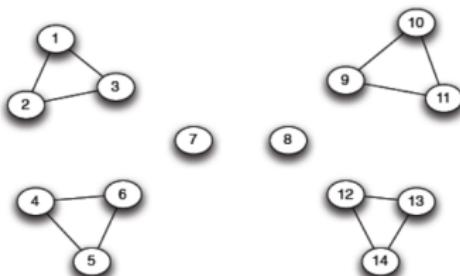
Connected components are communities

Gives a hierarchical decomposition of the network

Successively remove edges of highest betweenness, breaking up the network into separate components

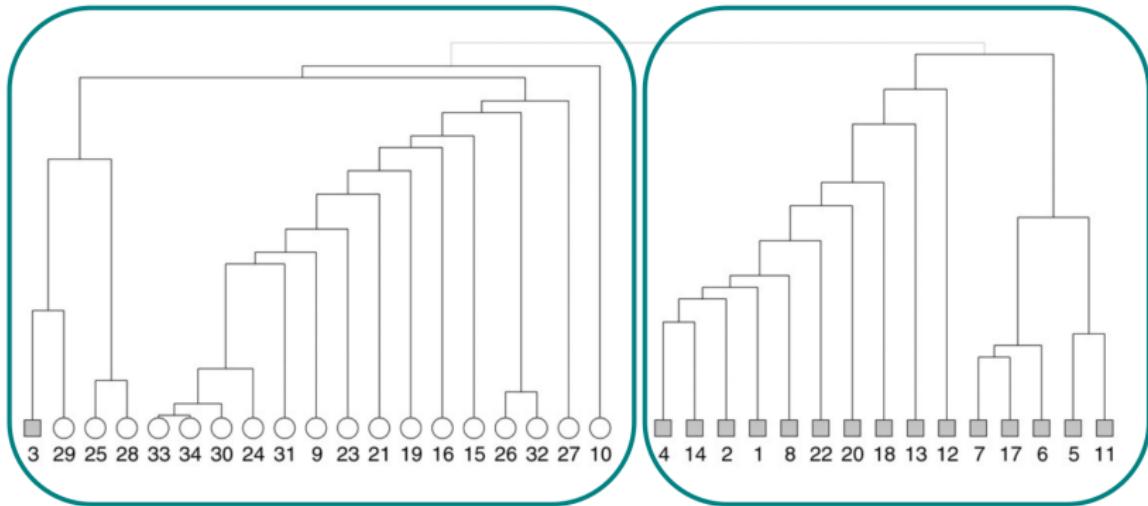


(a) Step 1



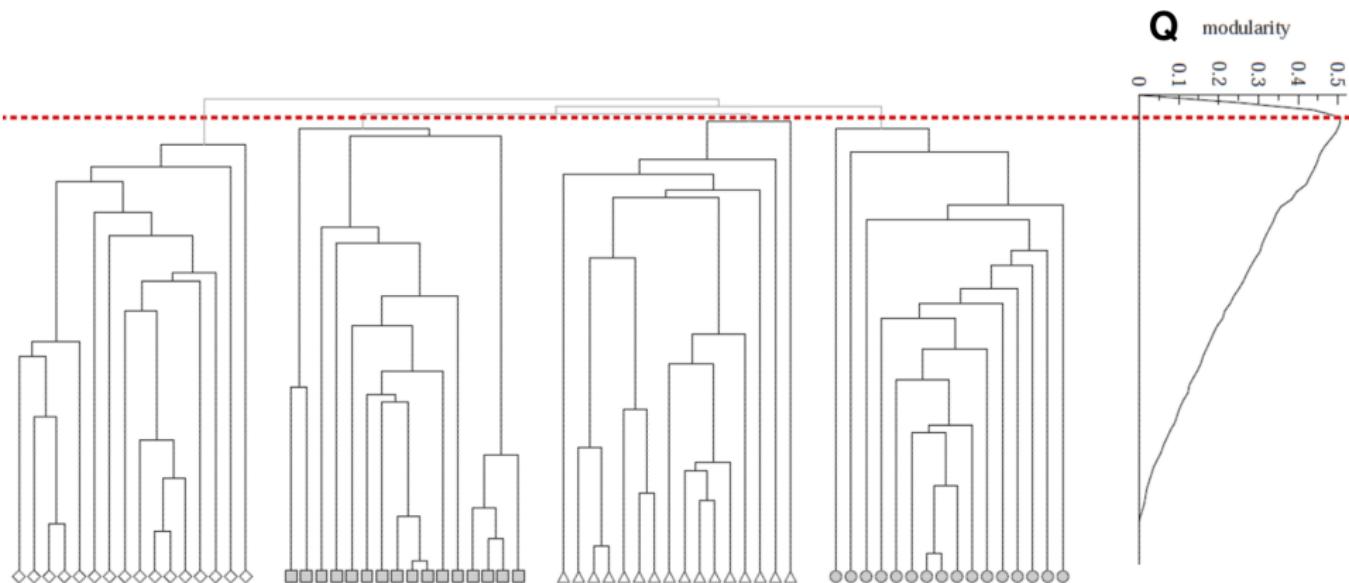
(b) Step 2

Example: Karate club — betweenness clustering



How many clusters?

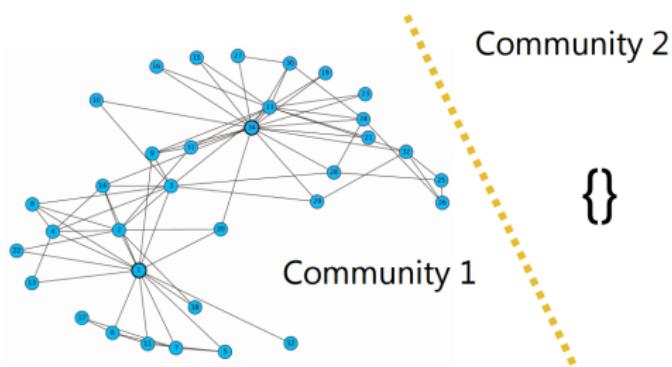
Modularity — a measure of how well a network is partitioned into communities



Graph cuts

Cut the network into two partitions such that the number of edges crossed by the cut is minimal

Degenerate solution

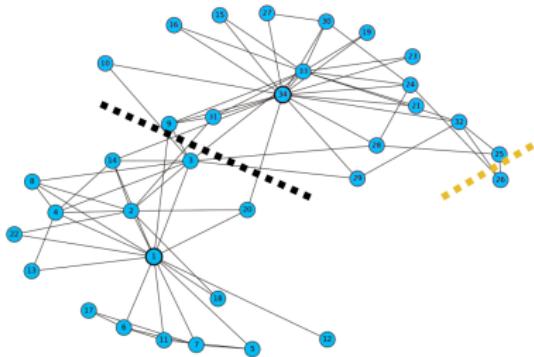


Graph cuts

Want a cut that favors large communities over small ones

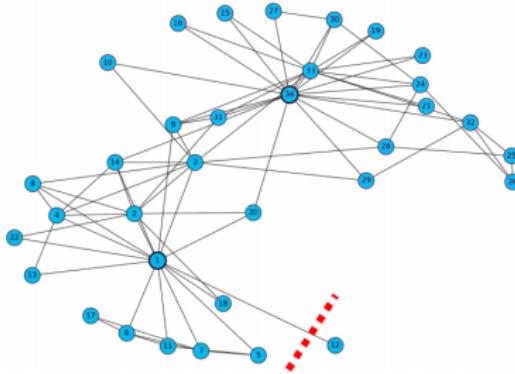
$$\text{Ratio Cut}(C) = \frac{1}{|C|} \sum_{c \in C} \frac{\text{cut}(c, \bar{c})}{|c|}$$

#of edges that separate c from the rest of the network
Proposed set of communities
size of this community



$$\text{Ratio Cut}(\text{---}) = \frac{1}{2} \left(\frac{3}{33} + \frac{3}{1} \right) = 1.54545$$

$$\text{Ratio Cut}(\text{...}) = \frac{1}{2} \left(\frac{9}{16} + \frac{9}{18} \right) = 0.53125$$



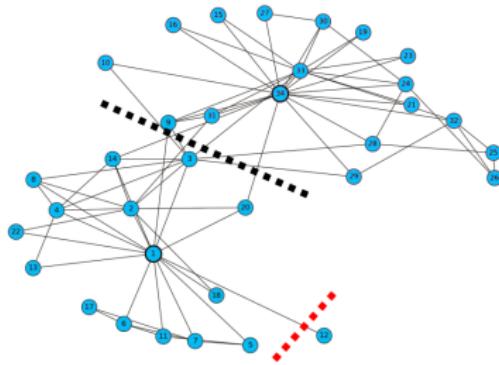
$$\text{Ratio Cut}(\text{---}) = \frac{1}{2} \left(\frac{1}{33} + \frac{1}{1} \right) = 0.51515$$

Normalized graph cuts

Rather than counting all nodes equally in a community, we should give additional weight to “influential”, or high-degree nodes

$$\text{Normalized Cut}(C) = \frac{1}{|C|} \sum_{c \in C} \frac{\text{cut}(c, \bar{c})}{\sum \text{degrees in } c}$$

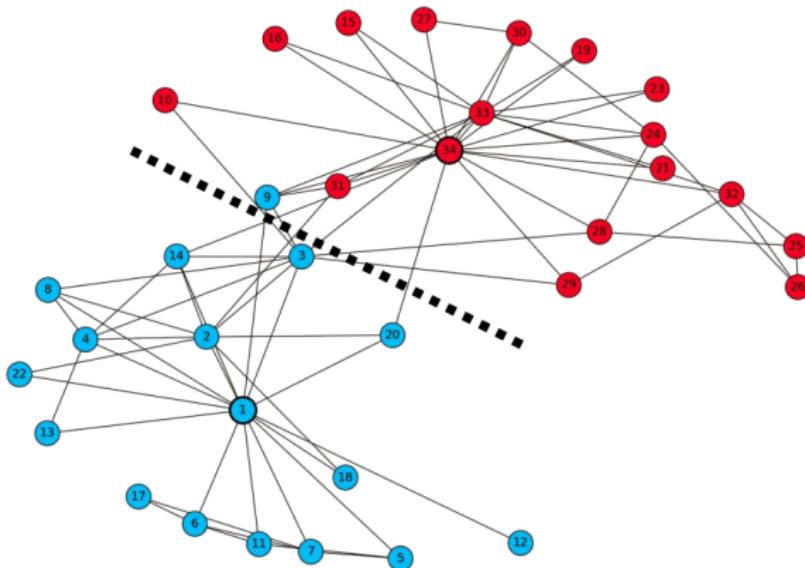
nodes of high degree will have more influence in the denominator



$$\text{Norm. Cut}(\text{---}) = \frac{1}{2}\left(\frac{1}{155} + \frac{1}{1}\right) = 0.50322$$

$$\text{Norm. Cut}(\cdot\cdot) = \frac{1}{2}\left(\frac{9}{76} + \frac{9}{80}\right) = 0.11546$$

Example: Karate club — graph cut



.. \swarrow = Optimal cut

Red/blue = actual split