CSS 2013 day1

Network Representation and Description

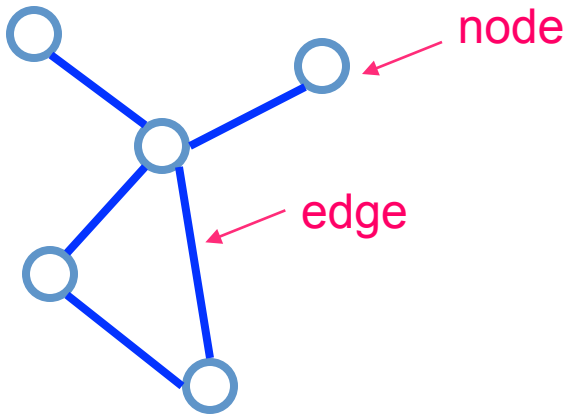
Lexing Xie
Research School of Computer Science

Lecture slides credit: Lada Adamic, Univ. Michigan
Jure Leskovec, Stanford University

What are networks?

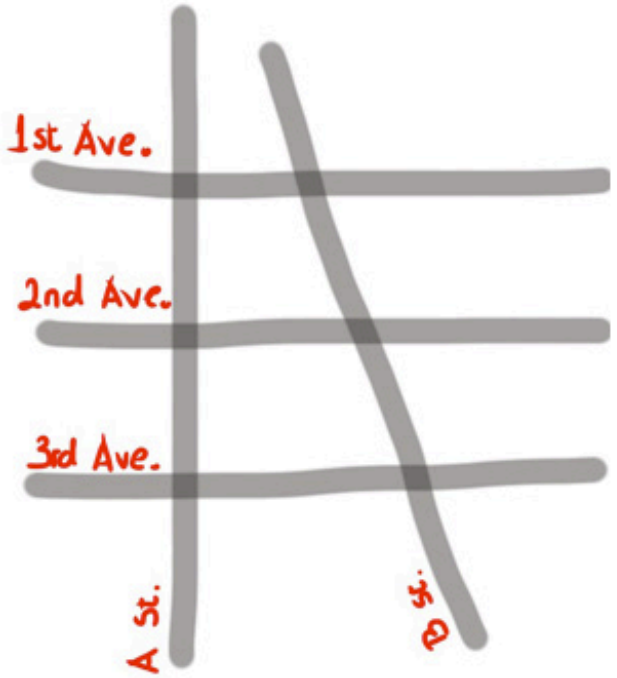
- Networks are sets of nodes connected by edges.

"Network" ≡ "Graph"



points	lines	
vertices	edges, arcs	math
nodes	links	computer science
sites	bonds	physics
actors	ties, relations	sociology

The attached image shows 5 streets (A and B streets, and 1st, 2nd, and 3rd Avenue). How can a network be constructed from these streets?
(Check all that apply)



- [ ] Roads (A St., B St., 1st Ave, ...) are nodes and an edge is drawn between every pair of roads that intersect.
- [ ] Intersections are nodes (e.g. A St. and 1st Ave, B St. and 2nd Ave), and an edge is drawn between any two intersections that are directly connected by a segment of street with no intervening intersections.
- [ ] Street blocks are nodes (e.g. the block between A and B, and 2nd and 3rd), and blocks that are adjacent (i.e. across the street from each other) have edges.

Submit

Skip

https://www.coursera.org/course/sna

Topics for this 1.5 hours

- Are nodes connected through the network?
- How far apart are they?
- Are some nodes more important due to their position in the network?
- Is the network composed of communities?

Network elements: edges

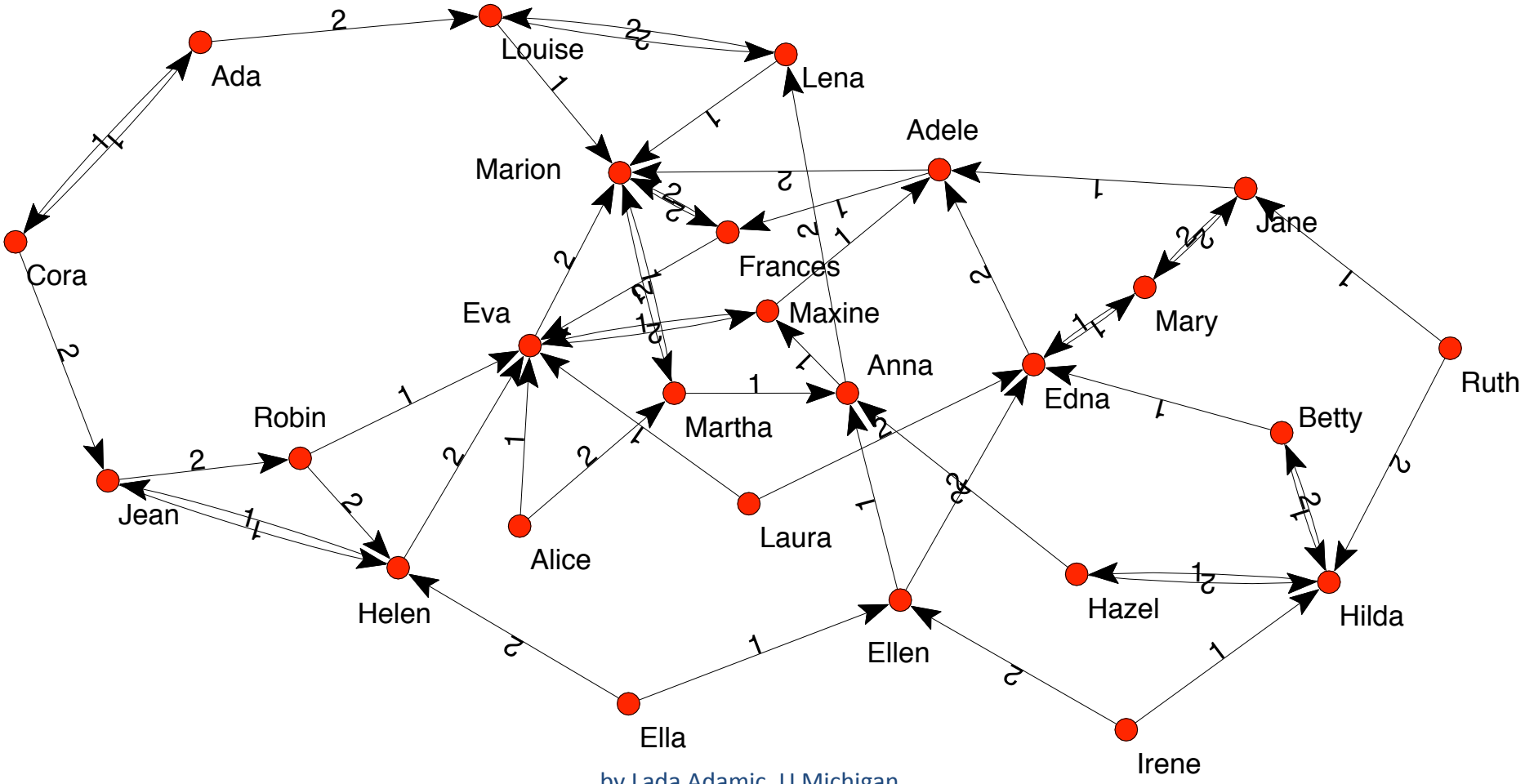
- Directed (also called arcs, links)
  - A -> B
    - A likes B, A gave a gift to B, A is B's child
- Undirected
  - A <-> B or A – B
    - A and B like each other
    - A and B are siblings
    - A and B are co-authors

Edge attributes

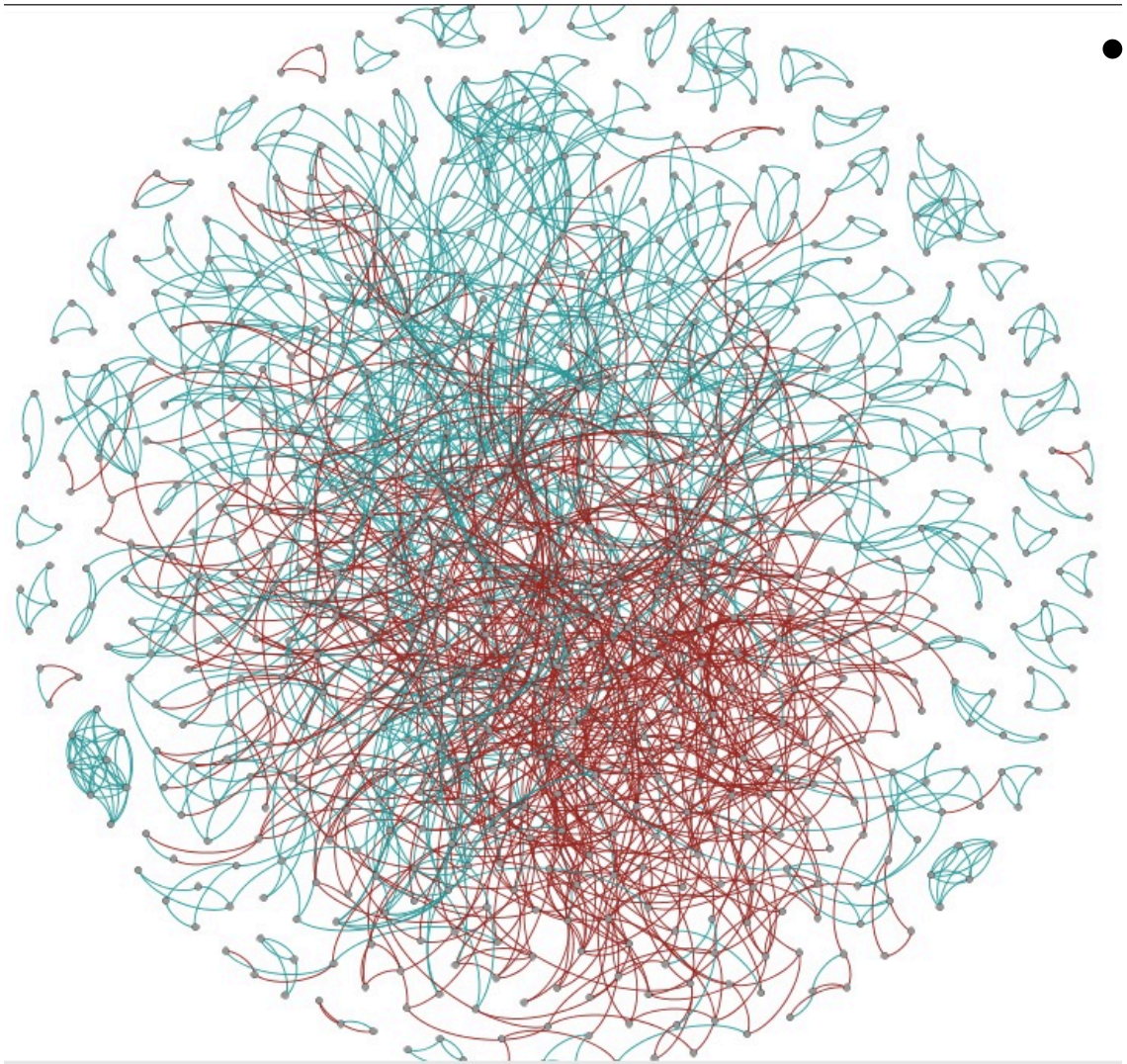
- Examples
  - weight (e.g. frequency of communication)
  - ranking (best friend, second best friend…)
  - type (friend, relative, co-worker)
  - properties depending on the structure of the rest of the graph: e.g. betweenness

Directed networks

- girls' school dormitory dining-table partners, 1st and 2nd choices
  (Moreno, *The sociometry reader*, 1960)



Positive and negative weights



- e.g. one person trusting/distrusting another
  - Research challenge: How does one 'propagate' negative feelings in a social network? Is my enemy's enemy my friend?

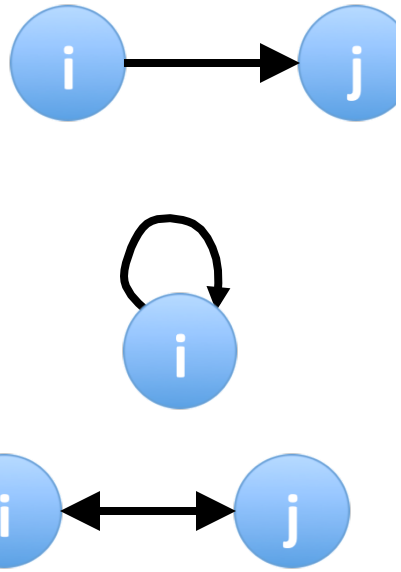
sample of positive & negative ratings from Epinions network
by Lada Adamic, U Michigan

Data representation

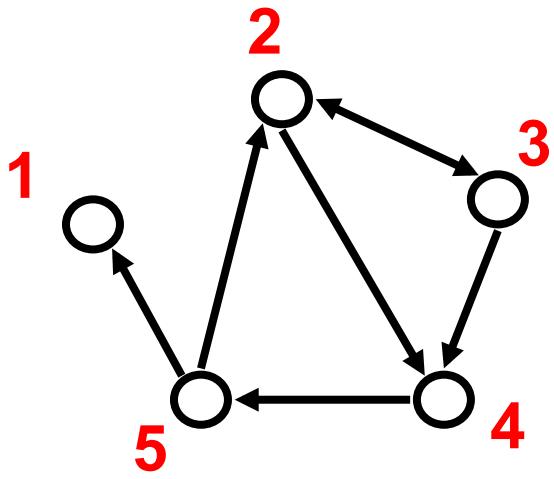
- adjacency matrix
- edgelist
- adjacency list

Adjacency matrices

- Representing edges (who is adjacent to whom) as a matrix
  - A_{ij} = 1 if node *i* has an edge to node *j*
    = 0 if node *i* does not have an edge to j
  - A_{ii} = 0 unless the network has self-loops
  - $A_{ij} = A_{ji}$ if the network is undirected, or if *i* and *j* share a reciprocated edge



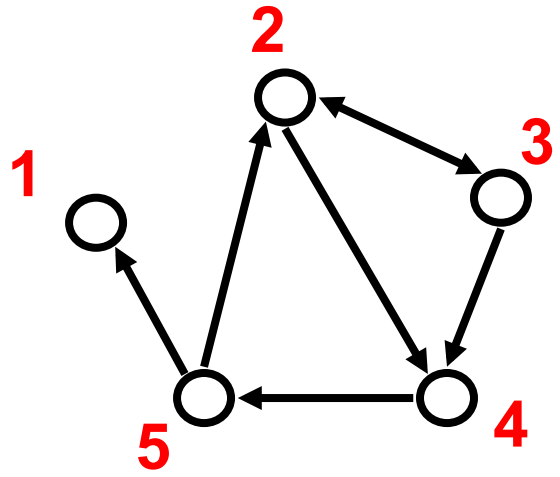
Example adjacency matrix



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

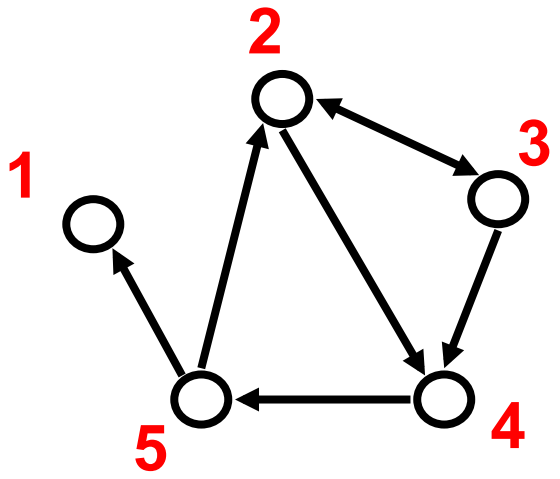
Edge list

- Edge list
  - 2, 3
  - 2, 4
  - 3, 2
  - 3, 4
  - 4, 5
  - 5, 2
  - 5, 1



Adjacency lists

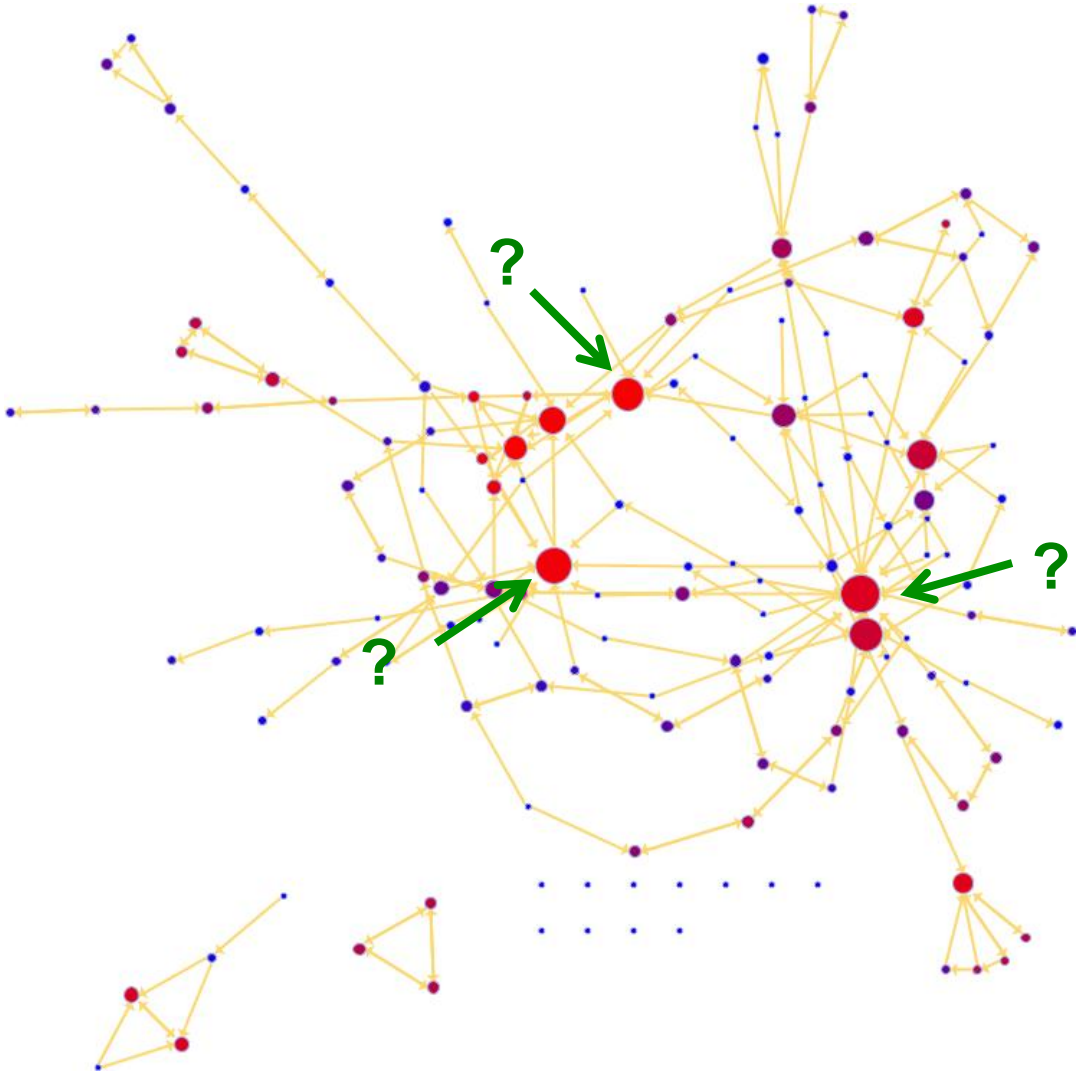
- Adjacency list
  - is easier to work with if network is
    - large
    - sparse
  - quickly retrieve all neighbors for a node
    - 1:
    - 2: 3 4
    - 3: 2 4
    - 4: 5
    - 5: 1 2



Computing metrics

- degree & degree distribution
- connected components

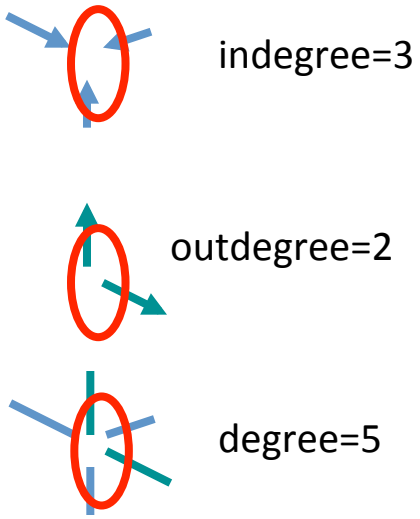
Degree: which node has the most edges?

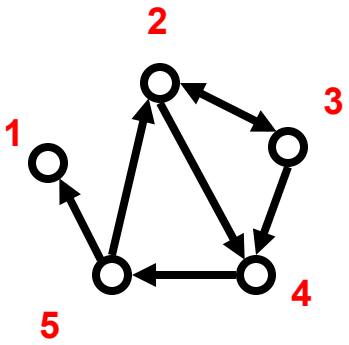


by Lada Adamic, U Michigan

Node degrees

- Node network properties
  - from immediate connections
    - indegree
      how many directed edges (arcs) are incident on a node
    - outdegree
      how many directed edges (arcs) originate at a node
    - degree (in or out)
      number of edges incident on a node
  - from the entire graph
    - centrality (betweenness, closeness)





Node degree from matrix values

- Outdegree = $\sum_{j=1}^{n} A_{ij}$

A =

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

example: outdegree for node 3 is 2, which we obtain by summing the number of non-zero entries in the 3rd *row*

$$\sum_{j=1}^{n} A_{3j}$$

- Indegree = $\sum_{i=1}^{n} A_{ij}$

A =

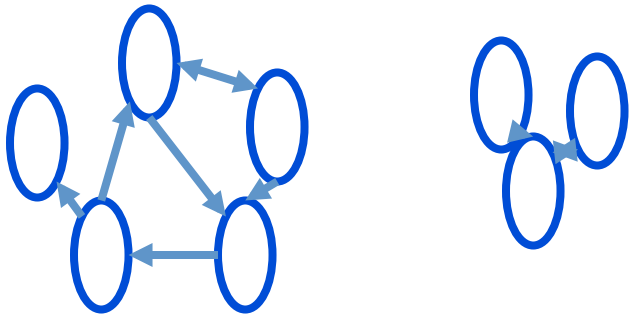
$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

example: the indegree for node 3 is 1, which we obtain by summing the number of non-zero entries in the 3rd *column*

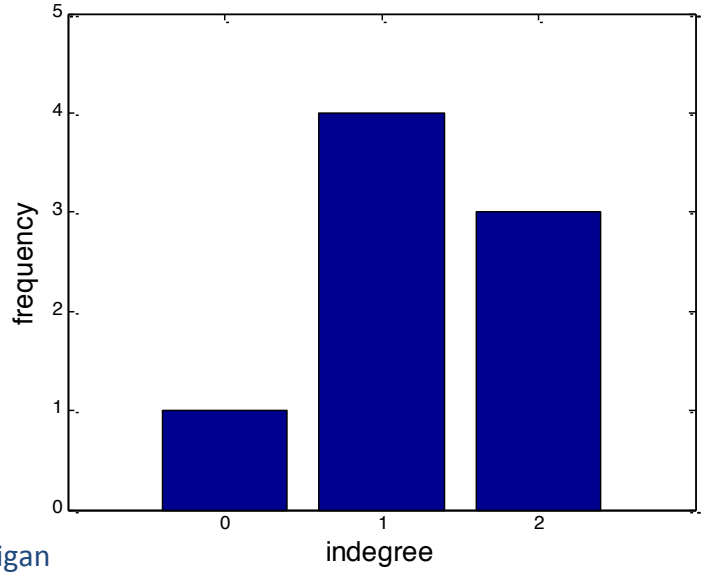
$$\sum_{i=1}^{n} A_{i3}$$

Network metrics: degree sequence and degree distribution

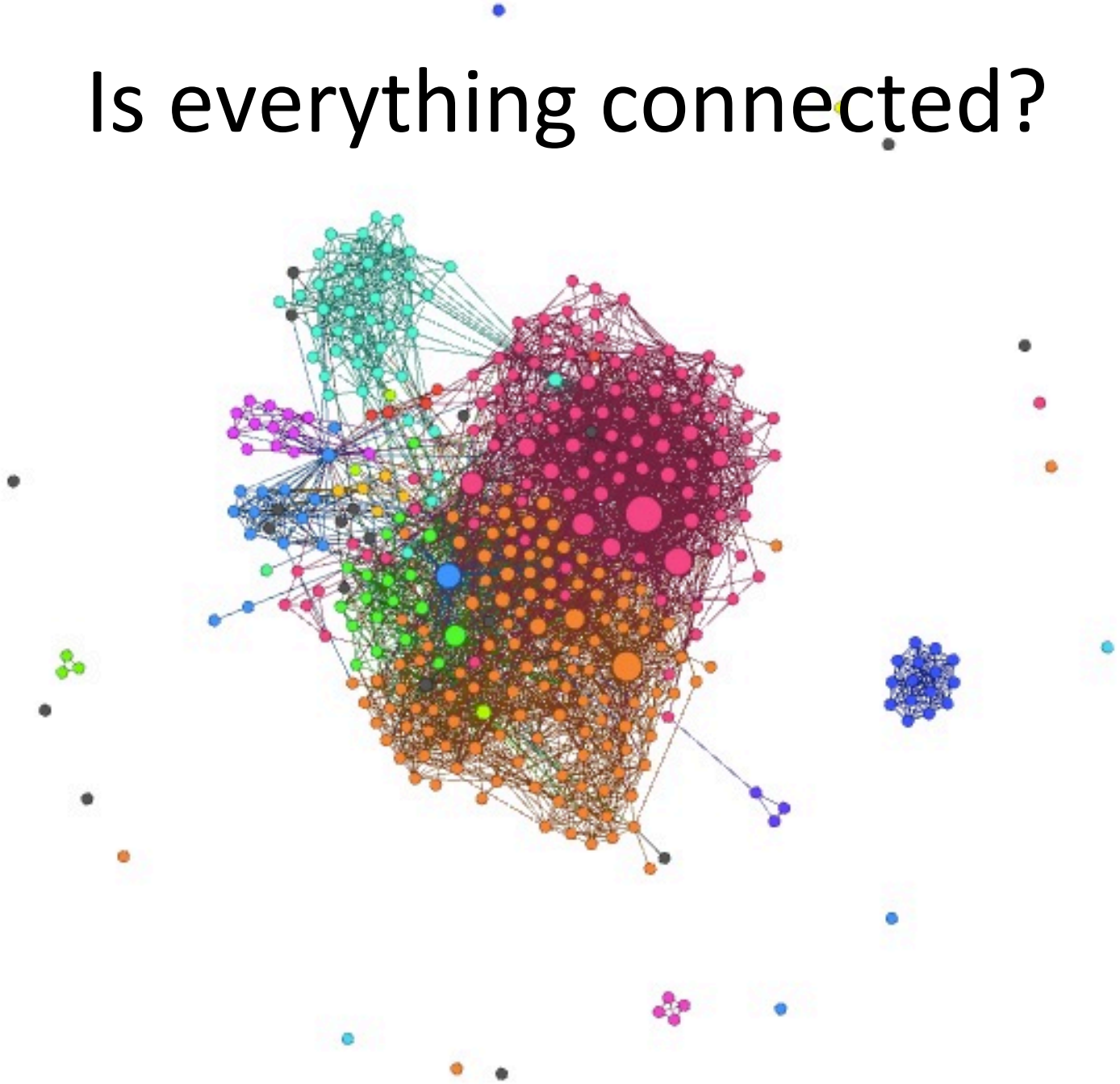
- Degree sequence: An ordered list of the (in,out) degree of each node
  - In-degree sequence:
    - [2, 2, 2, 1, 1, 1, 1, 0]
  - Out-degree sequence:
    - [2, 2, 2, 2, 1, 1, 1, 0]
  - (undirected) degree sequence:
    - [3, 3, 3, 2, 2, 1, 1, 1]



- Degree distribution: A frequency count of the occurrence of each degree
  - In-degree distribution:
    - [(2,3) (1,4) (0,1)]
  - Out-degree distribution:
    - [(2,4) (1,3) (0,1)]
  - (undirected) distribution:
    - [(3,3) (2,2) (1,3)]



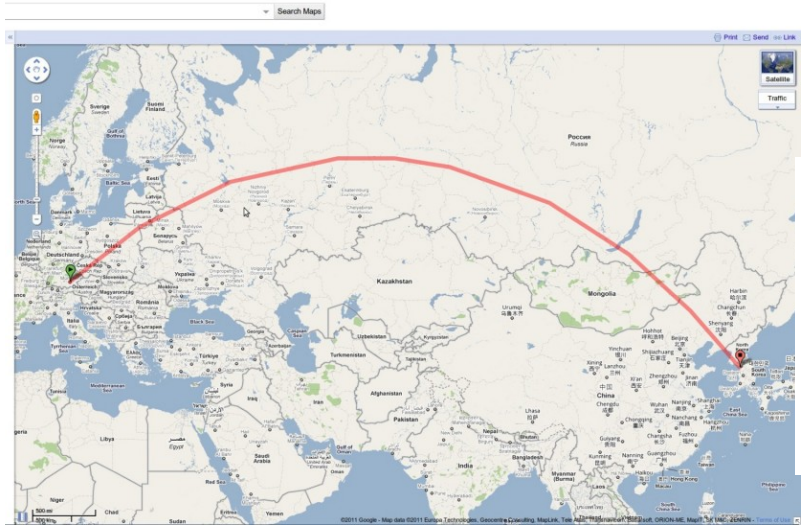
Is everything connected?

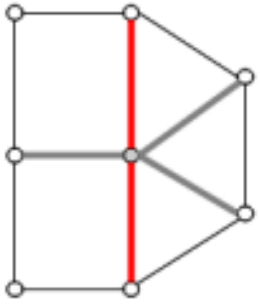


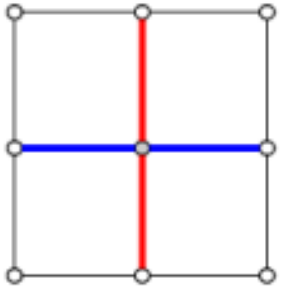
by Lada Adamic, U Michigan

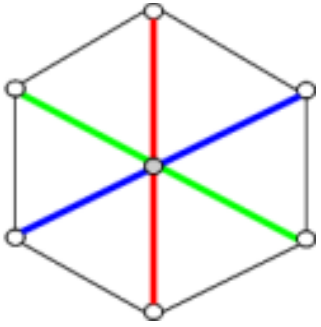
Distances in a Network

- Path: a walk (i1,i2,... iK) with each node ik distinct
- Cycle: a walk where i1 = iK
- Geodesic: a shortest path between two nodes

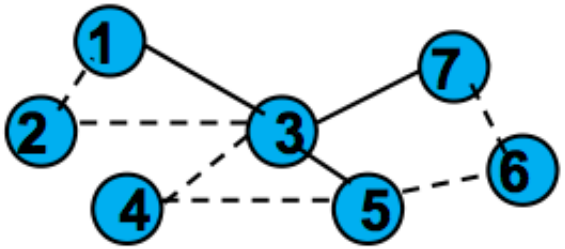




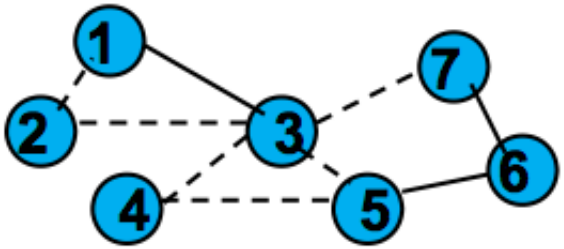




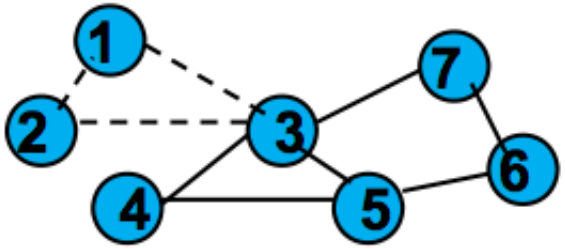
Paths, Walks, Cycles...



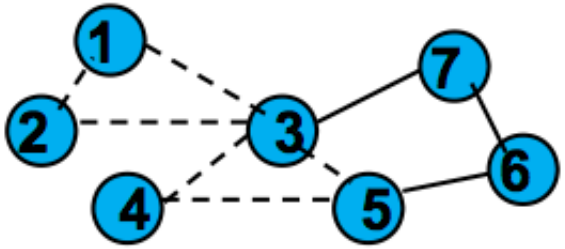
Path (and a walk) from 1 to 7:
1, 2, 3, 4, 5, 6, 7



Walk from 1 to 7 that is not a path:
1, 2, 3, 4, 5, 3, 7



Simple Cycle (and a walk)
from 1 to 1: 1, 2, 3, 1



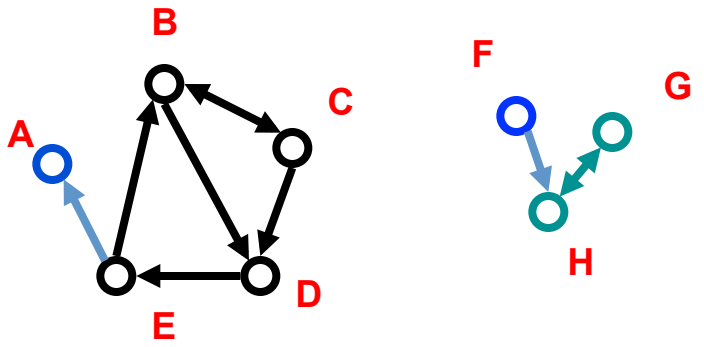
Cycle (and a walk) from 1 to 1:
1, 2, 3, 4, 5, 3, 1

Slide by Matthew Jackson, Stanford

Connected components

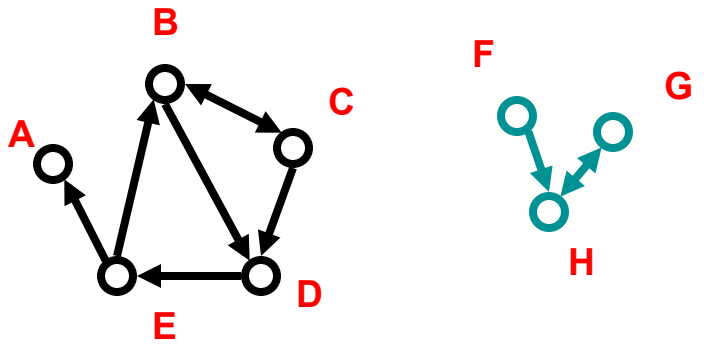
- Strongly connected components
  - Each node within the component can be reached from every other node in the component by following directed links

  - Strongly connected components
    - B C D E
    - A
    - G H
    - F



- Weakly connected components: every node can be reached from every other node by following links in either direction

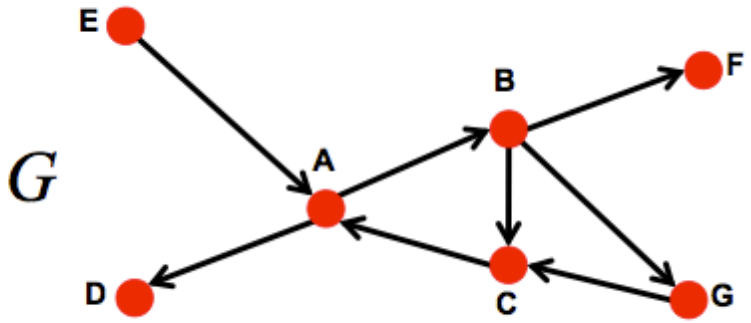
  - Weakly connected components
    - A B C D E
    - G H F



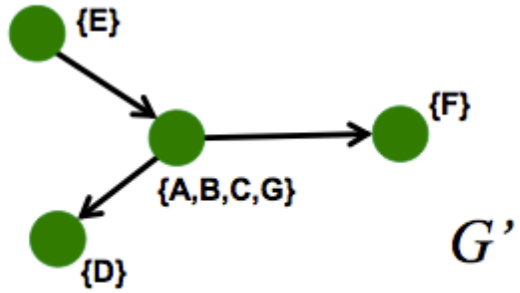
- In undirected networks one talks simply about ‘connected components’

Strongly Connected Component (SCC)

- **Fact:** Every directed graph is a DAG on its SCCs
  - (1) SCCs partitions the nodes of G
    - Each node is in exactly one SCC
  - (2) If we build a graph G' whose nodes are SCCs, and with an edge between nodes of G' if there is an edge between corresponding SCCs in G, then G' is a DAG

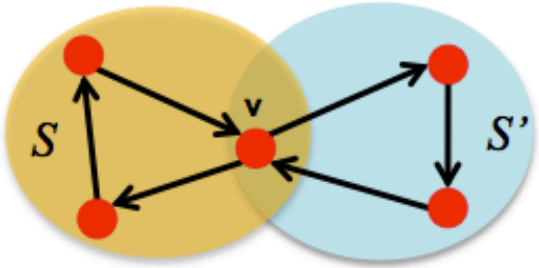


(1) Strongly connected components of graph G: {A,B,C,G}, {D}, {E}, {F}
(2) G' is a DAG:

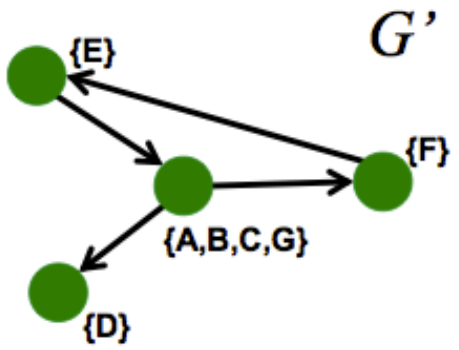


Proof (*)

- **Why is (1) true?** SCCs partitions the nodes of G.
  - Suppose node v is a member of 2 SCCs S and S'.
  - Then $S \cup S'$ is one large SCC:



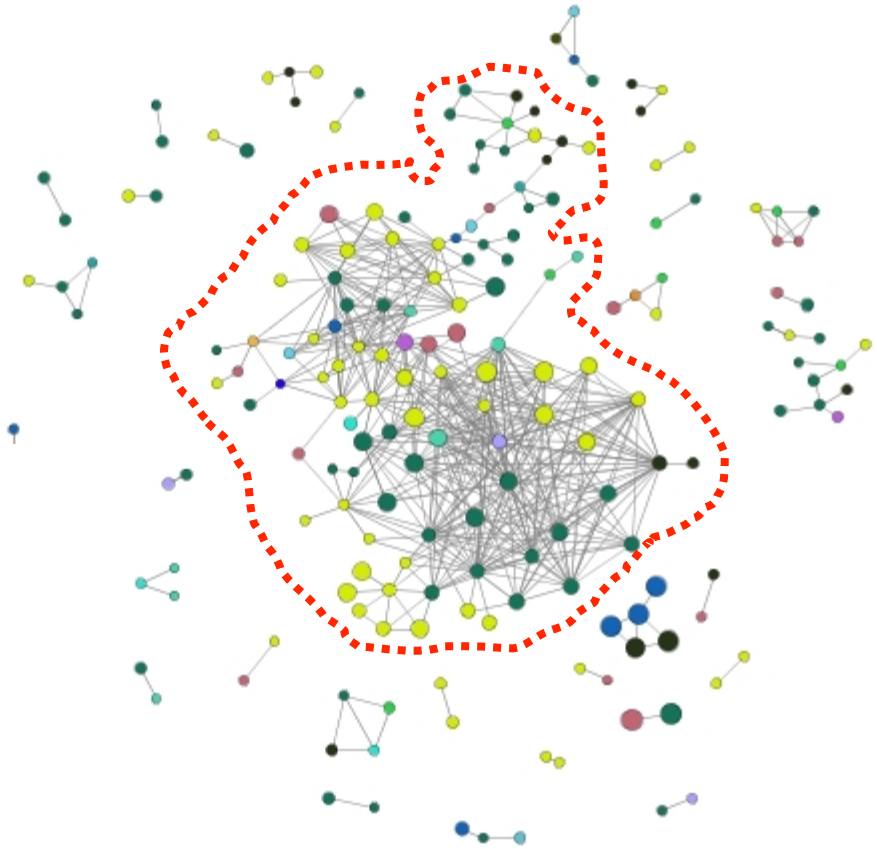
- **Why is (2) true?** G' (graph of SCCs) is a DAG
  - If G' is not a DAG, then we have a directed cycle.
  - Now all nodes on the cycle are mutually reachable, and all are part of the same SCC.



Now {A,B,C,G,E,F} is a SCC

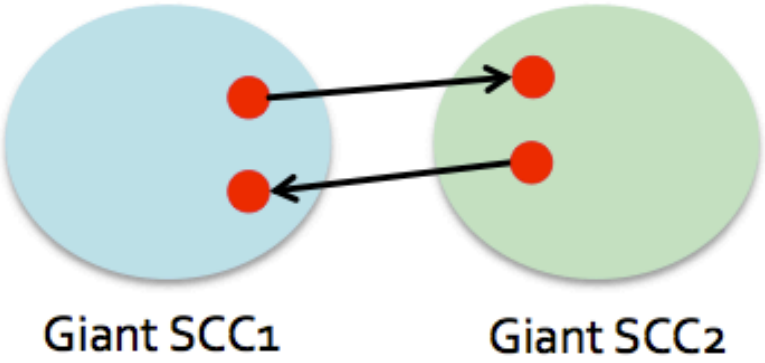
Giant component

- if the largest component encompasses a significant fraction of the graph, it is called the **giant component**

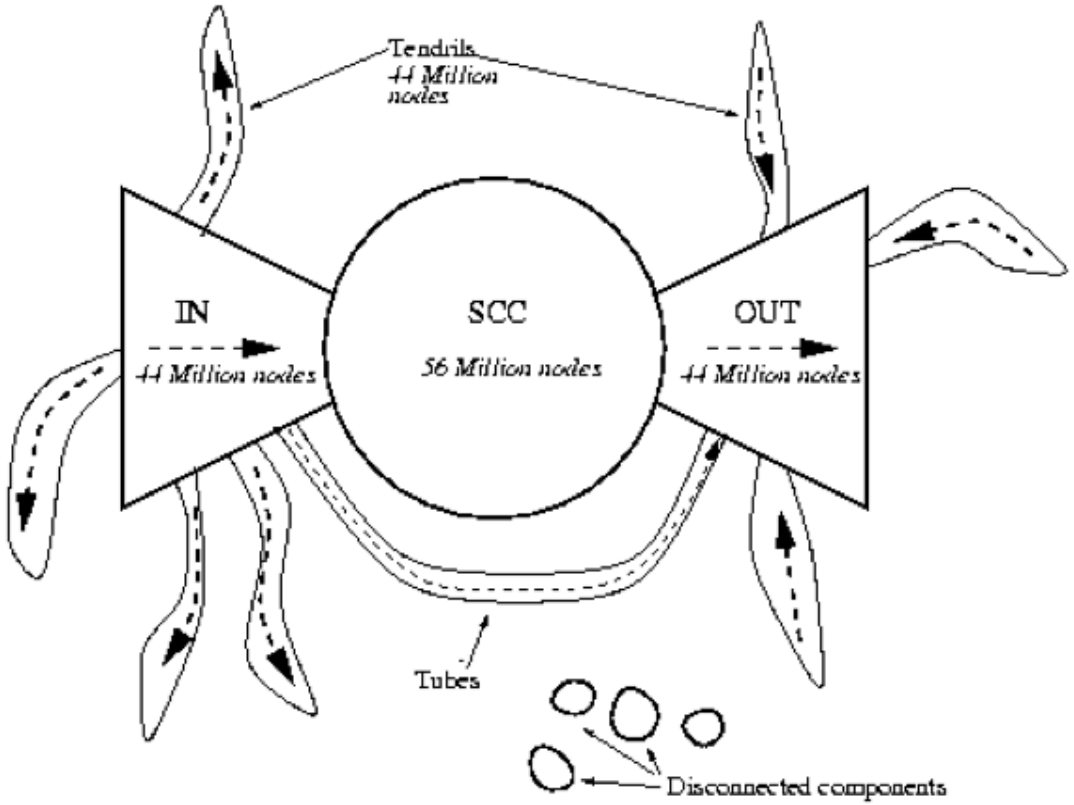


by Lada Adamic, U Michigan

- There is a giant SCC
- There won't be 2 giant SCCs: Why not? (*)
  - Just takes 1 page from one SCC to link to the other SCC
  - If the components have millions of pages the likelihood of this is very large



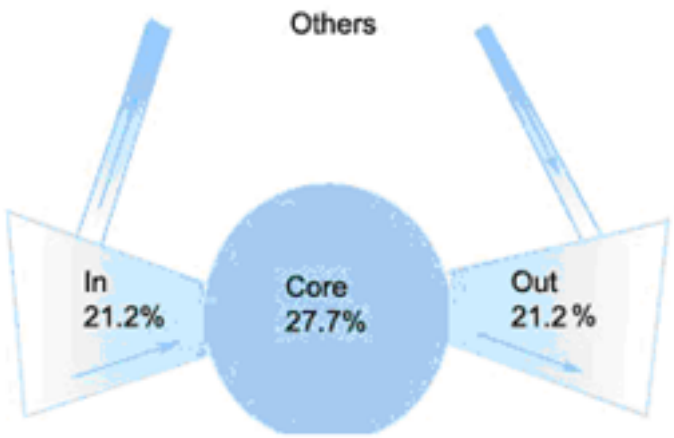
Bow-Tie Structure of the Web



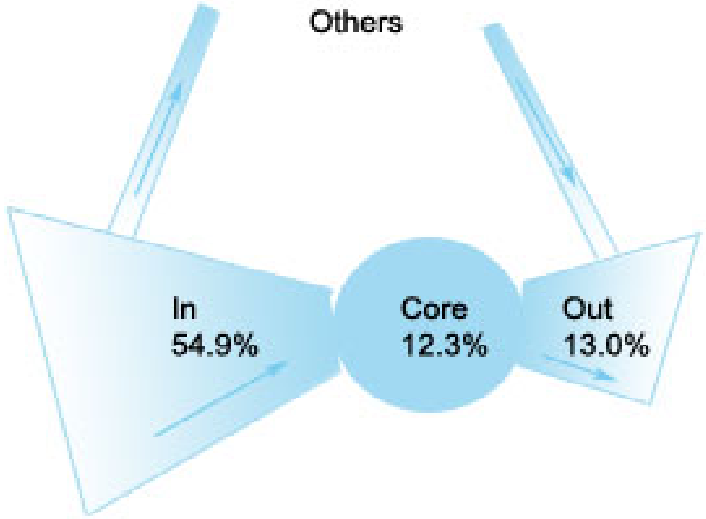
- 250 million pages, 1.5 billion links

Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. 2000. Graph structure in the Web. Comput. Netw. 33, 1-6 (June 2000), 309-320.

Not Everyone Asks/Replies



The Web is a bow tie

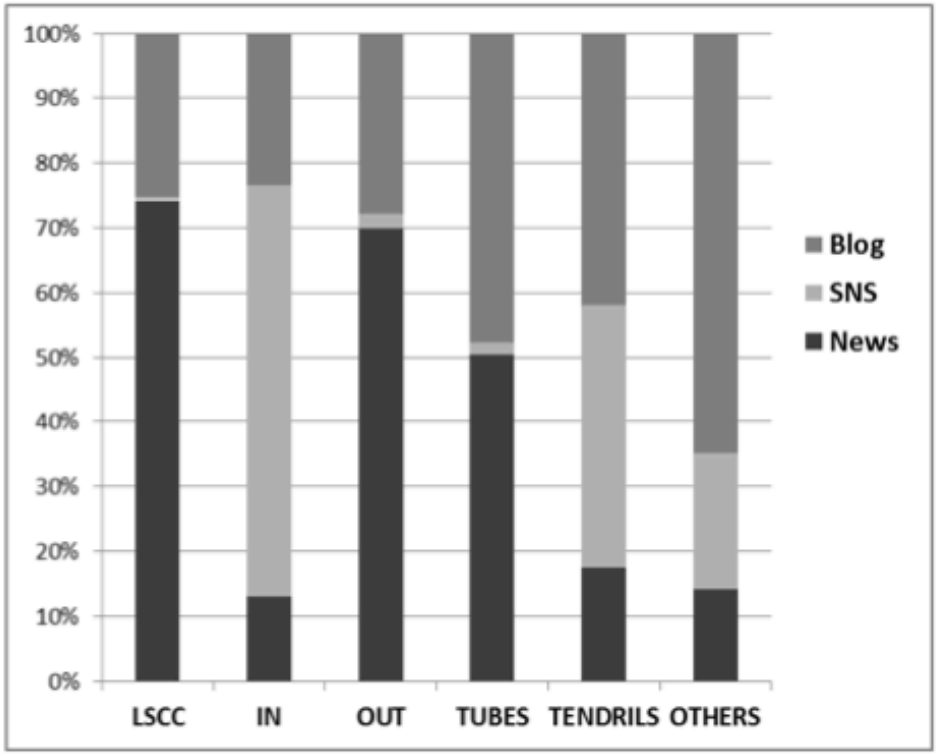


The Java Forum network is an uneven bow tie

- Core: A strongly connected component, in which everyone asks and answers
- IN: Mostly askers.
- OUT: Mostly Helpers

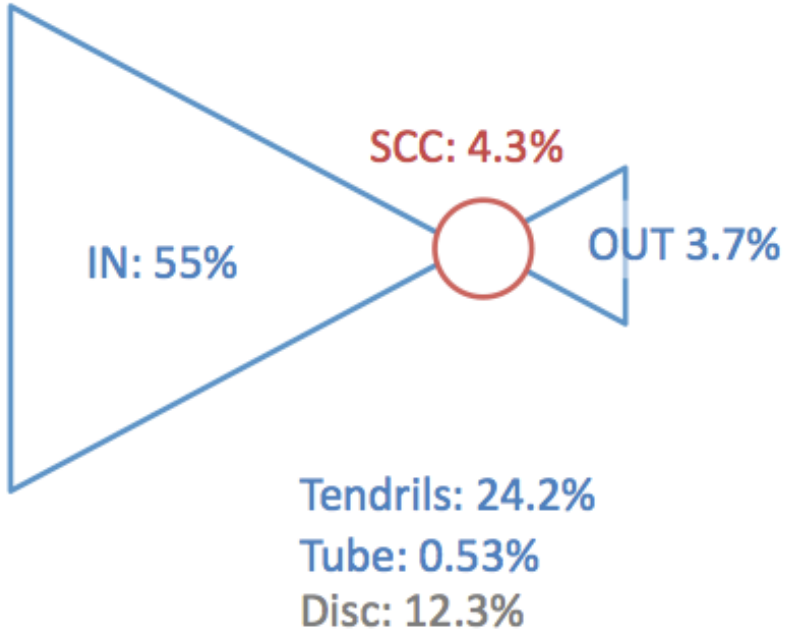
Jun Zhang, Mark S. Ackerman, and Lada Adamic. 2007. Expertise networks in online communities: structure and algorithms. In Proceedings of the 16th international conference on World Wide Web (WWW '07)., 221-230.

Event Media as "Skewed Bowtie"



Proportions of users by media types

~285K news, blog and social media users, authoring ~2M documents on the events in Jan-Feb 2011.



Moreover, 1% of the users, consisting of the reciprocal core of SCC, authored 43% of all documents.

Minkyoung Kim, Lexing Xie, Peter Christen, Event Diffusion Patterns in Social Media (2012) Intl. Conf. on Weblogs and Social Media (ICWSM), 8 pages, Dublin, Ireland, June 2012

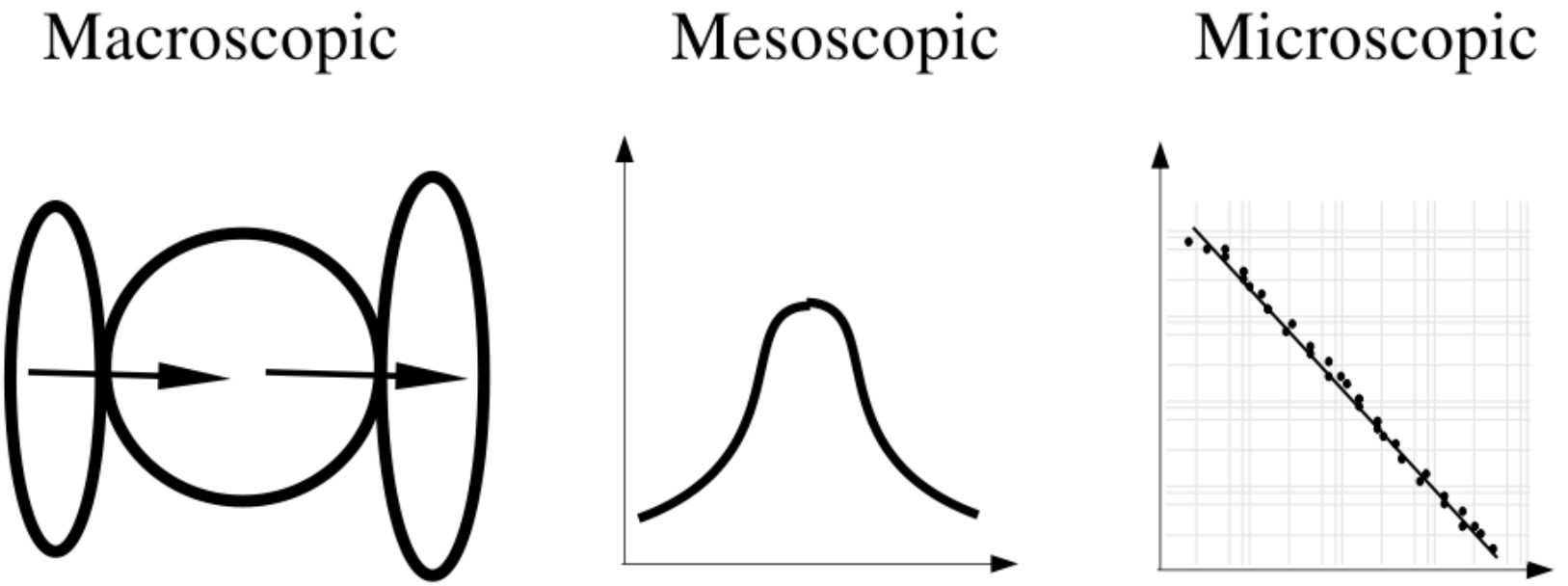


Figure 11.5: Levels of link-based analysis [92].

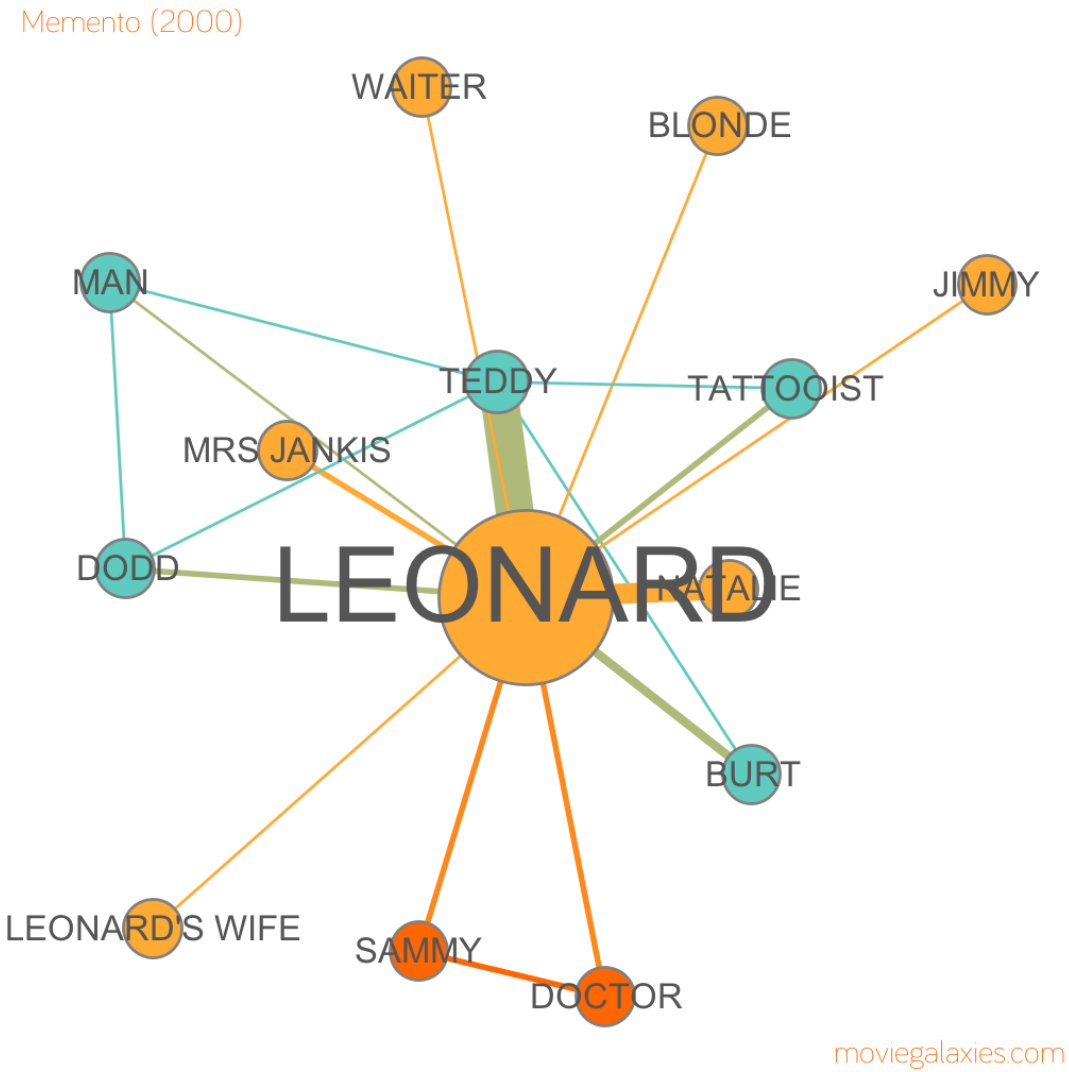
Modern Information Retrieval: The Concepts and Technology behind Search
Ricardo Baeza-Yates, Berthier Ribeiro-Neto
Addison-Wesley Professional; 2 edition (February 10, 2011)

Recap

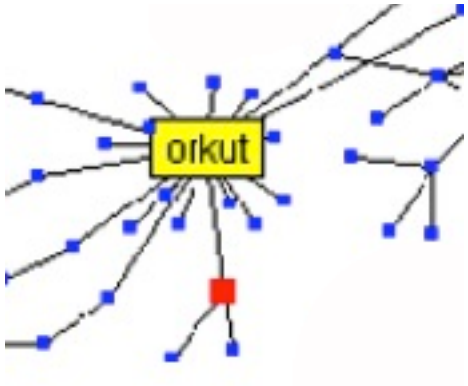
- Networks can be represented as matrices
- Useful network metrics:
  - degree and degree distribution
  - connected components
    - strong
    - weak
    - Giant

- After the break: distance and centrality

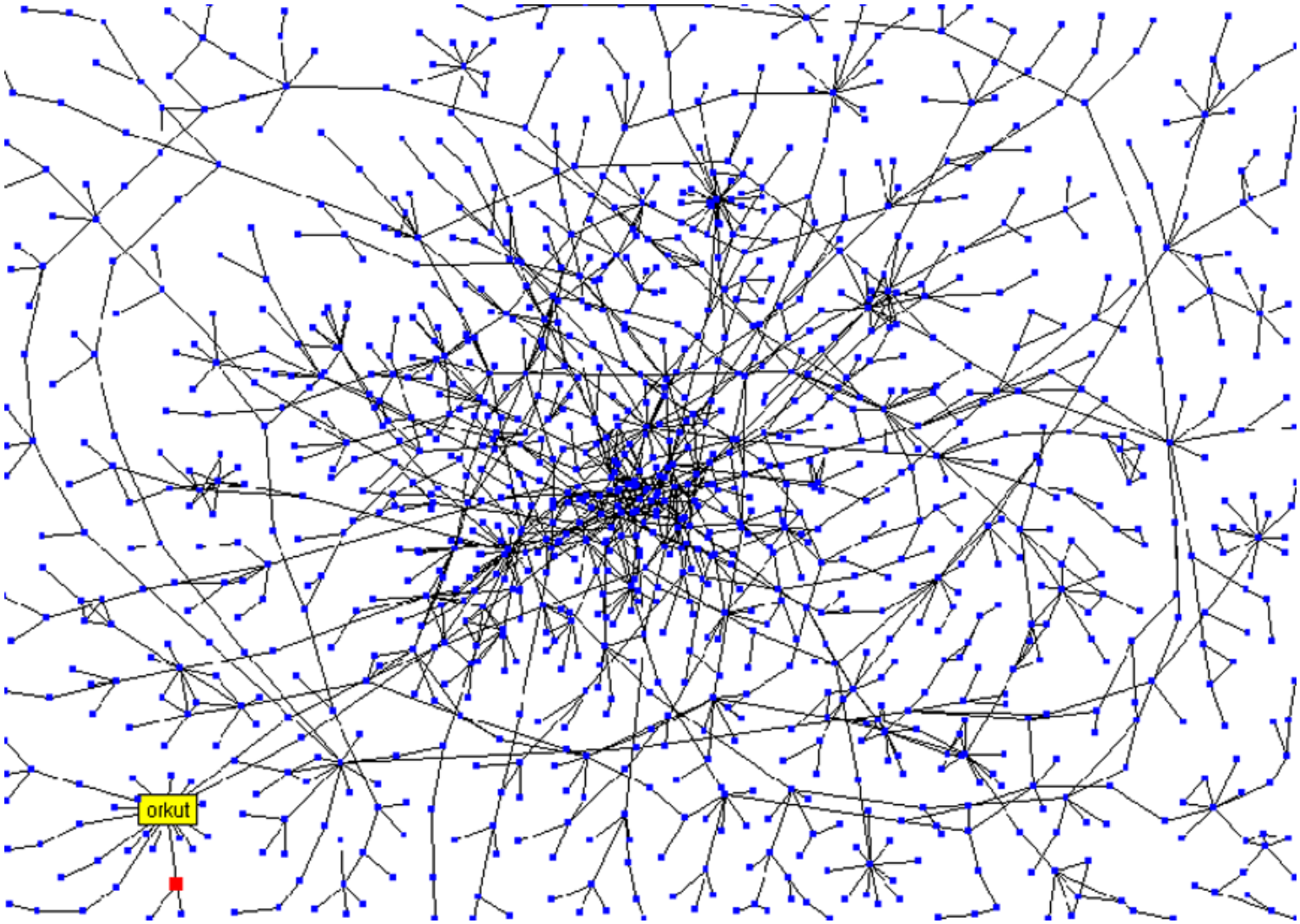
Who is the Center of a network?



is counting the edges enough?



Stanford Social Web (ca. 1999)

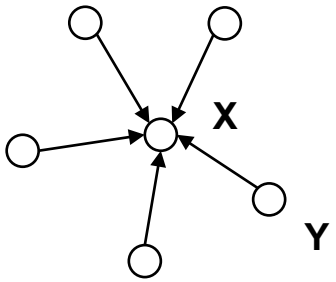


network of personal homepages at Stanford

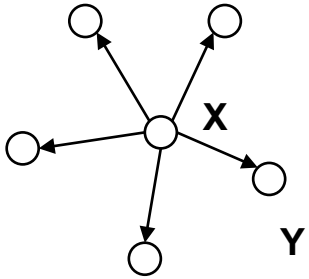
by Lada Adamic, U Michigan

different notions of centrality

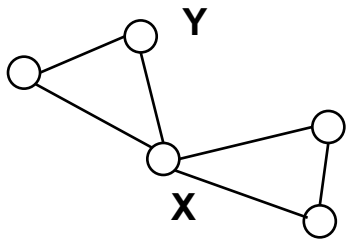
In each of the following networks, X has higher centrality than Y according to a particular measure



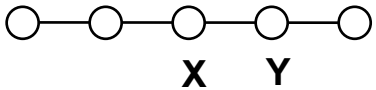
indegree



outdegree

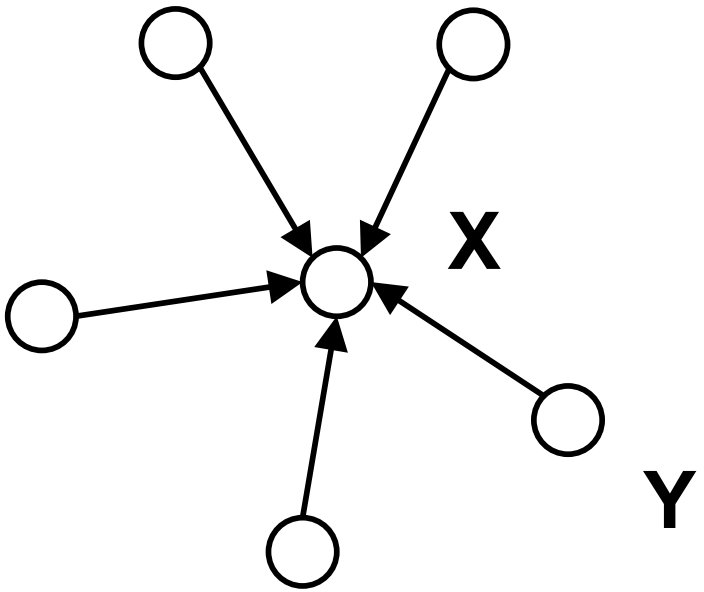


betweenness



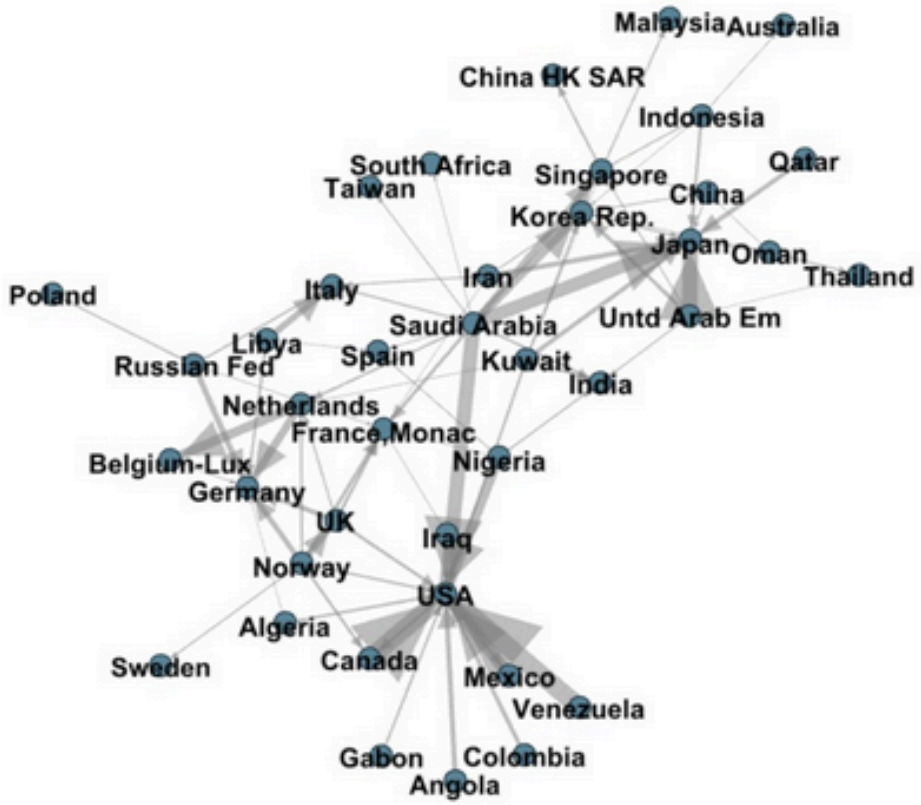
closeness

review: indegree



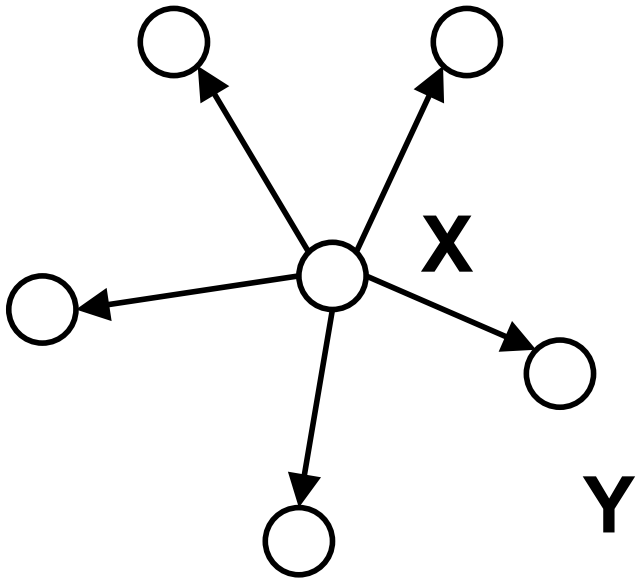
Which countries have high indegree (import petroleum and petroleum products from many others)

- [ ] Saudi Arabia
- [ ] Japan
- [ ] Iraq
- [ ] USA
- [ ] Venezuela



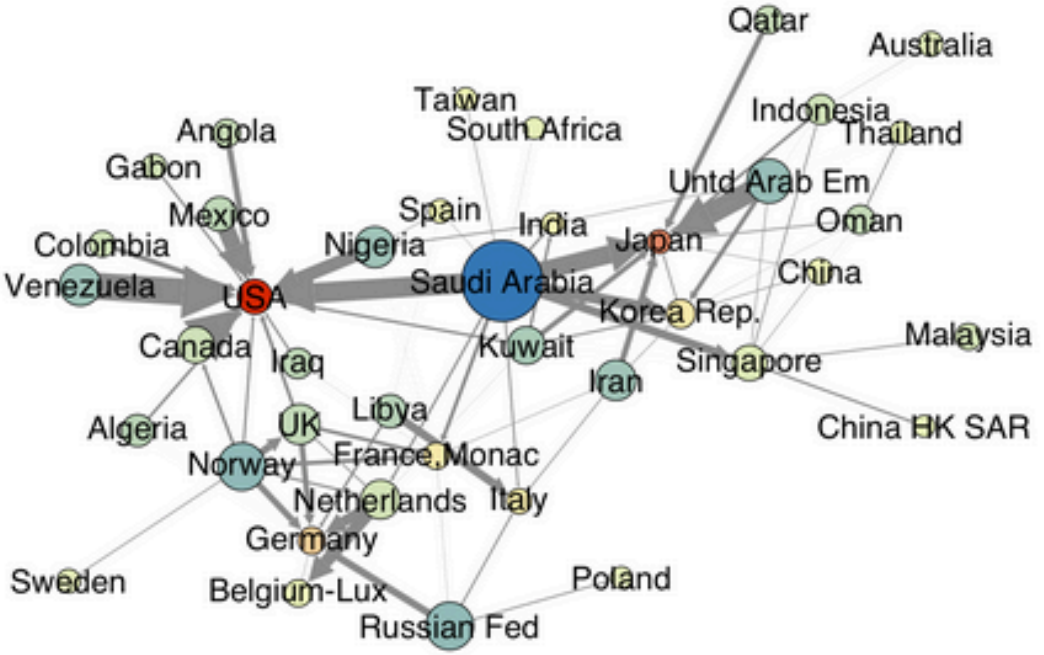
trade in petroleum and petroleum products, 1998, source: NBER-United Nations Trade Data

review: outdegree

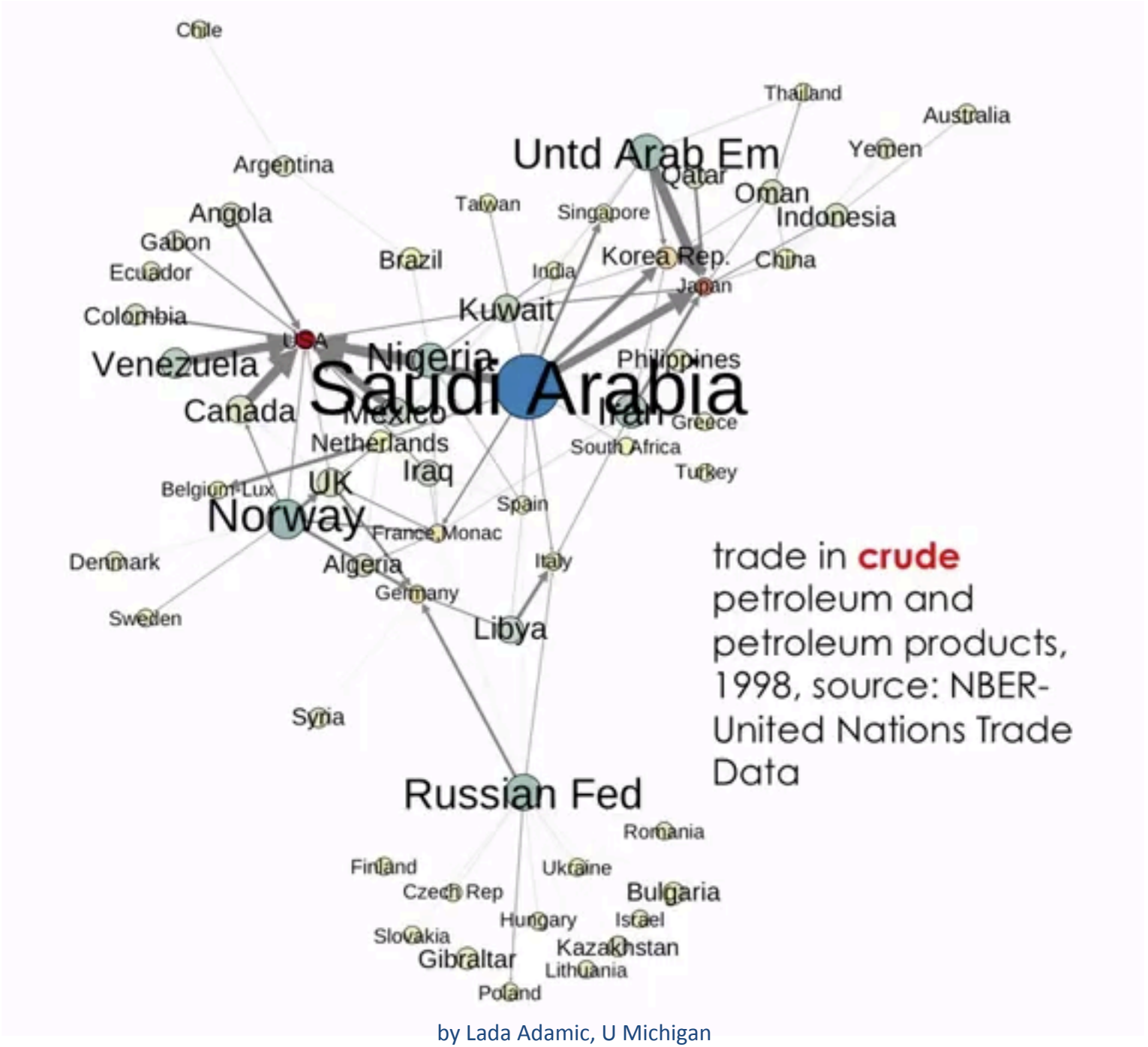


Which country has low outdegree but exports a significant quantity (thickness of the edges represents $$ value of export) of petroleum products

- [ ] Saudi Arabia
- [ ] Japan
- [ ] Iraq
- [ ] USA
- [ ] Venezuela



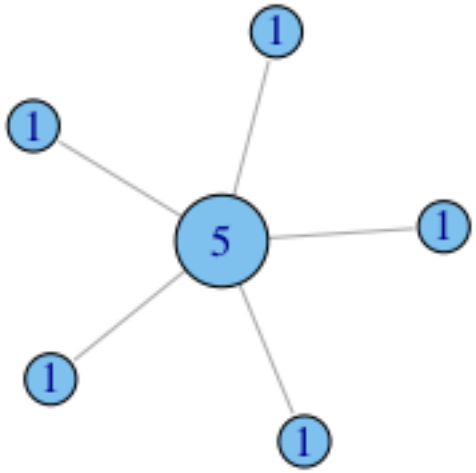
trade in petroleum and petroleum products, 1998, source: NBER-United Nations Trade Data



by Lada Adamic, U Michigan

putting numbers to it

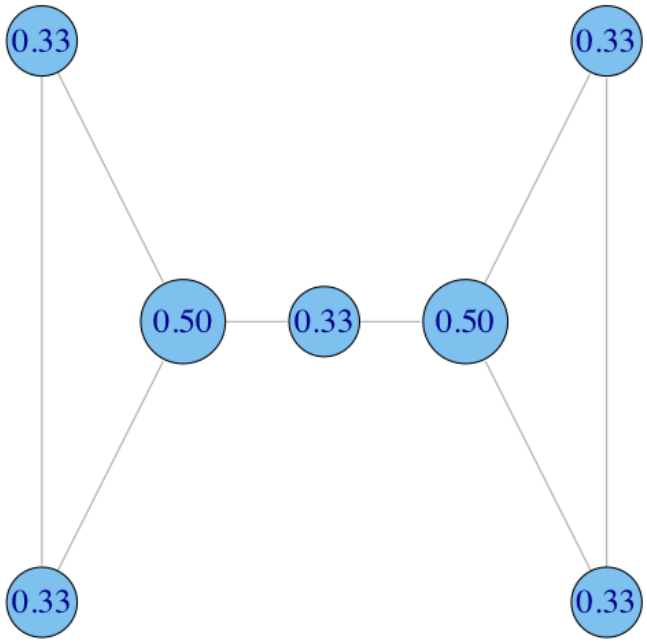
Undirected degree, e.g. nodes with more friends are more central.

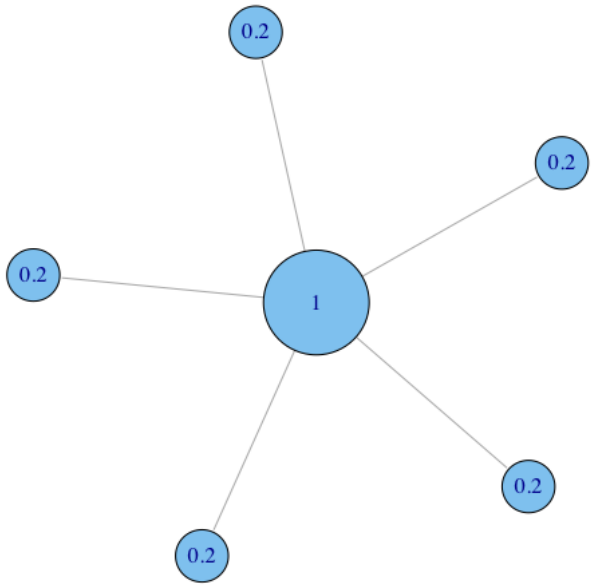


Assumption: the connections that your friend has don't matter, it is what they can do directly that does (e.g. go have a beer with you, help you build a deck...)

normalization

divide degree by the max. possible, i.e. (N-1)





centralization: skew in distribution

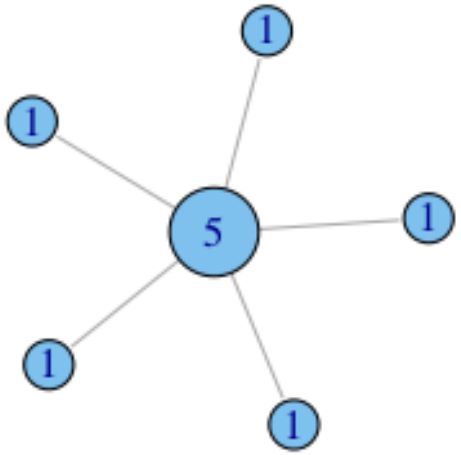
How much variation is there in the centrality scores among the nodes?

Freeman's general formula for centralization (can use other metrics, e.g. gini coefficient or standard deviation):

maximum value in the network

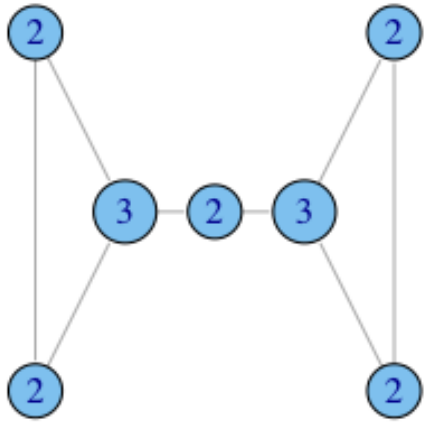
$$C_D = \frac{\sum_{i=1}^{g} [C_D(n^*) - C_D(i)]}{[(N-1)(N-2)]}$$

degree centralization examples



$C_D = 1.0$

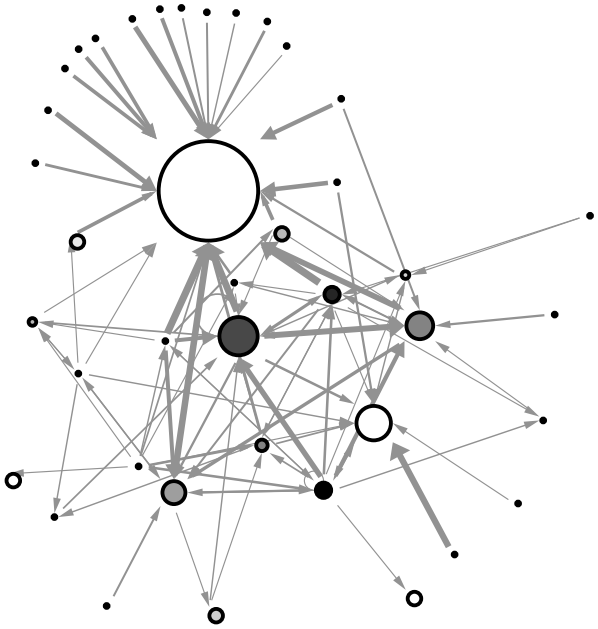
$C_D = 0.167$



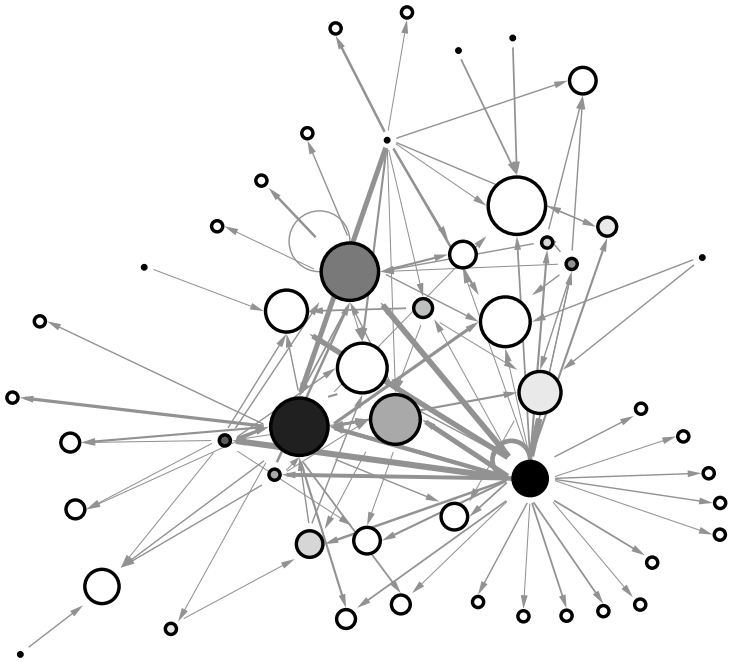
$C_D = 0.167$

real-world examples

example financial trading networks



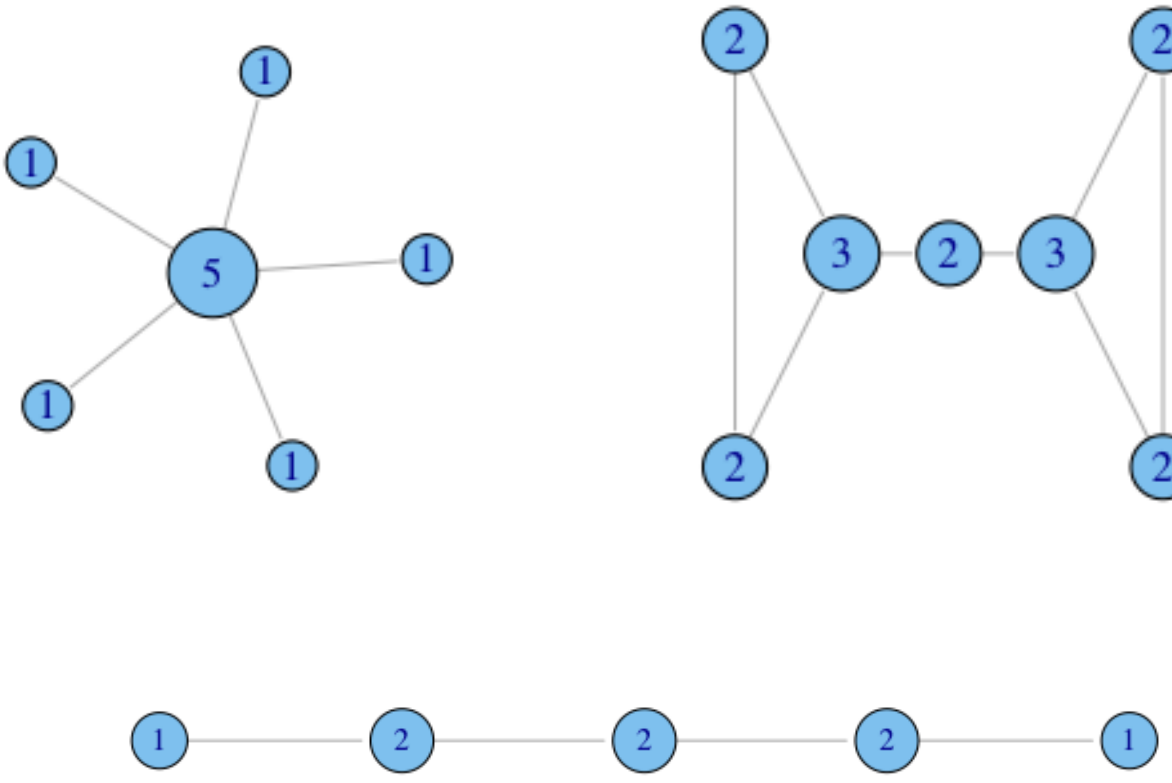
high in-centralization:
one node buying from many others



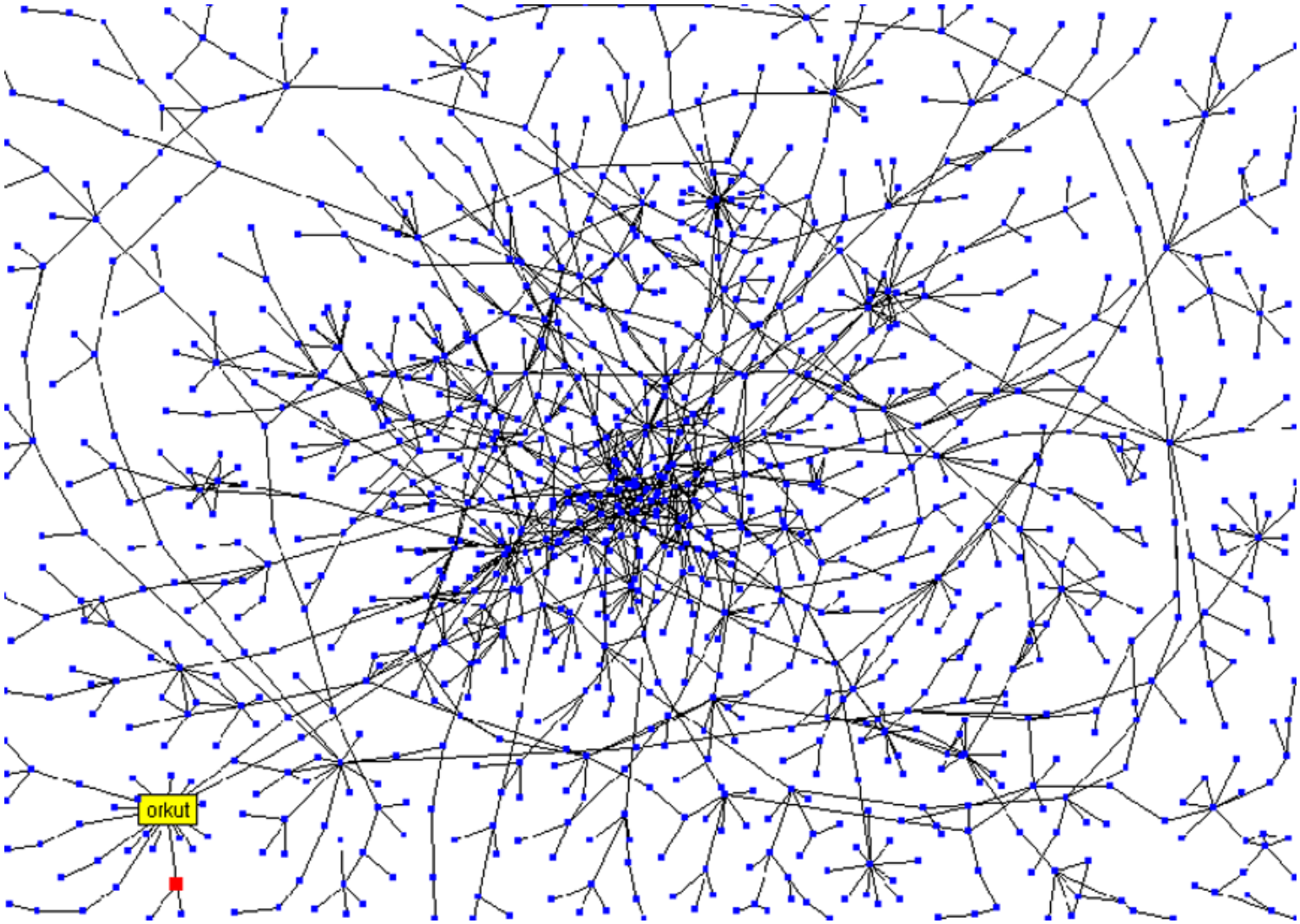
low in-centralization:
buying is more evenly distributed

what does degree not capture?

In what ways does degree fail to capture centrality in the following graphs?

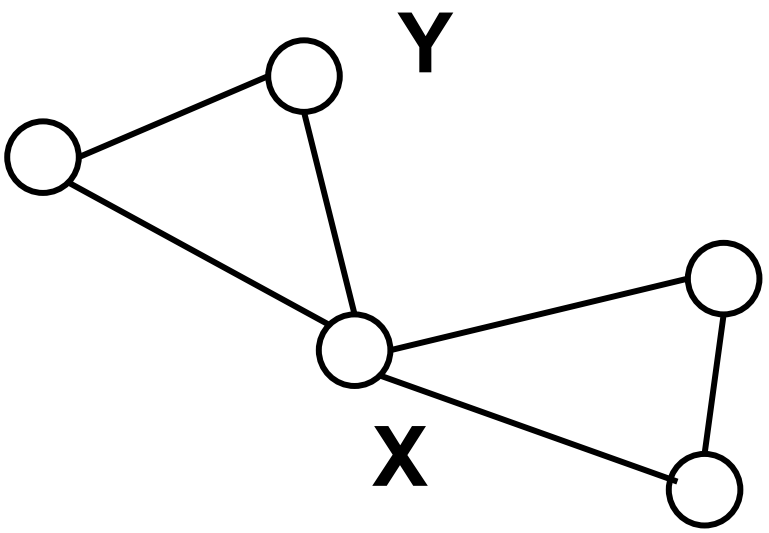


Stanford Social Web (ca. 1999)



network of personal homepages at Stanford

Brokerage not captured by degree

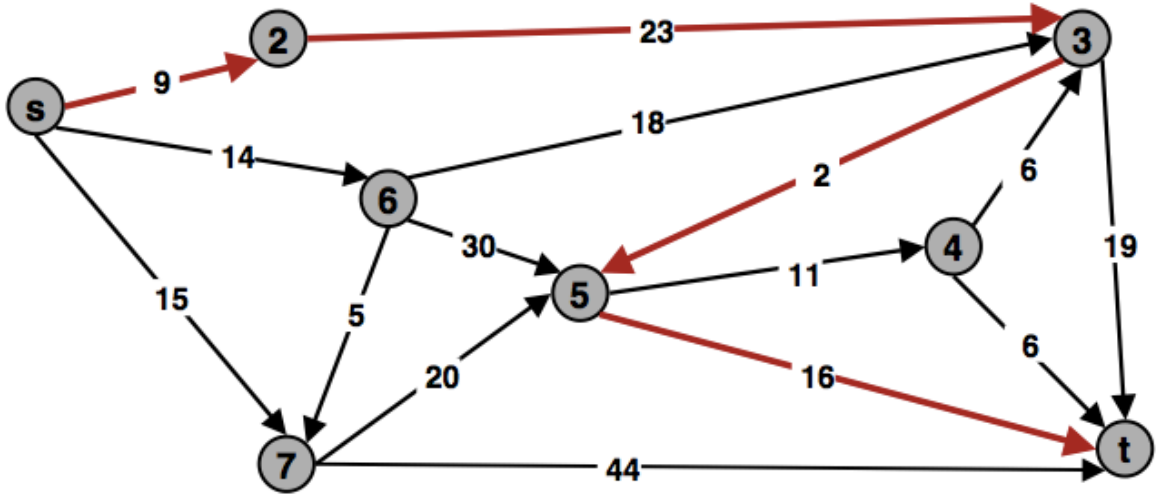


Review: shortest path in a network

Shortest path network: (V, E, s, t, c) .

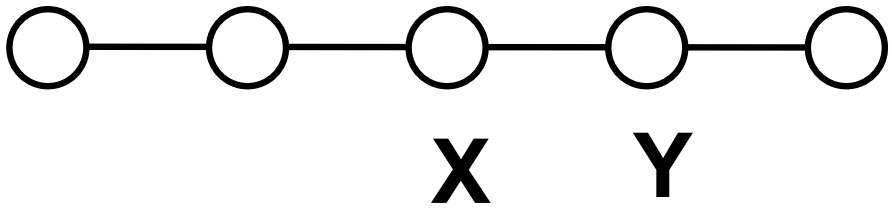
- **Directed graph (V, E).**
- **Source s ∈ V, sink t ∈ V.**
- **Arc costs c(v, w).**
- **Cost of path = sum of arc costs in path.**

Cost of path s - 2 - 3 - 5 - t
= 9 + 23 + 2 + 16
= 48.



betweenness: capturing brokerage

- intuition: how many pairs of individuals would have to go through you in order to reach one another in the minimum number of hops?



betweenness: definition

$$C_B(i) = \sum_{j<k} g_{jk}(i) / g_{jk}$$

Where g_{jk} = the number of shortest paths connecting *jk*
$g_{jk}(i)$ = the number that actor *i* is on.

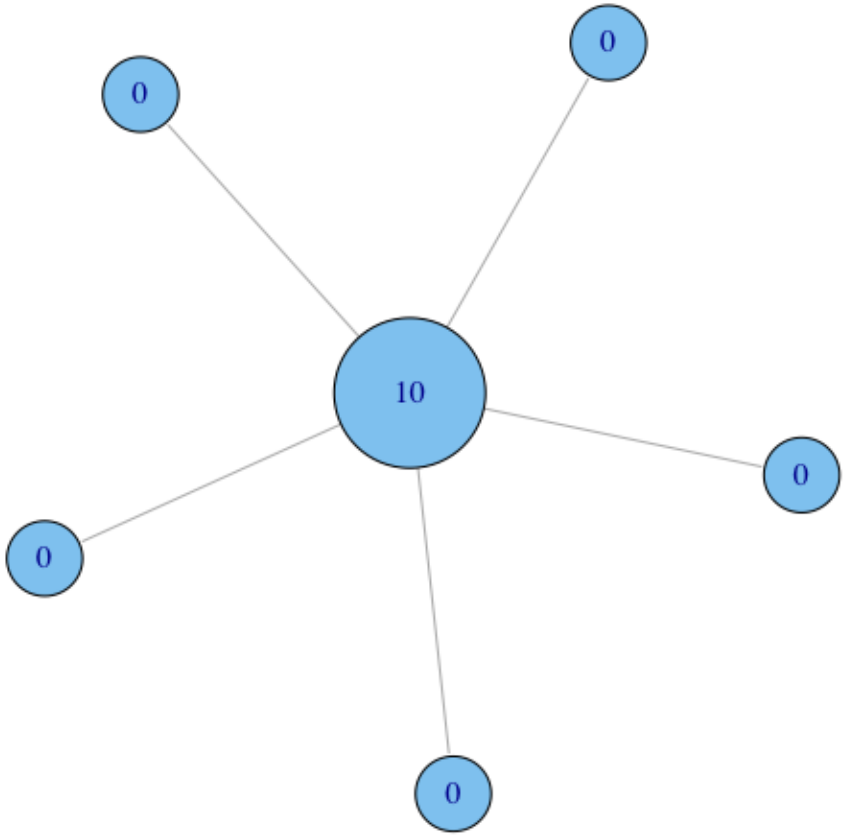
Usually normalized by:

$$C'_B(i) = C_B(i\ ) / [(n-1)(n-2)/2]$$

number of pairs of vertices excluding the vertex itself

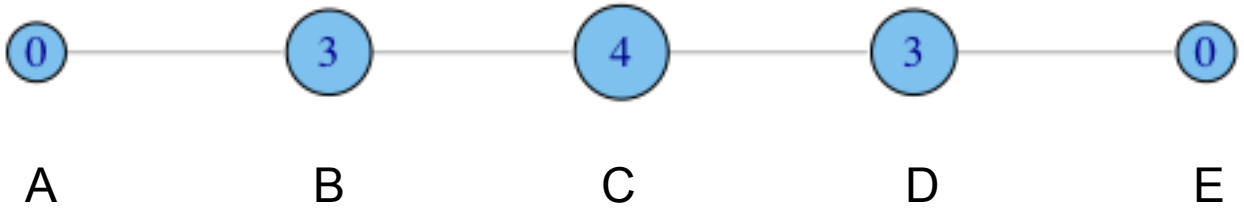
betweeness on toy network

- non-normalized version:



betweeness on toy networks

- non-normalized version:

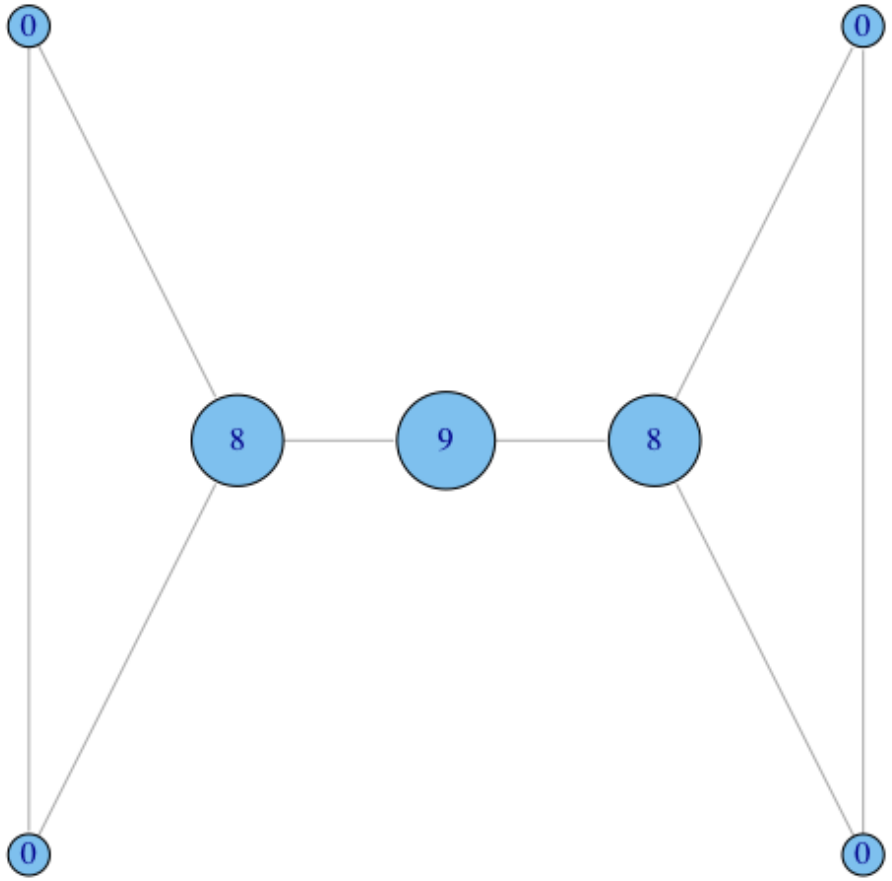


- A lies between no two other vertices
- B lies between A and 3 other vertices: C, D, and E
- C lies between 4 pairs of vertices (A,D),(A,E),(B,D),(B,E)

- note that there are no alternate paths for these pairs to take, so C gets full credit

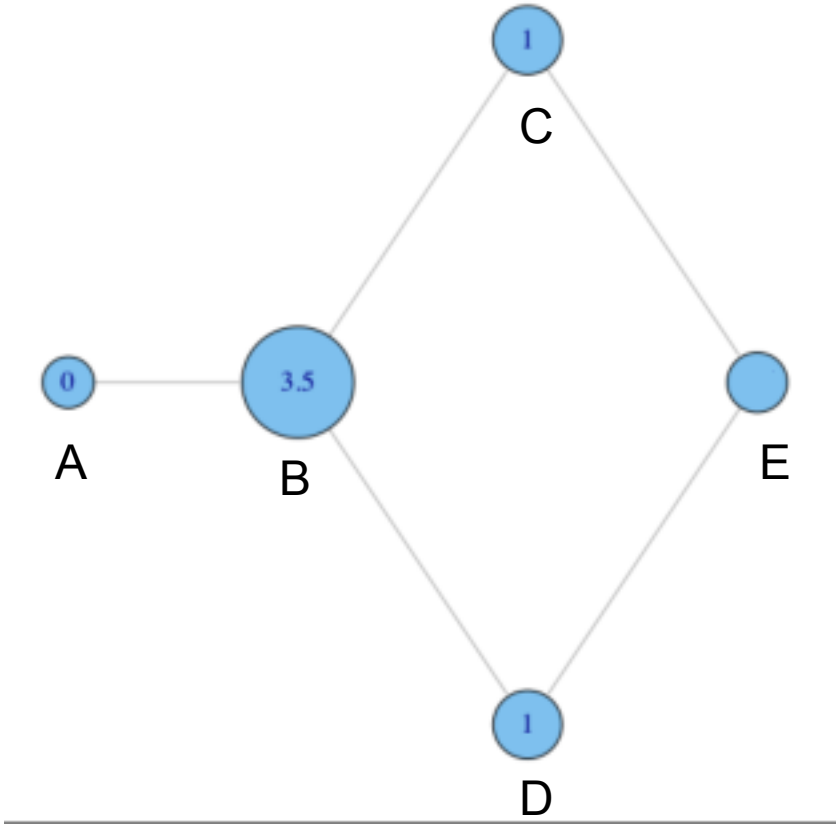
betweenness on toy networks

- non-normalized version:



betweenness on toy networks

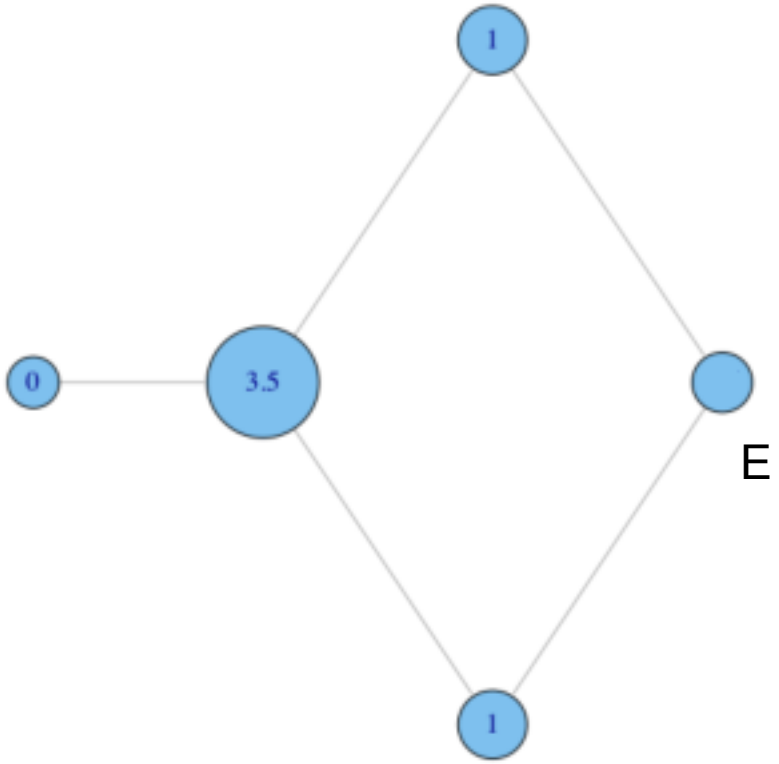
- non-normalized version:



- why do C and D each have betweenness 1?
- They are both on shortest paths for pairs (A,E), and (B,E), and so must share credit:
  - ½+½ = 1

Quiz Question

- What is the betweenness of node E?



a) 0.5

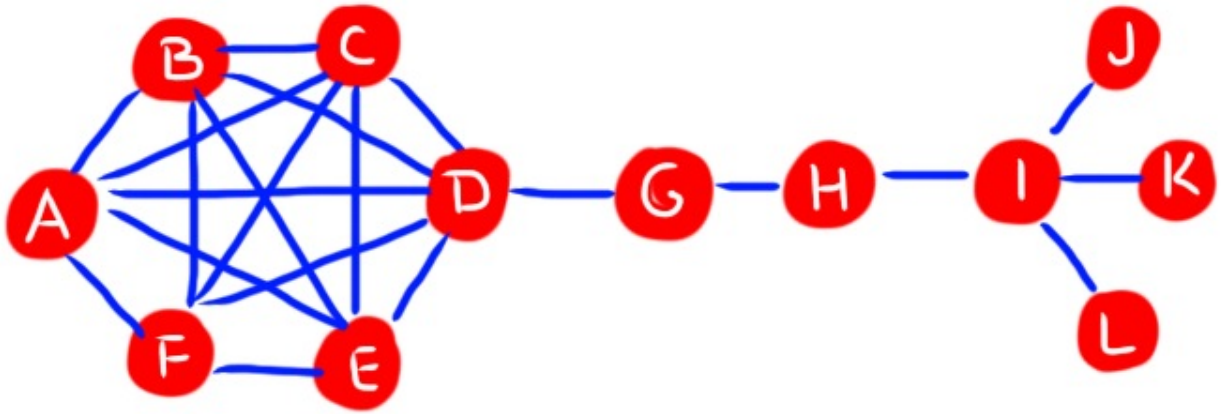
b) 1

c) 1.5

d) 2

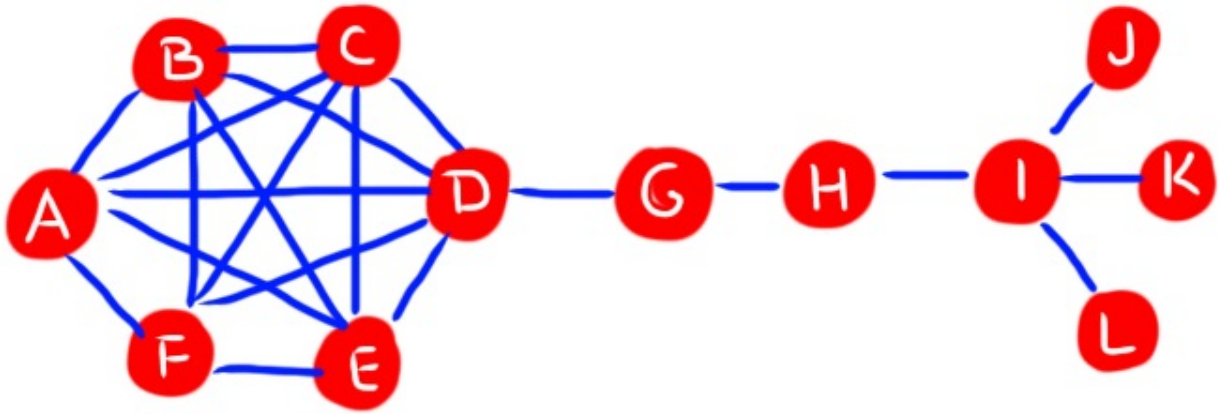
Quiz Q:

- Find a node that has high betweenness but low degree



Quiz Q:

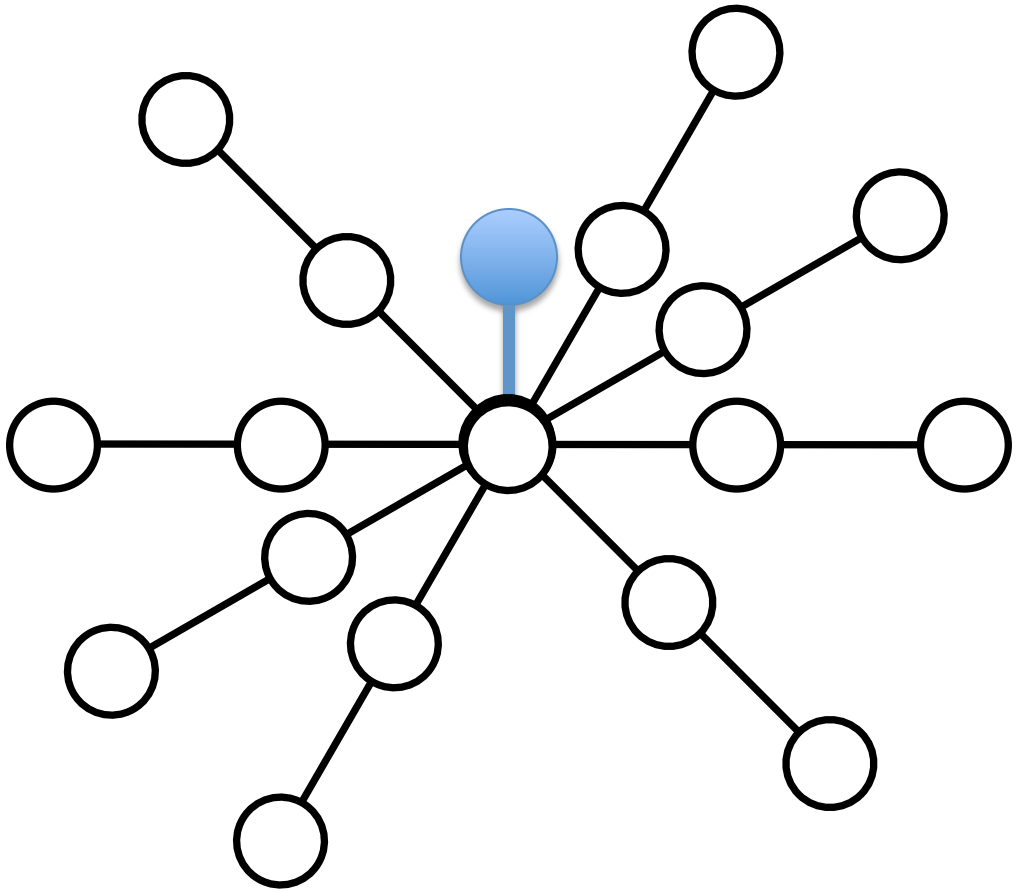
- Find a node that has low betweenness but high degree



closeness

- What if it's not so important to have many direct friends?
- Or be "between" others
- But one still wants to be in the "middle" of things, not too far from the center

need not be in a brokerage position



closeness: definition

Closeness is based on the length of the average shortest path between a node and all other nodes in the network

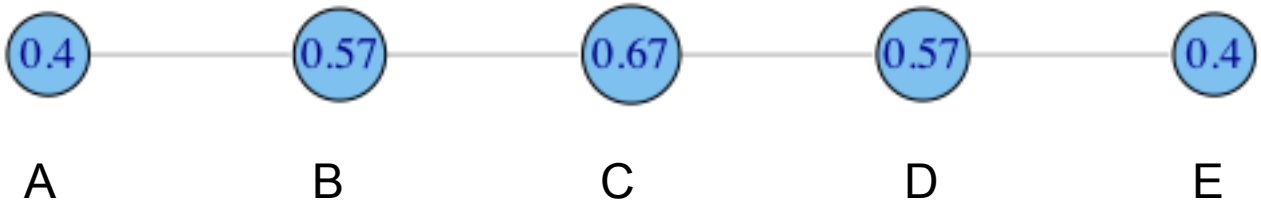
Closeness Centrality:

$$C_c(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$$

Normalized Closeness Centrality

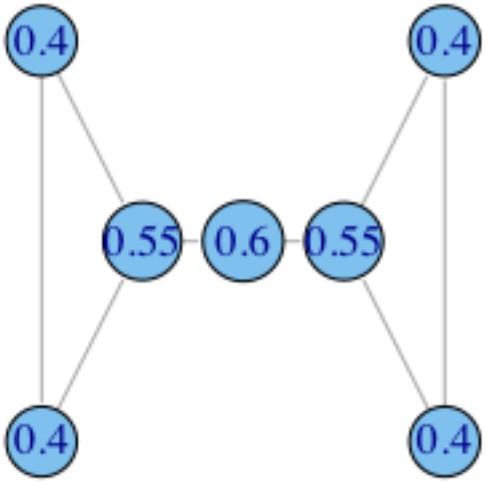
$$C_C^{'}(i) = (C_C(i)) * (N-1)$$

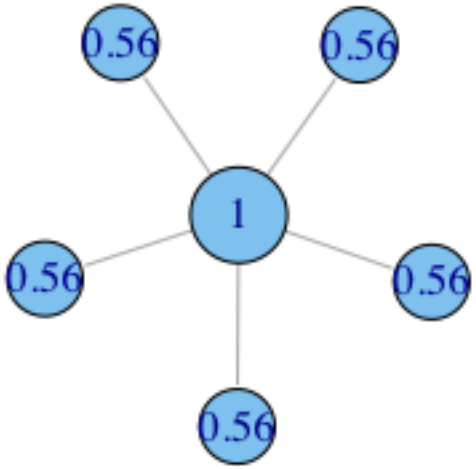
closeness: toy example



$$C_c^{'}(A) = \left[\frac{\sum_{j=1}^{N} d(A,j)}{N-1} \right]^{-1} = \left[\frac{1+2+3+4}{4} \right]^{-1} = \left[\frac{10}{4} \right]^{-1} = 0.4$$

closeness: more toy examples





Quiz Q:

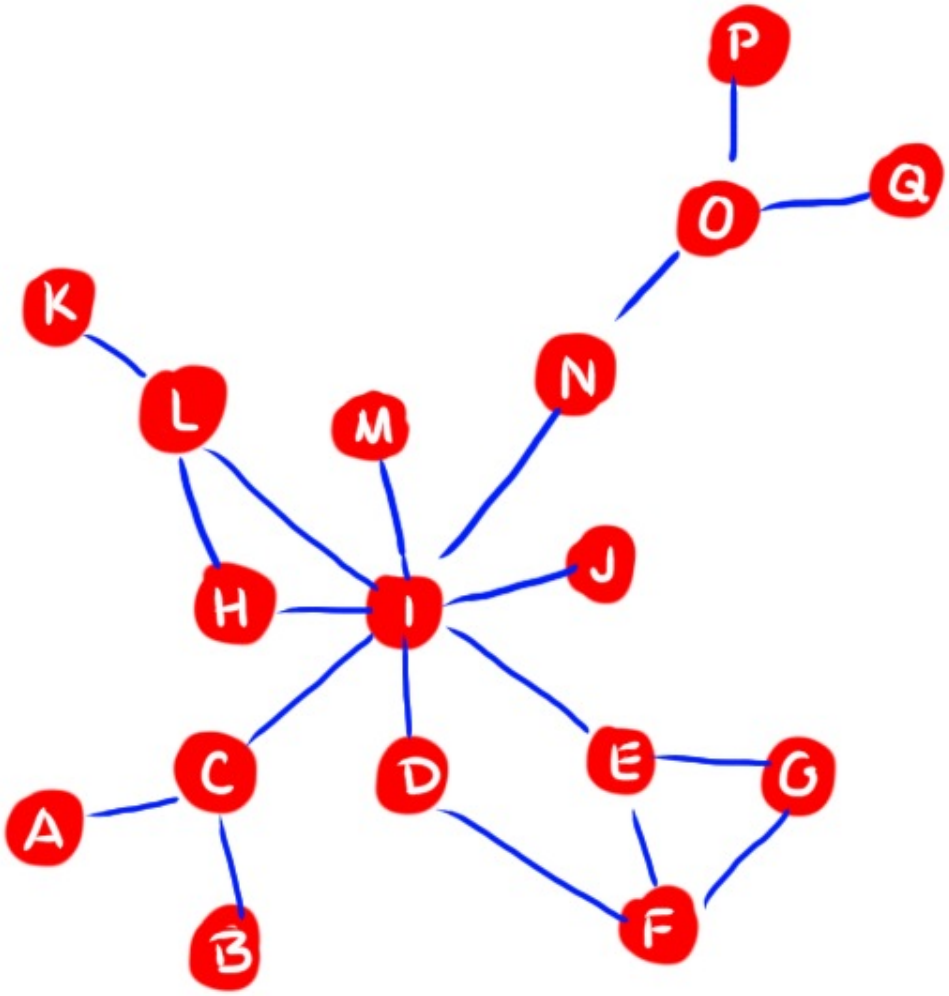
Which node has relatively high degree but low closeness?

a) I

b) J

c) E

d) O



What else can shortest-path be used for?

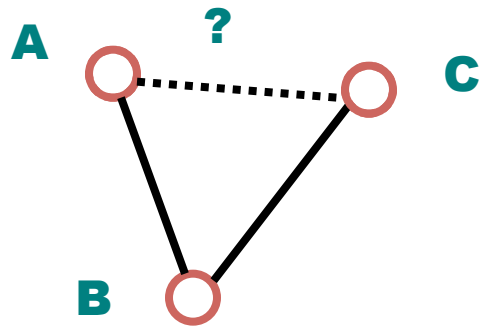
- What is the radius of a network?
- Define the diameter of a network?

- … you will see this in the lab session this afternoon.

Transitivity, triadic closure, clustering

- Transitivity:
  - if A is connected to B and B is connected to C
    what is the probability that A is connected to C?

  - my friends' friends are likely to be my friends



Clustering

- Global clustering coefficient

  3 x number of triangles in the graph

  number of connected triples of vertices

$$C = \frac{\text{3 x number of triangles in the graph}}{\text{number of connected triples}}$$

Question: How long will be a naïve algorithm take to compute clustering coefficient?
O(n), O(nlogn), O(n^2), O(n^3), … ?

Local clustering coefficient (Watts&Strogatz 1998)

- For a vertex i
  - The fraction pairs of neighbors of the node that are themselves connected
  - Let n_i be the number of neighbors of vertex *i*

$$C_i = \frac{\text{\# of connections between i's neighbors}}{\text{max \# of possible connections between i's neighbors}}$$

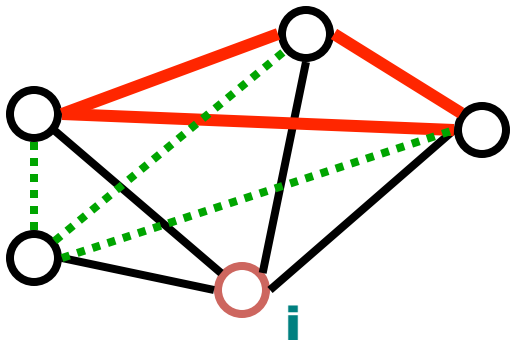
$$C_{i\ \text{directed}} = \frac{\text{\# directed connections between i's neighbors}}{n_i * (n_i - 1)}$$

$$C_{i\ \text{undirected}} = \frac{\text{\# undirected connections between i's neighbors}}{n_i * (n_i - 1)/2}$$

Local clustering coefficient (Watts&Strogatz 1998)

- Average over all n vertices

$$C = \frac{1}{n}\sum_i C_i$$



link present

link absent

$n_i = 4$

max number of connections:

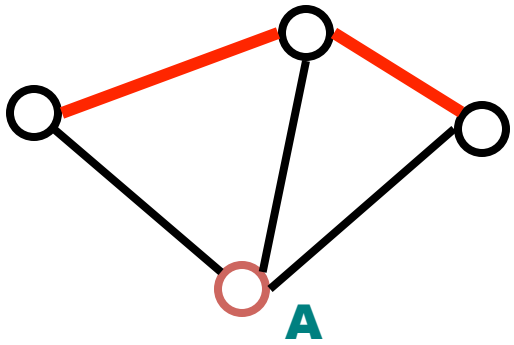
4*3/2 = 6

3 connections present

$C_i = 3/6 = 0.5$

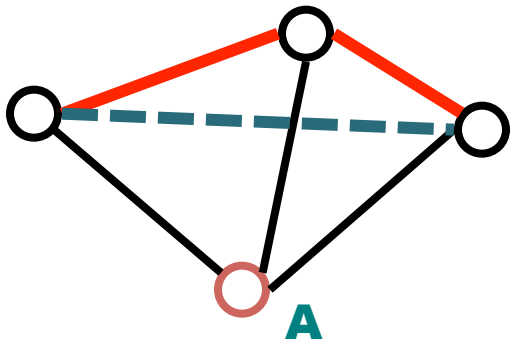
Quiz Q:

- The clustering coefficient for vertex A is:



Explanation

- $n_i = 3$
- there are 2 connections present out of max of 3 possible
- $C_i = 2/3$

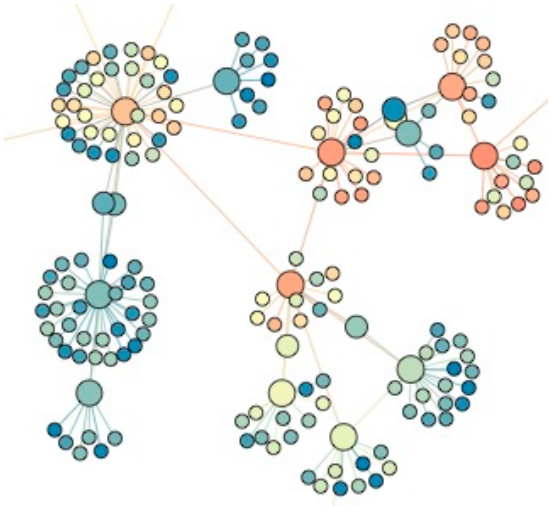


Network Description: so far

- Representing a network as a graph

- Connected components: strong, weak, …
- Centrality: degree, betweeness, closeness, … (and many more)
- Clustering coefficient and triadic closure

- Up next: Is the network composed of communities?



Finding Communities

Lexing Xie
Research School of Computer Science

Lecture slides credit: Lada Adamic, Univ. Michigan
Jure Leskovec, Stanford University

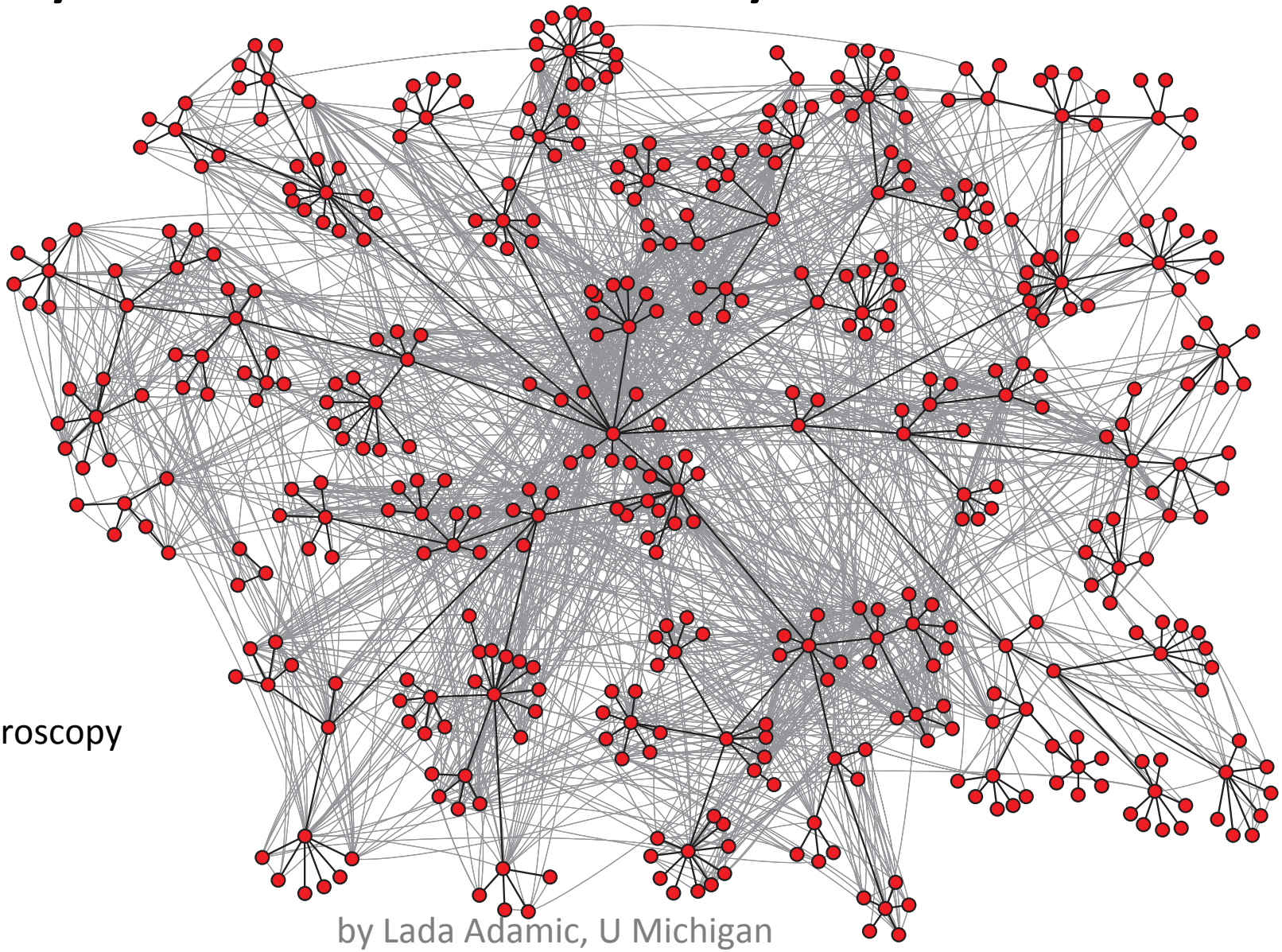
Outline

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

Why do it?

- Discover communities of practice
- Measure isolation of groups
- Understand opinion dynamics / adoption

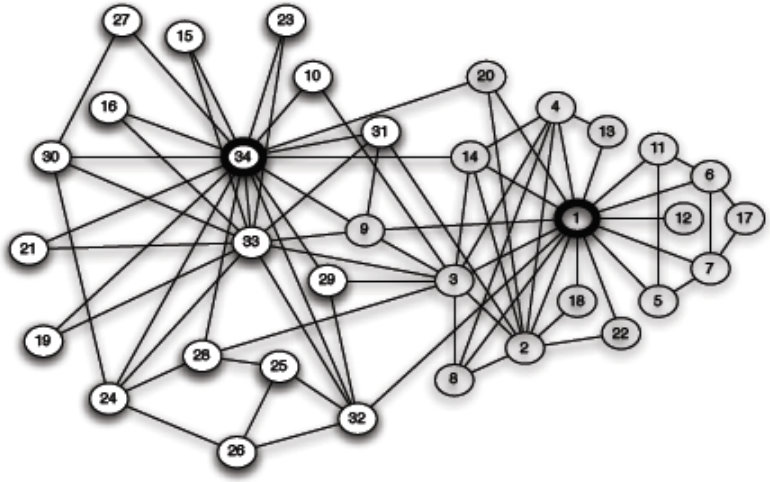
Why look for community structure?



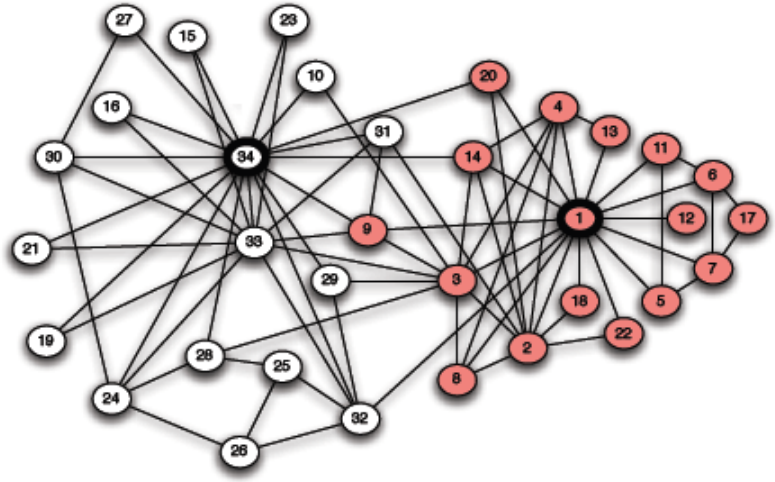
example:
email spectroscopy

by Lada Adamic, U Michigan

Zachary Karate Club



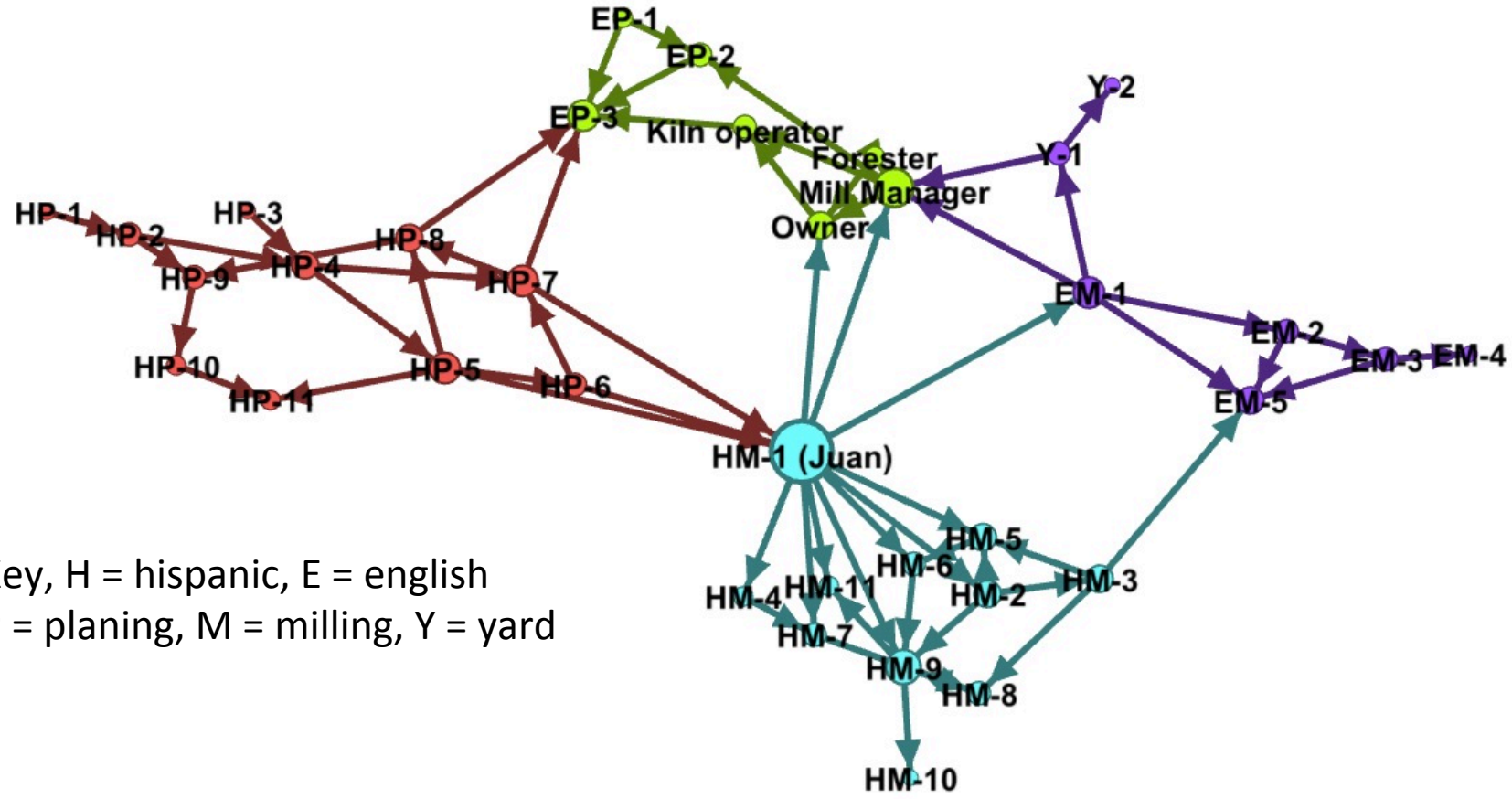
(a) *Karate club network*



(b) *After a split into two clubs*

source:Easley/Kleinberg

Why look for community structure?



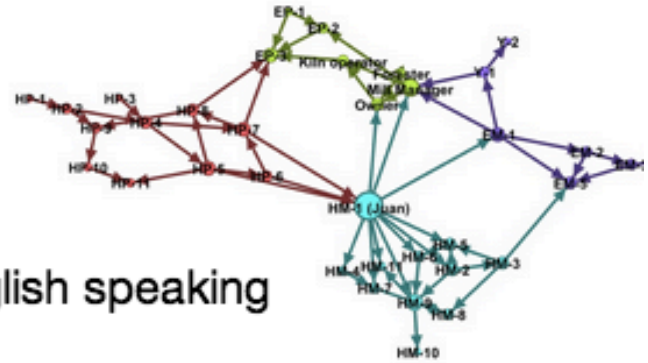
Key, H = hispanic, E = english
P = planing, M = milling, Y = yard

Sawmill network: source Exploratory Social Network Analysis with Pajek
by Lada Adamic, U Michigan

Quiz Q:

- The management at the sawmill was having difficulty persuading the workers to adopt a new plan, even though everyone would benefit. In particular the Hispanic workers (H) were reluctant to agree. The management called in a sociologist who mapped out who talked to whom regularly. Then they suggested that the management talk to Juan and have him talk to the Hispanic workers. It was a success, promptly everyone was on board with the new plan. Why?

Why did getting Juan on board with the plan help resolve the conflict?



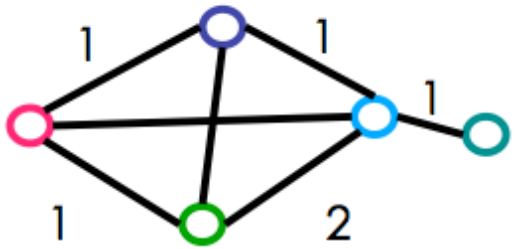
- [ ] Juan is a broker between the Spanish and English speaking communities.
- [ ] Strong community structure can impede information flow and enable opinions to stay rooted within groups.
- [ ] Juan has more social ties with the workers than the management does.
- [ ] Juan's ego network has high constraint.

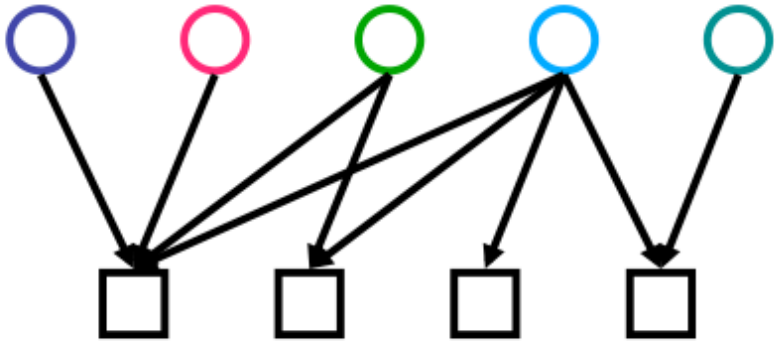
What makes a community?

- 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

Affiliation Networks

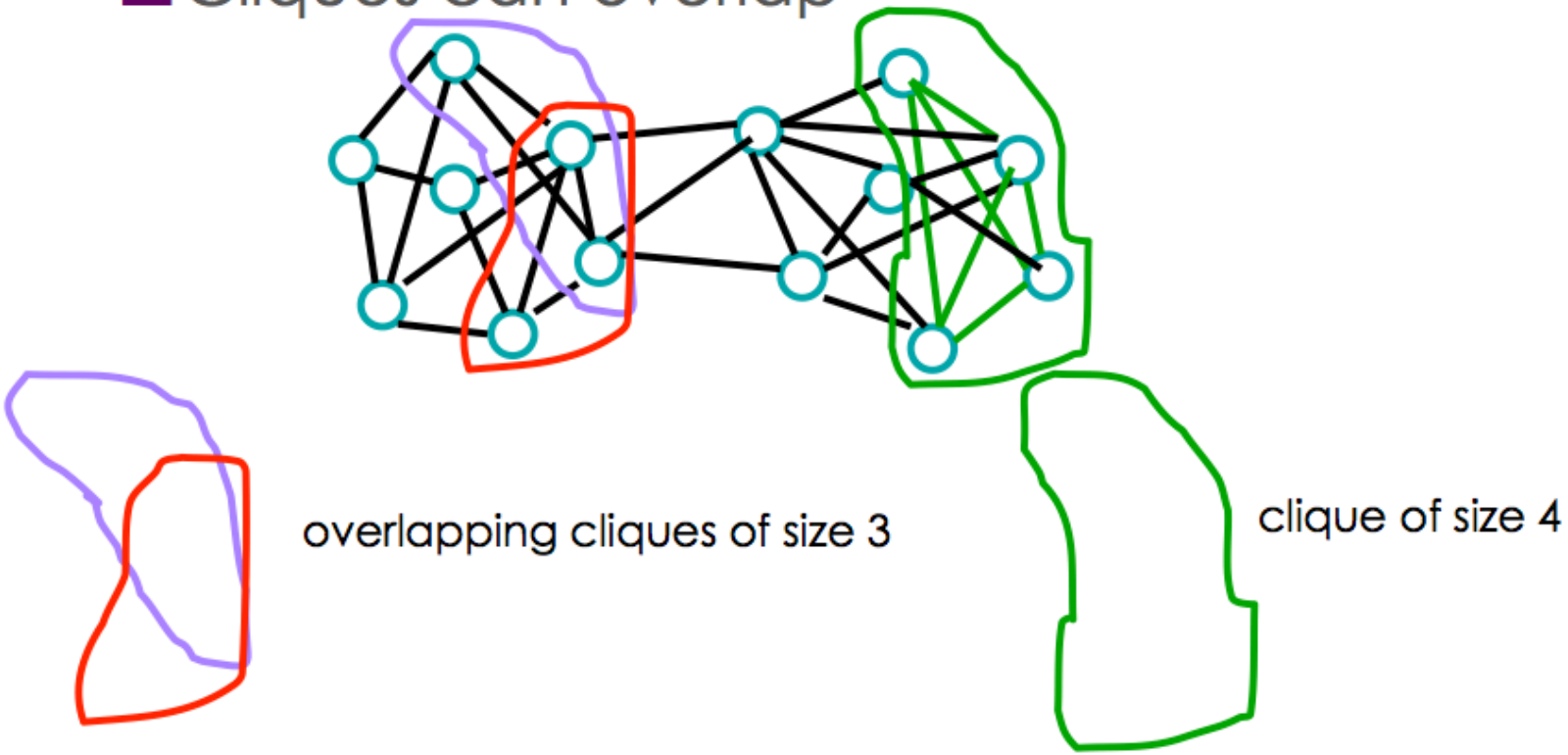
- otherwise known as
  - membership network
    - e.g. board of directors
  - hypernetwork or hypergraph
  - bipartite graphs
  - interlocks





Cliques

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

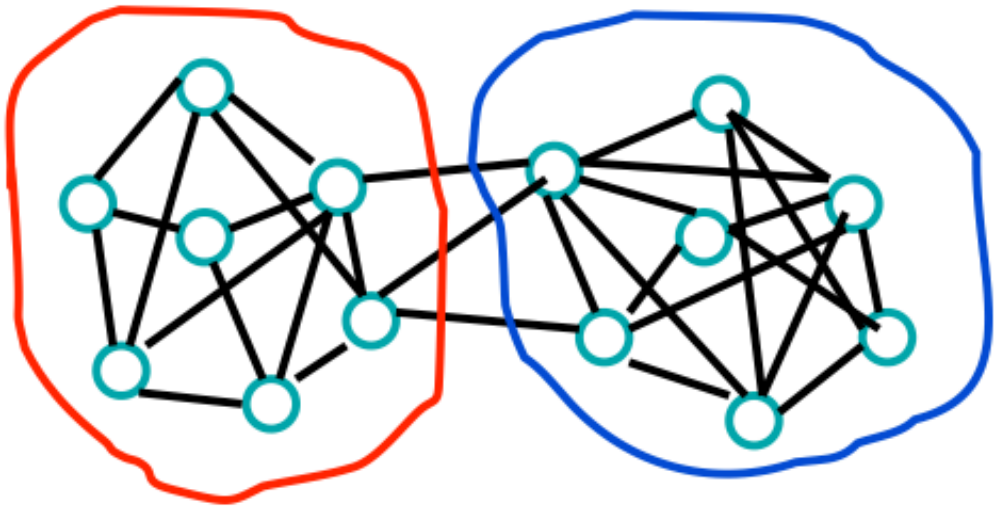


Cliques

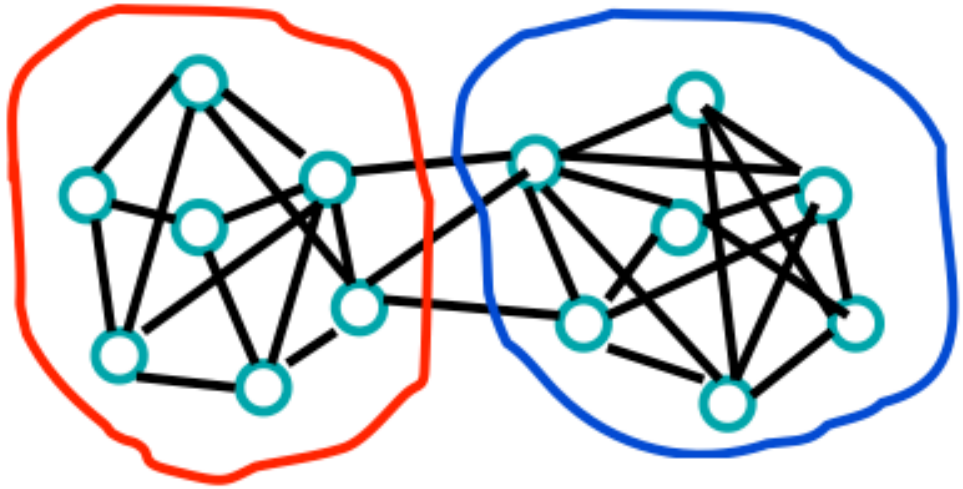
- Not robust
  - one missing link can disqualify a clique
- Not interesting
  - everybody is connected to everybody else
  - no core-periphery structure
  - no centrality measures apply
- How cliques overlap can be more interesting than that they exist

k-cores: similar idea, less stringent

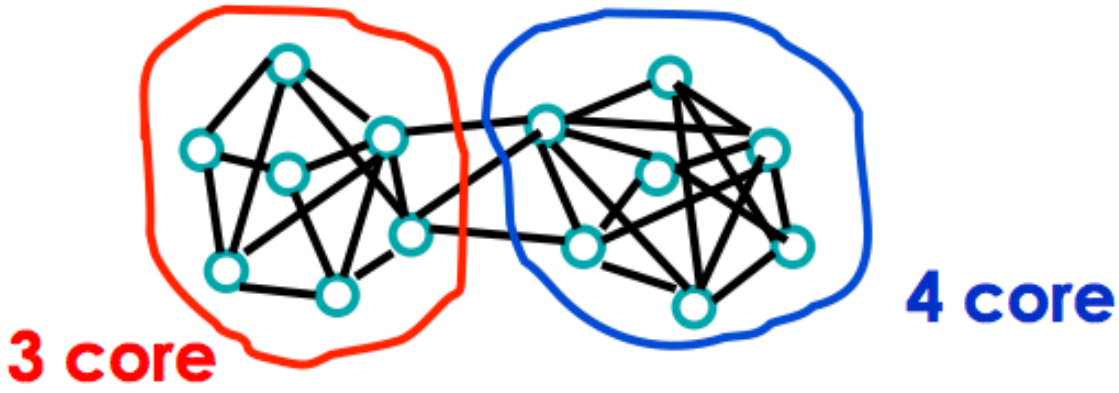
- Each node within a group is connected to k other nodes in the group



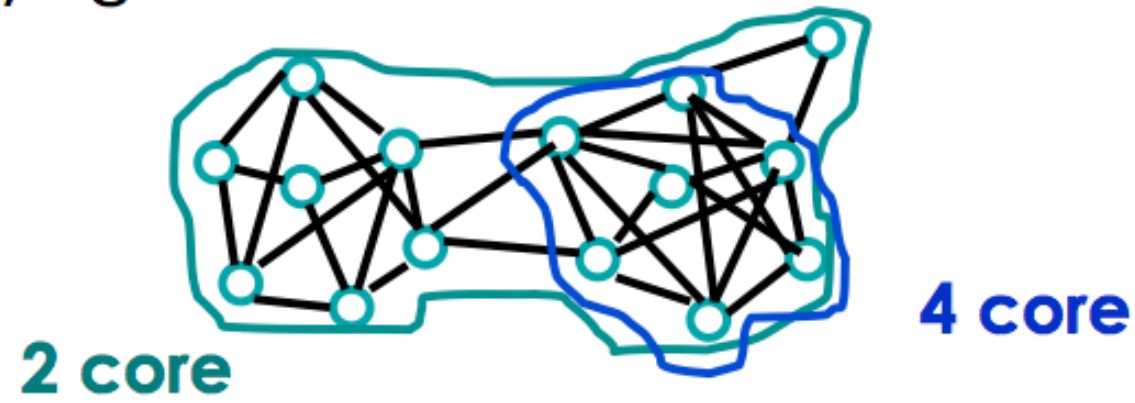
- What is the “k” for the core circled in red?
- What is the “k” for the core circled in blue?



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

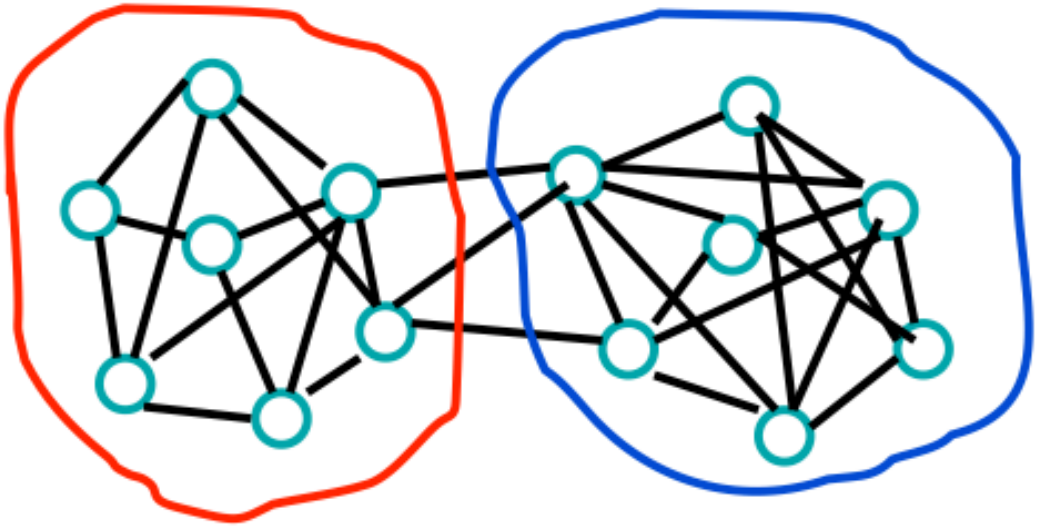


- but even this is too stringent of a requirement for identifying natural communities



Use reachability and diameter?

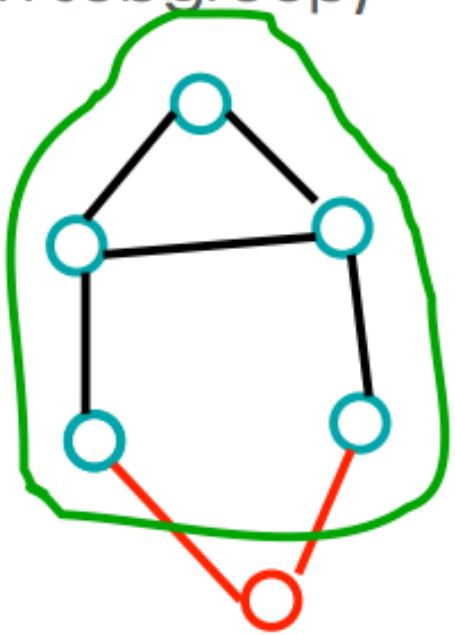
- n – cliques
  - maximal distance between any two nodes in subgroup is n



2-cliques

- theoretical justification
  - information flow through intermediaries

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



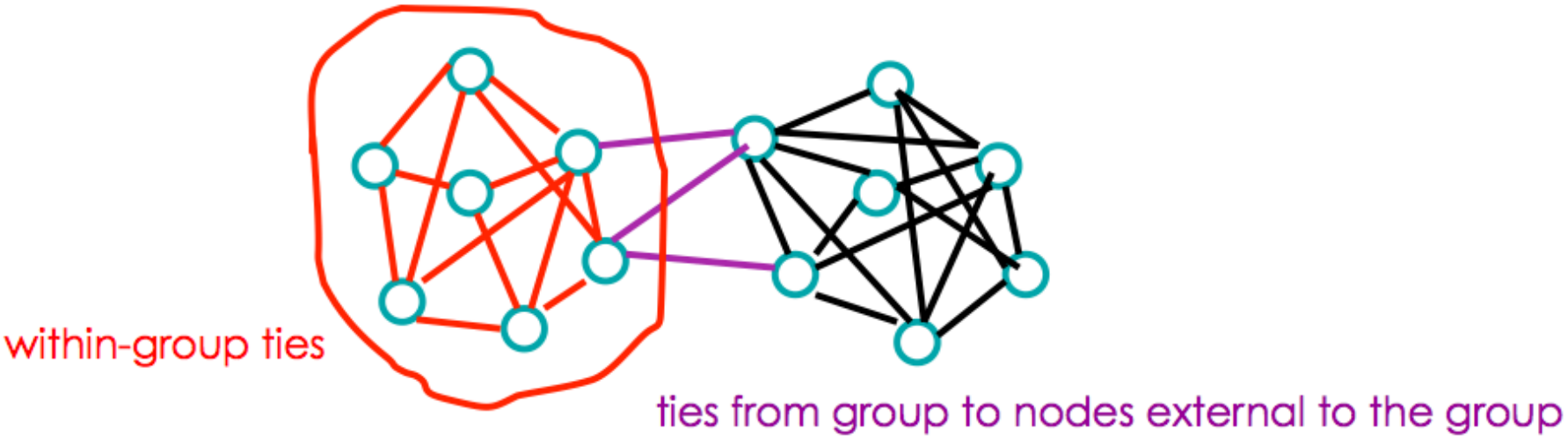
2 – clique
diameter = 3

path outside the 2-clique

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

p-clique: fraction 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.

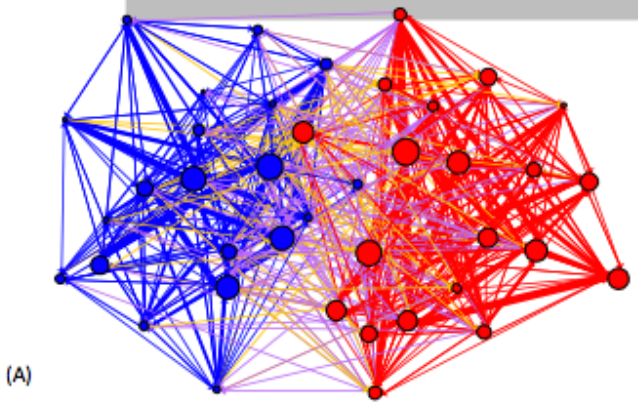


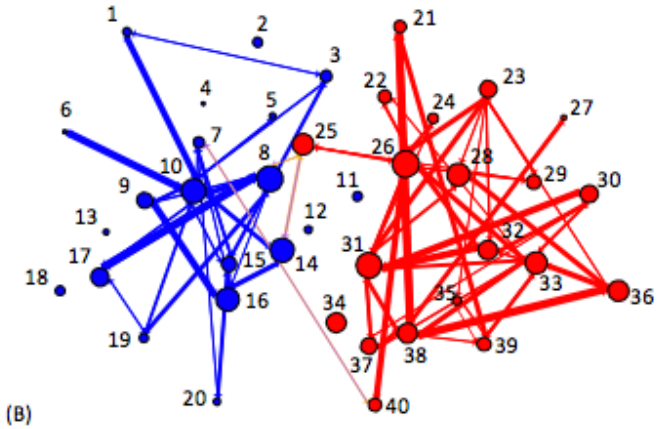
cohesion in directed & weighted networks

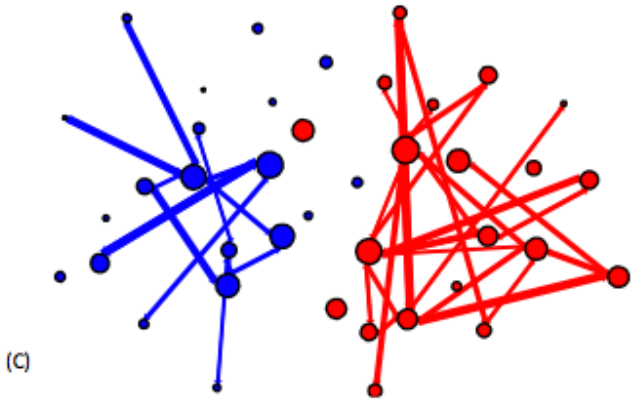
- something we've already learned how to do:
  - find strongly connected components

- keep only a subset of ties before finding connected components
  - reciprocal ties
  - edge weight above a threshold

Example: political blogs
(Aug 29th – Nov 15th, 2004)







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 JawaReport
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 INDC Journal
34 Real Clear Politics
35 Winds of Change
36 Allahpundit
37 Michelle Malkin
38 WizBang
39 Dean's World
40 Volokh

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

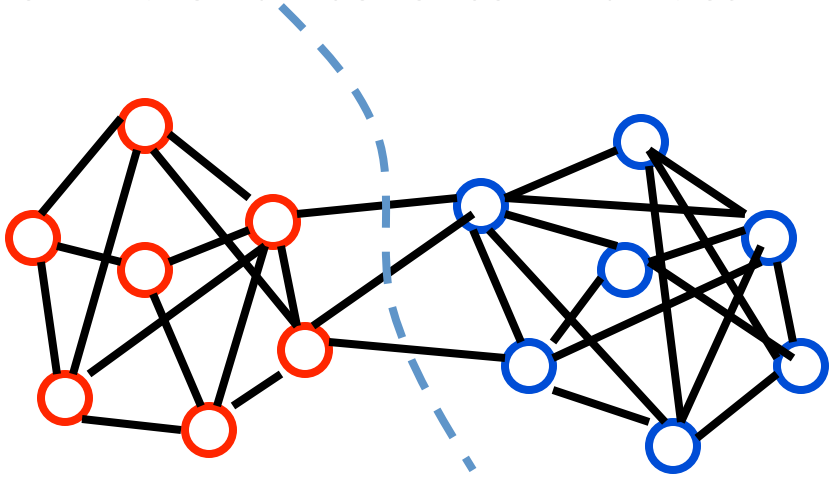
C) edges further limited to those exceeding 25 combined citations

only 15% of the citations bridge communities

source: Adamic & Glance, LinkKDD200

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



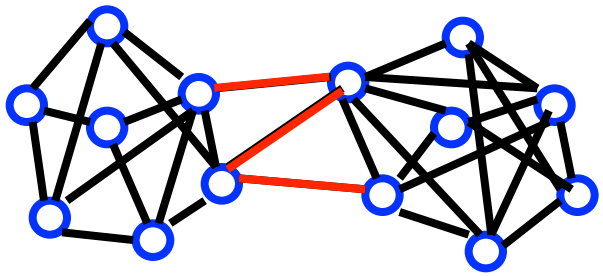
- Without “looking”, can we discover community structure in an automated way?

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

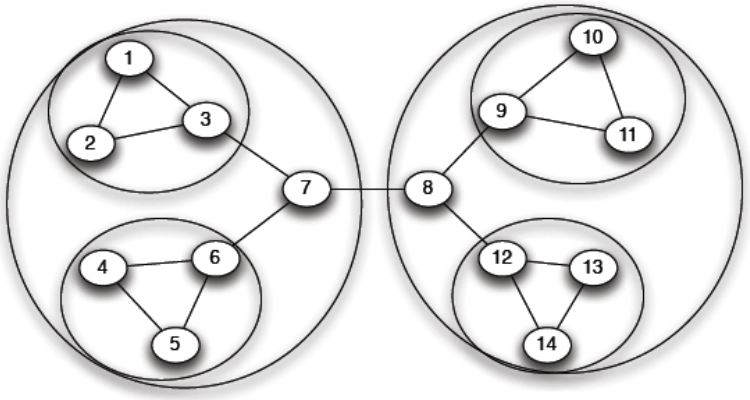
betweenness clustering algorithm

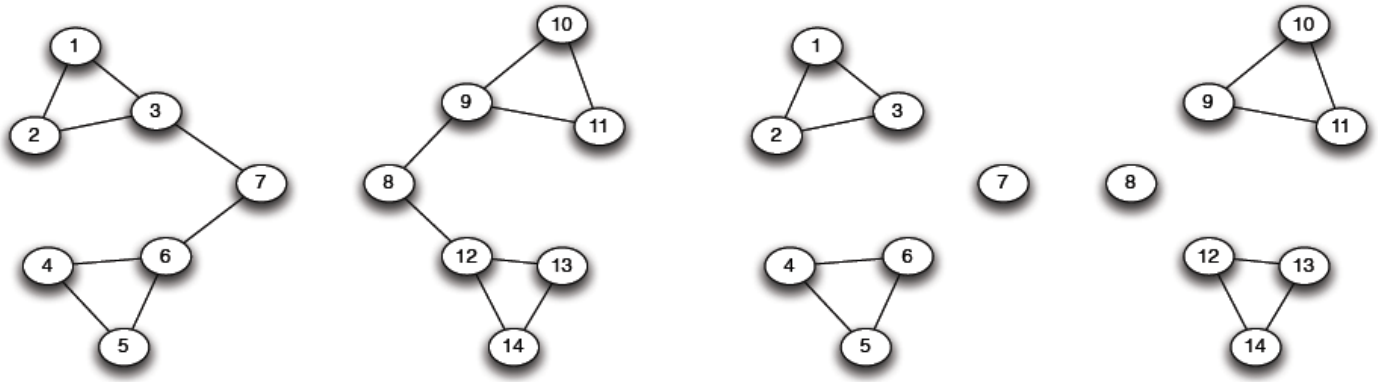


by Lada Adamic, U Michigan

betweenness clustering:

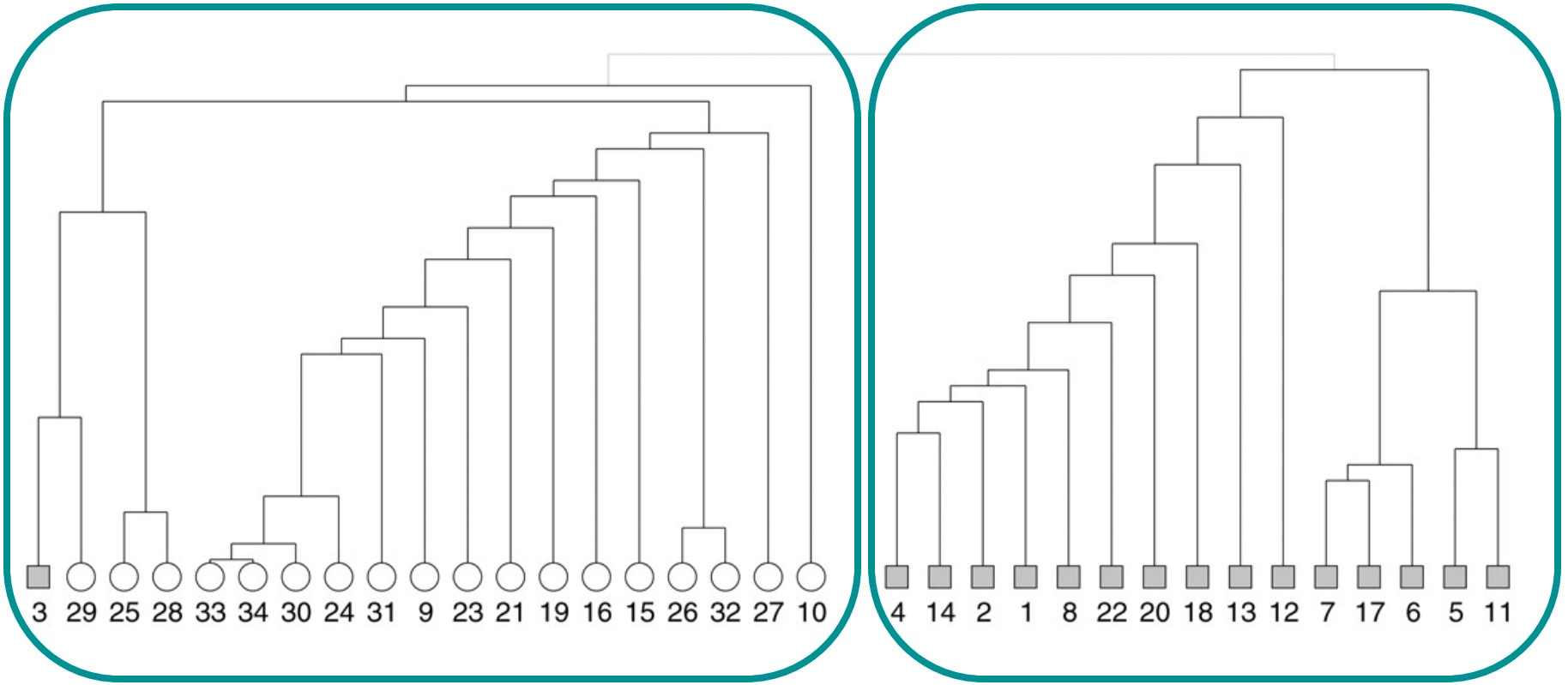
- successively remove edges of highest betweenness (the bridges, or local bridges), breaking up the network into separate components





(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

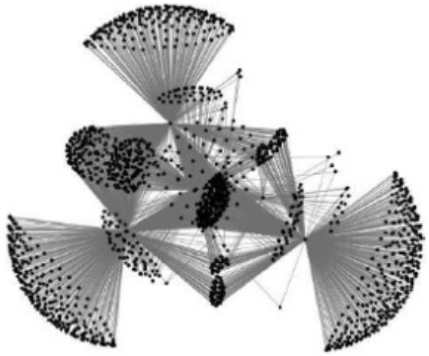
probability of an edge between two vertices is proportional to their degrees

- 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 greedy search



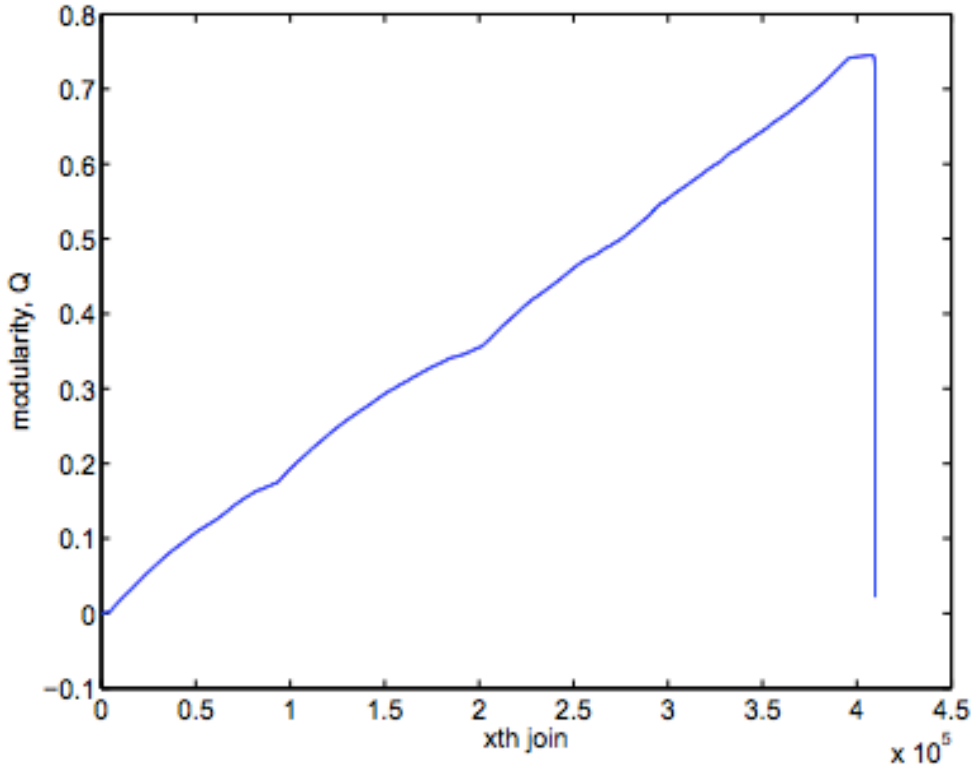


FIG. 1: The modularity Q over the course of the algorithm (the x axis shows the number of joins). Its maximum value is $Q = 0.745$, where the partition consists of 1684 communities.

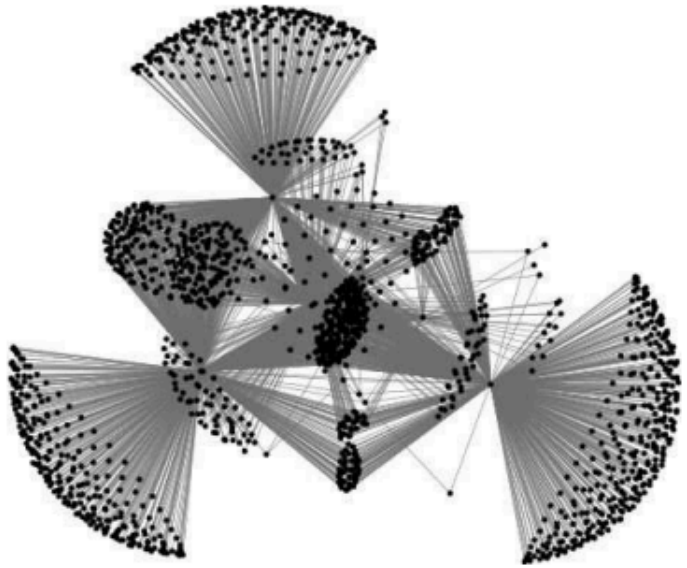
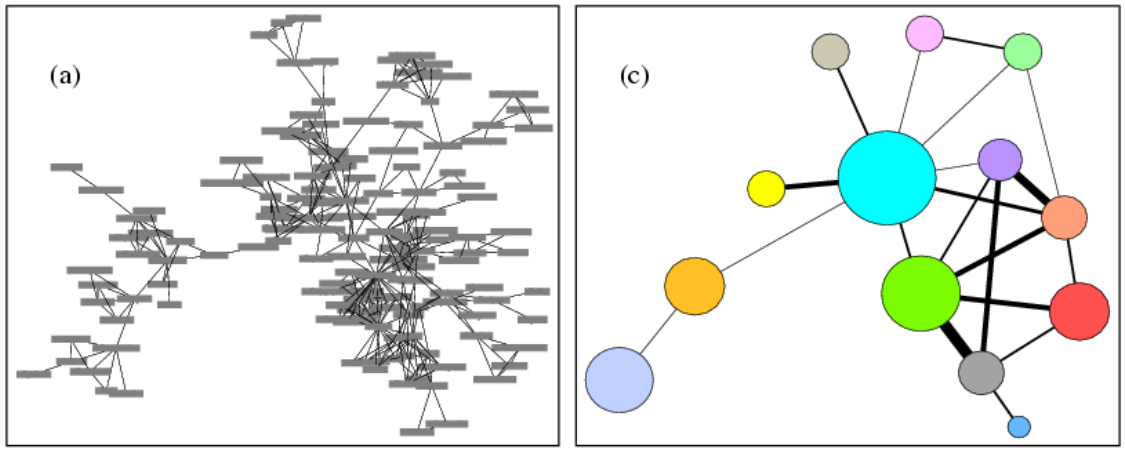
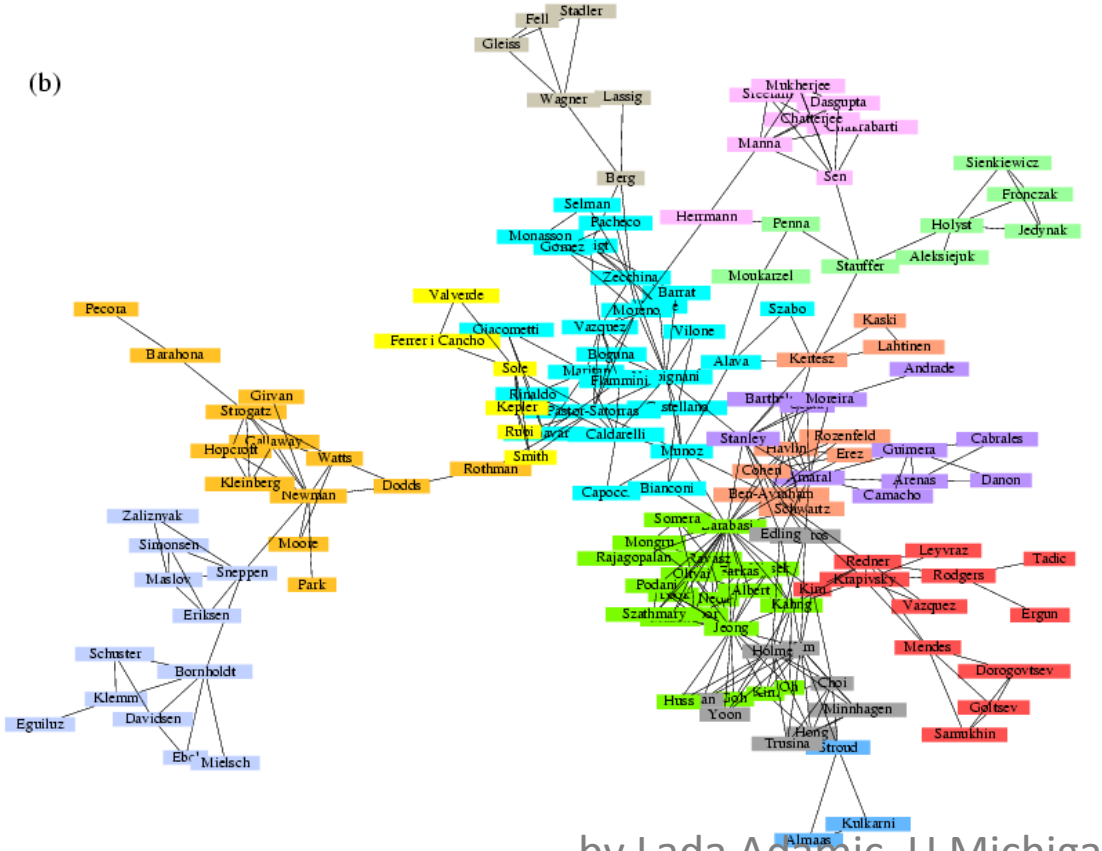


FIG. 2: A visualization of the community structure at maximum modularity. Note that the some major communities have a large number of "satellite" communities connected only to them (top, lower left, lower right). Also, some pairs of major communities have sets of smaller communities that act as "bridges" between them (e.g., between the lower left and lower right, near the center).

Finding community structure in very large networks

Authors: Aaron Clauset, M. E. J. Newman, Cristopher Moore, 2004





modularity can help us visualize large networks

by Lada Adamic, U Michigan

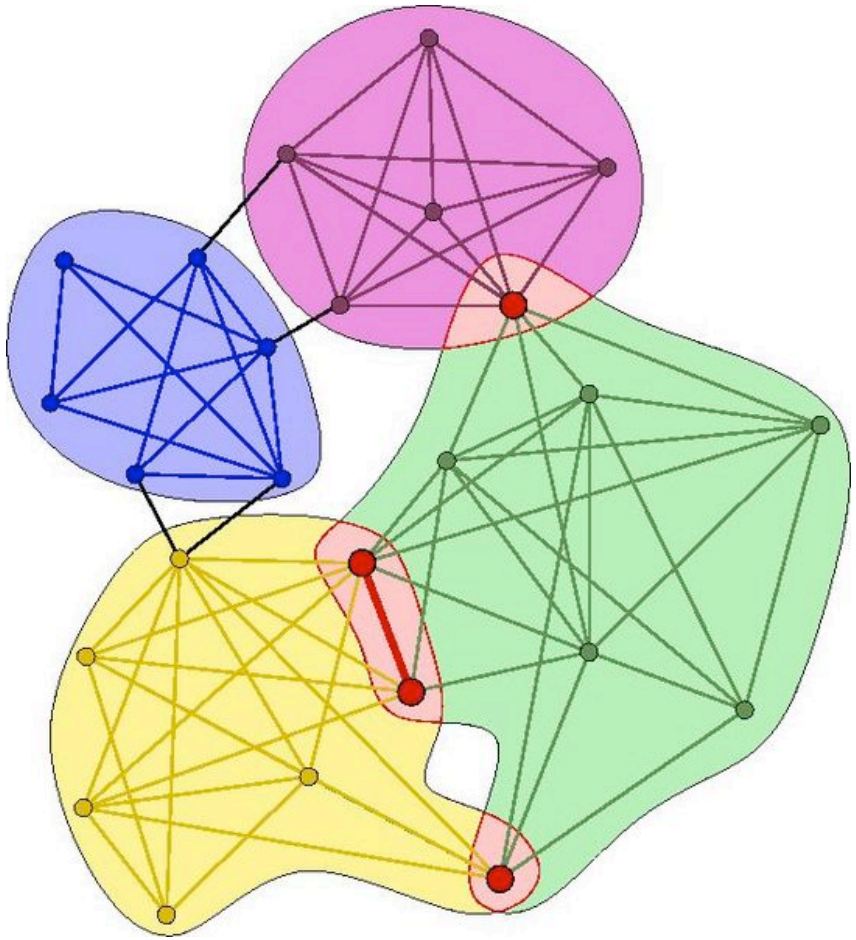
What if communities overlap?

- 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)



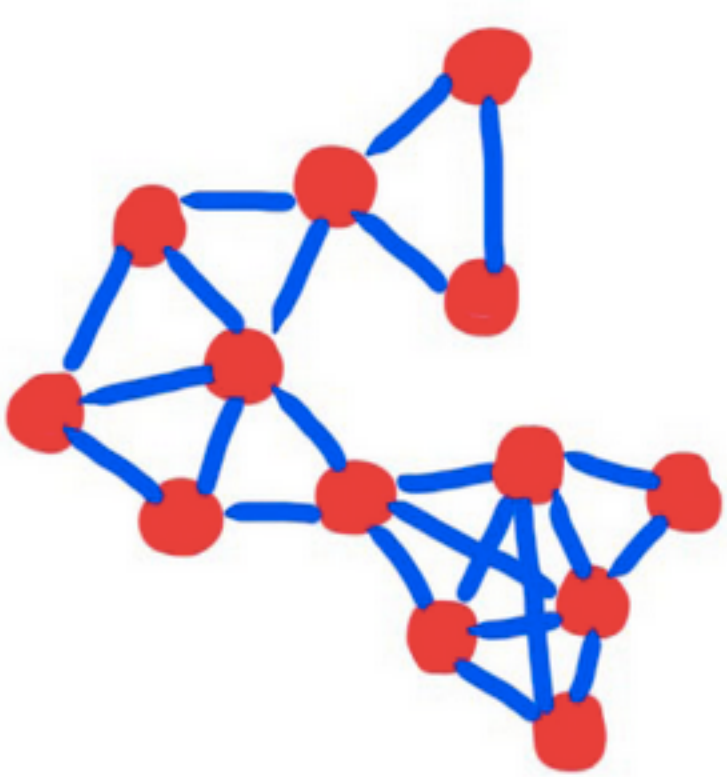
If you were to run a clique-percolation algorithm on this network using 3-cliques (triangles), you would find how many communities?

a) 0

b) 1

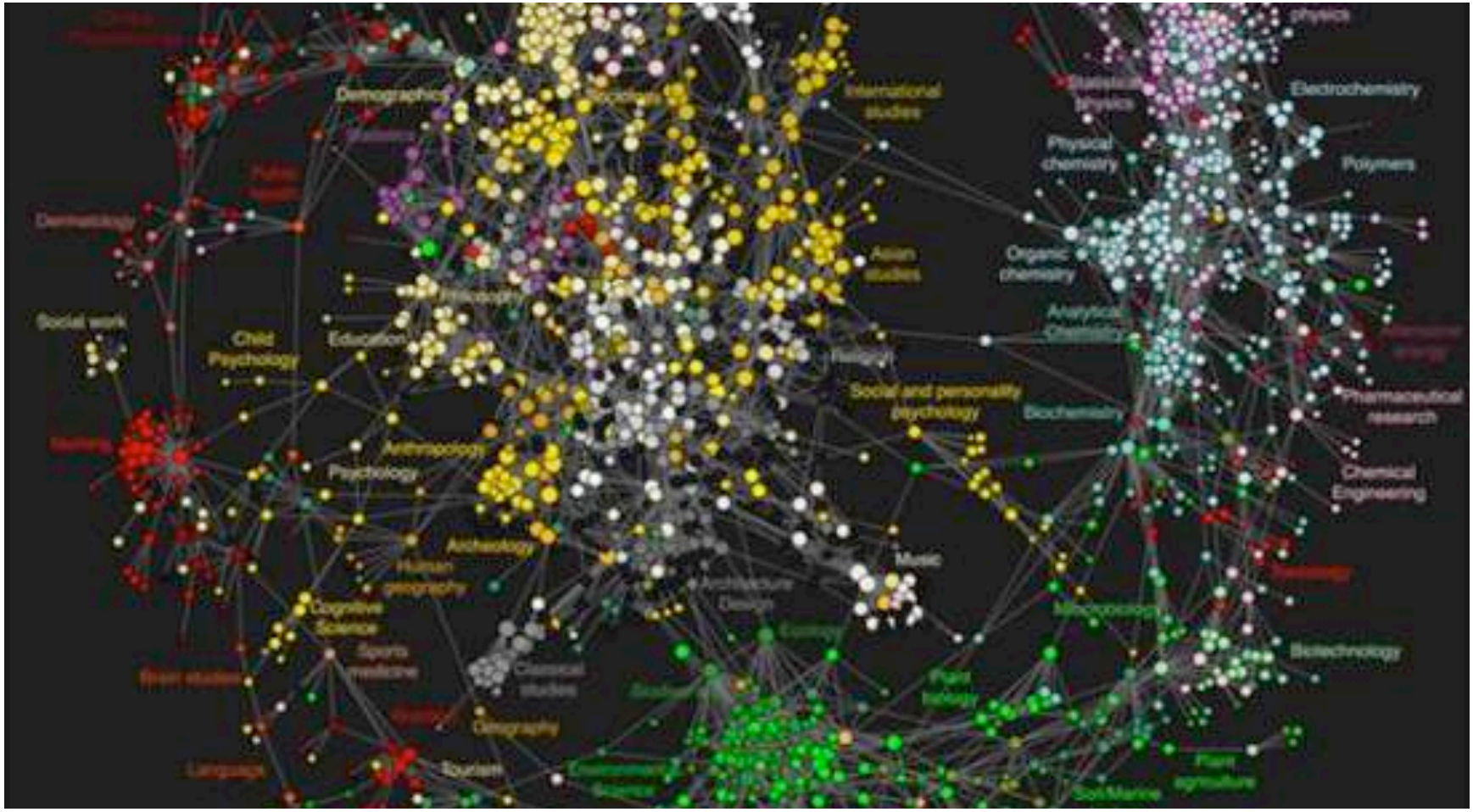
c) 2

d) 3



high-res maps of science

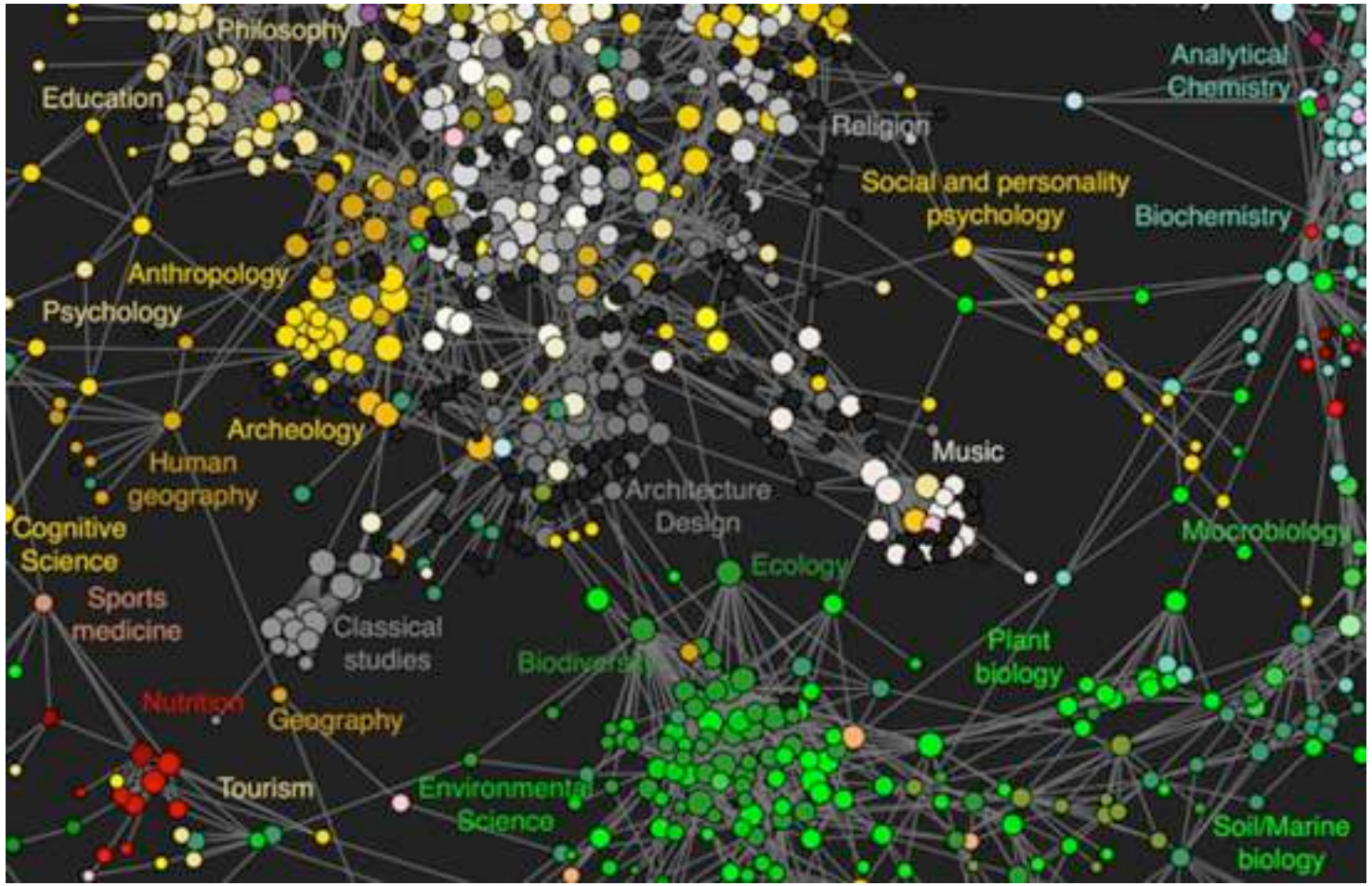
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004803



by Lada Adamic, U Michigan

high-res maps of science

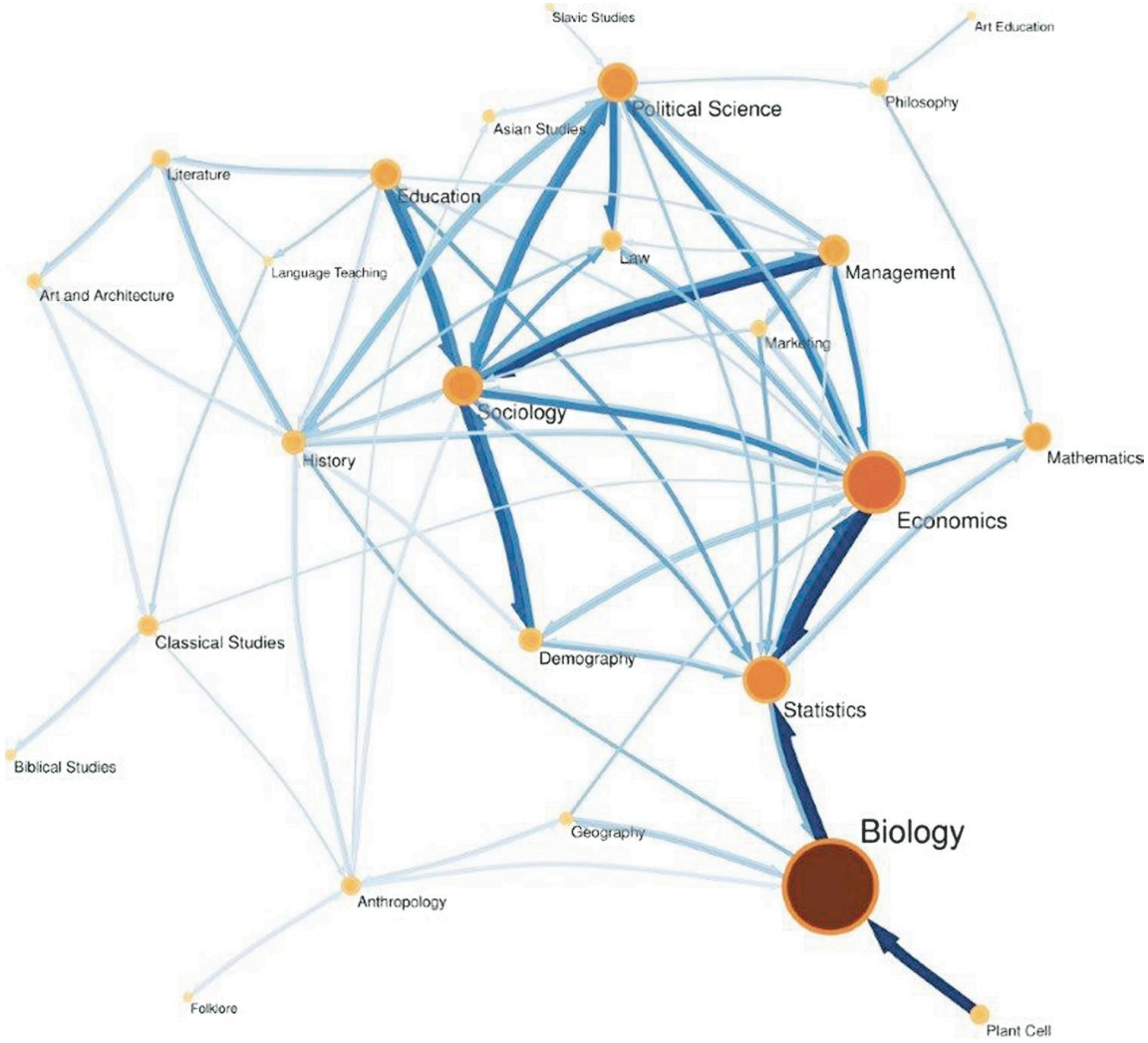
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0004803



by Lada Adamic, U Michigan

high-res maps of science

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



by Lada Adamic, U Michigan

Community Finding: 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