Online Social Networks and Media

Chapter 3, from D. Easley and J. Kleinberg book

Section 10.2.4, from A. Rajaraman, J. Ullman, J. Leskovec

BETWEENNESS AND GRAPH PARTITIONING

Centrality Measures

- Not all nodes are equally important
- Centrality Analysis:
  - Find out the most important nodes in one network

- Commonly-used Measures
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality
  - *Eigenvector Centrality*

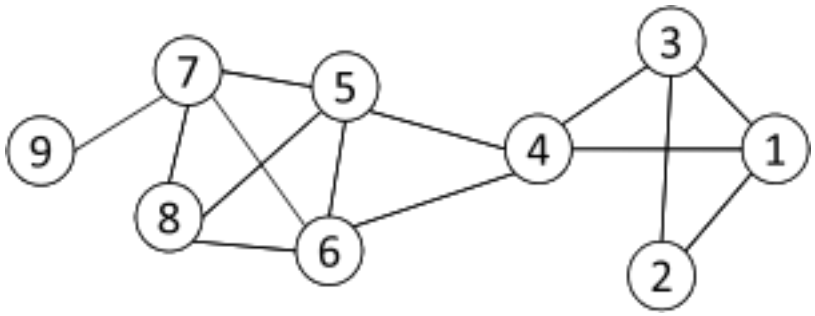
Degree Centrality

- The importance of a node is determined by the number of nodes adjacent to it
  - The larger the degree, the more import the node is
  - Only a small number of nodes have high degrees in many real-life networks
- Degree Centrality

$$C_D(v_i) = d_i = \sum_j A_{ij}$$

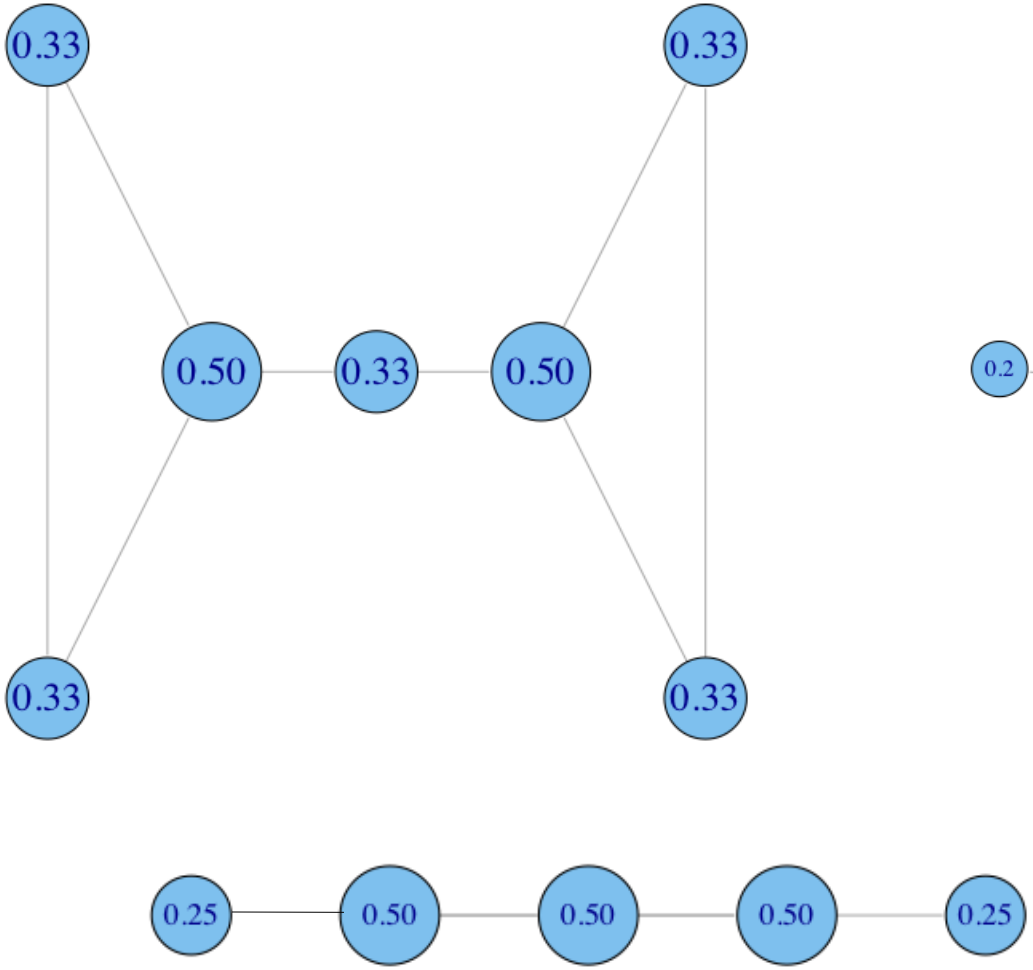
- Normalized Degree Centrality:

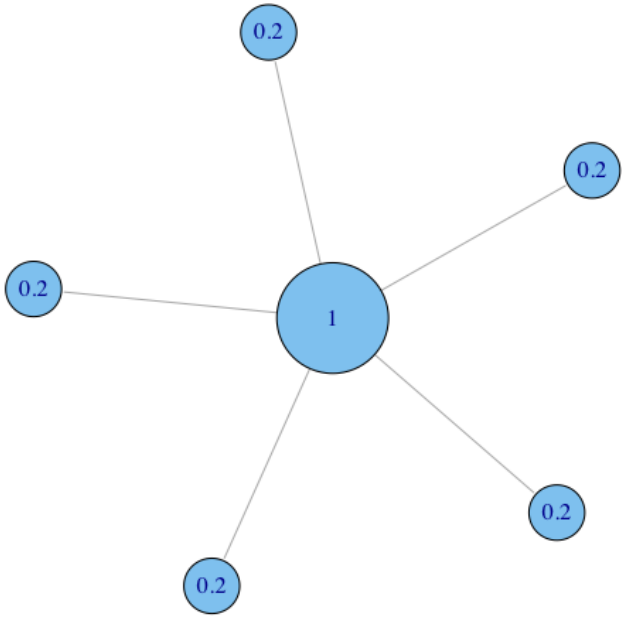
$$C'_D(v_i) = d_i/(n-1)$$



For node 1, degree centrality is 3; Normalized degree centrality is 3/(9-1)=3/8.

Degree Centrality (normalized)

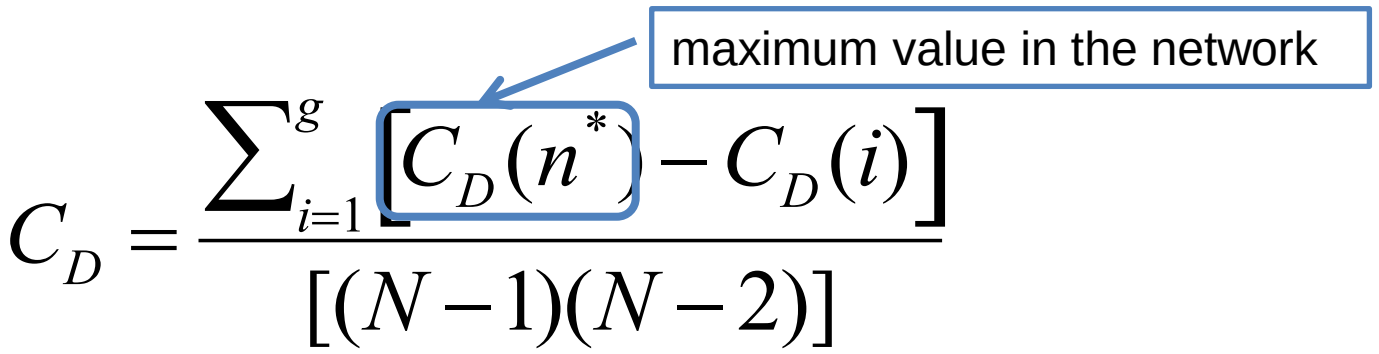




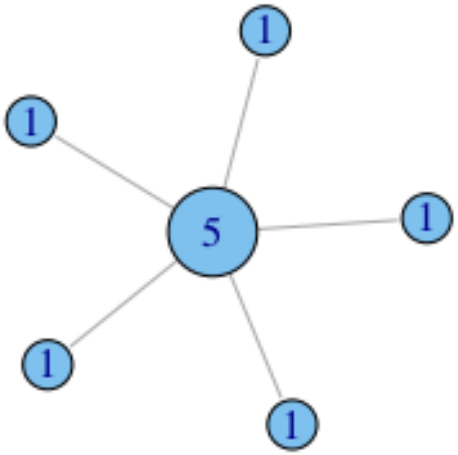
Degree Centralization: how equal are the nodes?

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):

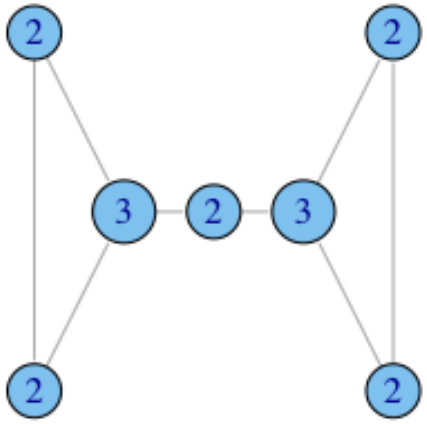


Degree Centralization



C_D = 1.0

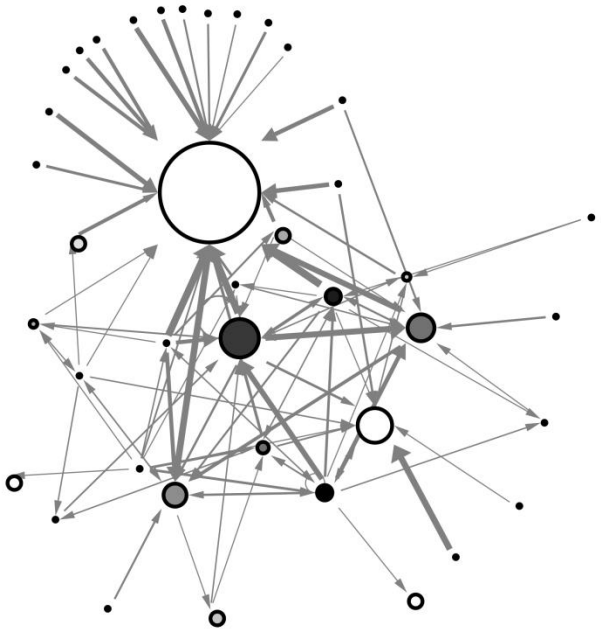
C_D = 0.167



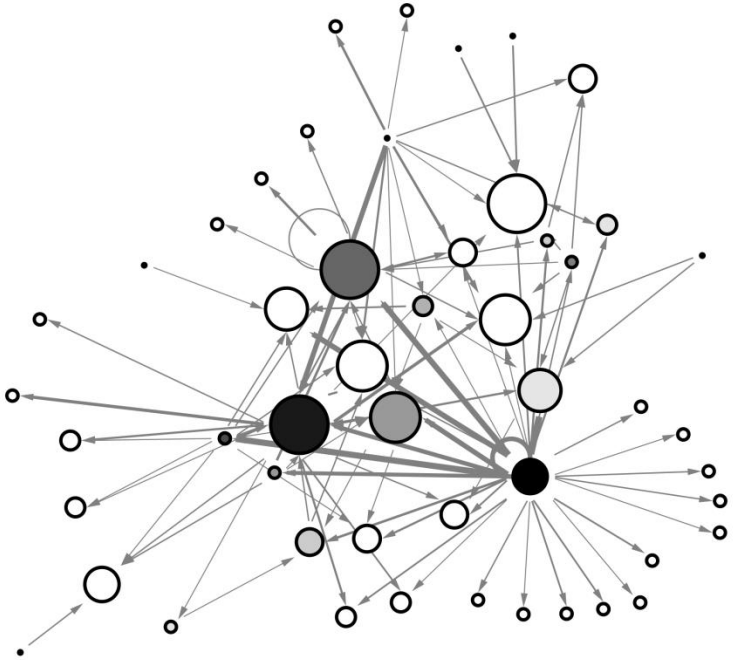
C_D = 0.167

Degree Centralization

example financial trading networks



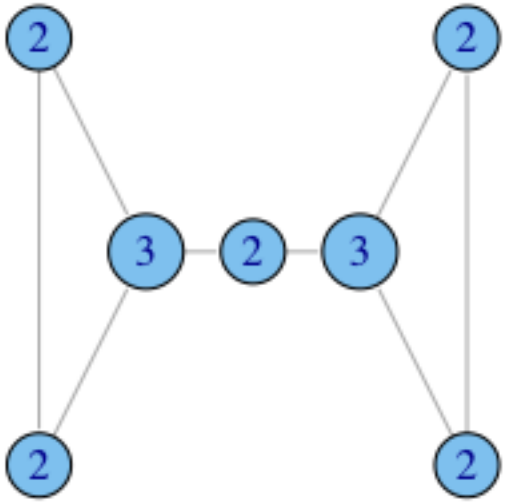
high centralization: one node trading with many others



low centralization: trades are more evenly distributed

when degree isn't everything

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



Closeness Centrality

- "Central" nodes are important, as they can reach the whole network more quickly than non-central nodes
- Importance measured by how close a node is to other nodes

- Average Distance: $$D_{avg}(v_i) = \frac{1}{n-1} \sum_{j \neq i}^{n} g(v_i, v_j)$$

- Closeness Centrality

$$C_C(v_i) = \left[\frac{1}{n-1} \sum_{j \neq i}^{n} g(v_i, v_j) \right]^{-1} = \frac{n-1}{\sum_{j \neq i}^{n} g(v_i, v_j)}$$

Closeness Centrality Example

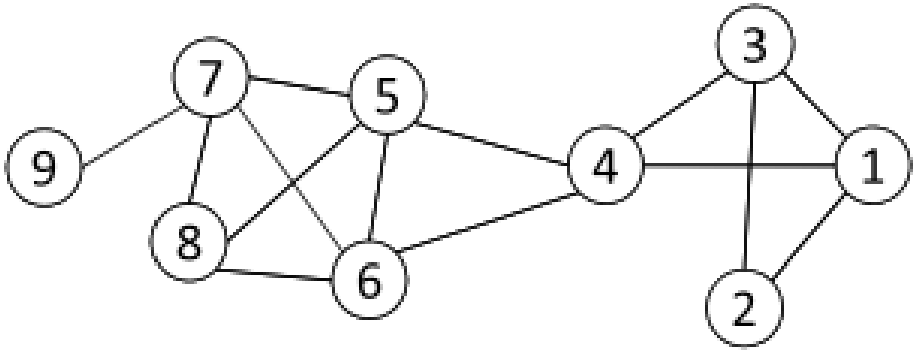


Table 2.1: Pairwise geodesic distance

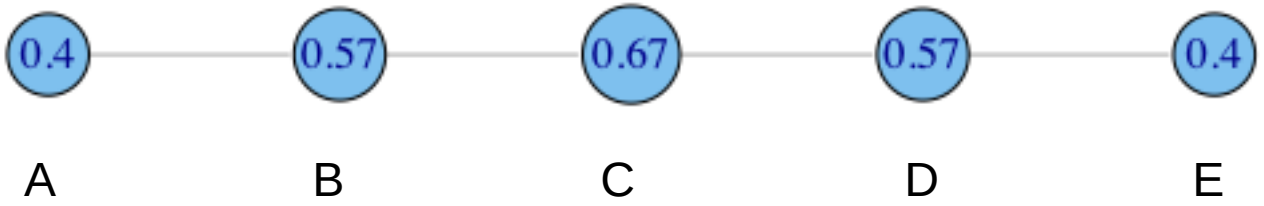
Node	1	2	3	4	5	6	7	8	9
1	0	1	1	1	2	2	3	3	4
2	1	0	1	2	3	3	4	4	5
3	1	1	0	1	2	2	3	3	4
4	1	2	1	0	1	1	2	2	3
5	2	3	2	1	0	1	1	1	2
6	2	3	2	1	1	0	1	1	2
7	3	4	3	2	1	1	0	1	1
8	3	4	3	2	1	1	1	0	2
9	4	5	4	3	2	2	1	2	0

$$C_C(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = 8/17 = 0.47,$$

$$C_C(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = 8/13 = 0.62.$$

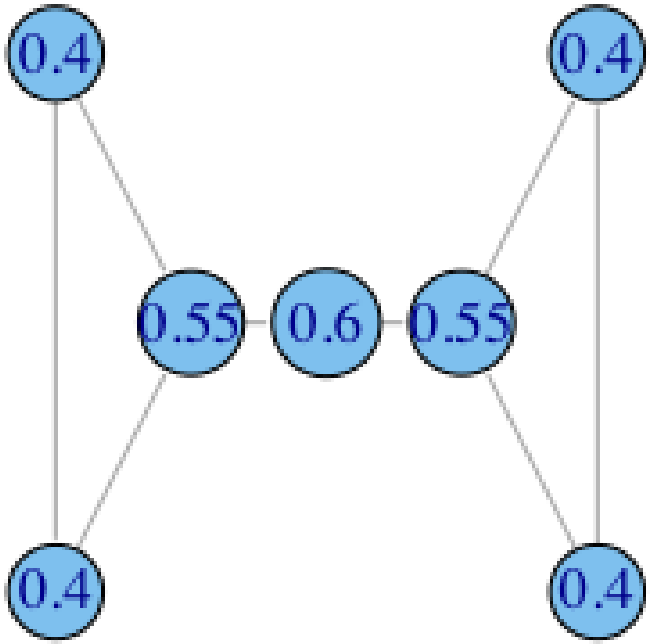
Node 4 is more central than node 3

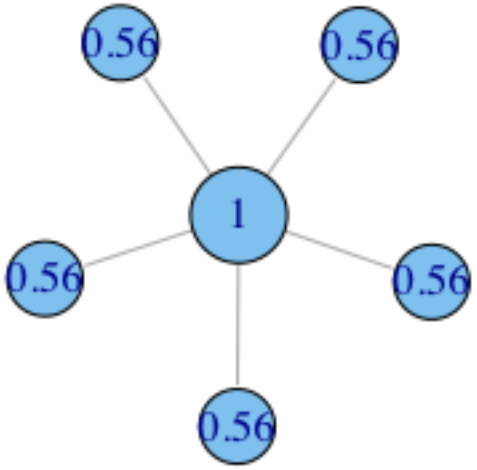
Closeness Centrality Example



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

Closeness Centrality Example





Betweenness Centrality

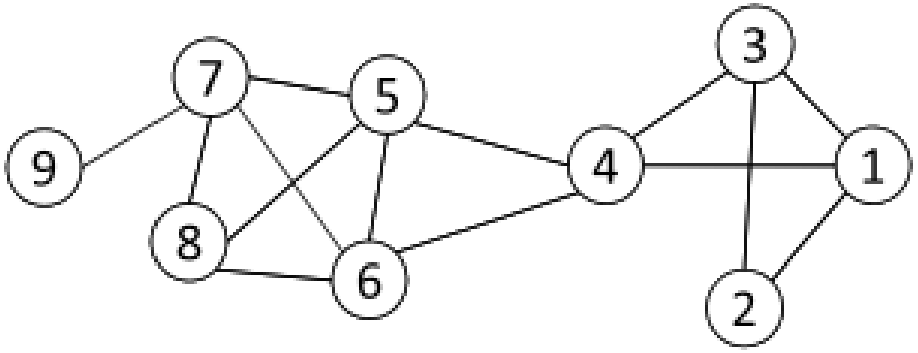
- Node betweenness counts the number of shortest paths that pass one node
- Nodes with high betweenness are important in communication and information diffusion

- Betweenness Centrality $$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

σ_{st} : The number of shortest paths between s and t

$\sigma_{st}(v_i)$: The number of shortest paths between s and t that pass v_i

Betweenness Centrality Example



$$C_B(4) = 15$$

Table 2.2: $\sigma_{st}(4)/\sigma_{st}$

	$s=1$	$s=2$	$s=3$
$t=5$	1/1	2/2	1/1
$t=6$	1/1	2/2	1/1
$t=7$	2/2	4/4	2/2
$t=8$	2/2	4/4	2/2
$t=9$	2/2	4/4	2/2

What's the betweenness centrality for node 5?

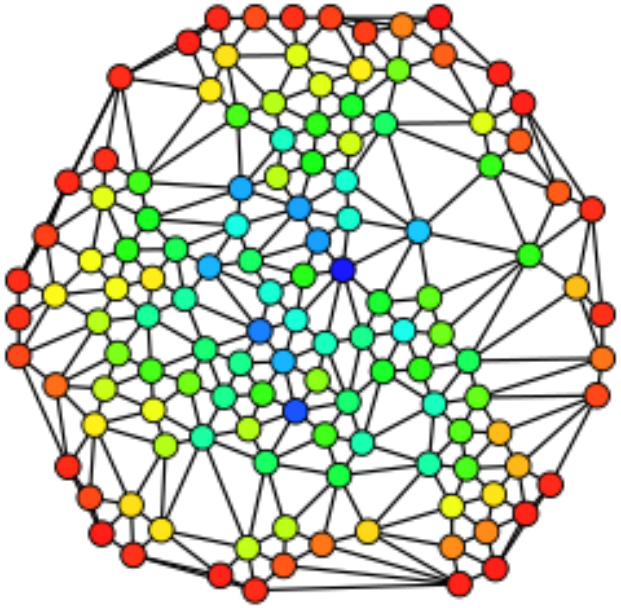
σ_{st} : The number of shortest paths between s and t

$\sigma_{st}(v_i)$: The number of shortest paths between s and t that pass v_i

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

Betweenness Centrality Example

Blue (max)
Red (0)



Eigenvector Centrality

- One's importance is determined by one's friends
- If one has many important friends, one should be important as well.

$$C_E(v_i) \propto \sum_{v_j \in N_i} A_{ij} C_E(v_j)$$

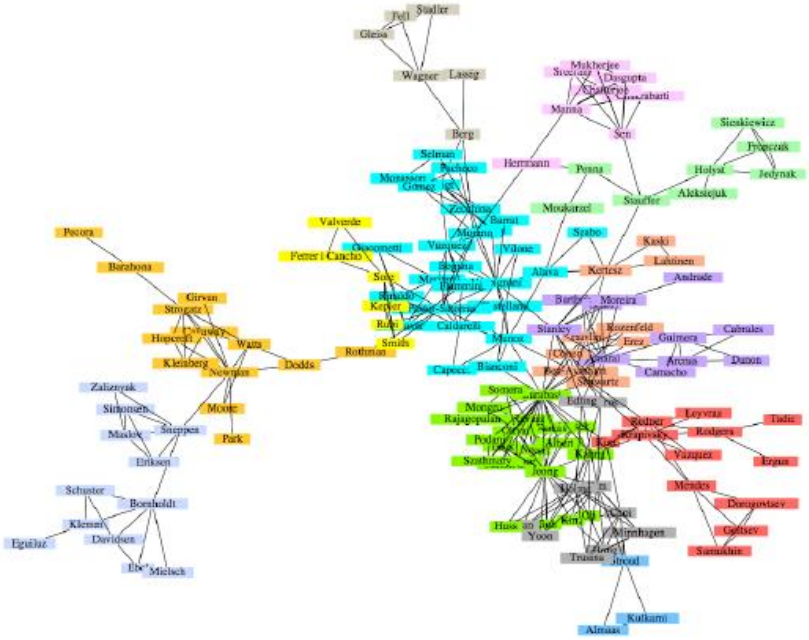
$$\mathbf{x} \propto A\mathbf{x} \quad \Rightarrow \quad A\mathbf{x} = \lambda \mathbf{x}.$$

- The centrality corresponds to the top eigenvector of the adjacency matrix A.
- A variant of this eigenvector centrality is the PageRank score.

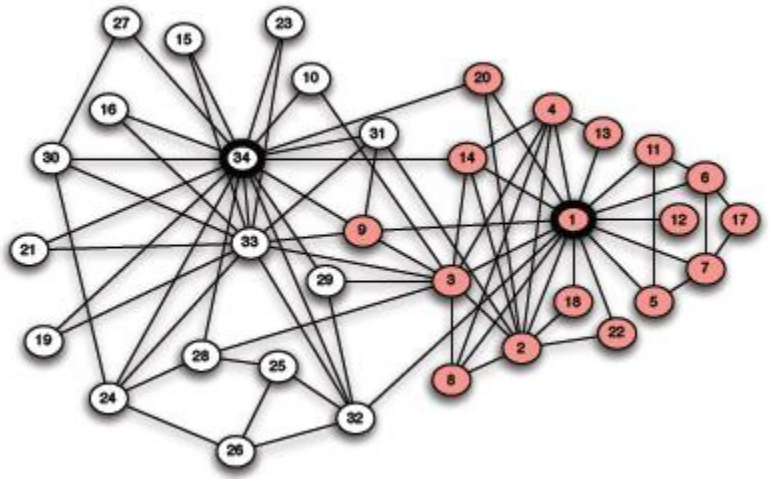
Betweenness and Graph Partitioning

Graph partitioning:

Given a network dataset, how to identity *densely connected groups* of nodes within it



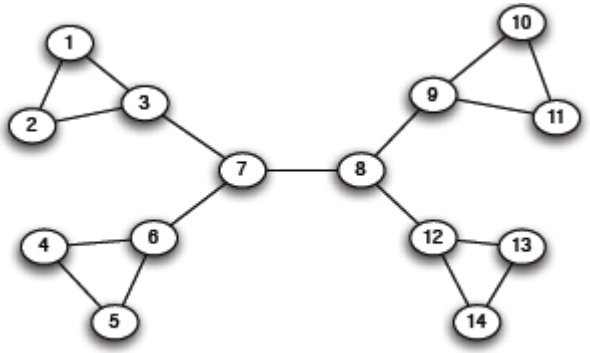
Co-authorship network of physicists and applied mathematicians

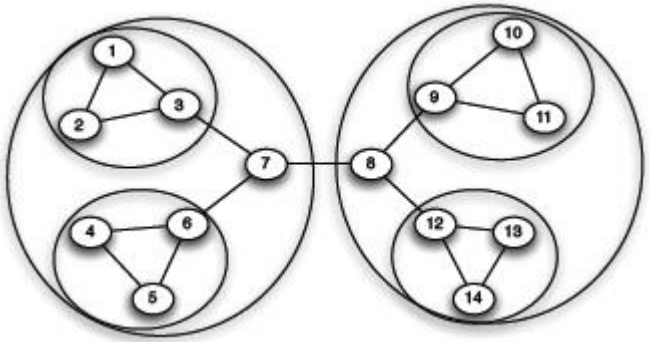


Karate club

Betweenness and Graph Partitioning

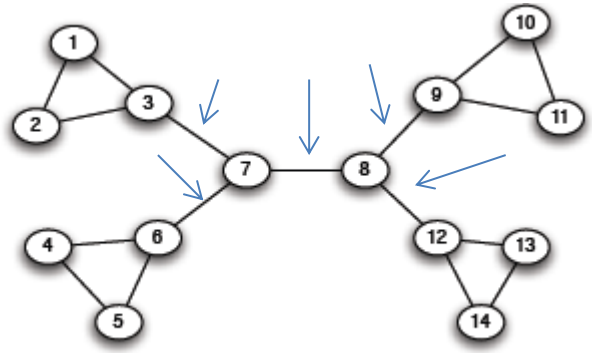
- Divisive methods: try to identify and remove the "spanning links" between densely-connected regions
- Agglomerative methods: Find nodes that are likely to belong to the same region and merge them together (bottom-up)





Girvan and Newman

- Divisive method

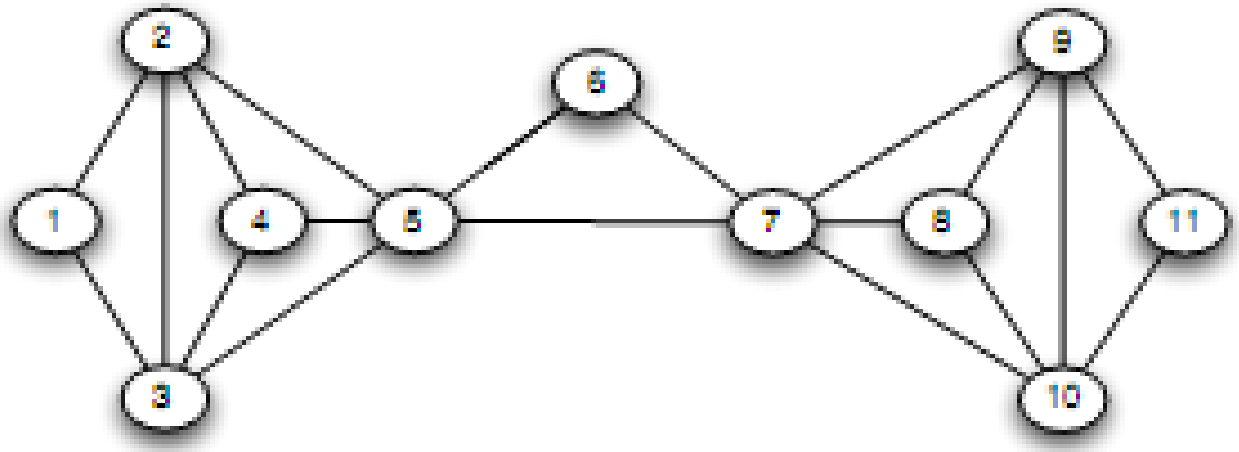


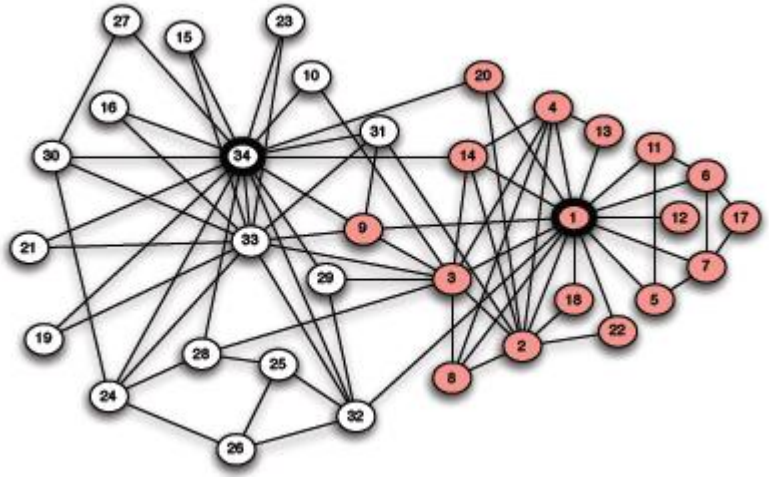
Finding bridges and local bridges?

Which one to choose?

Girvan and Newman

There is no local bridge

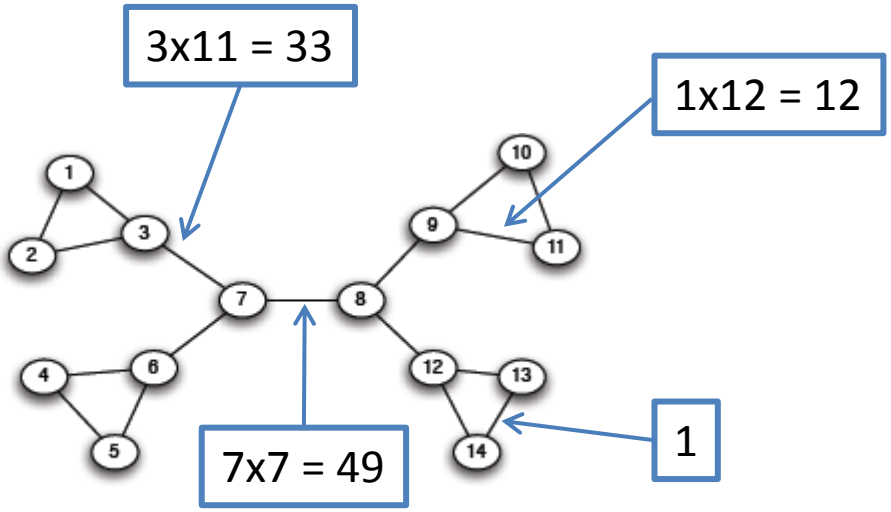




Edge Betweenness

Betweenness of an edge (a, b): number of pairs of nodes x and y such that the edge (a, b) lies on the shortest path between x and y - since there can be several such shortest paths edge (a, b) is credited with the fraction of those shortest paths that include (a, b).

$$\mathrm{bt}(a,b) = \sum_{x,y} \frac{\#shortest\_paths(x,y)through(a,b)}{\#shortest\_paths(x,y)}$$



Edges that have a high probability to occur on a randomly chosen shortest path between two randomly chosen nodes have a high betweenness.

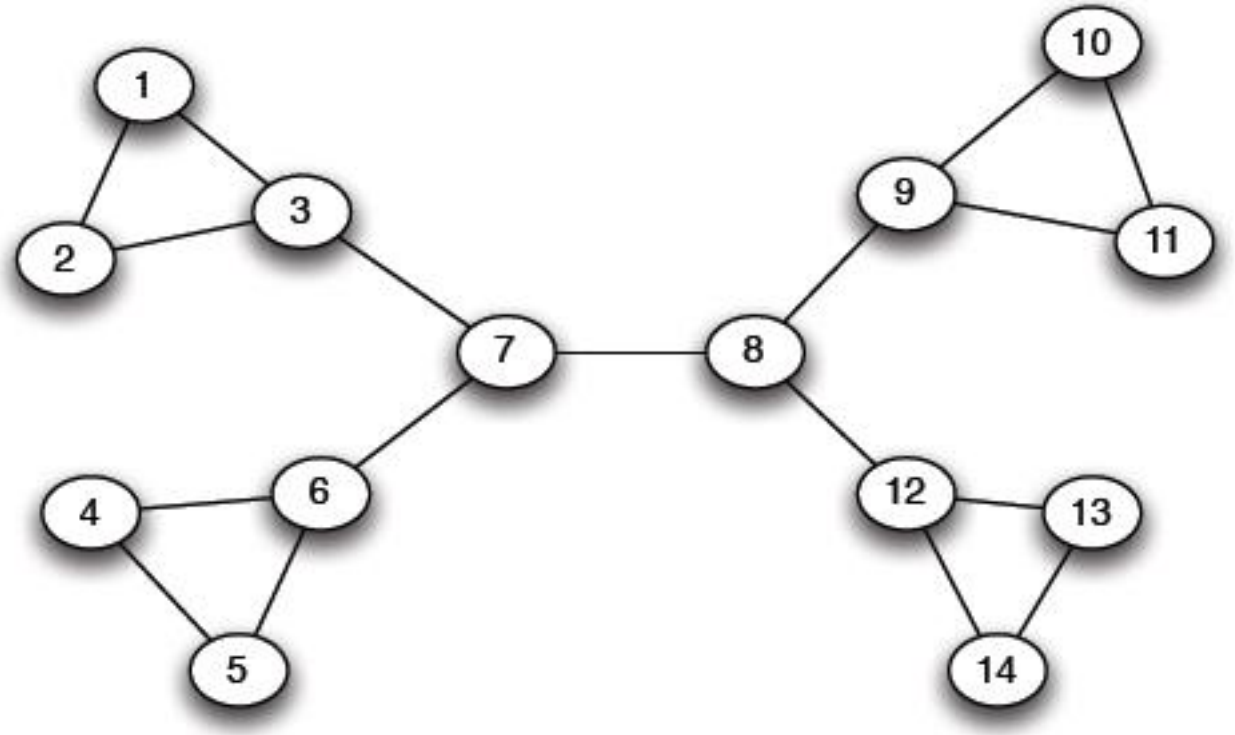
Traffic (unit of flow)

Girvan and Newman

1. The betweenness of all existing edges in the network is calculated first.
2. The edge with the highest betweenness is removed.
   If this separates the graph -> partition.
3. The betweenness of all edges affected by the removal is recalculated.

Steps 2 and 3 are repeated until no edges remain.

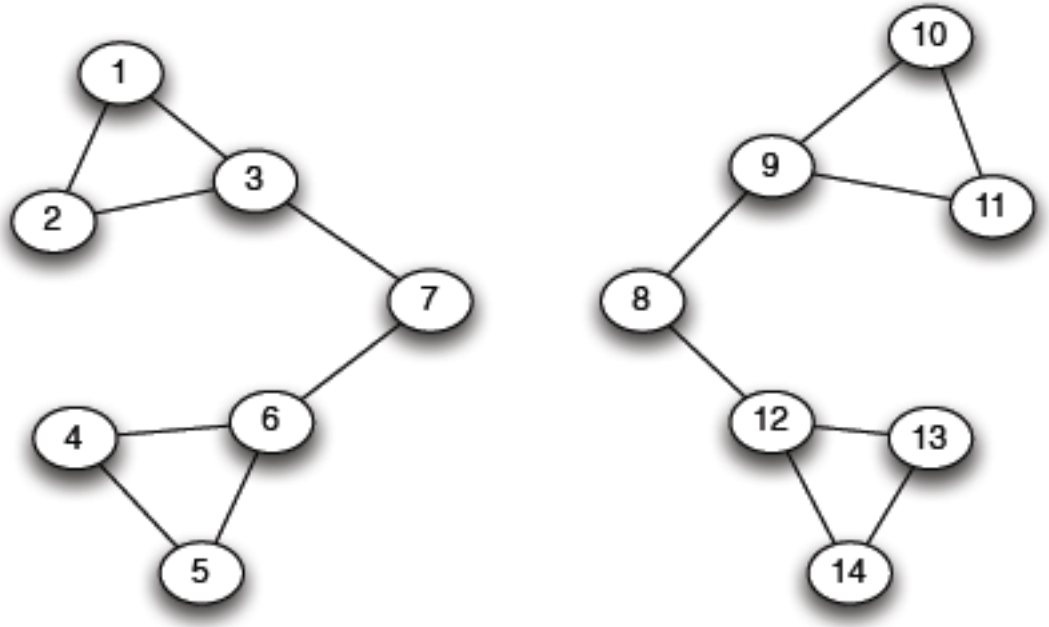
Girvan Newman method: An example



Betweenness(7, 8)= 7x7 = 49 Betweenness(1, 3) = 1X12=12

Betweenness(3, 7)=Betweenness(6-7)=Betweenness(8, 9) = Betweenness(8, 12)= 3X11=33

Girvan Newman method: An example

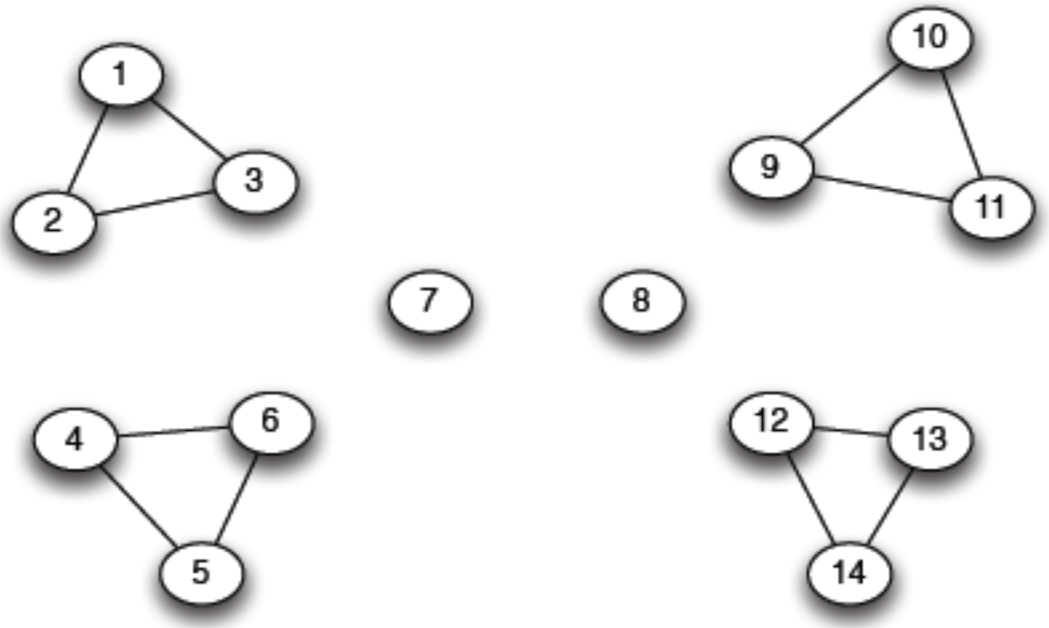


(a) *Step 1*

Betweenness(1, 3) = 1X5=5

Betweenness(3,7)=Betweenness(6,7)=Betweenness(8-9) = Betweenness(8,12)= 3X4=12

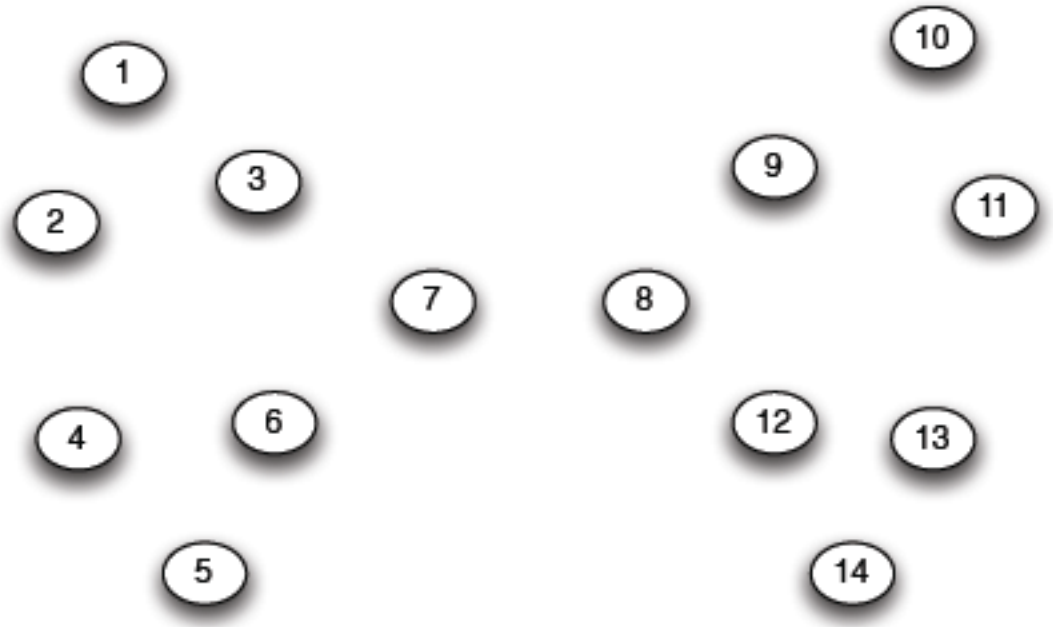
Girvan Newman method: An example



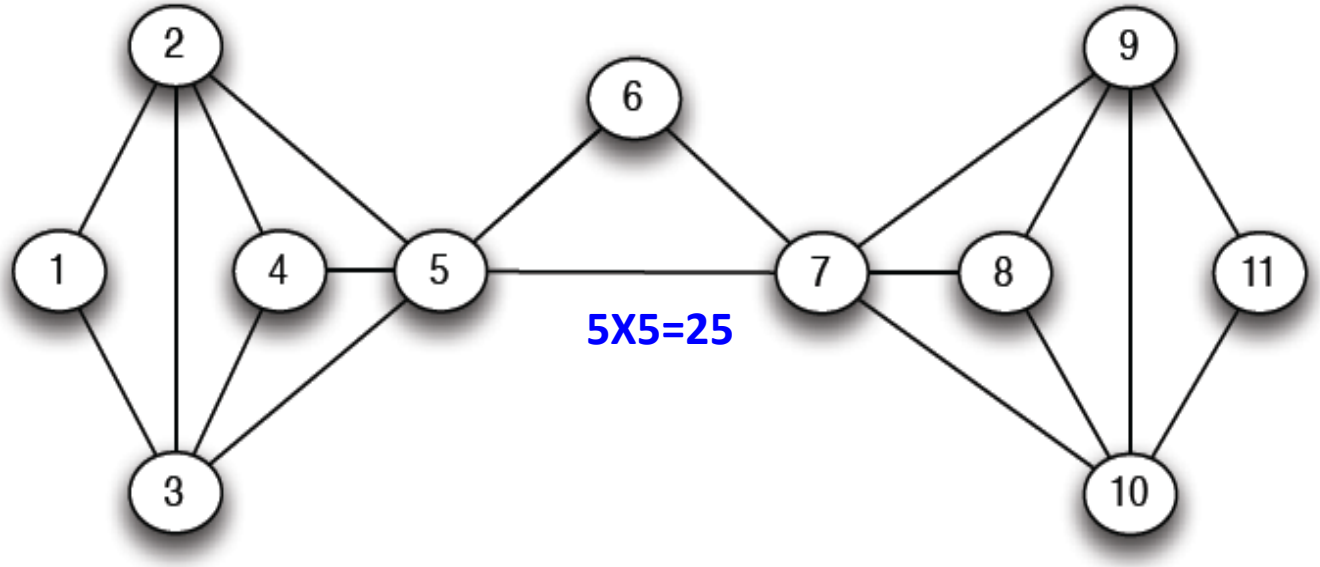
(b) *Step 2*

Betweenness of every edge = 1

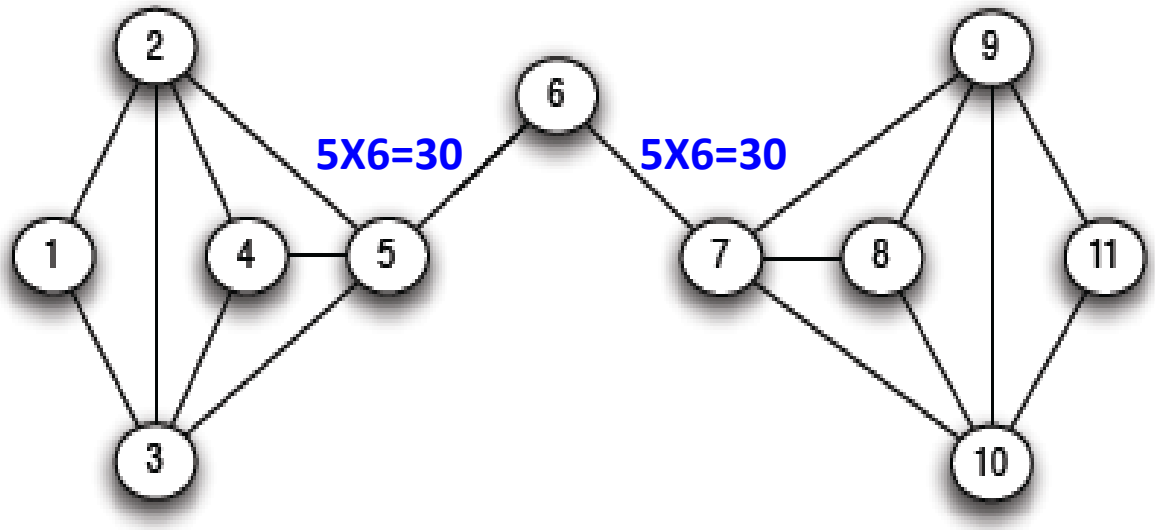
Girvan Newman method: An example



Another example

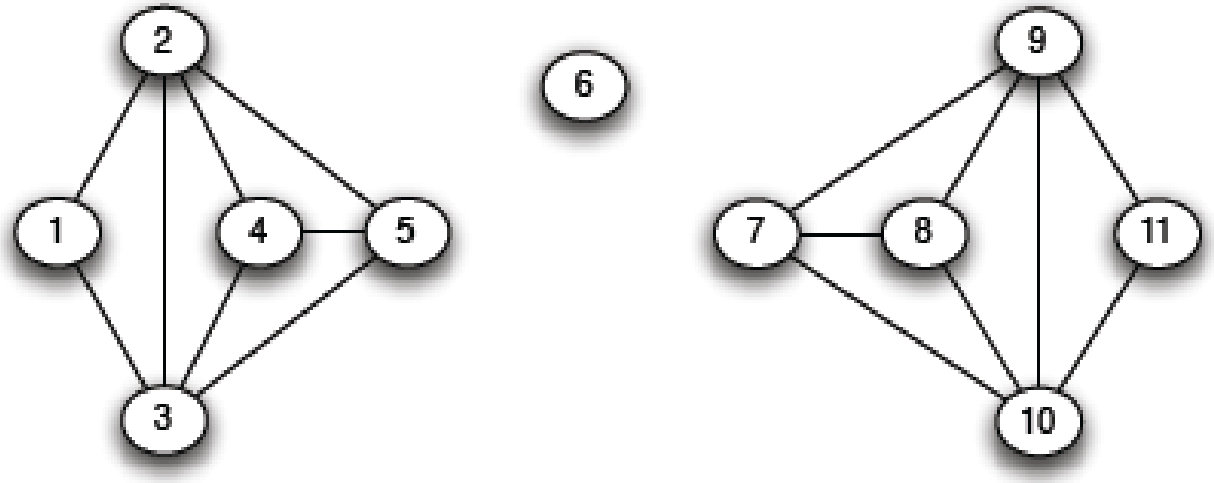


Another example



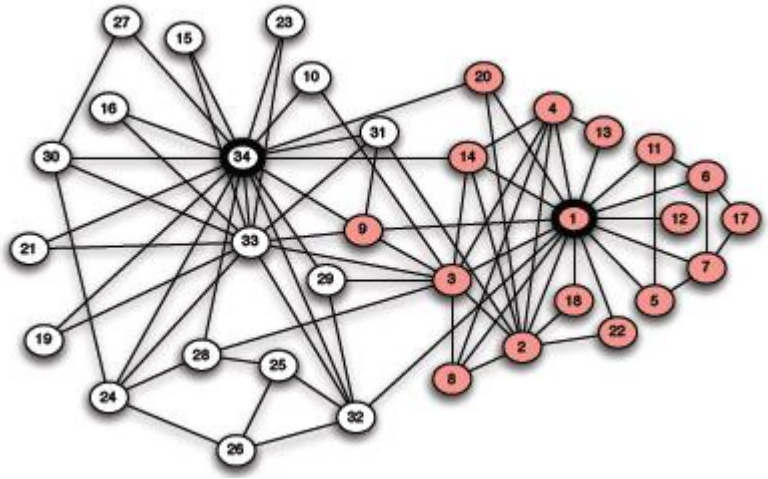
(a) *Step 1*

Another example



(b) *Step 2*

Girvan and Newman



34 president

1 instructor

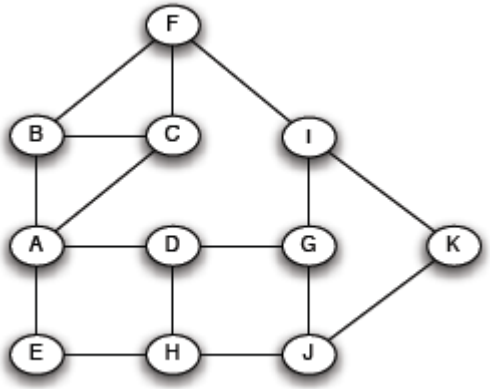
Correct but node 9 (attached it to 34) – why? 3 weeks away from getting a black belt

Minimum cut approach – the same outcome

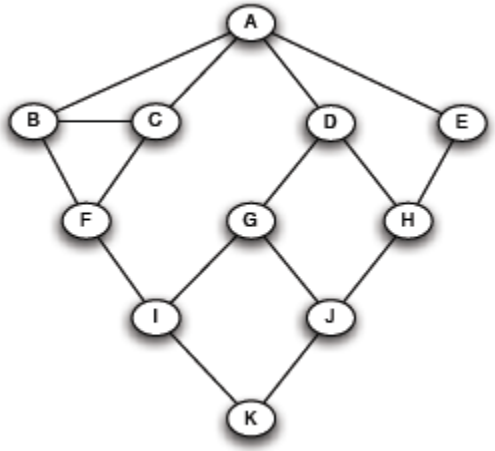
Computing Betweenness

1. Perform a BFS starting from A
2. Determine the shortest path from A to each other node
3. Based on these numbers, determine the amount of flow from A to all other nodes that uses each edge

Step 1



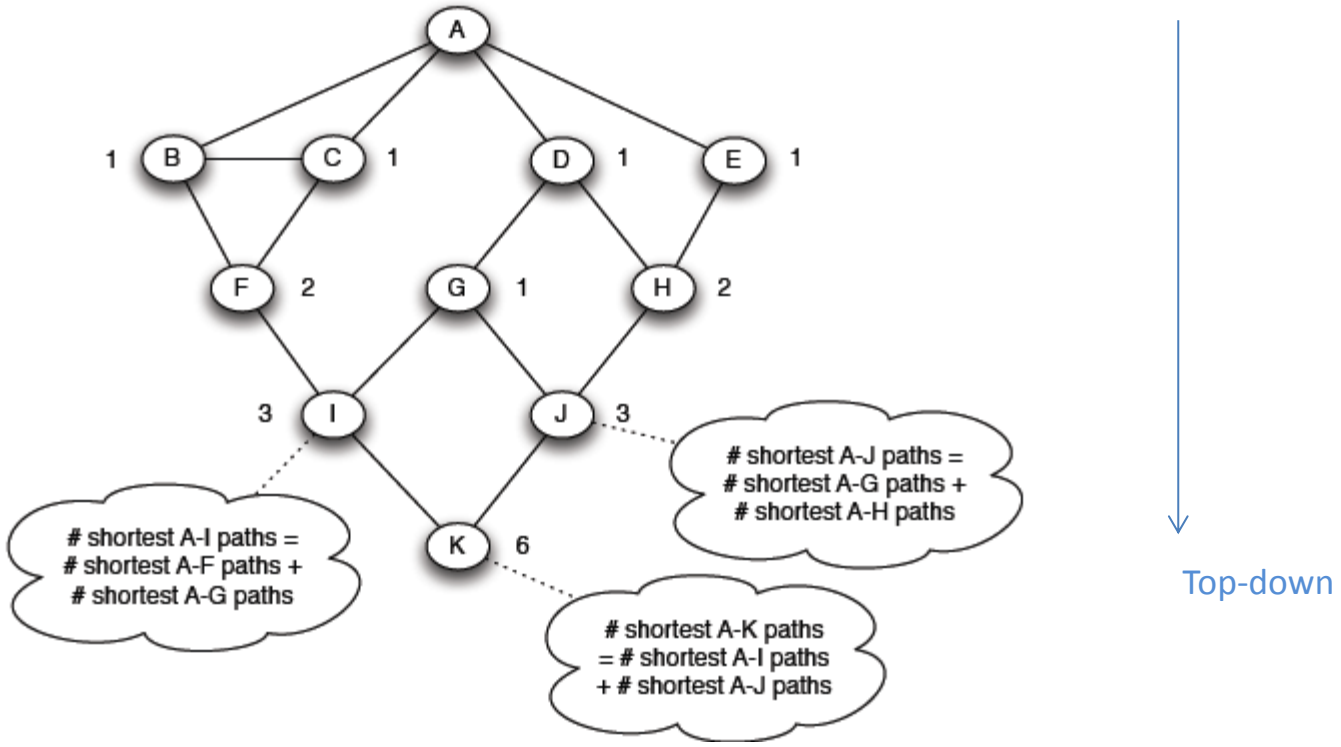
Initial network



BFS on A

Step 2

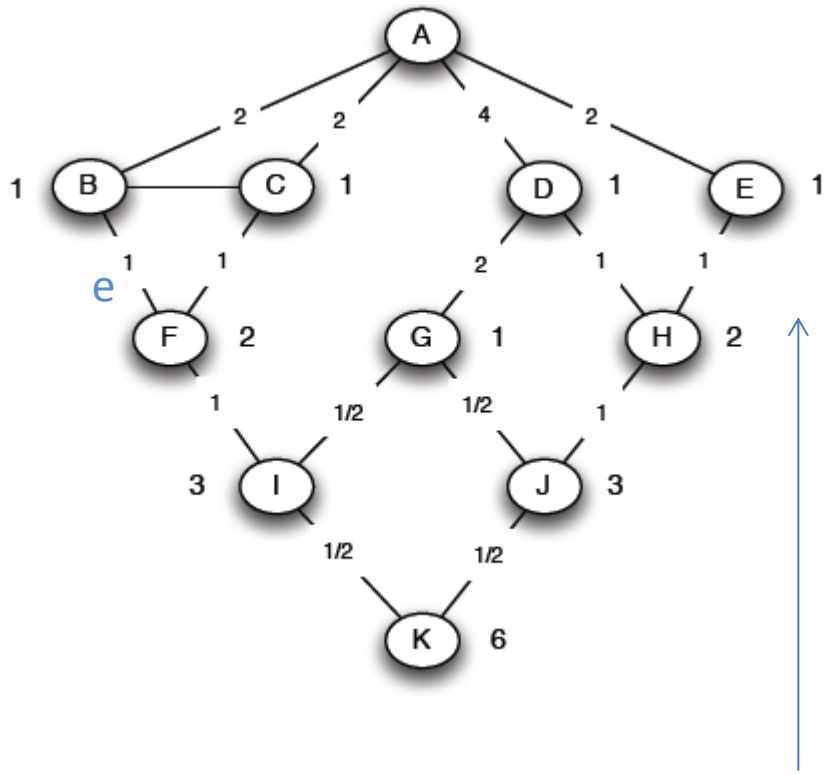
Count how many shortest paths from A to a specific node



Step 3

For *each edge e*: calculate the sum *over all nodes Y* of the fraction of shortest paths *from the root A to Y* that go through e.

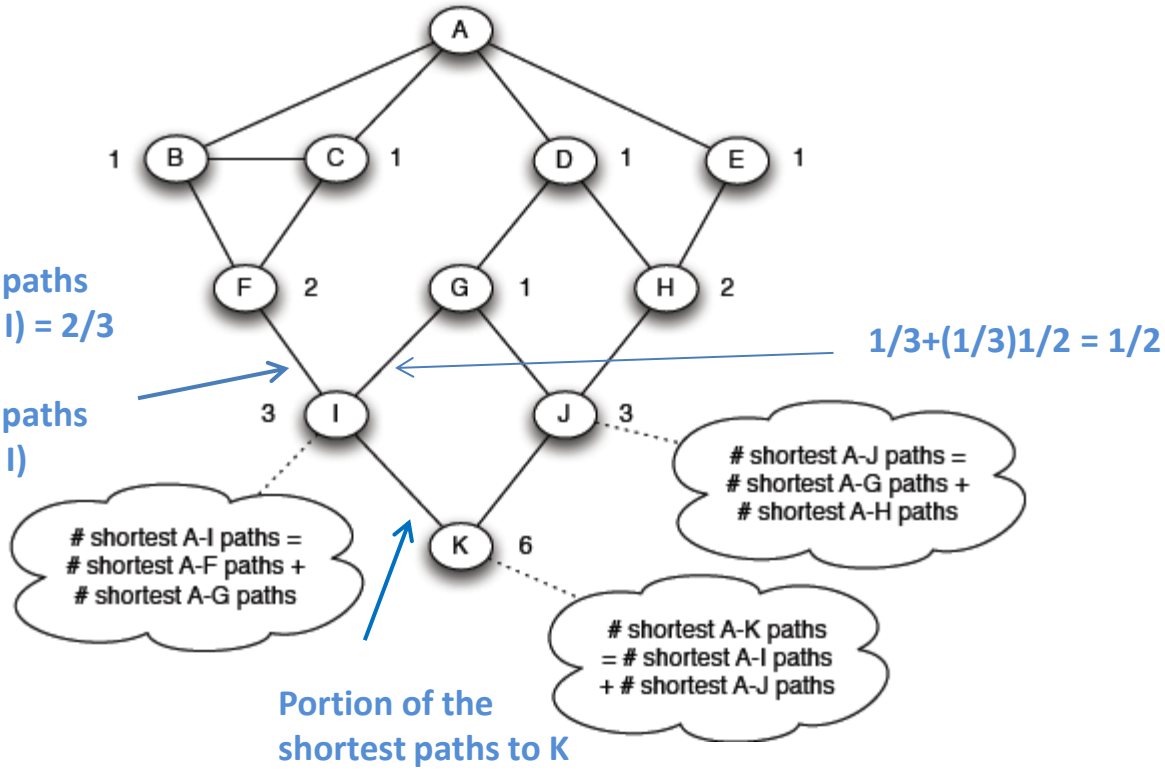
Each edge (X, Y) participates in the shortest-paths from the root to Y and to nodes (at levels) below Y -> Bottom up calculation



Step 3

Count the flow through each edge

$$credit(e) = \sum_{X,Y} \frac{|\,shortest - path(X,Y) \text{through } e\,|}{|\,shortest\_path(X,Y)\}\,|}$$

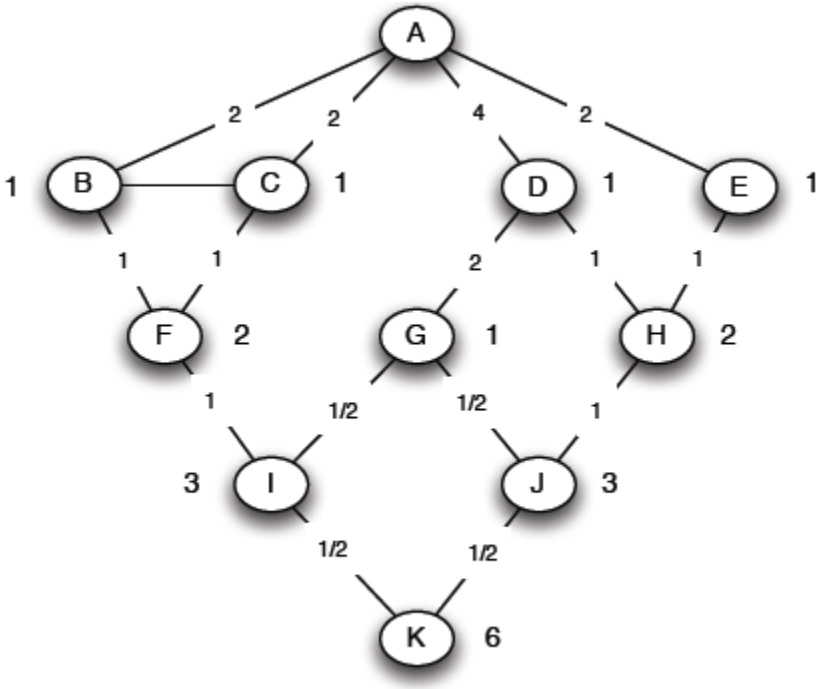


Portion of the shortest paths to I that go through (F, I) = 2/3
+
Portion of the shortest paths to K that go through (F, I)
(1/2)(2/3) = 1/3
= 1

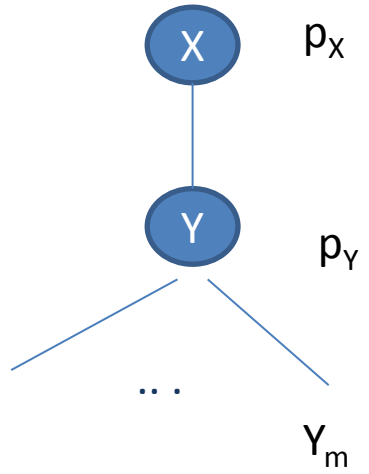
1/3+(1/3)1/2 = 1/2

Portion of the shortest paths to K that go through (I, K)
= 3/6 = 1/2

Step 3: formula



(X, Y)



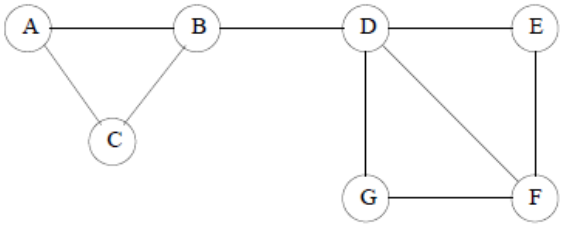
$$flow(X,Y) = p_X / p_Y \sum_{Y_i childof Y} (p_X / p_Y)\, flow(Y, Y_i)$$

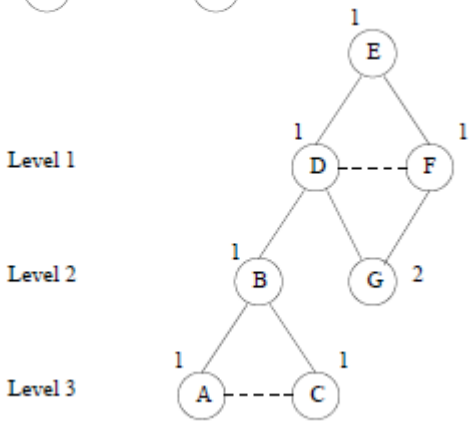
Computing Betweenness

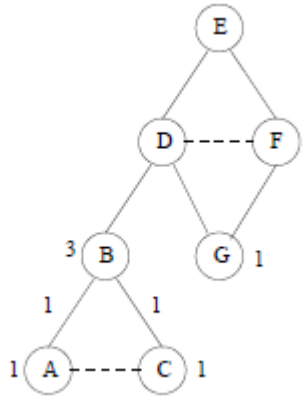
Repeat the process for all nodes

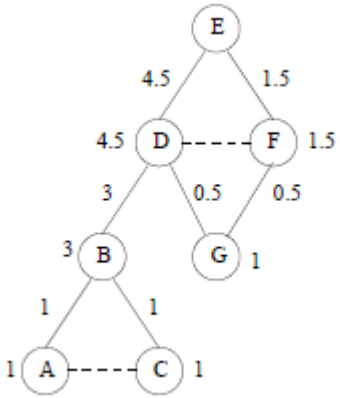
Sum over all BFSs

Example

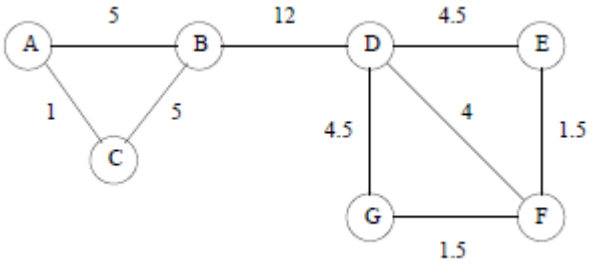

A
B
D
E
C
G
F

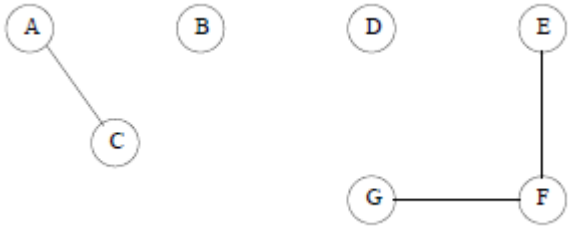

Level 1
Level 2
Level 3
1
E
1
1
D
F
1
1
B
G
2
1
1
A
C


E
D
F
3
B
G
1
1
1
1
A
C
1


E
4.5
1.5
4.5
D
F
1.5
3
0.5
0.5
3
B
G
1
1
1
1
A
C
1

Example





Computing Betweenness

Issues

- Test for connectivity?
- Re-compute all paths, or only those affected
- Parallel computation
- “Sampling

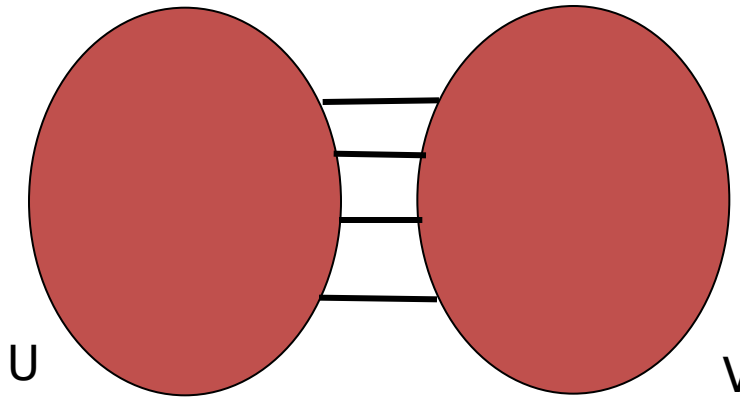
Other approaches to graph partitioning

- The general problem
  - Input: a graph G=(V,E)
    - edge (u,v) denotes similarity between u and v
    - weighted graphs: weight of edge captures the degree of similarity
  - Partitioning as an optimization problem:
    - Partition the nodes in the graph such that nodes within clusters are well interconnected (high edge weights), and nodes across clusters are sparsely interconnected (low edge weights)
    - most graph partitioning problems are NP hard

Measuring connectivity

- What does it mean that a set of nodes are well or sparsely interconnected?

- min-cut: the min number of edges such that when removed cause the graph to become disconnected
  - small min-cut implies sparse connectivity
  - $\min_{U} E(U, V-U) = \sum_{i \in U} \sum_{j \in V-U} A[i, j]$



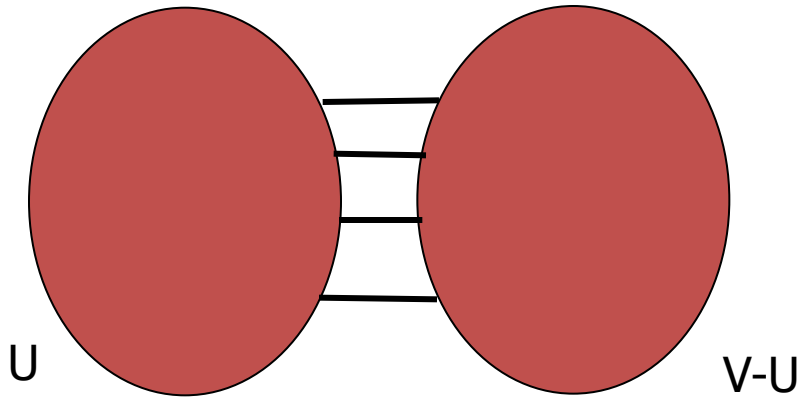
This problem can be solved in polynomial time

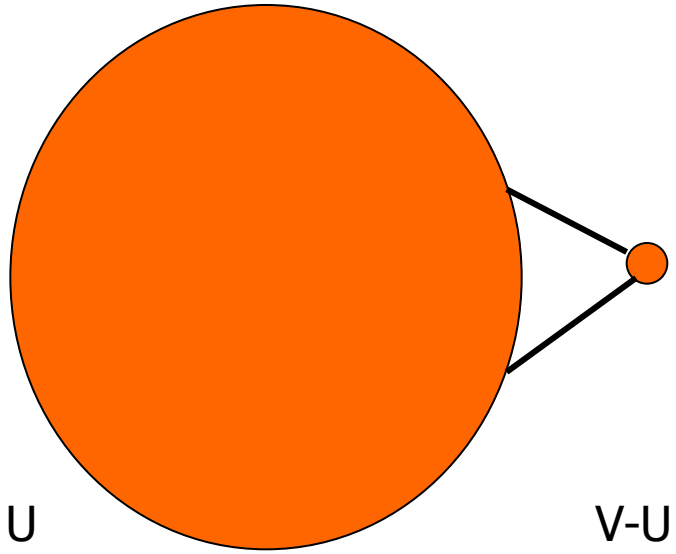
Min-cut/Max-flow algorithm

Measuring connectivity

- What does it mean that a set of nodes are well interconnected?

- min-cut: the min number of edges such that when removed cause the graph to become disconnected
  - not always a good idea!





A bad example

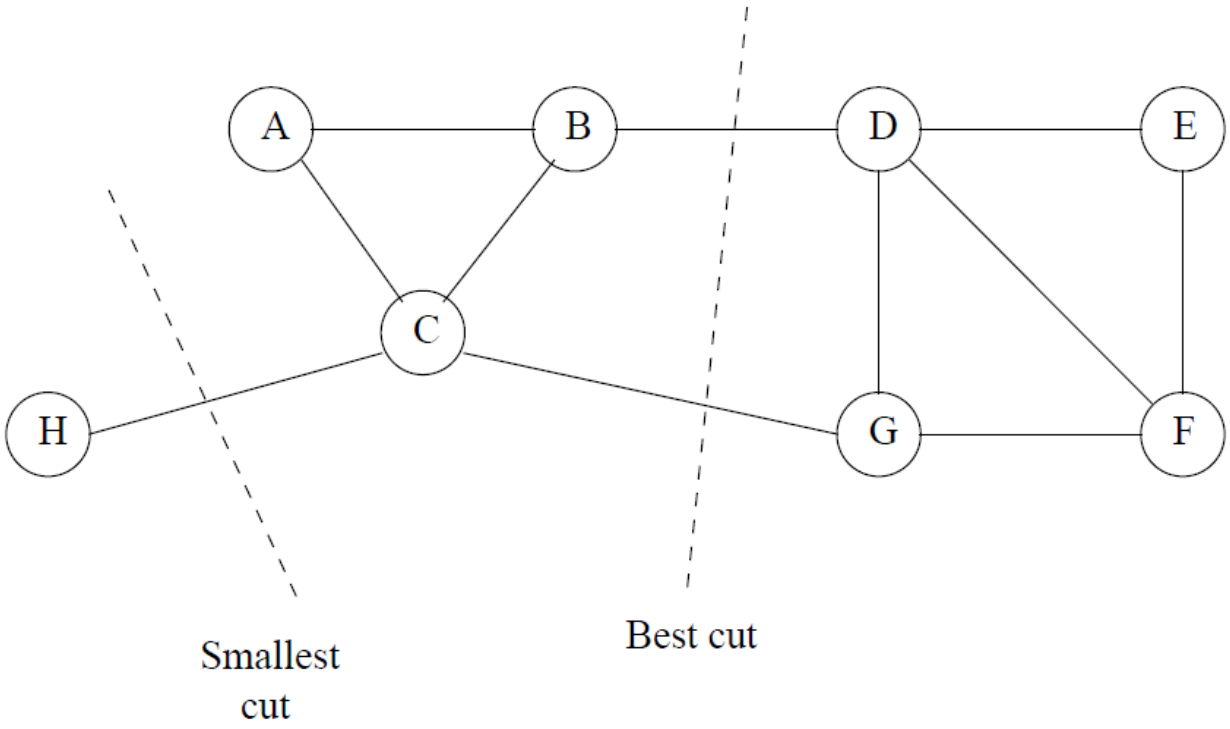


Figure 10.11: The smallest cut might not be the best cut

Graph expansion

- Normalize the cut by the size of the smallest component
- Cut ratio:

$$\alpha = \frac{E(U, V-U)}{\min\{|U|, |V-U|\}}$$

- Graph expansion:

$$\alpha(G) = \min_{U} \frac{E(U, V-U)}{\min\{|U|, |V-U|\}}$$

- Other Normalized Cut Ratio:

$$\beta = \frac{E(U,V-U)}{Vol(U)} + \frac{E(U,V-U)}{Vol(V-U)}$$

Vol(U) = number of edges with one endpoint in U
= total degree of nodes in U

Spectral analysis

- The Laplacian matrix $L = D - A$ where
  - A = the adjacency matrix
  - $D = \text{diag}(d_1, d_2, \ldots, d_n)$
    - d_i = degree of node i

- Therefore
  - $L(i,i) = d_i$
  - $L(i,j) = -1$, if there is an edge (i,j)

Laplacian Matrix properties

- The matrix L is symmetric and positive semi-definite
  - all eigenvalues of L are positive

- The matrix L has 0 as an eigenvalue, and corresponding eigenvector $w_1 = (1,1,\ldots,1)$
  - $\lambda_1 = 0$ is the smallest eigenvalue

The second smallest eigenvalue

- The second smallest eigenvalue (also known as Fielder value) λ_2 satisfies

$$\lambda_2 = \min_{x \perp w_1, \|x\|=1} x^T L x$$

- The eigenvector for eigenvalue λ_2 is called the Fielder vector. It minimizes

$$\lambda_2 = \min_{x \neq 0} \sum_{(i,j) \in E} (x_i - x_j)^2 \quad \text{where} \quad \sum_i x_i = 0$$

Spectral ordering

- The values of x minimize

$$\min_{x \neq 0} \sum_{(i,j) \in E} (x_i - x_j)^2 \qquad \sum_i x_i = 0$$

- For weighted matrices

$$\min_{x \neq 0} \sum_{(i,j)} A[i,j](x_i - x_j)^2 \qquad \sum_i x_i = 0$$

- The ordering according to the x_i values will group similar (connected) nodes together

- Physical interpretation: The stable state of springs placed on the edges of the graph

Spectral partition

- Partition the nodes according to the ordering induced by the Fielder vector
- If $u = (u_1, u_2, \ldots, u_n)$ is the Fielder vector, then split nodes according to a threshold value s
  - bisection: s is the median value in u
  - ratio cut: s is the value that minimizes α
  - sign: separate positive and negative values ($s=0$)
  - gap: separate according to the largest gap in the values of u
- This works well (provably for special cases)

Fielder Value

- The value λ_2 is a good approximation of the graph expansion

$$\frac{\alpha(G)^2}{2d} \leq \lambda_2 \leq 2\alpha(G)$$

d = maximum degree

$$\frac{\lambda_2}{2} \leq \alpha(G) \leq \sqrt{\lambda_2(2d-\lambda_2)}$$

- For the minimum ratio cut of the Fielder vector we have that

$$\frac{\alpha^2}{2d} \leq \lambda_2 \leq 2\alpha(G)$$

- If the max degree d is bounded we obtain a good approximation of the minimum expansion cut

Chapter 4, from D. Easley and J. Kleinberg book

NETWORKS AND SURROUNDING CONTEXTS

Introduction

Surrounding context: factors other than node and edges that affect how the network structure evolves

Homophily: people tend to be similar to their friends

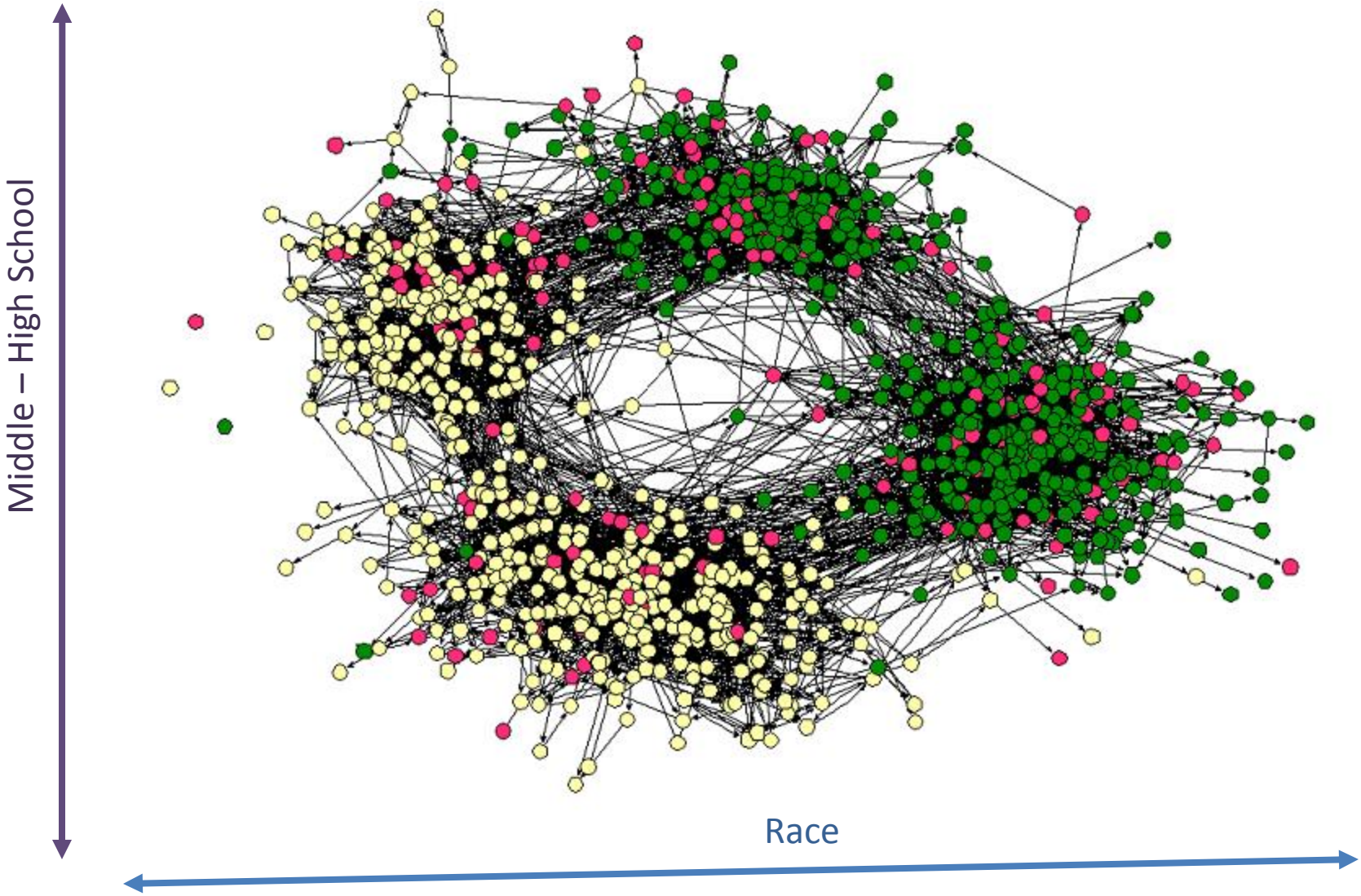
Αριστοτέλης love those who are like themselves

Πλάτωνα Όμοιος ομοίω αεί πελάζει (similarity begets friendship)

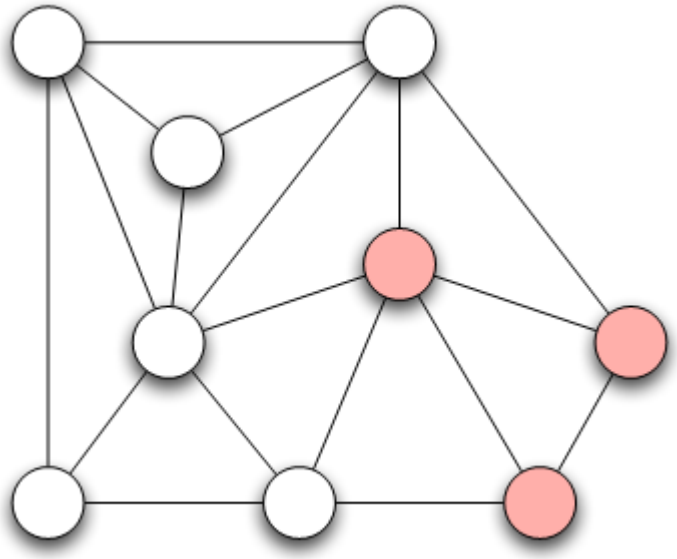
Birds of a feather flock together

Factors intrinsic to the network (introduced by a common friend) and contextual factors (eg attend the same school)

Homophily



Measuring Homophily



If the fraction of cross-gender edges is significantly less than expected, then there is evidence for homophily

gender male with probability p
gender female with probability q

Probability of cross-gender edge?

$$\frac{\#cross\_gender\_edges}{\#edges} << 2pq$$

Measuring Homophily

- “significantly” less than
- Inverse homophily
- Characteristics with more than two values:
  - Number of heterogeneous edges (edge between two nodes that are different)

Mechanisms Underlying Homophily: Selection and Social Influence

Selection: tendency of people to form friendships with others who are like then

Socialization or Social Influence: the existing social connections in a network are influencing the individual characteristics of the individuals

Social Influence as the inverse of Selection

Mutable & immutable characteristics

The Interplay of Selection and Social Influence

Longitudinal studies in which the social connections and the behaviors within a group are tracked over a period of time

Why?

- Study teenagers, scholastic achievements/drug use (peer pressure and selection)
- Relative impact?
- Effect of possible interventions (example, drug use)

The Interplay of Selection and Social Influence

Christakis and Fowler on obesity, 12,000 people over a period of 32-years

People more similar on obesity status to the network neighbors than if assigned randomly

Why?
(i) Because of selection effects, choose friends of similar obesity status,
(ii) Because of confounding effects of homophily according to other characteristics that correlate with obesity
(iii) Because changes in the obesity status of person's friends was exerting an influence that affected her

(iii) As well -> "contagion" in a social sense

Tracking Link Formation in Online Data: interplay between selection and social influence

- Underlying social network
- Measure for behavioral similarity

Wikipedia
Node: Wikipedia editor who maintains a user account and user talk page
Link: if they have communicated with one writing on the user talk page of the other

Editor's behavior: set of articles she has edited

$$\frac{|N_A \cap N_B|}{|N_A \cup N_B|}$$

Neighborhood overlap in the bipartite affiliation network of editors and articles consisting only of edges between editors and the articles they have edited

FACT: Wikipedia editors who have communicated are significantly more similar in their behavior than pairs of Wikipedia editors who have not (homomphily), **why?**
Selection (editors form connections with those have edited the same articles) vs Social Influence (editors are led to the articles of people they talk to)

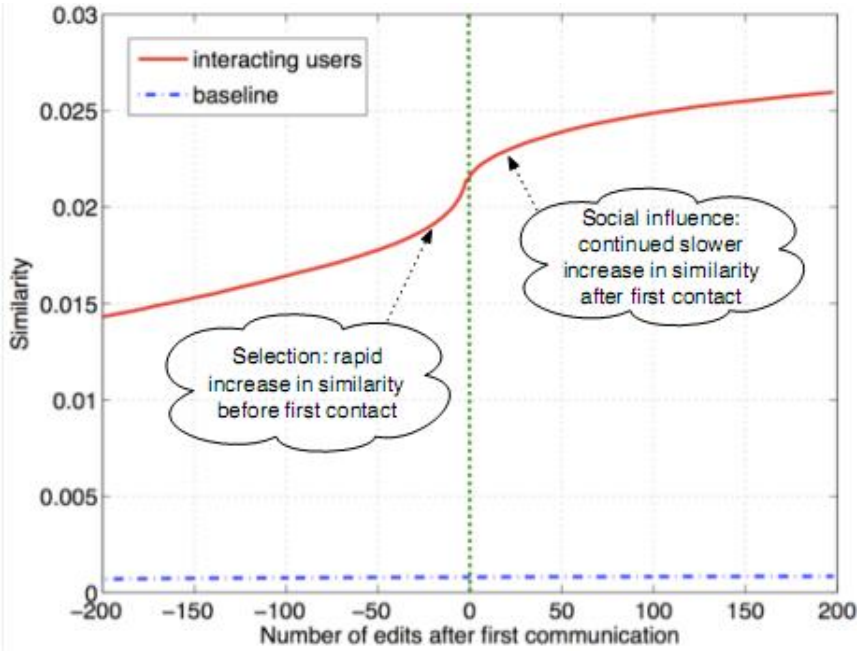
Tracking Link Formation in Online Data: interplay between selection and social influence

Actions in Wikipedia are time-stamped

For each pair of editors A and B who have ever communicated,

- Record their similarity over time
- Time 0 when they first communicated -- Time moves in discrete units, advancing by one "tick" whenever either A or B performs an action on Wikipedia
- Plot one curve for each pair of editors

Average, single plot: average level of similarity relative to the time of first interaction



Similarity is clearly increasing both before and after the moment of first interaction (both selection and social influence)

Not symmetric around time 0 (particular role on similarity): Significant increase before they meet

Blue line shows similarity of a random pair (non-interacting)