Strength of Weak Ties and Community Structure in Networks

CS224W: Social and Information Network Analysis

Jure Leskovec, Stanford University

http://cs224w.stanford.edu

Networks: Flow of Information

- How information flows through the network?
- How different nodes play structurally distinct roles in this process?
- How different links (short range vs. long range) play different roles in diffusion?

[Granovetter '73]

Strength of Weak Ties

- **How people find out about new jobs?**
  - Mark Granovetter, part of his PhD in 1960s
  - People find the information through personal contacts
- **But:** Contacts were often **acquaintances** rather than close friends
  - **This is surprising:**
    - One would expect your friends to help you out more than casual acquaintances when you are between the jobs
- **Why is it that distance acquaintances are most helpful?**

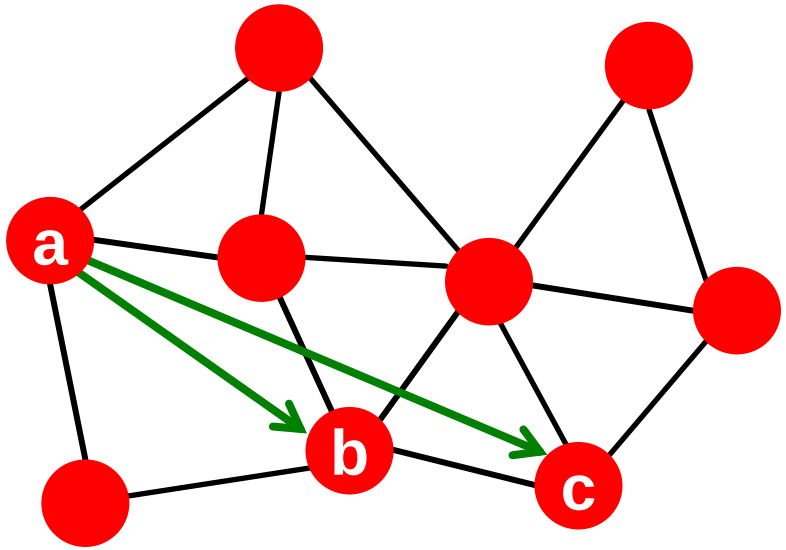
[Granovetter '73]

Granovetter's Answer

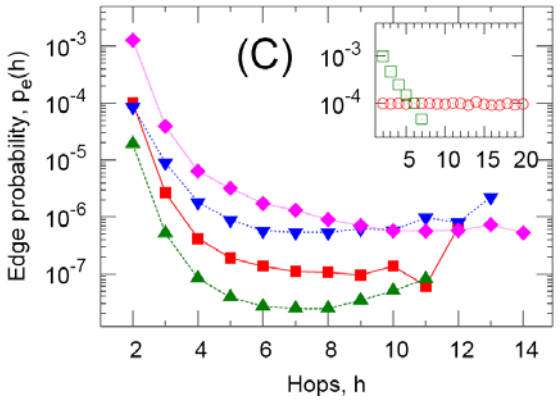
- **Two perspectives on friendships:**
  - **Structural:**
    - Friendships span different portions of the network
  - **Interpersonal:**
    - Friendship between two people is either **strong** or **weak**

Triadic Closure

- **Which edge is more likely a-b or a-c?**



- **Triadic closure:** If two people in a network have a friend in common there is an increased likelihood they will become friends themselves



Triadic Closure

- **Triadic closure == High clustering coefficient**

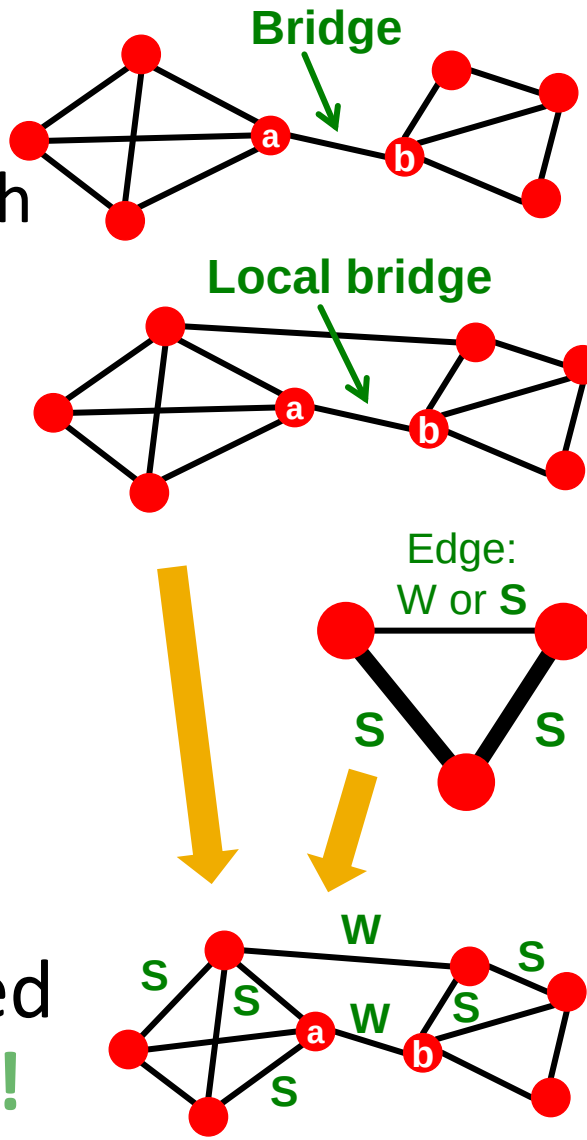
Reasons for triadic closure:

- If B and C have a friend A in common, then:
  - B is **more likely to meet** C
    - (since they both spend time with A)
  - B and C **trust** each other
    - (since they have a friend in common)
  - A has **incentive** to bring B and C together
    - (as it is hard for A to maintain two disjoint relationships)

- **Empirical study by Bearman and Moody:**
  - Teenage girls with low clustering coefficient are more likely to contemplate suicide

Granovetter's Explanation

- Define: **Bridge edge**
  - If removed, it disconnects the graph
- Define: **Local bridge**
  - Edge not in a triangle
- **Two types of edges:**
  - **Strong** (friend) and **weak ties** (acquaintance)
- **Strong triadic closure:**
  - Two strong ties imply a third edge
- If strong triadic closure is satisfied then **local bridges are weak ties!**



Tie strength in real data

- **For many years the Granovetter's theory was not tested**
- But, today we have large who-talks-to-whom graphs:
  - Email, Messenger, Cell phones, Facebook
- **Onnela et al. 2007:**
  - Cell-phone network of 20% of country's population
  - **Edge strength:** # phone calls

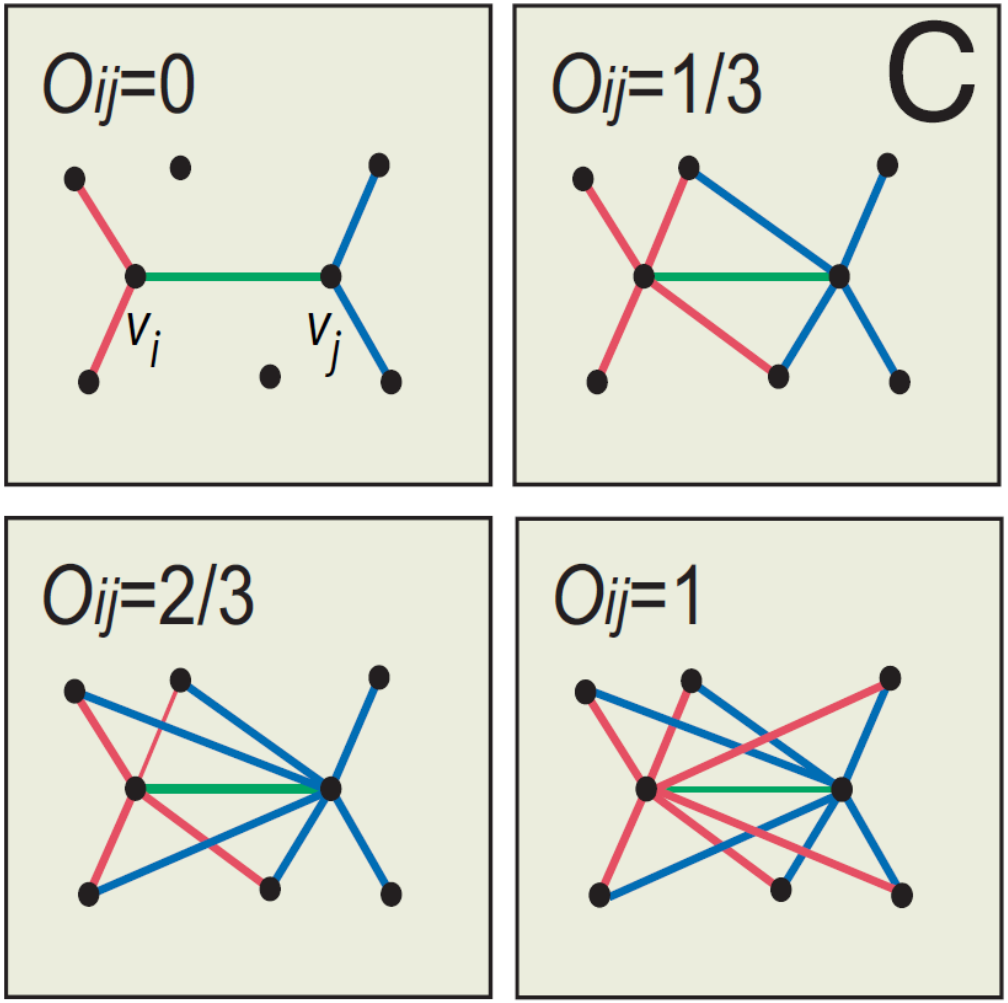
Neighborhood Overlap

- **Edge overlap:**

$$O_{ij} = \frac{N(i) \cap N(j)}{N(i) \cup N(j)}$$

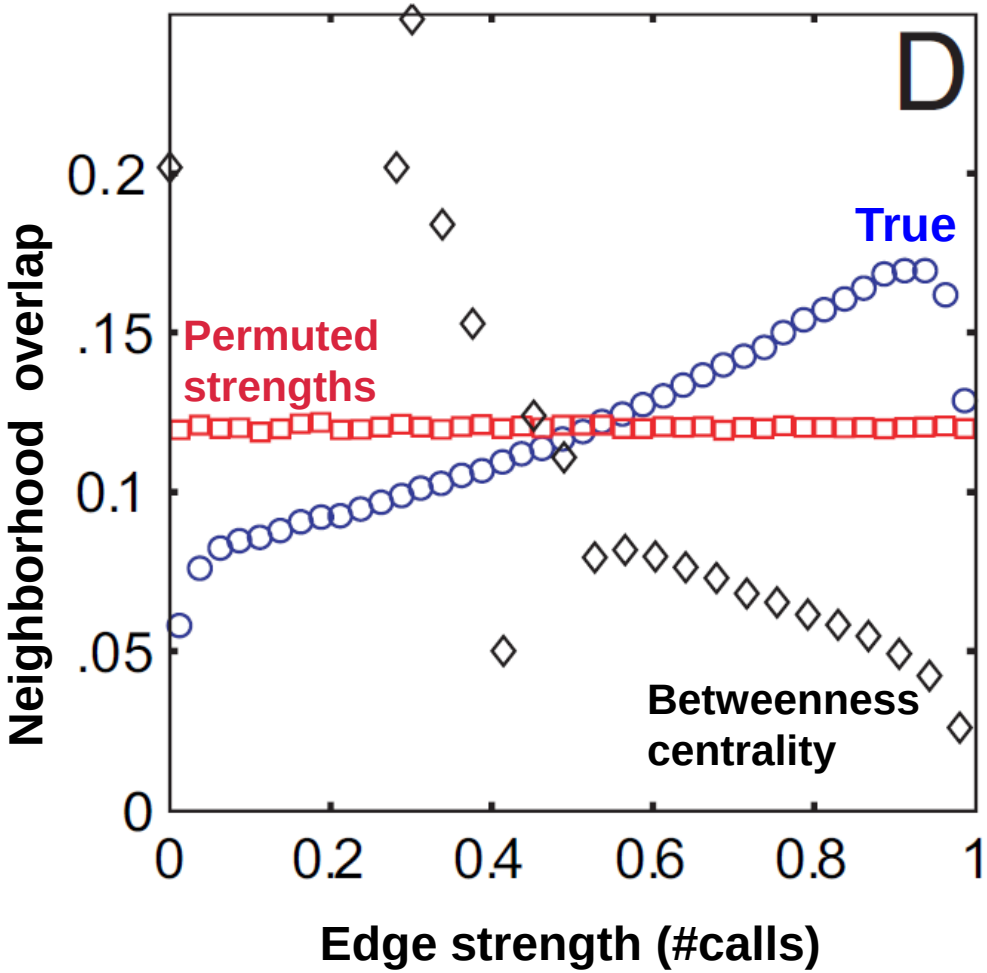
  - n(i) … set of neighbors of I

- Overlap $= 0$ when an edge is a **local bridge**

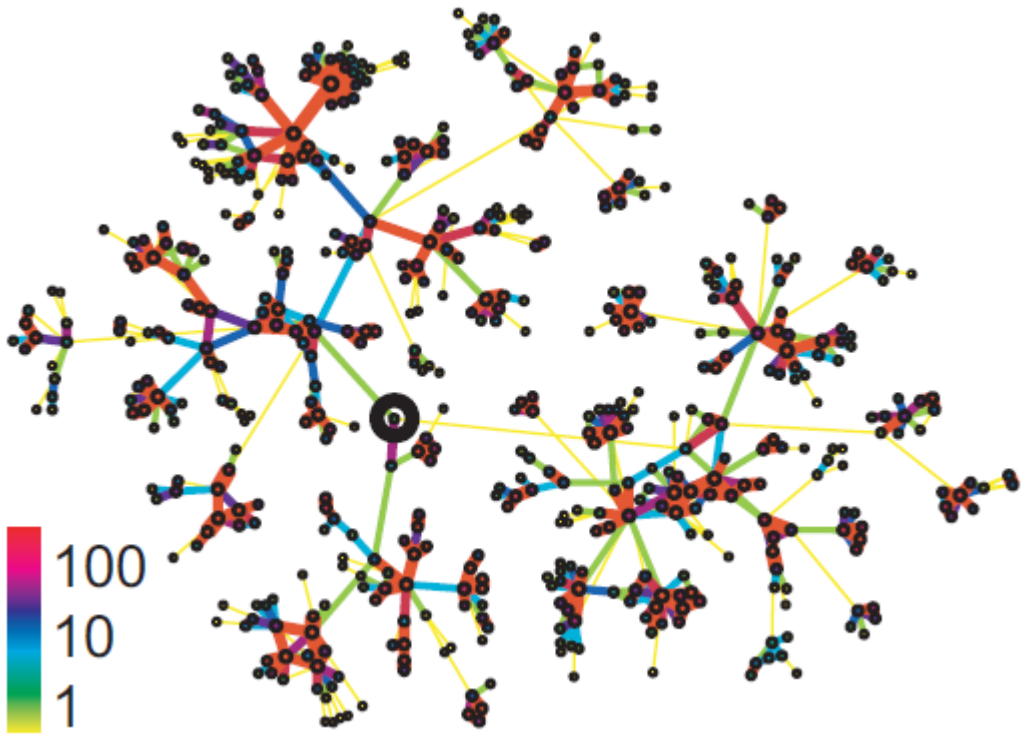


Phones: Edge Overlap vs. Strength

- **Cell-phone network**
- **Observation:**
  - Highly used links have high overlap!
- **Legend:**
  - **Permuted strengths:** Keep the network structure but randomly reassign edge strengths
  - **Betweenness centrality:** number of shortest paths going through an edge

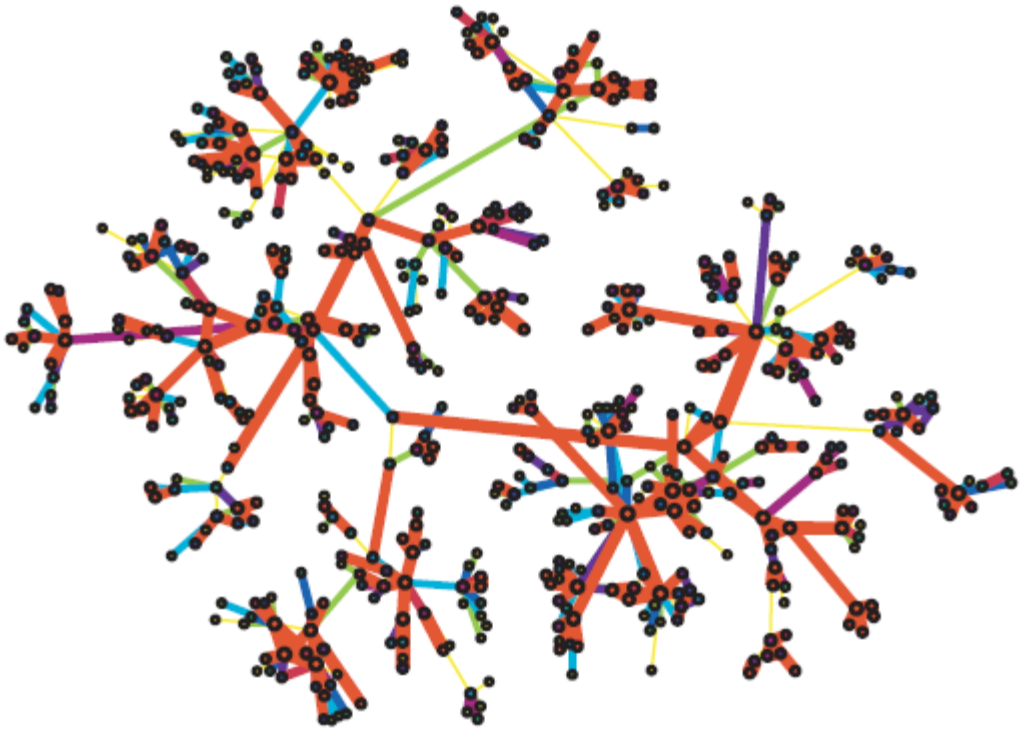


Real Network, Real Tie Strengths



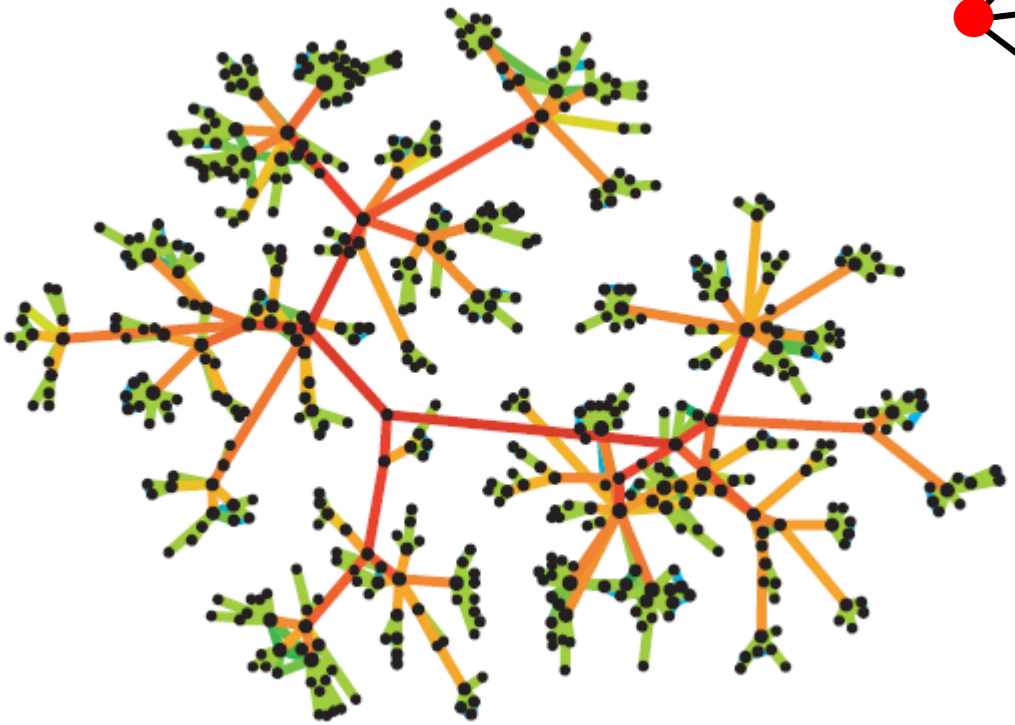
- **Real edge strengths in mobile call graph**
  - Strong ties are more embedded (have higher overlap)

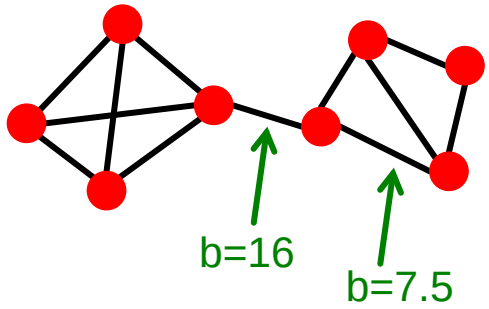
Real Net, Permuted Tie Strengths



- **Same network, same set of edge strengths**
  but now **strengths are randomly shuffled**

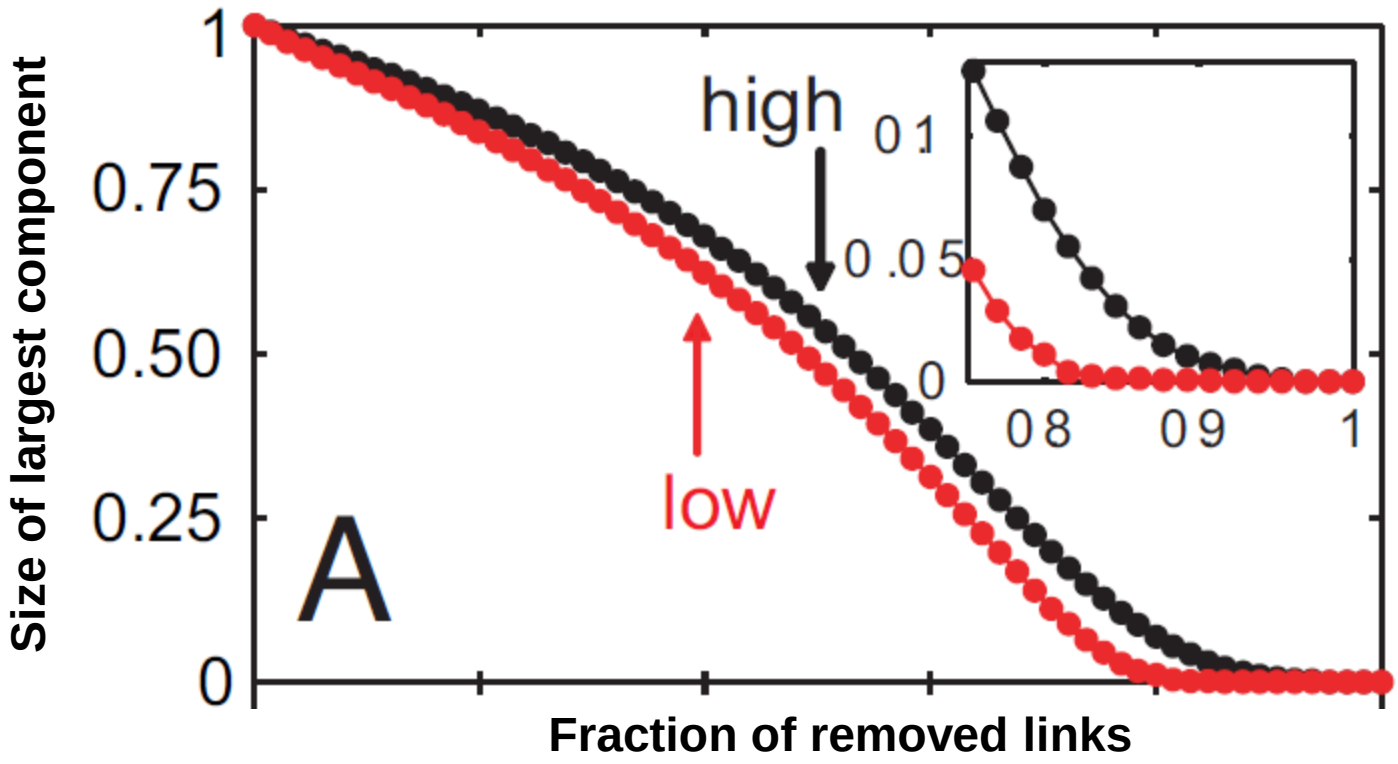
Edge Betweenness Centrality





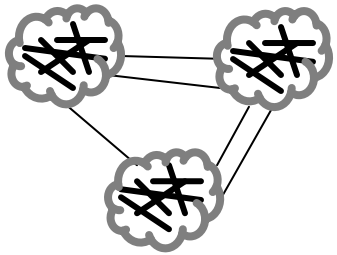
- Edges strength is labeled based on **betweenness centrality** (number of shortest paths passing through an edge)

Link Removal by Strength



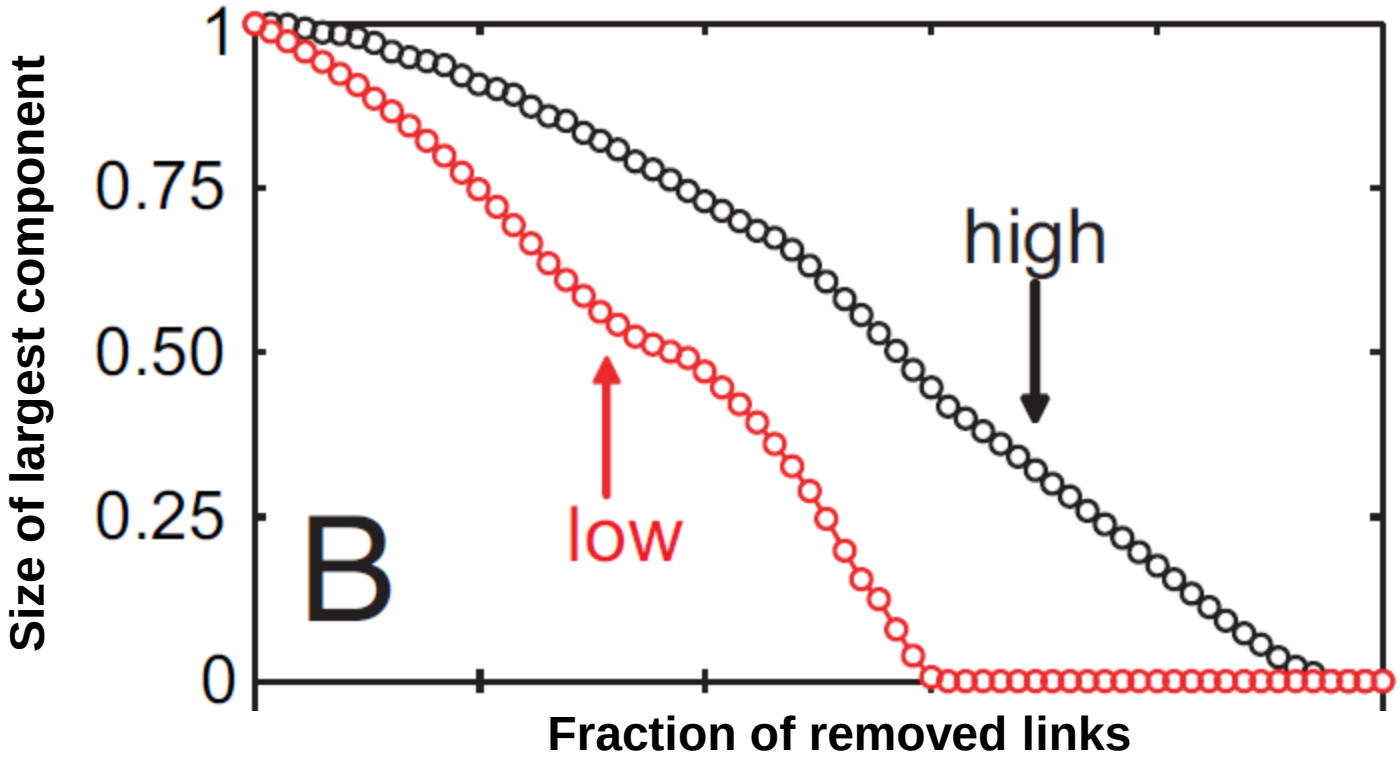
Low disconnects the network sooner

- Removing links by **strength (#calls)**
  - Low to high
  - High to low



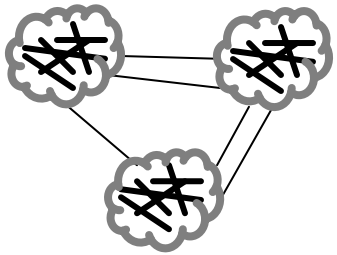
Conceptual picture of network structure

Link Removal by Overlap



Low disconnects the network sooner

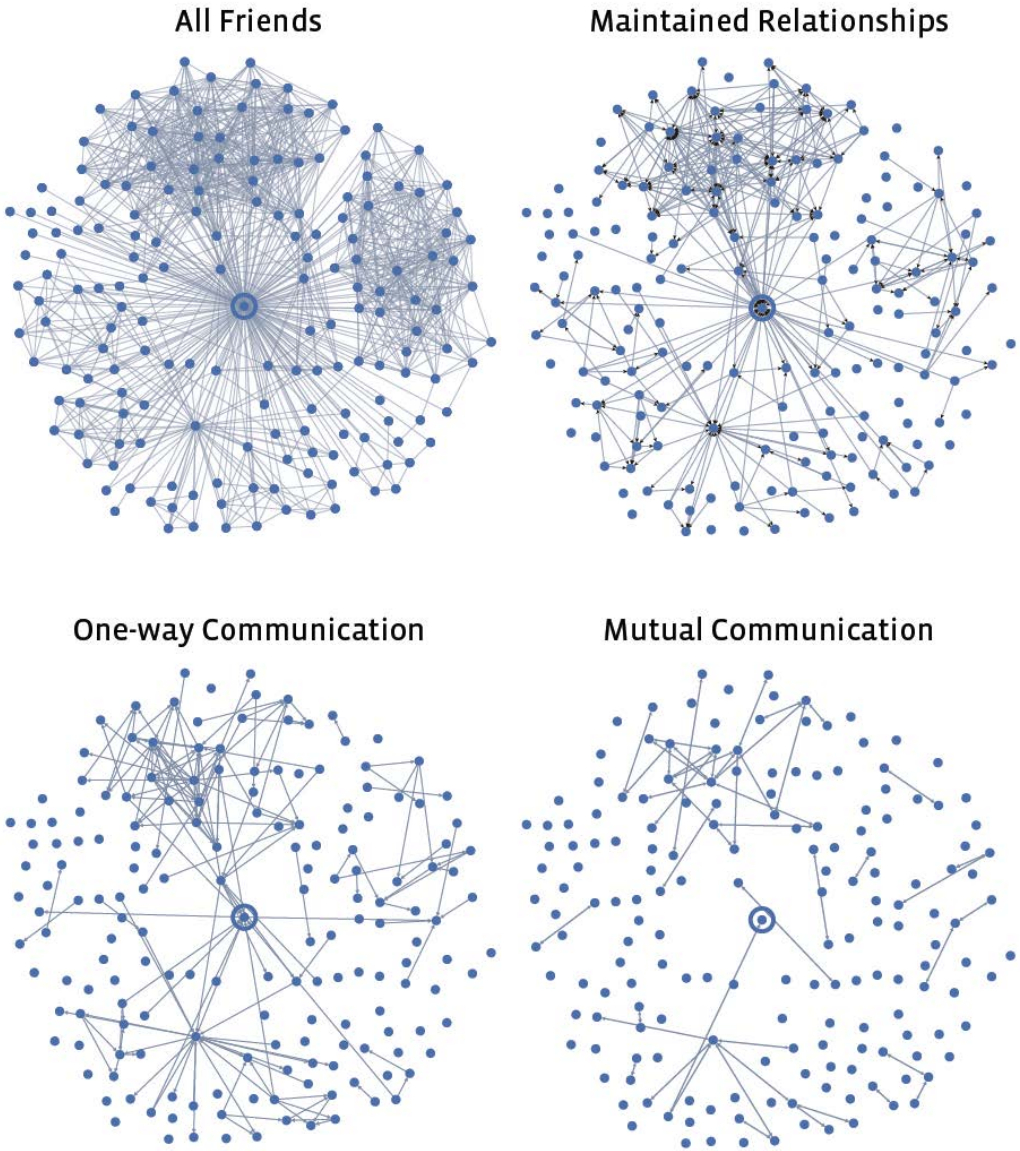
- Removing links based on **overlap**
  - Low to high
  - High to low



Conceptual picture of network structure

[Marlow et al. '09]

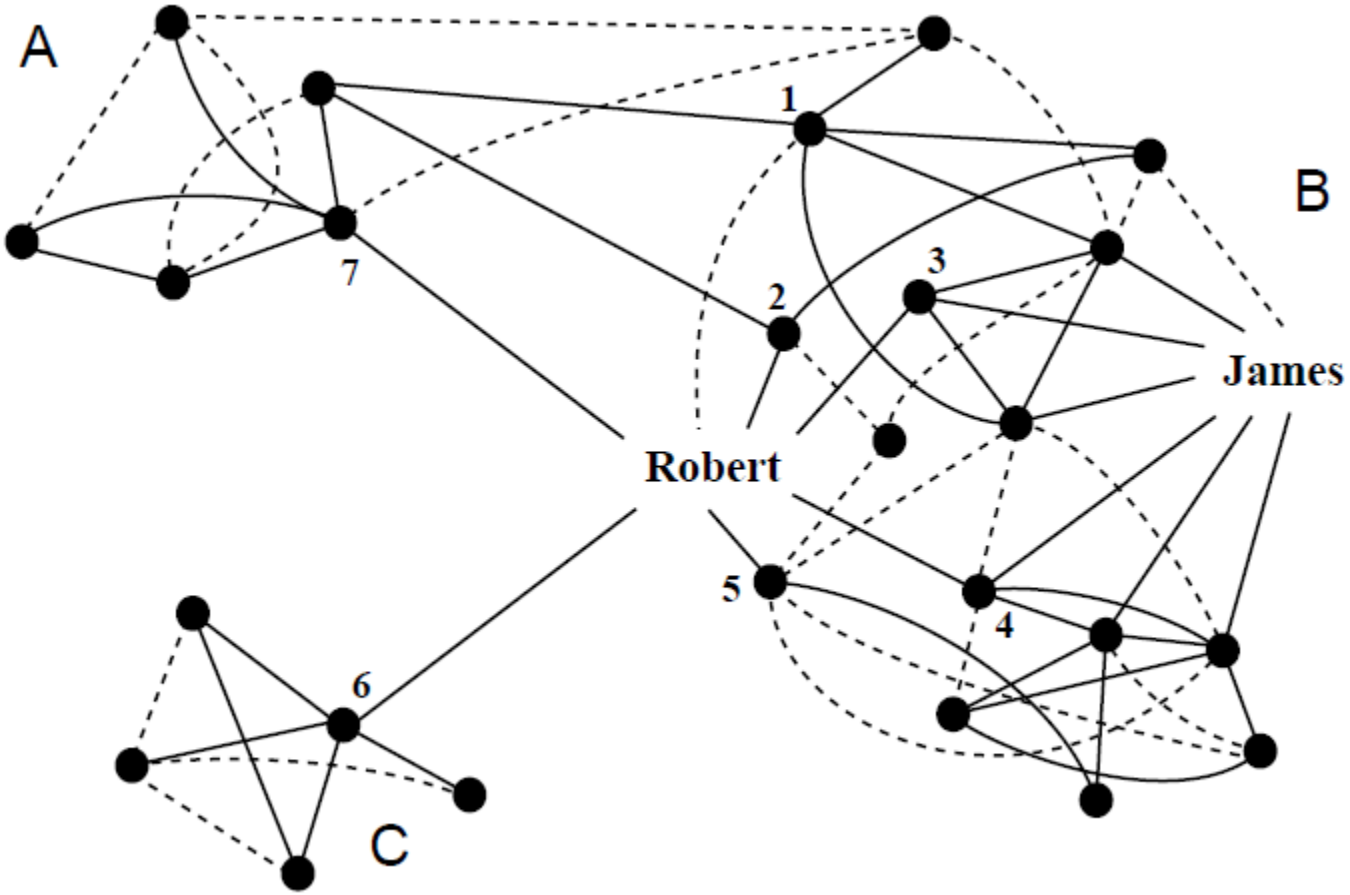
Another Example: Facebook



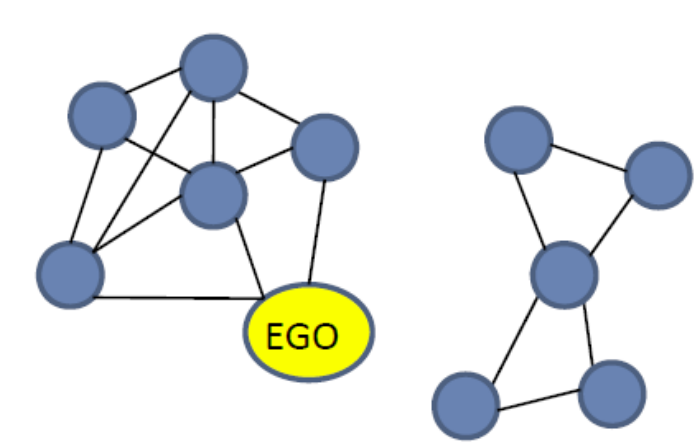
Small Detour: Structural Holes

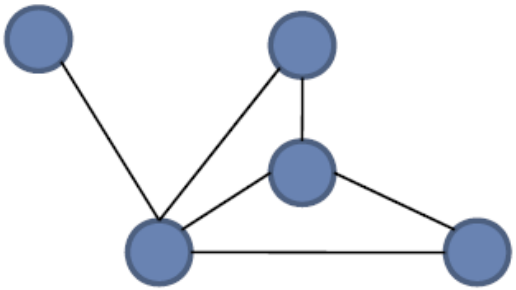
[Ron Burt]

Small Detour: Structural Holes

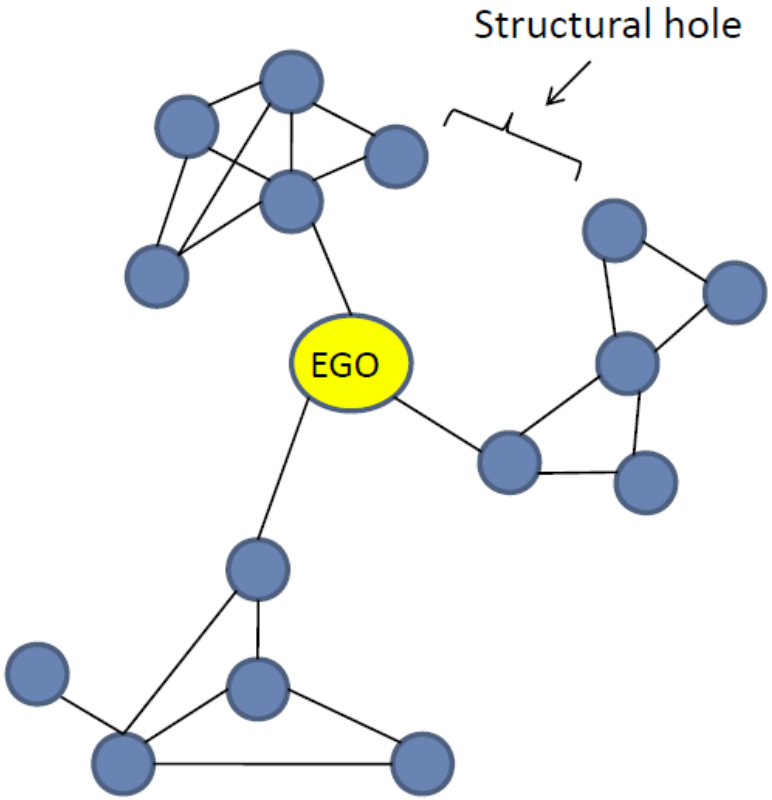


Structural Holes





Few structural holes

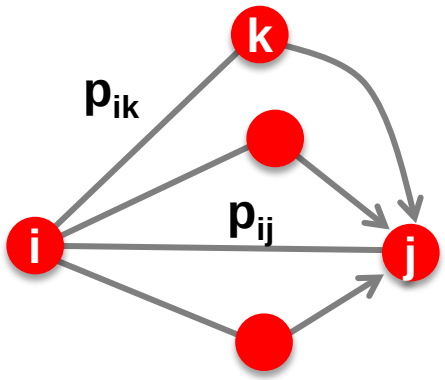


Many structural holes

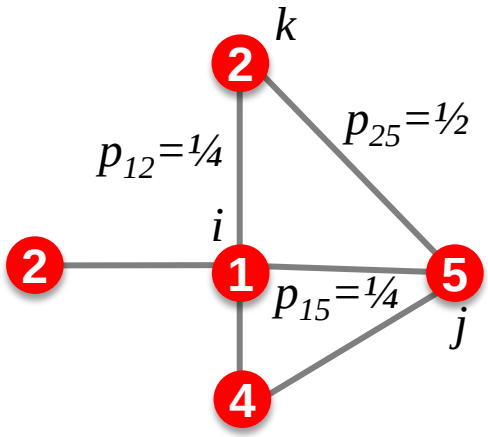
Structural Holes provide ego with access to novel information, power, freedom

Structural Holes: Network Constraint

- The "network constraint" measure [Burt]:
  - **To what extent are person's contacts redundant**



$p_{uv}=1/d_u$



- **Low**: disconnected contacts
- **High**: contacts that are close or strongly tied

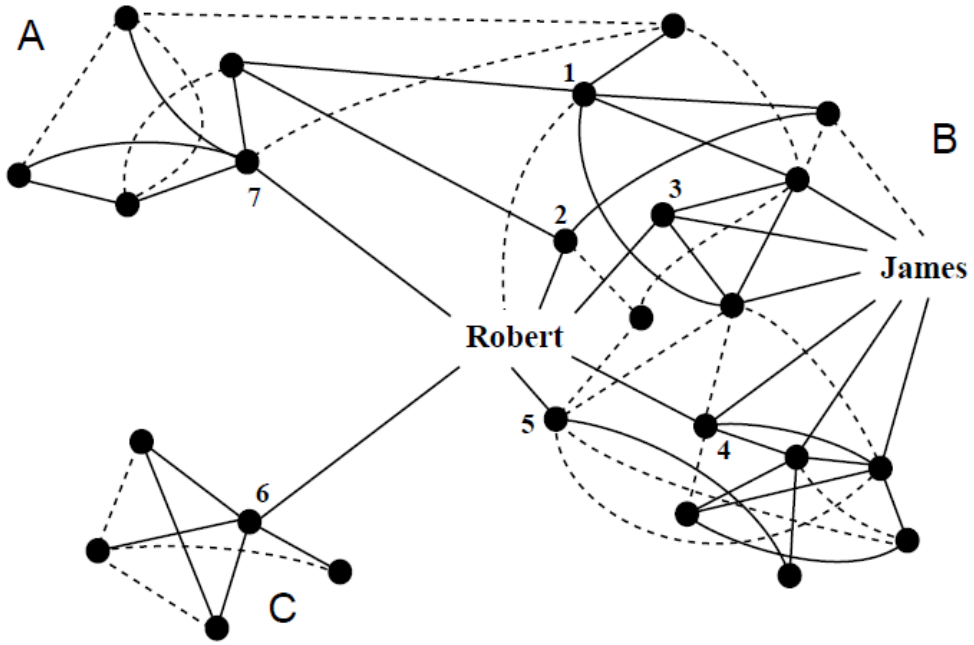
$$c_i = \sum_j c_{ij} = \sum_j \left[p_{ij} + \sum_k \left(p_{ik} p_{kj} \right) \right]^2$$

p_{uv} … prop. of u's "energy" invested in relationship with v

p

	1	2	3	4	5
1	.00	.25	.25	.25	.25
2	.50	.00	.00	.00	.50
3	1.0	.00	.00	.00	.00
4	.50	.00	.00	.00	.50
5	.33	.33	.00	.33	.00

Example: Robert vs. James

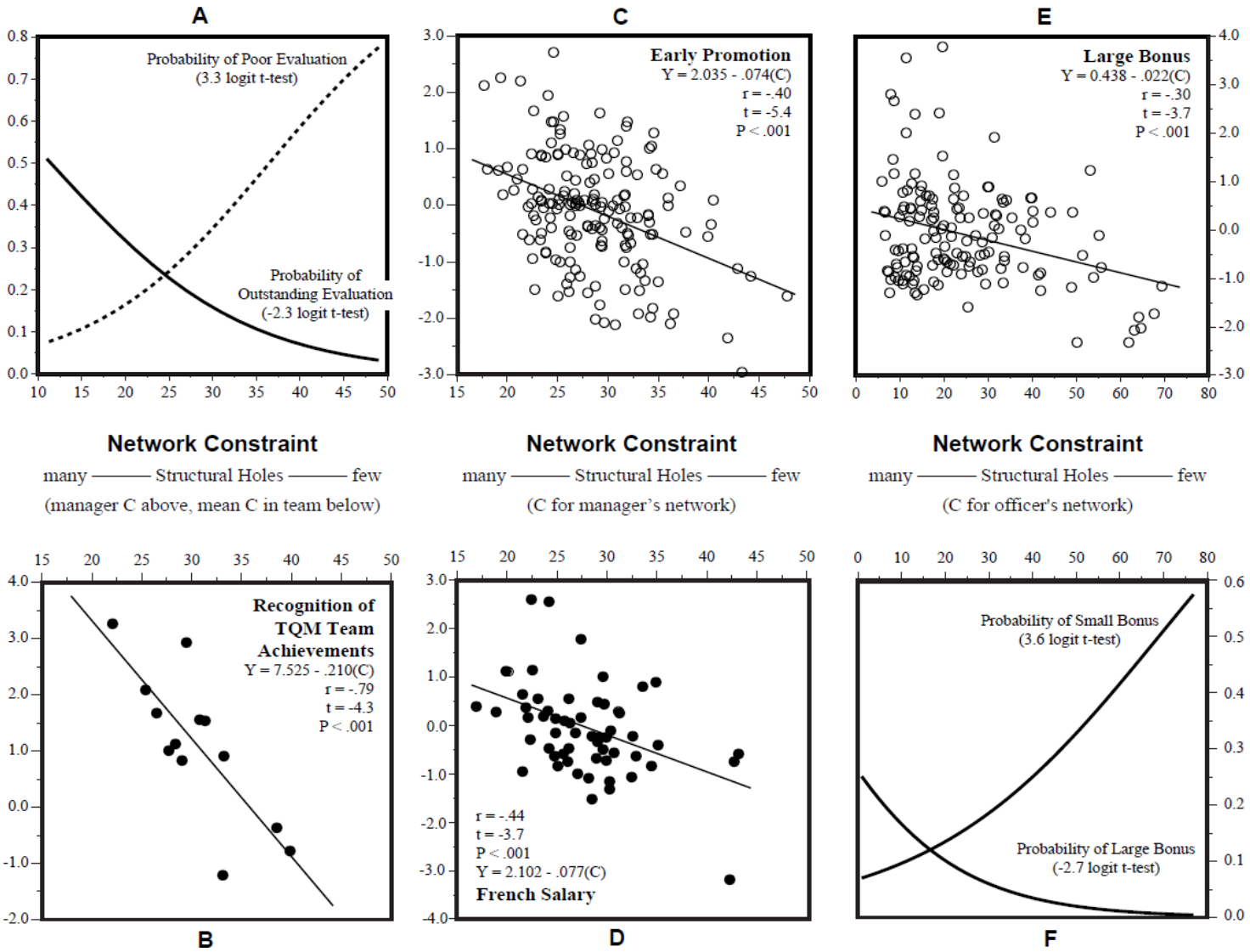


- **Network constraint:**
  - James: c_j=0.309
  - Robert: c_r=0.148

- **Constraint:** To what extent are person's contacts redundant
  - Low: disconnected contacts
  - High: contacts that are close or strongly tied

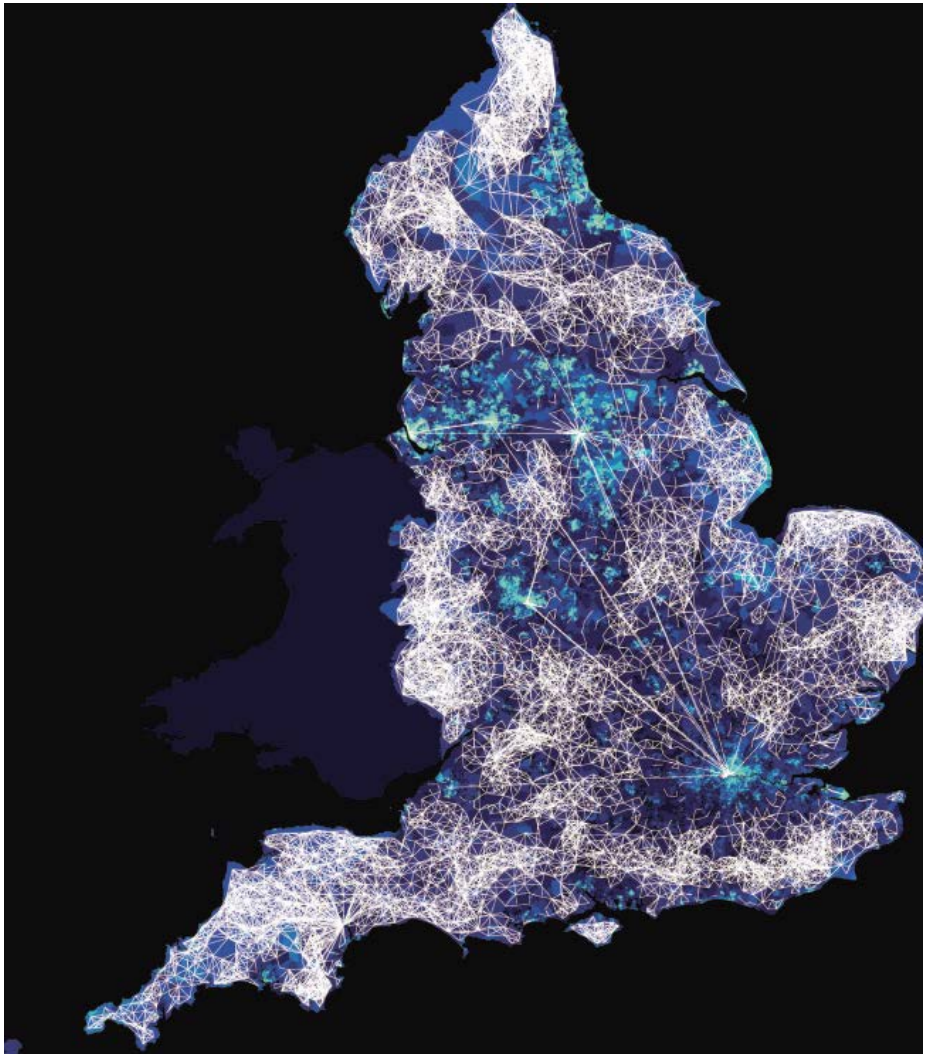
[Ron Burt]

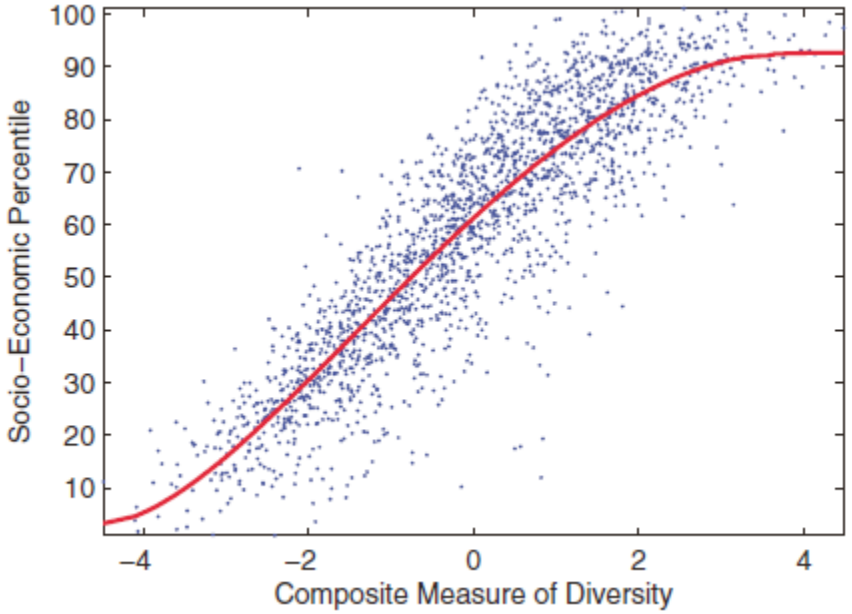
Spanning the Holes Matters



Diversity & Development

[Eagle-Macy, 2010]



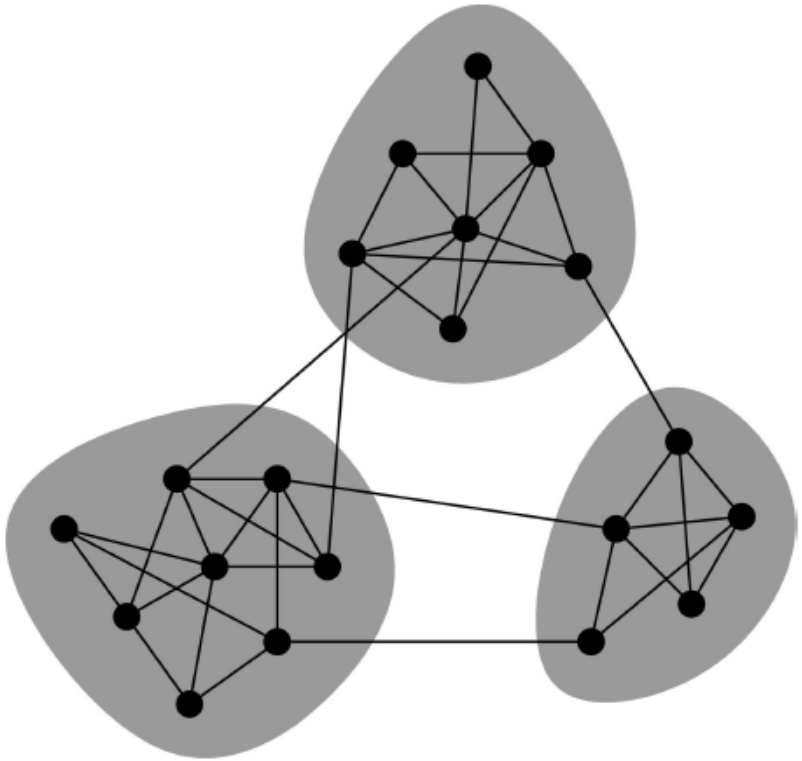


- **Measure of diversity:**
  - **≈ 1-c_i**
    - structural holes + entropy of edge strengths

Network Communities

Network Communities

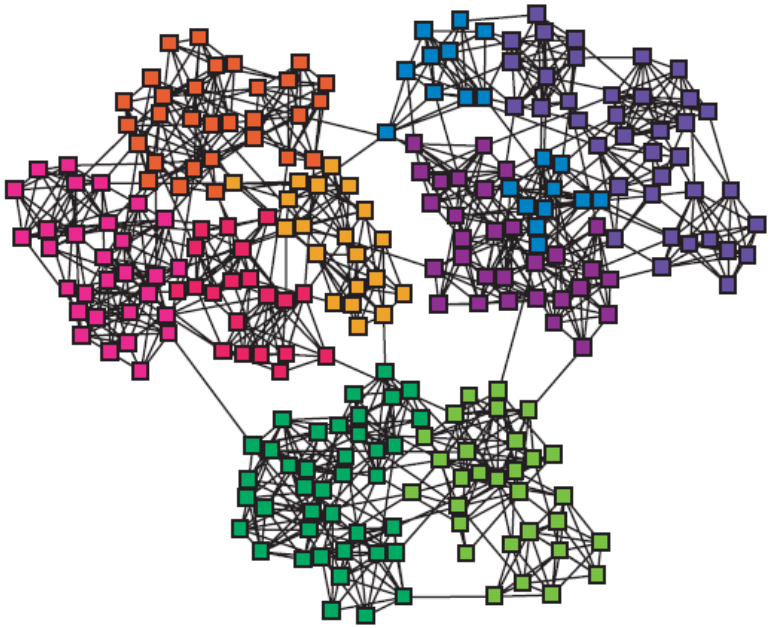
- Networks of **tightly connected groups**
- **Network communities:**
  - Sets of nodes with **lots** of connections **inside** and **few** to **outside** (the rest of the network)



Communities, clusters, groups, modules

Finding Network Communities

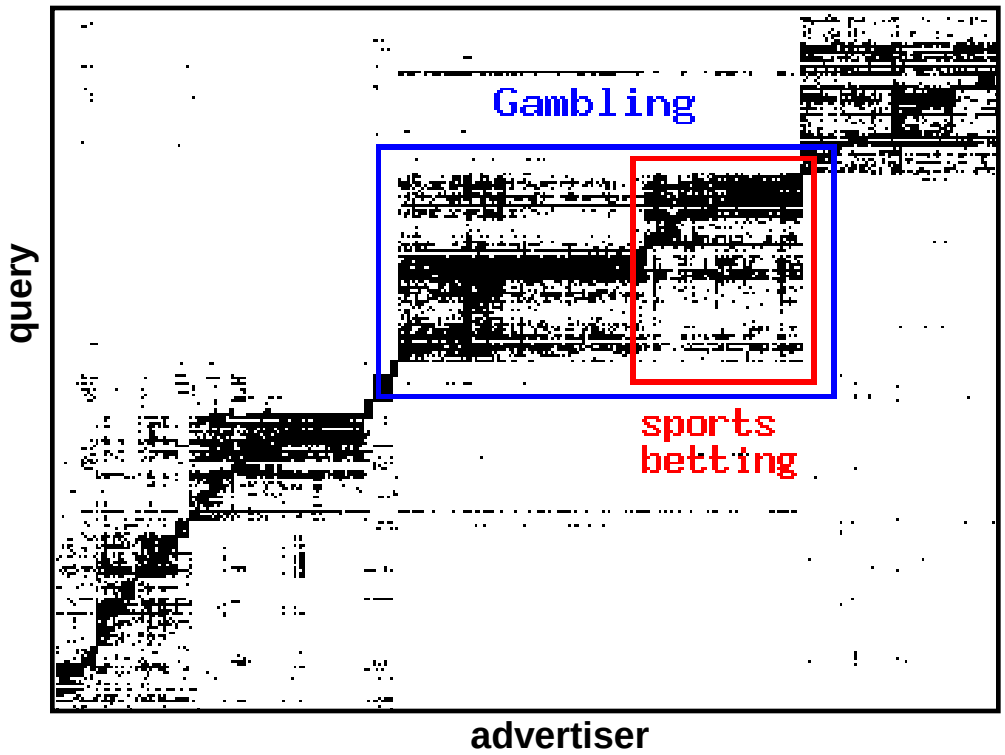
- **How to automatically find such densely connected groups of nodes?**
- Ideally such automatically detected clusters would then correspond to real groups
- For example:



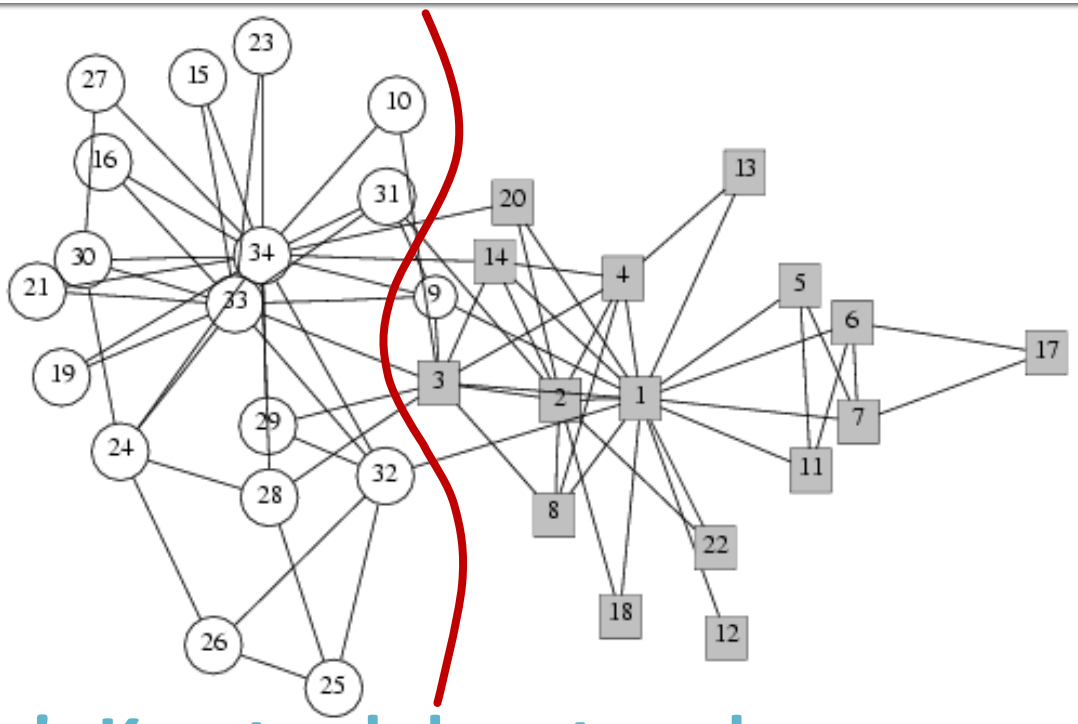
Communities, clusters, groups, modules

Micro-Markets in Sponsored Search

Find micro-markets by partitioning the "query x advertiser" graph:



Social Network Data

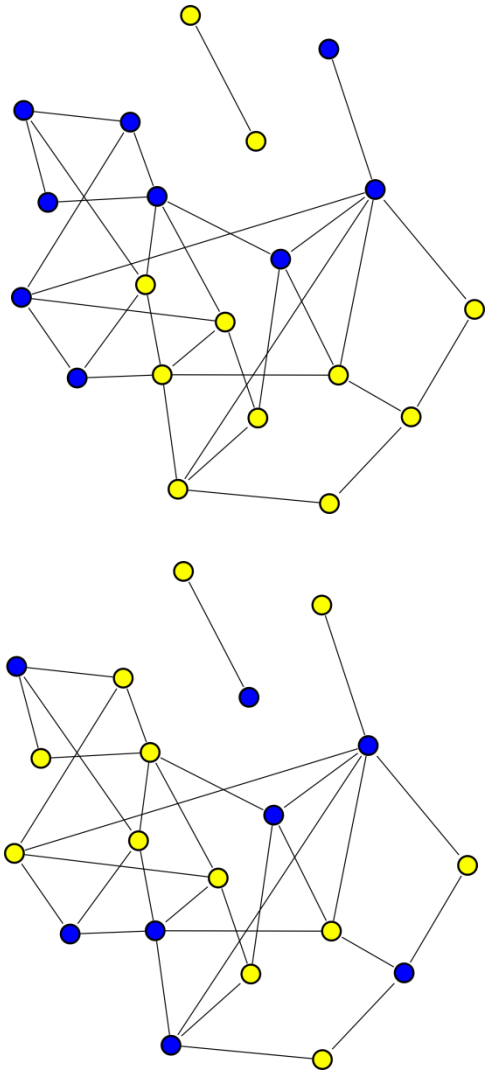


- **Zachary's Karate club network:**
  - Observe social ties and rivalries in a university karate club
  - During his observation, conflicts led the group to split
  - Split could be explained by a minimum cut in the network
- **Why would we expect such clusters to arise?**

[Backstrom et al. KDD '06]

Group Formation in Networks

- In a social network **nodes explicitly declare group membership:**
  - Facebook groups, Publication venue
- Can think of groups as **node colors**
- Gives **insights into social dynamics:**
  - Recruits friends? Memberships spread along edges
  - Doesn't recruit? Spread randomly
- **What factors influence a person's decision to join a group?**

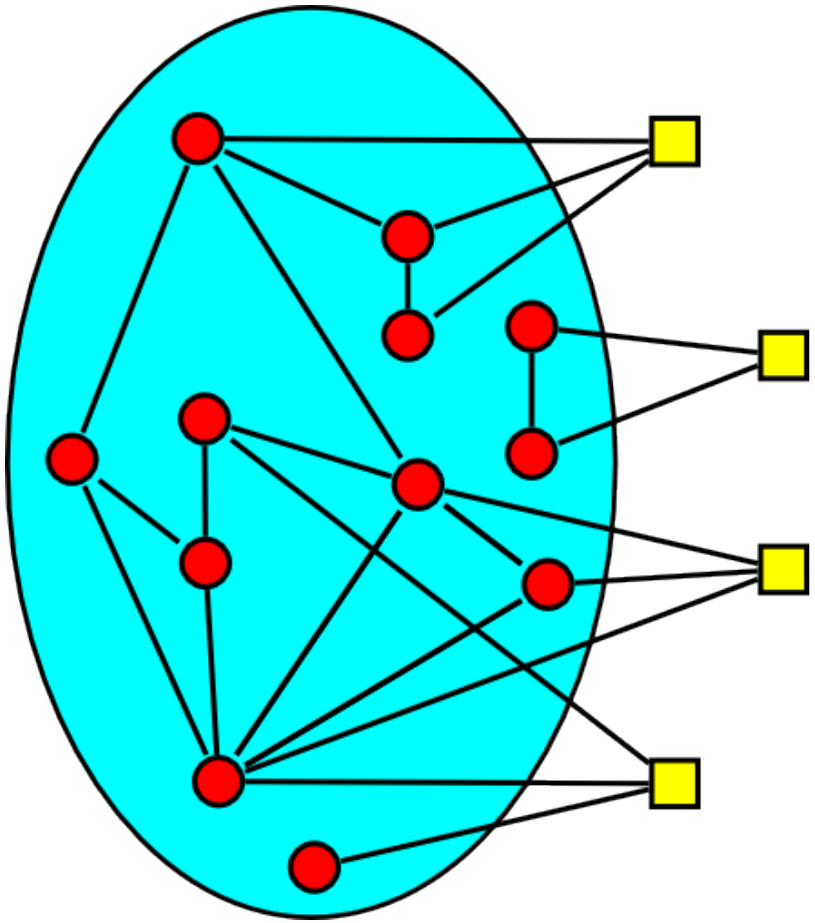


Group Growth as Diffusion

- **Analogous to diffusion**

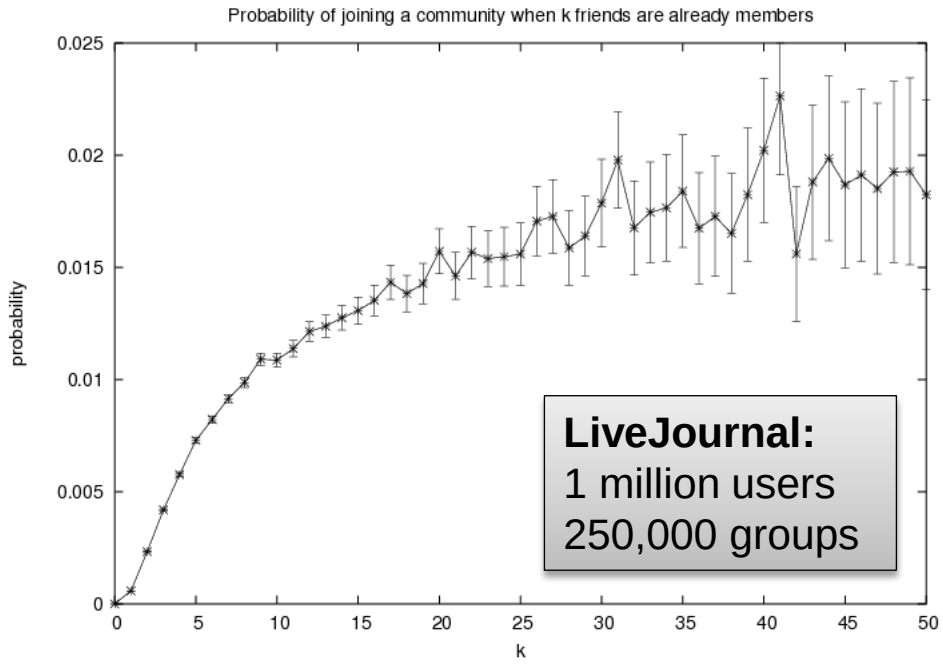
  Group memberships spread over the network:

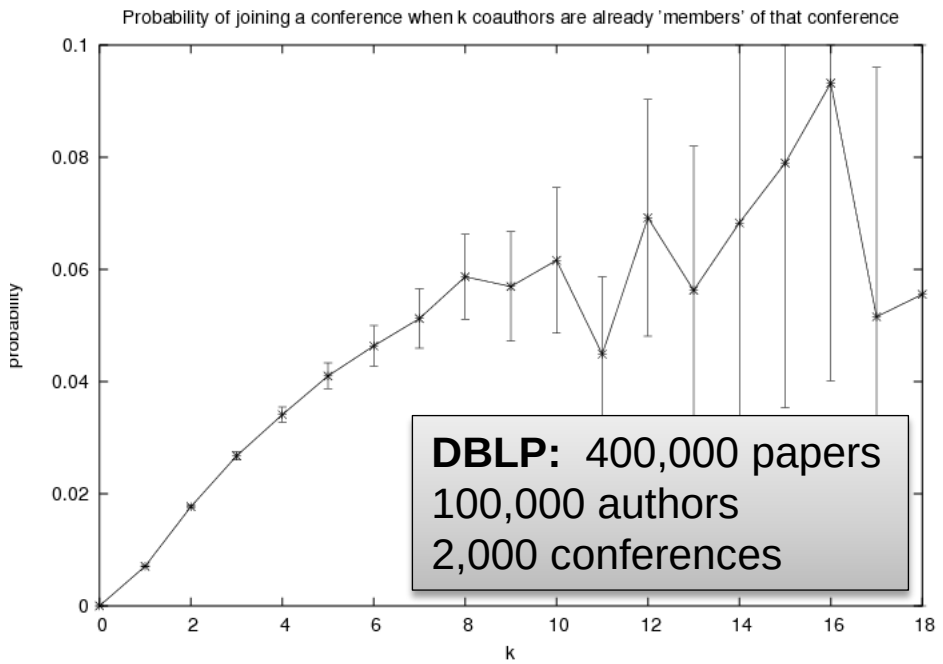
  - Red circles represent existing group members
  - Yellow squares may join
- **Question:**
  - How does prob. of joining a group depend on the number of friends already in the group?



[Backstrom et al. KDD '06]

P(join) vs. # friends in the group



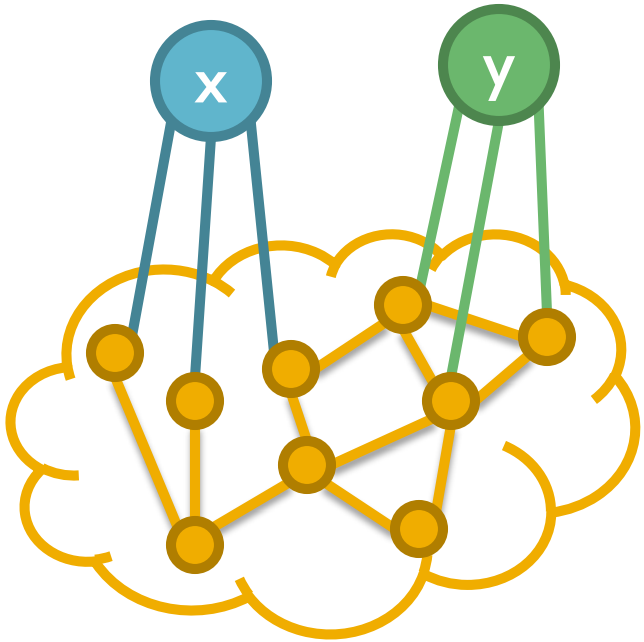


- **Diminishing returns:**
  - Probability of joining increases with the number of friends in the group
  - But increases get smaller and smaller

Groups: More Subtle Features

- **Connectedness of friends:**
  - *x* and *y* have three friends in the group
  - *x*'s friends are independent
  - *y*'s friends are all connected

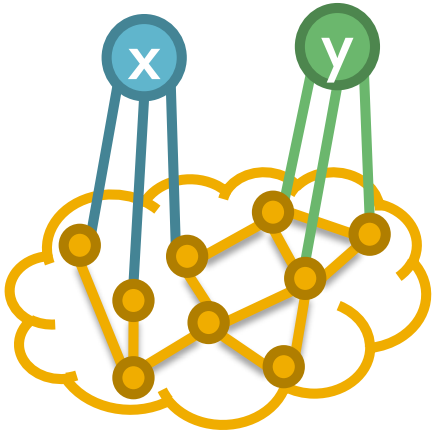
Who is more likely to join?



[Backstrom et al. KDD '06]

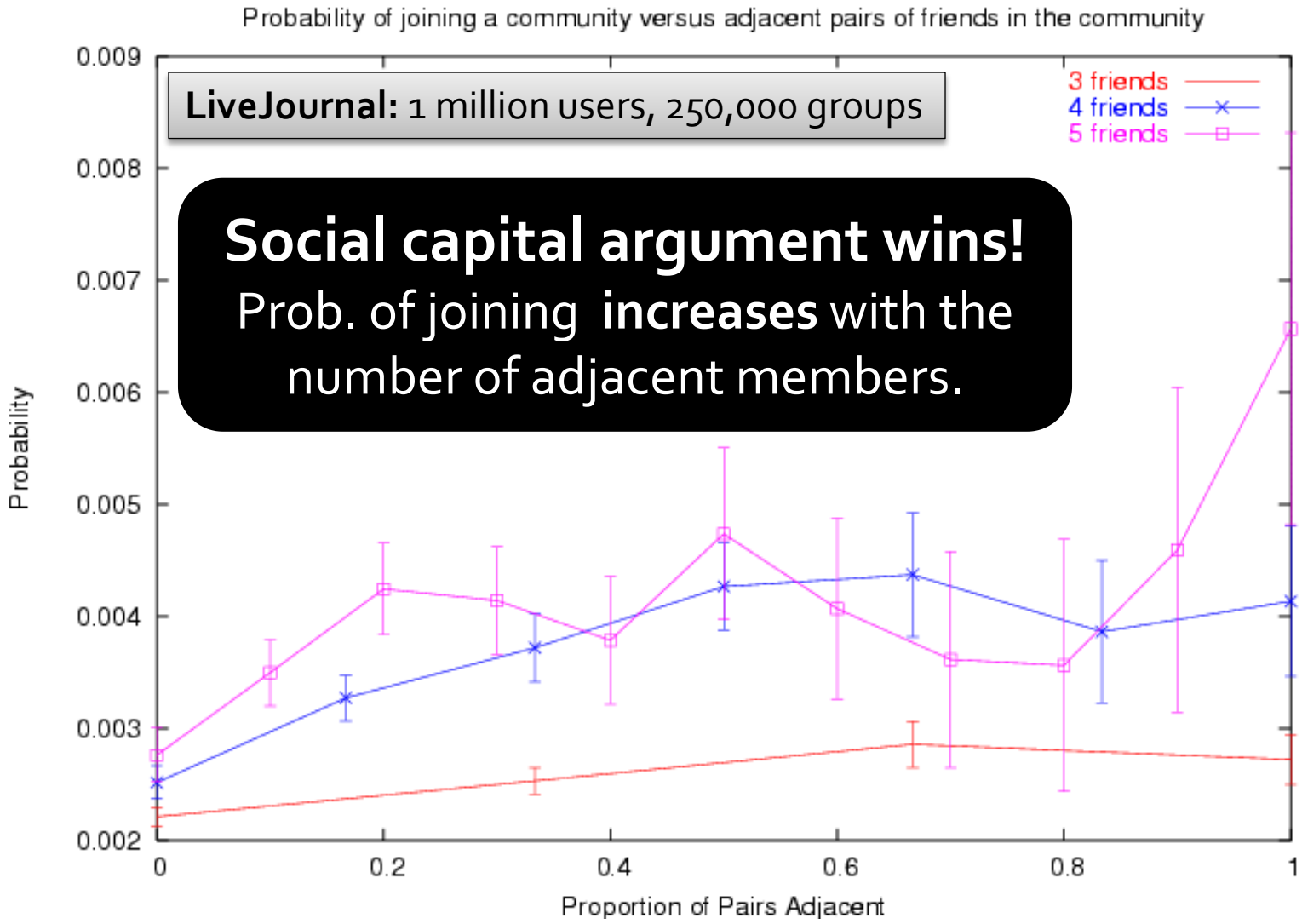
Connectedness of Friends

- **Competing sociological theories:**
  - Information argument [Granovetter '73]
  - Social capital argument [Coleman '88]



- **Information argument:**
  - Unconnected friends give independent support
- **Social capital argument:**
  - Safety/trust advantage in having friends who know each other

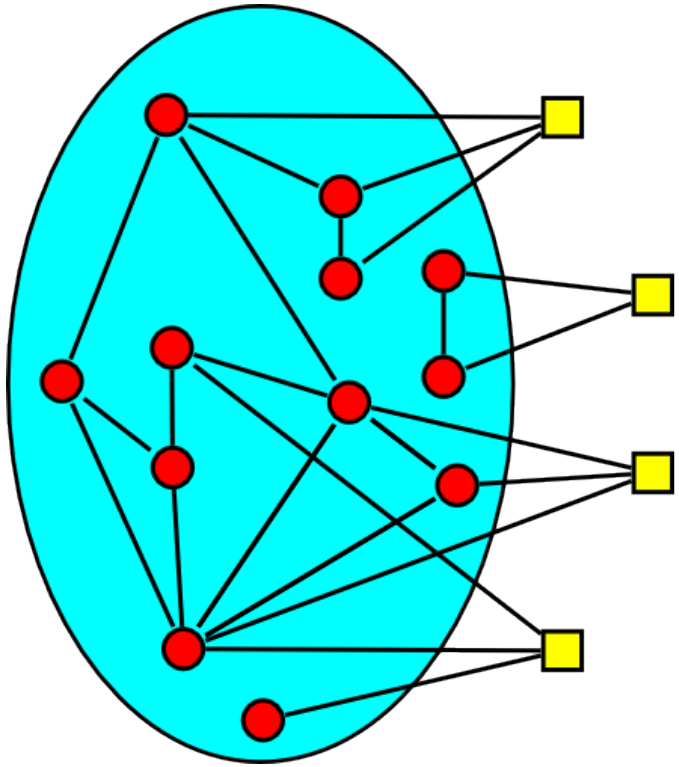
Connectedness of Friends



[Backstrom et al. KDD '06]

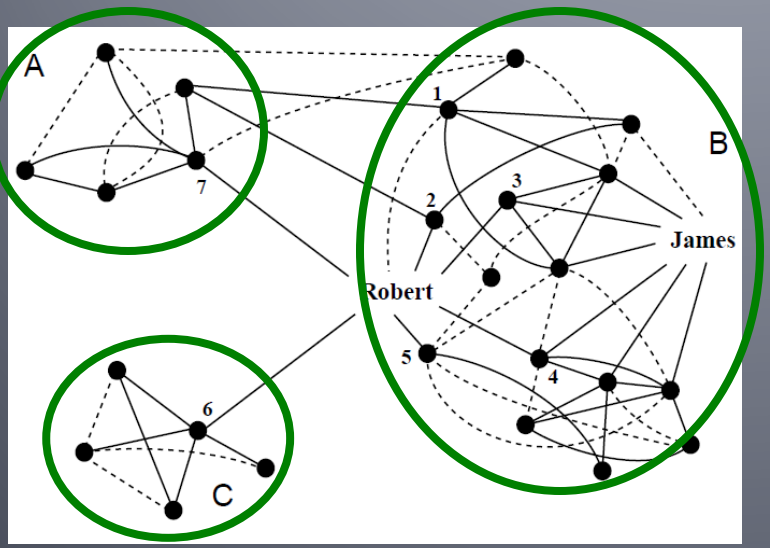
So, This Means That

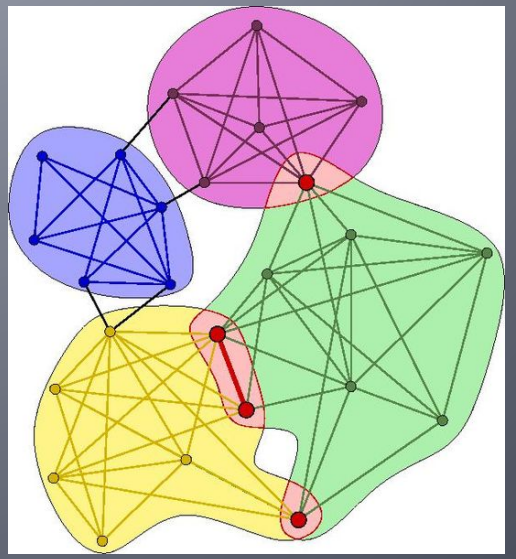
- **A person is more likely to join a group if**
  - she has more friends who are already in the group
  - friends have more connections between themselves
- **So, groups form clusters of tightly connected nodes**



Community Detection

How to find communities?

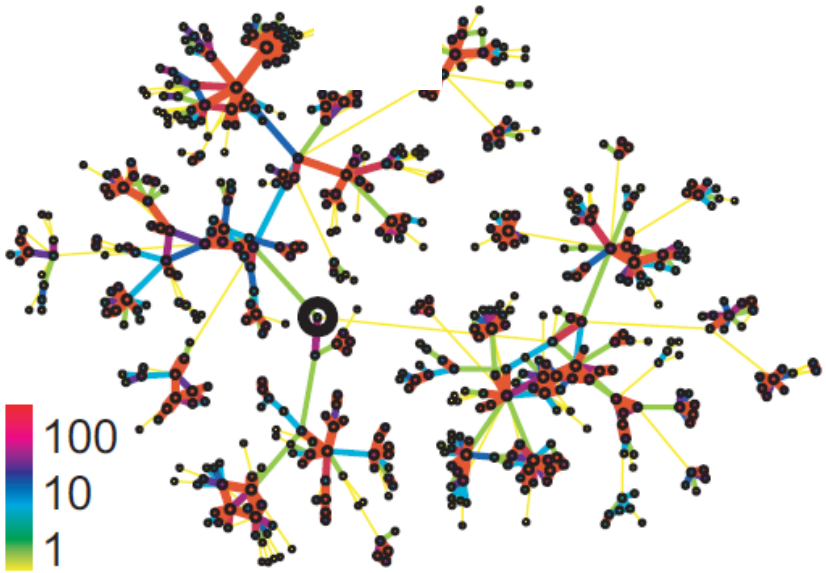




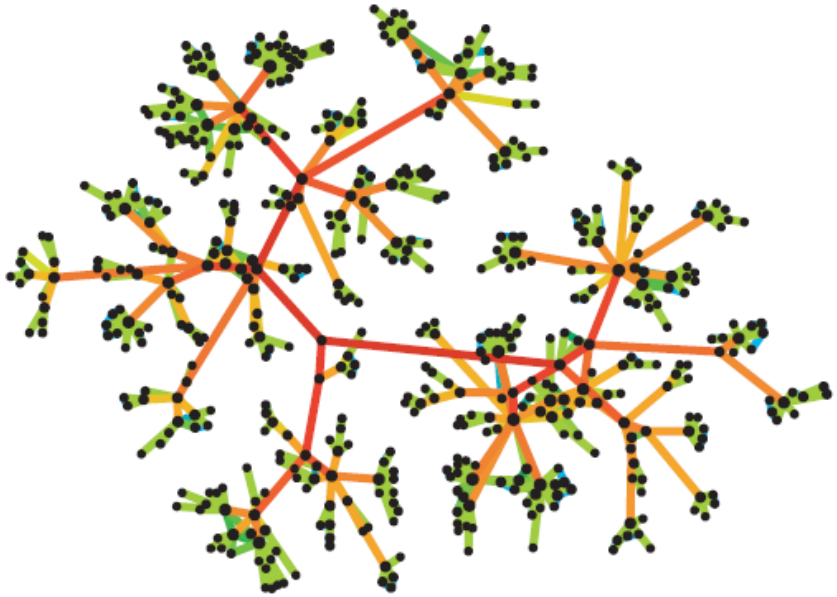
We will work with **undirected** (unweighted) networks

Method 1: Strength of Weak Ties

- **Intuition:**



Edge strengths (call volume) in real network



Edge betweenness in real network

[Girvan-Newman '02]

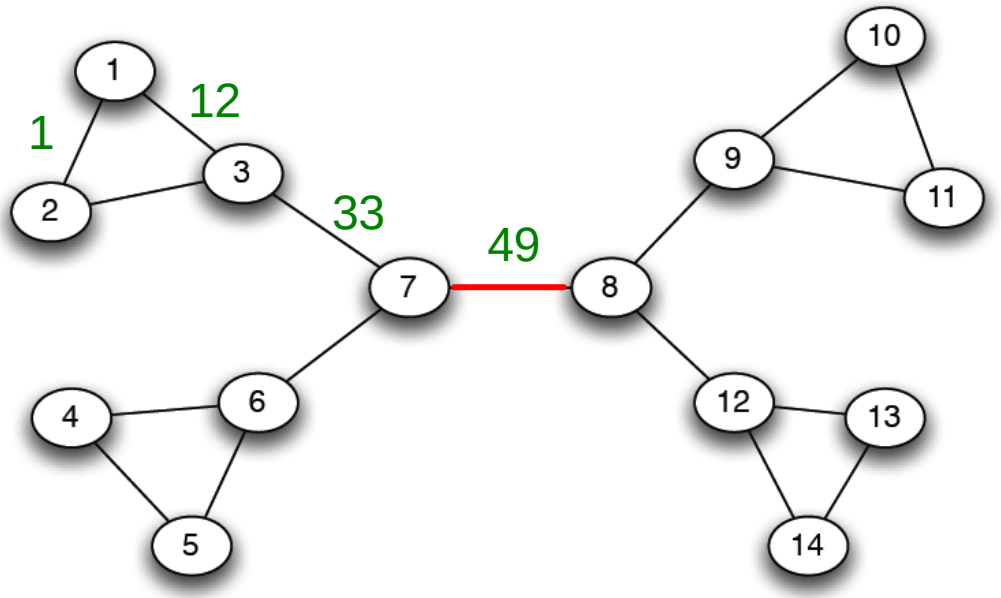
Method 1: Girvan-Newman

- Divisive hierarchical clustering based on the notion of edge **betweenness**:

  Number of shortest paths passing through the edge

- **Girvan-Newman Algorithm:**
  - **Undirected unweighted networks**
  - Repeat until no edges are left:
    - Calculate betweenness of edges
    - Remove edges with highest betweenness
  - Connected components are communities
  - Gives a hierarchical decomposition of the network

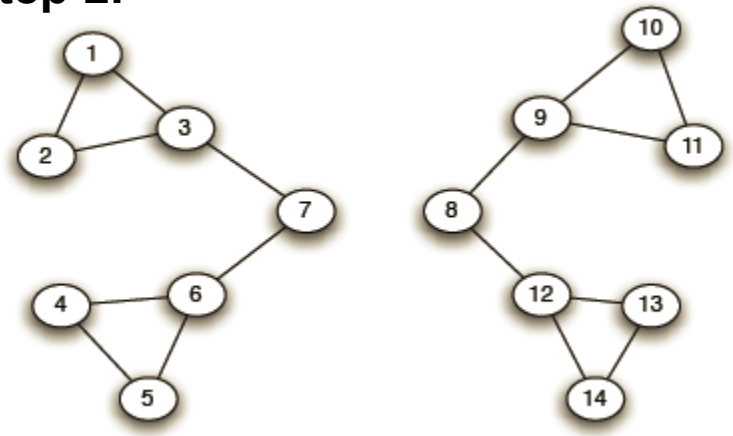
Girvan-Newman: Example



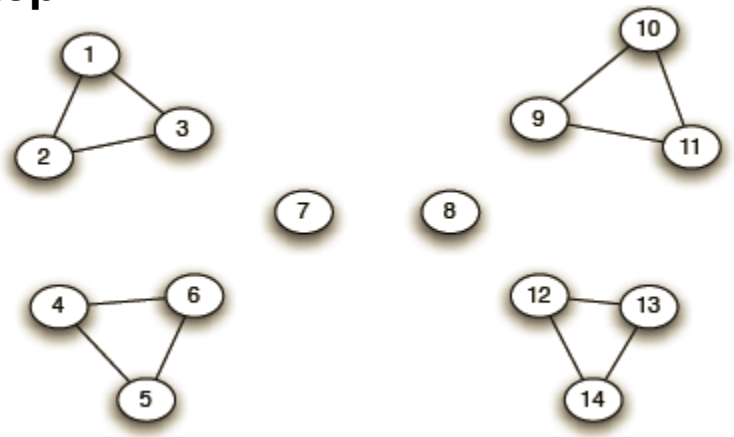
Need to re-compute betweenness at every step

Girvan-Newman: Example

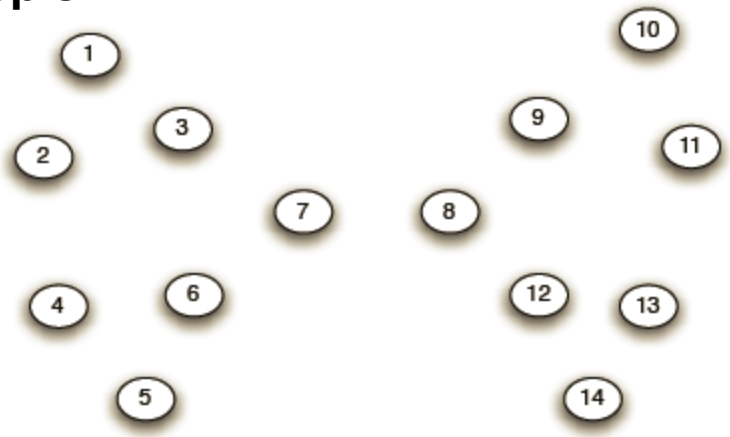
Step 1:



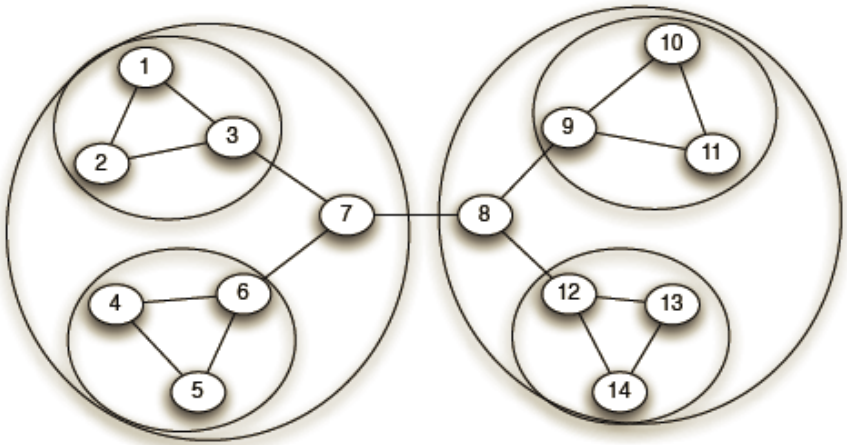
Step 2:



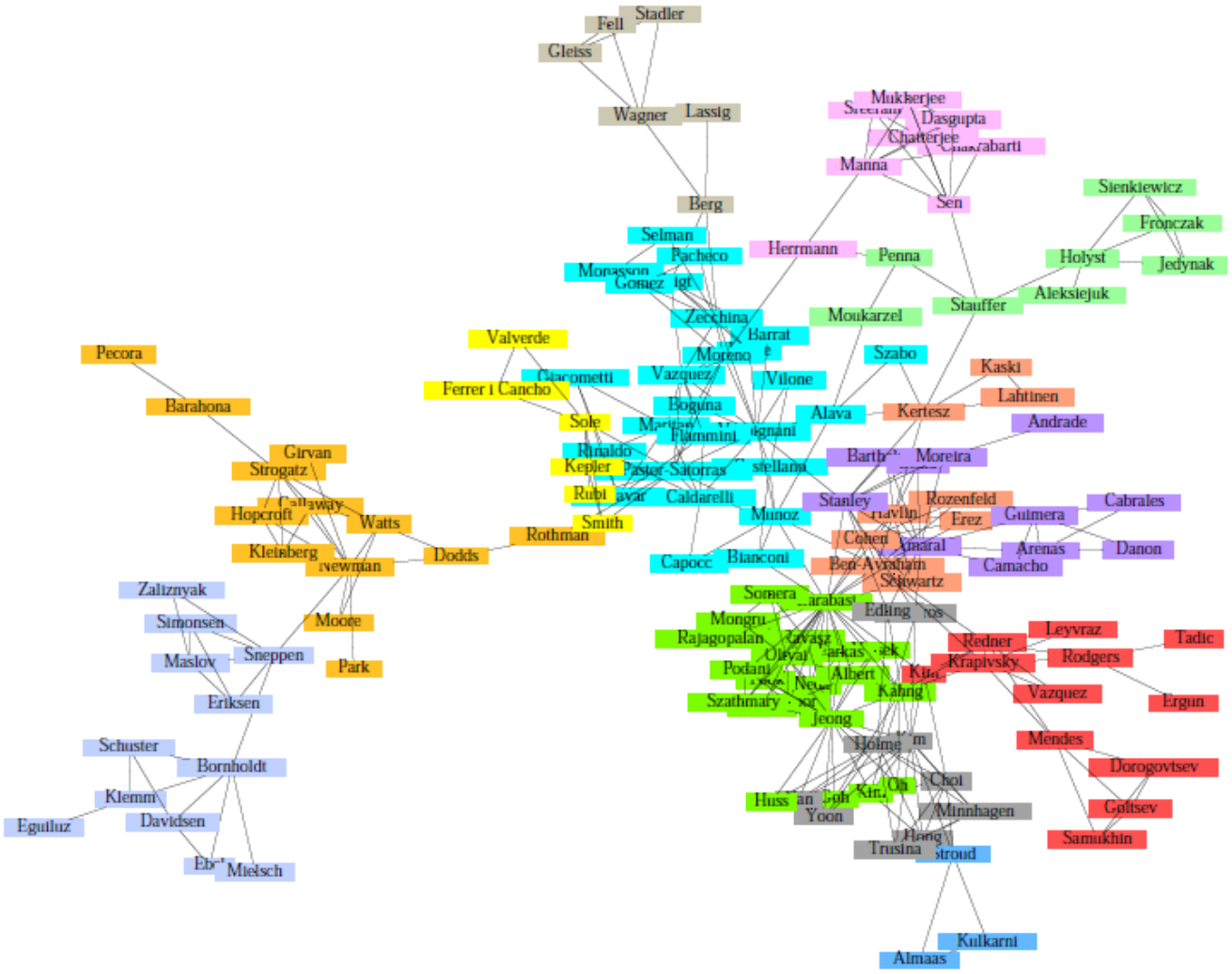
Step 3:



Hierarchical network decomposition:



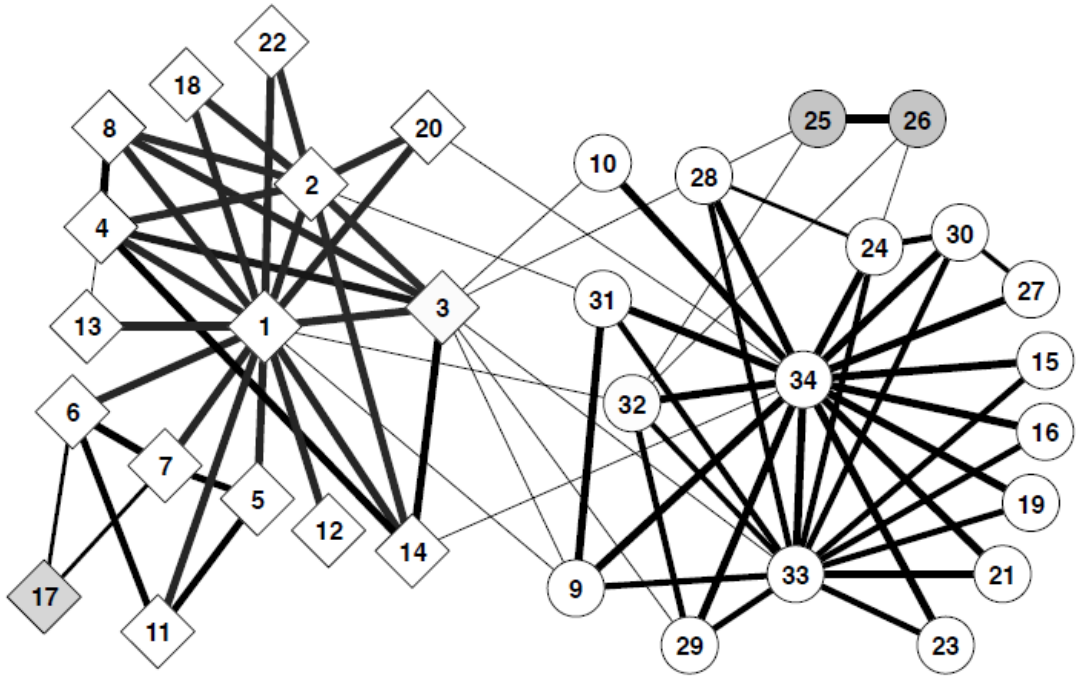
Girvan-Newman: Results

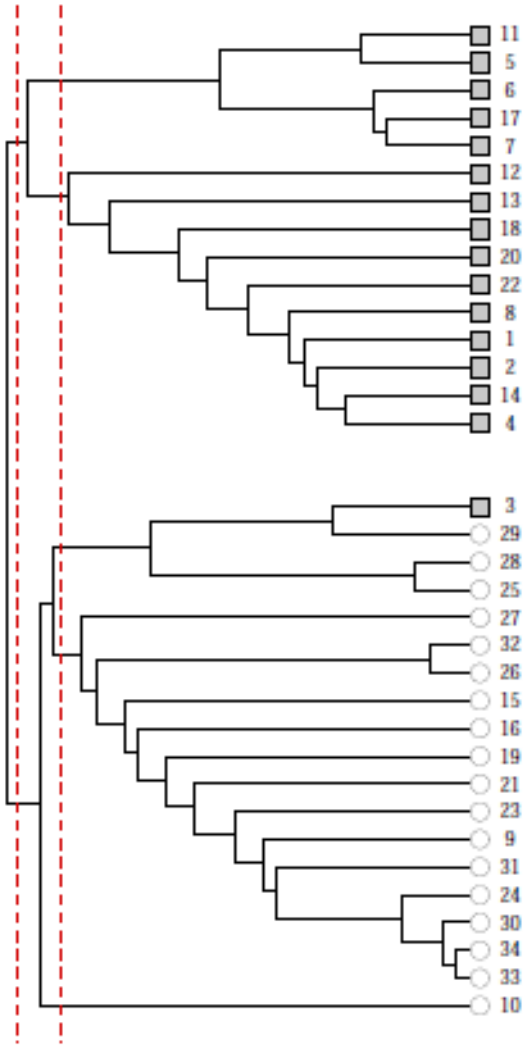


Communities in physics collaborations

Girvan-Newman: Results

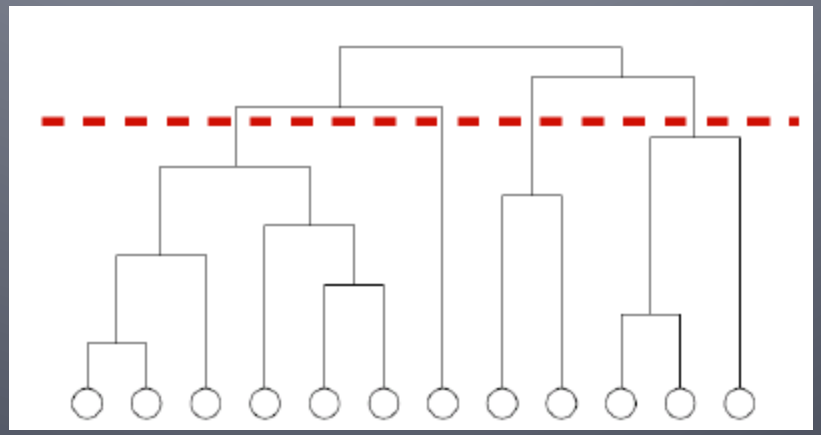
- **Zachary's Karate club:**
  hierarchical decomposition





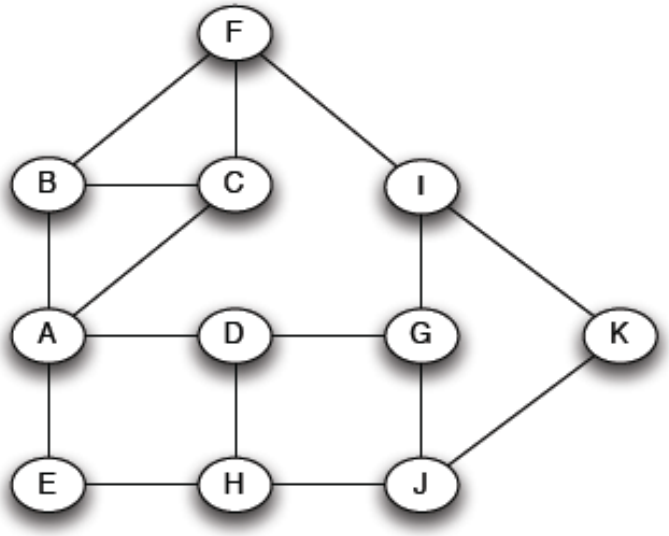
We need to resolve 2 questions

1. How to compute betweenees?
2. How to select the number of clusters?

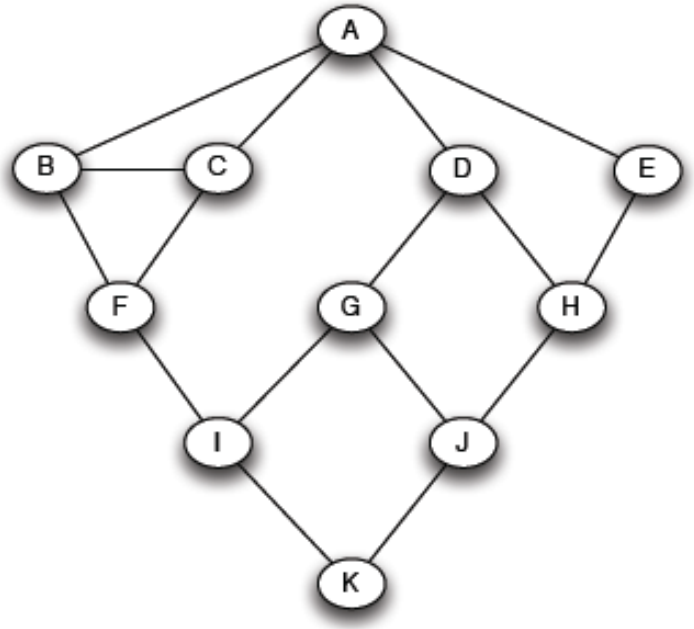


How to Compute Betweenness?

- Want to compute betweenness of paths starting at node A

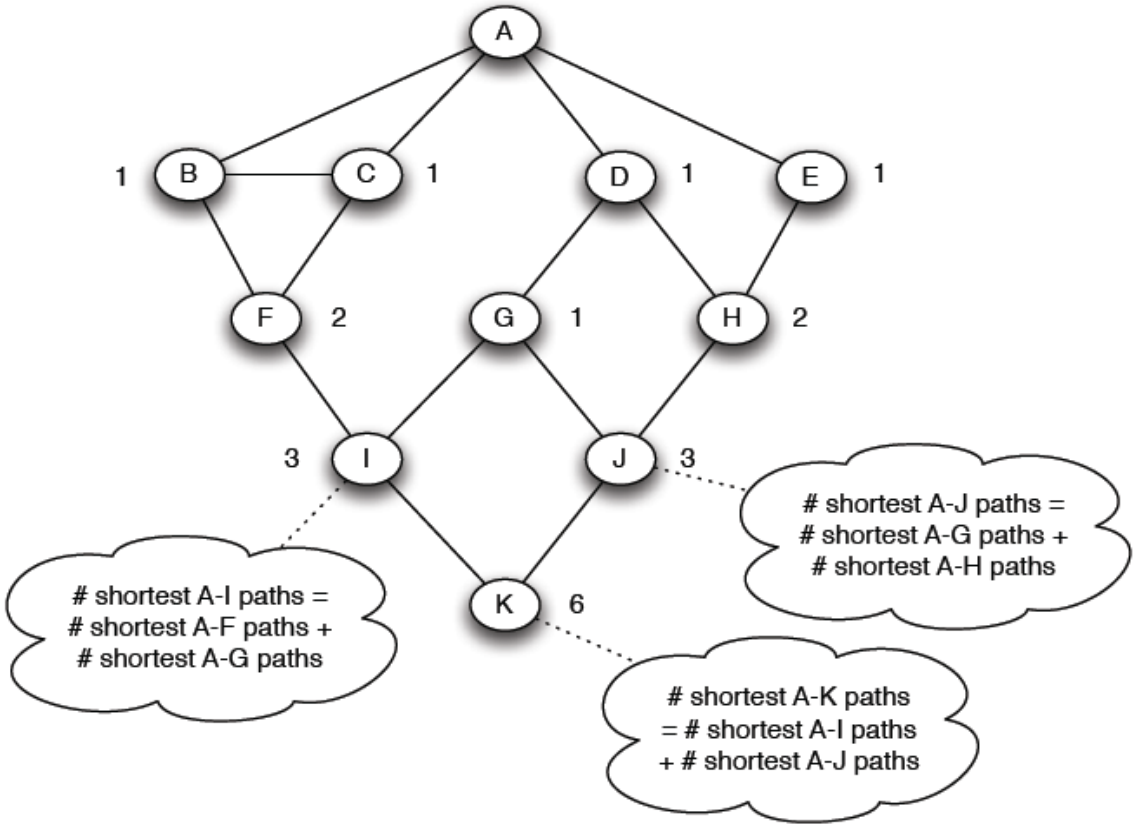


- Breath first search starting from A:



How to Compute Betweenness?

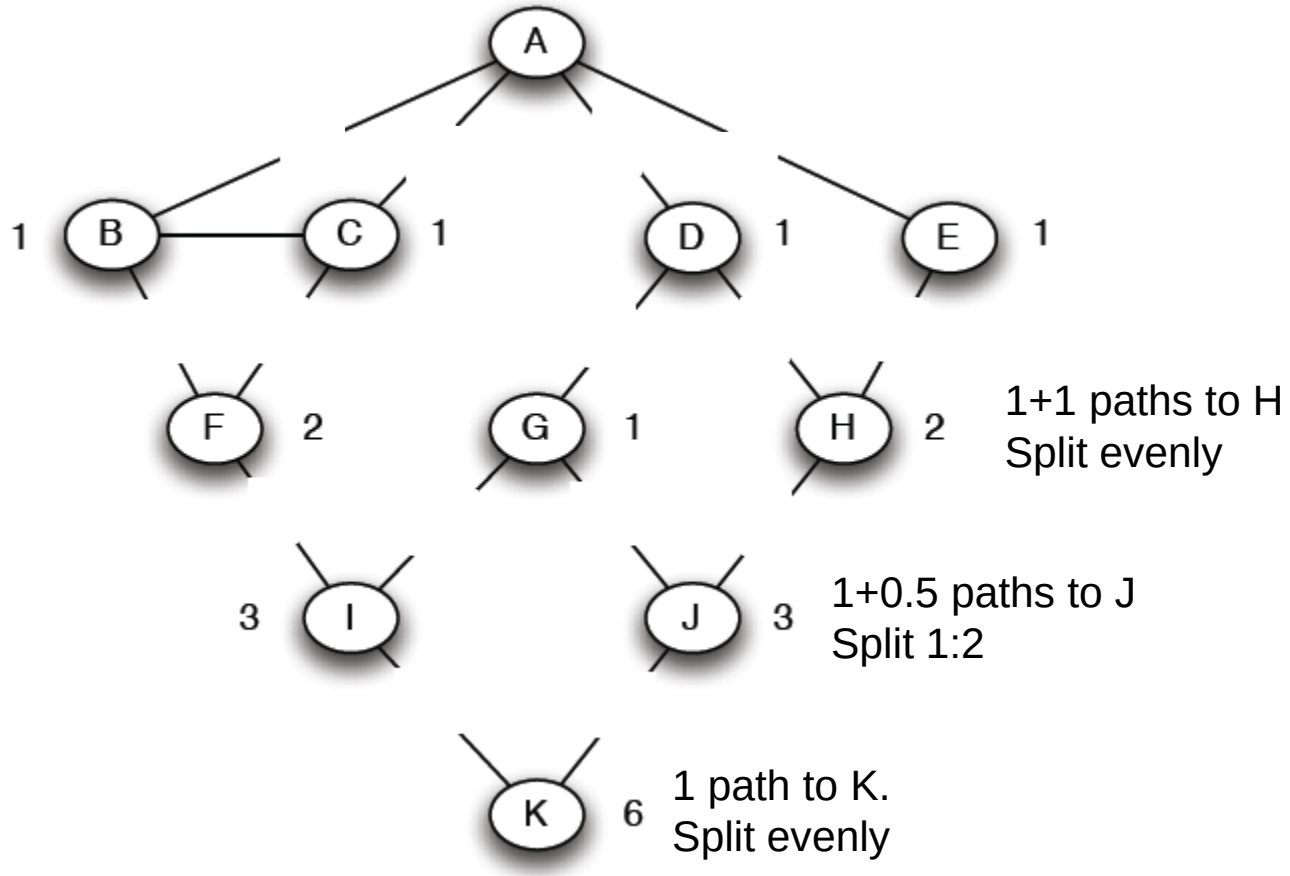
- **Count the number of shortest paths from A to all other nodes of the network:**



How to Compute Betweenness?

- **Compute betweenness by working up the tree:** If there are multiple paths count them fractionally

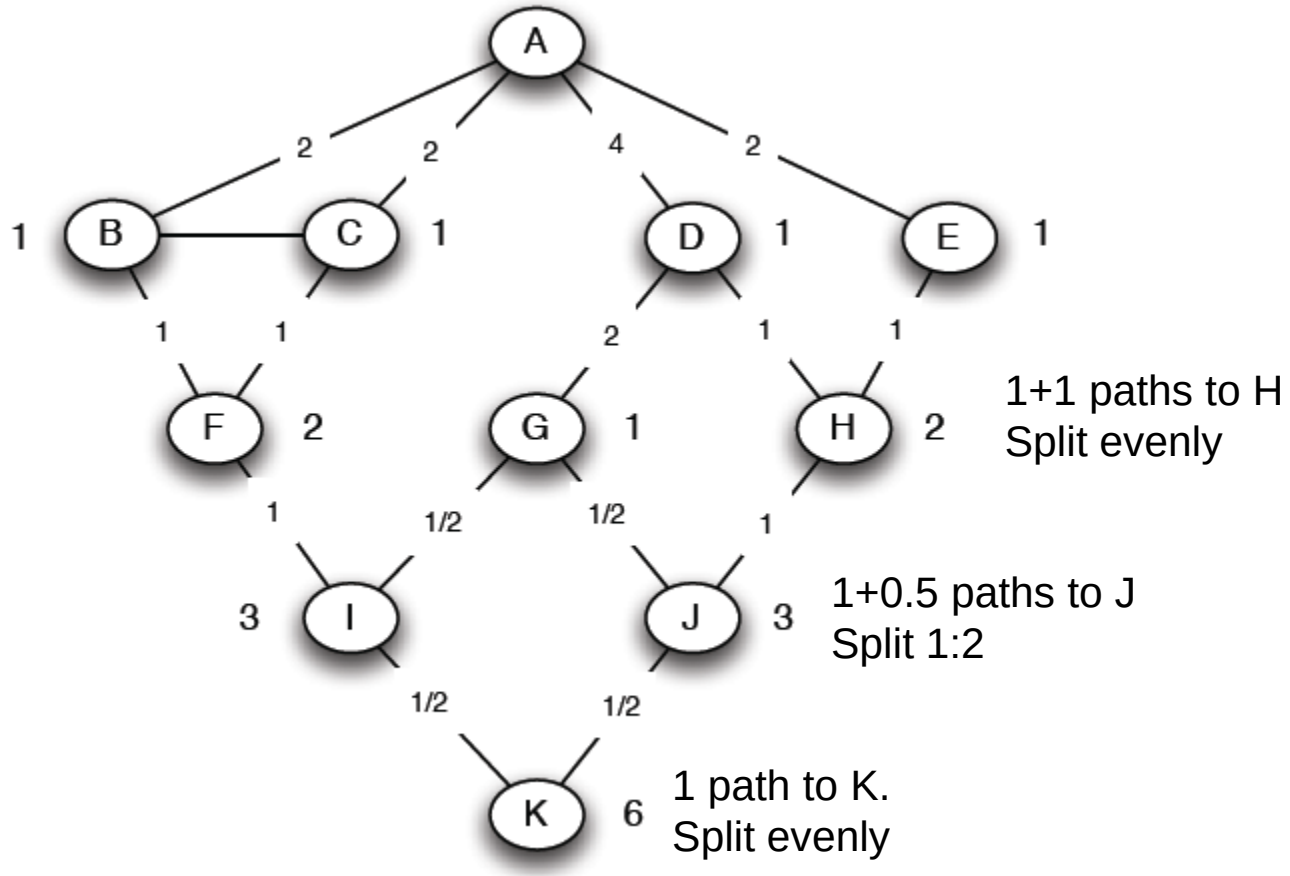
The algorithm:
- Add edge **flows**:
  -- node flow = 1+∑child edges
  -- split the flow up based on the parent value
- Repeat the BFS procedure for each starting node



How to Compute Betweenness?

- **Compute betweenness by working up the tree:** If there are multiple paths count them fractionally

The algorithm:
•Add edge **flows**:
-- node flow =
1+∑child edges
-- split the flow up based on the parent value
• Repeat the BFS procedure for each starting node



How to select number of clusters?

Define **modularity** to be

Q = (number of edges within groups) – (expected number within groups)

Actual number of edges between i and j is

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge } (i,j), \\ 0 & \text{otherwise.} \end{cases}$$

Expected number of edges between *i* and *j* is

$$\text{Expected number} = \frac{k_i k_j}{2m}.$$

m...number of edges

Modularity: Definition

- Q = (number of edges within groups) – (expected number within groups)
- Then:

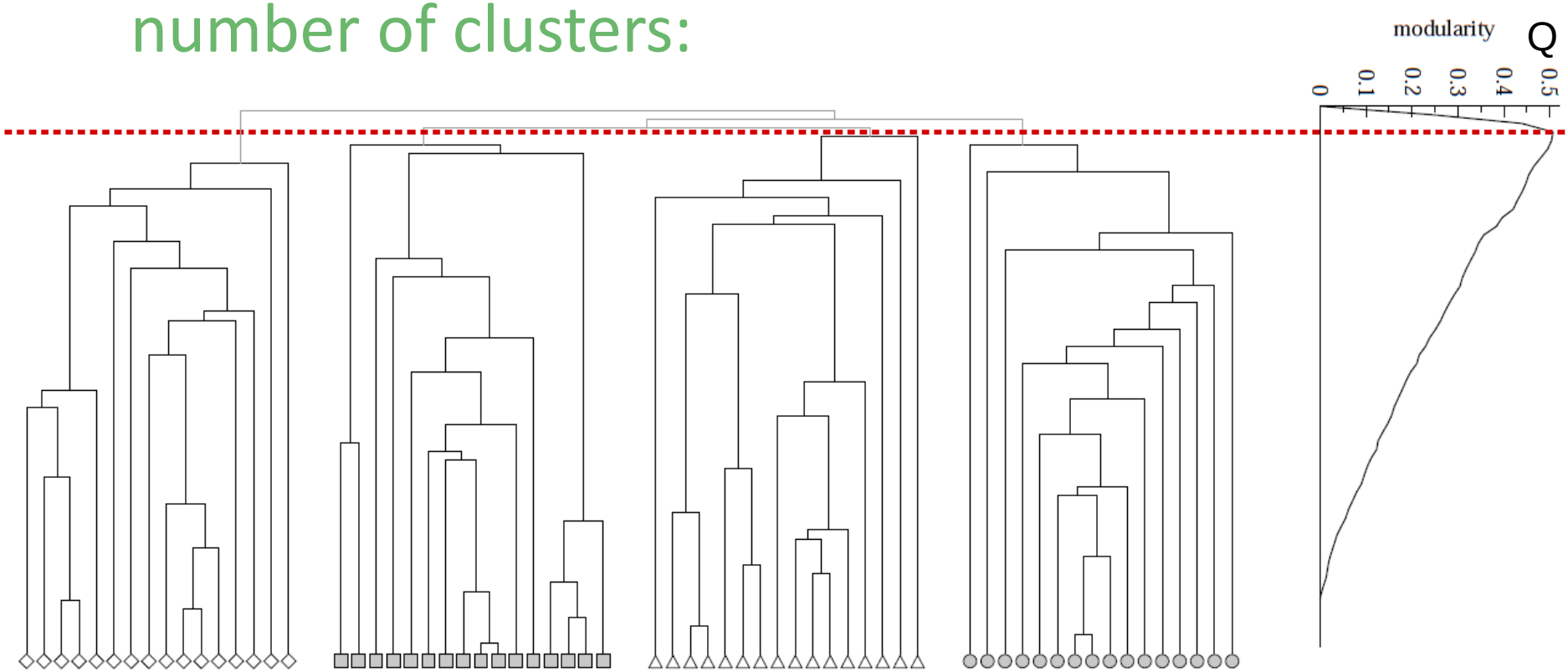
$$Q = \frac{1}{4m}\left[\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j)\right]$$

m … number of edges
A_{ij} … 1 if (i,j) is edge, else 0
k_i … degree of node i
c_i … group id of node i
$\delta(a, b)$ … 1 if a=b, else 0

- **Modularity lies in the range [−1,1]**
  - It is positive if the number of edges within groups exceeds the expected number
  - 0.3<Q<0.7 means significant community structure

Modularity: Number of clusters

- Modularity is useful for selecting the number of clusters:



Why not optimize modularity directly?