Finding communities and clusters in networks

CS 322: (Social and Information) Network Analysis

Jure Leskovec
Stanford University

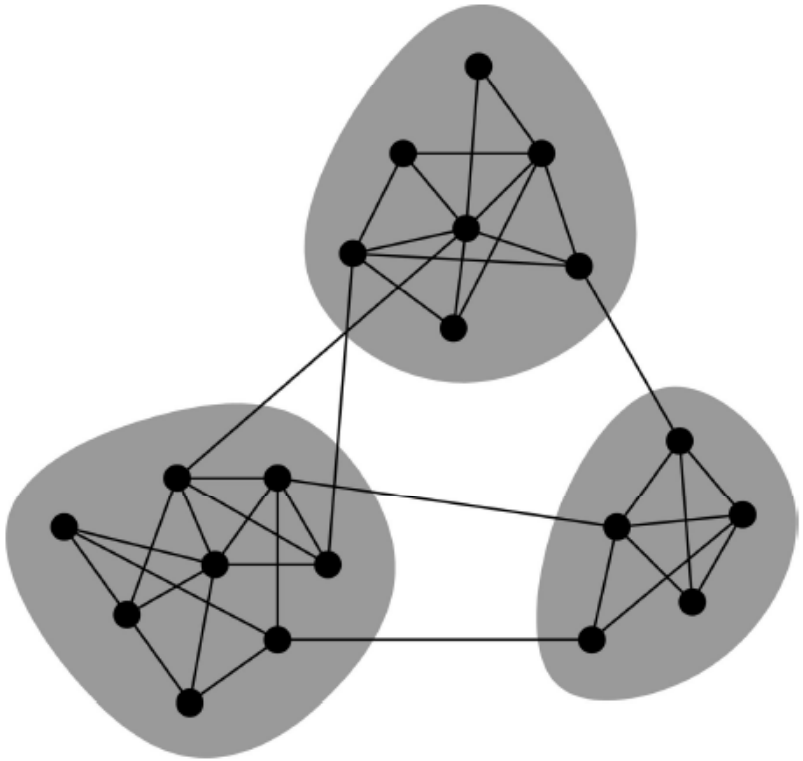
Announcements

Progress reports are due on Thursday!

- What do we expect from you?
  - About half of the work should be done
  - Milestone/progress report
  - Hand in a short write-up of your current results (what have you accomplished so far)
  - And a very briefly what further plans you have

Network communities

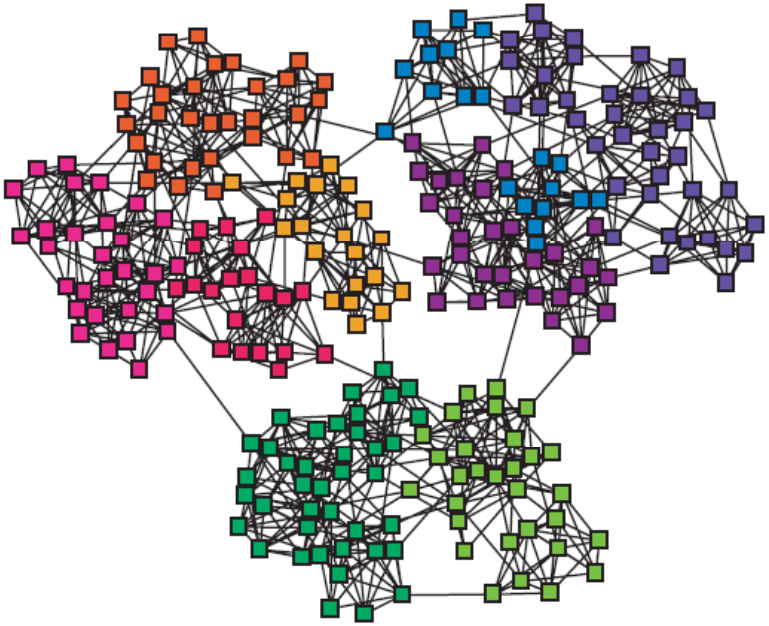
- Networks of **tightly connected groups**
- **Network communities:**
  - Sets of nodes with **lots** of connections **inside** and **few** to **outside** (the rest of the network)



Communities, clusters, groups, modules

Finding network communities

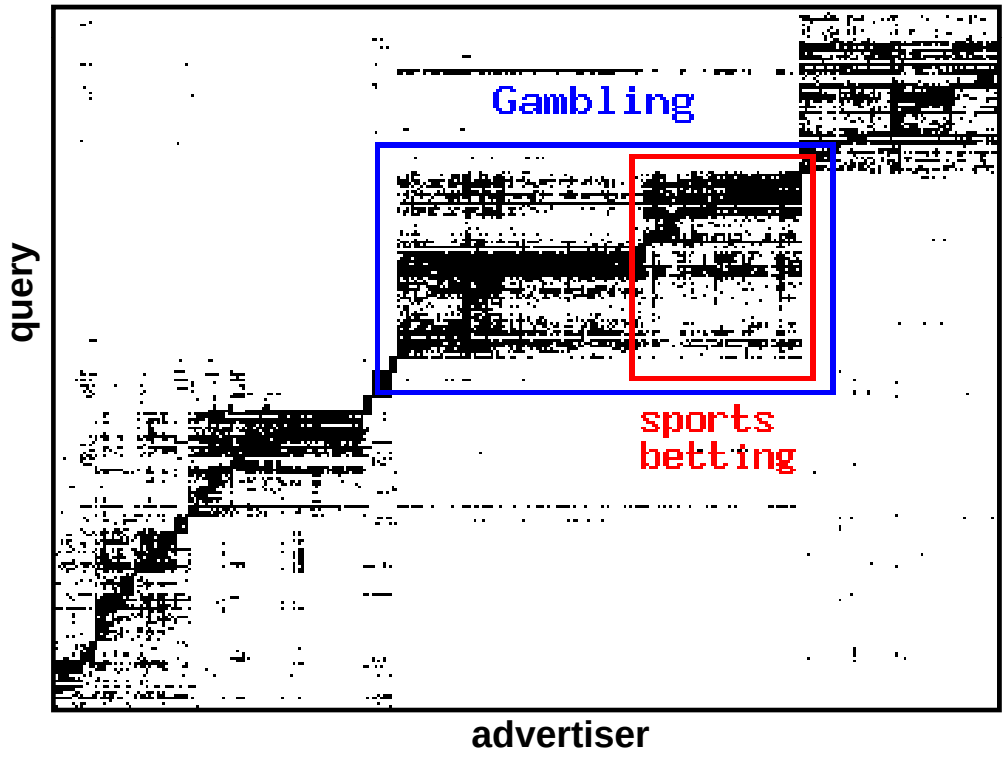
- How to automatically find such densely connected groups of nodes?
- Ideally such automatically detected clusters would then correspond to real groups
- For example:



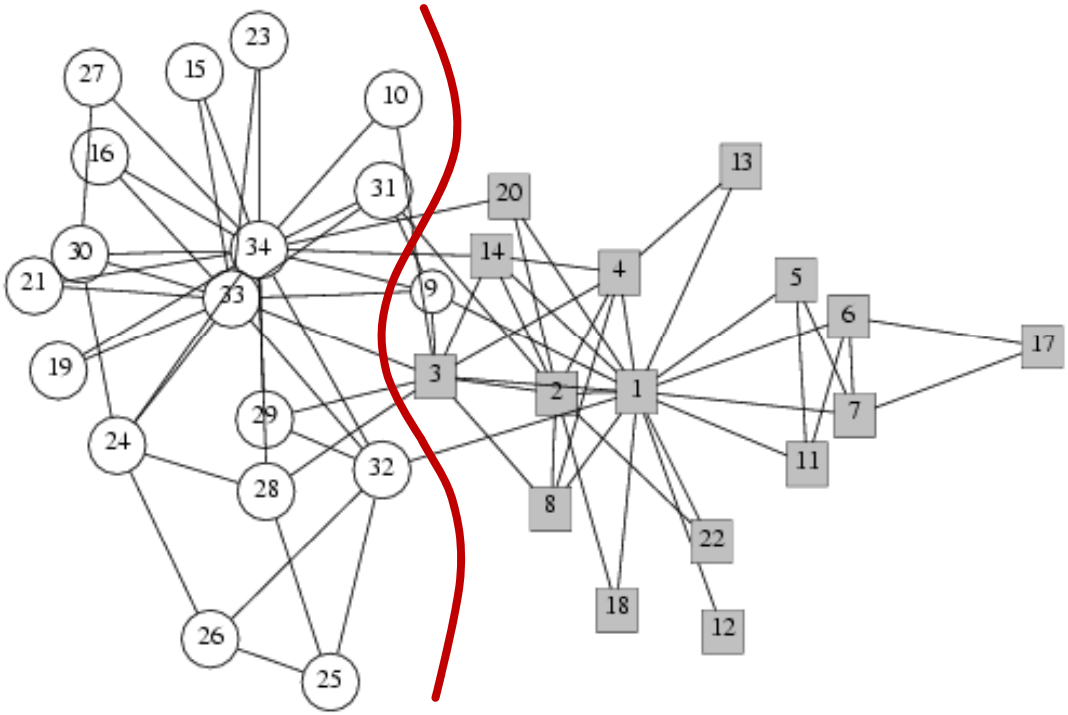
Communities, clusters, groups, modules

Micro-markets in sponsored search

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



Social Network Data

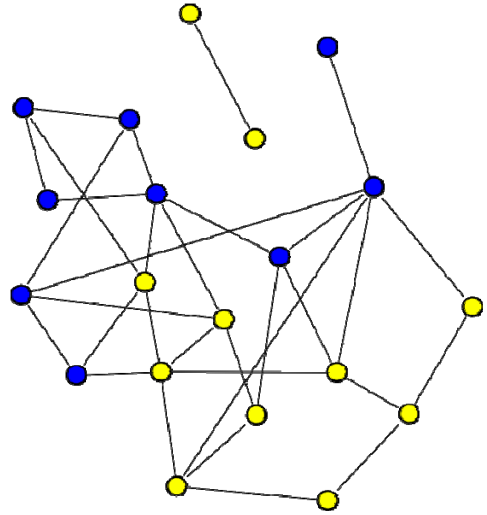


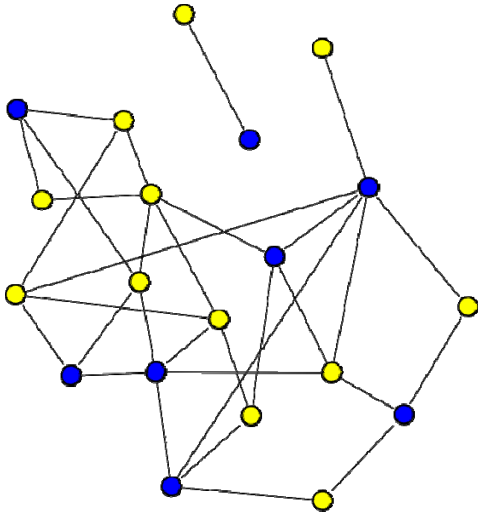
- Zachary's Karate club network:
  - Observe social ties and rivalries in a university karate club
  - During his observation, conflicts led the group to split
  - Split could be explained by a minimum cut in the network
- **Why would we expect such clusters to arise?**

[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?

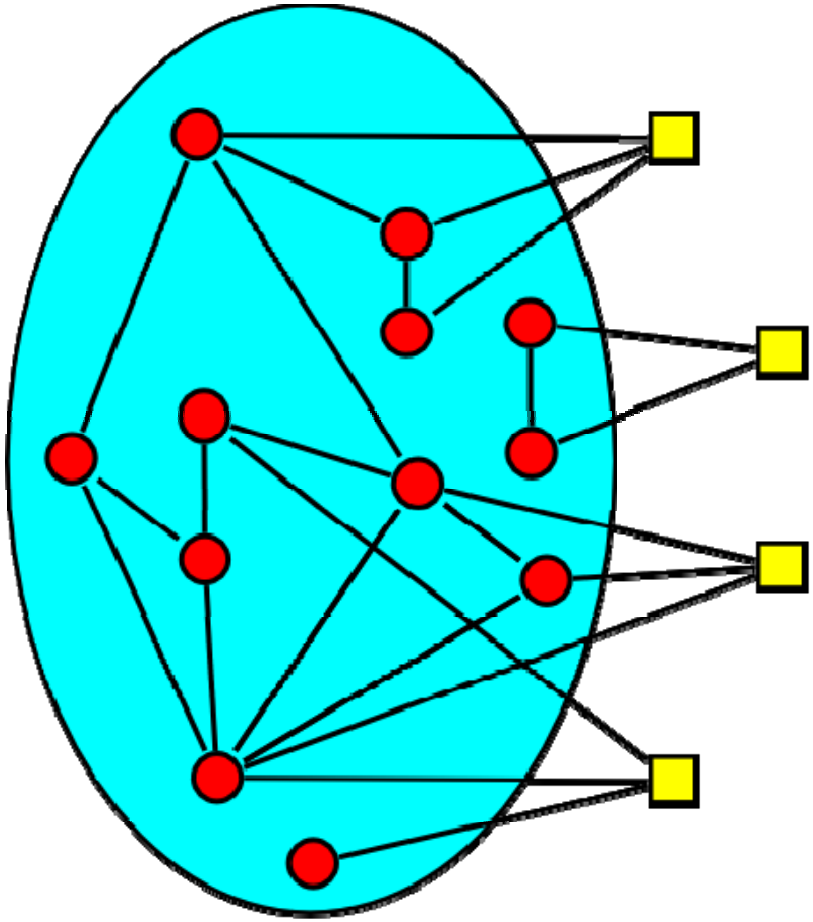




[Backstrom et al. KDD '06]

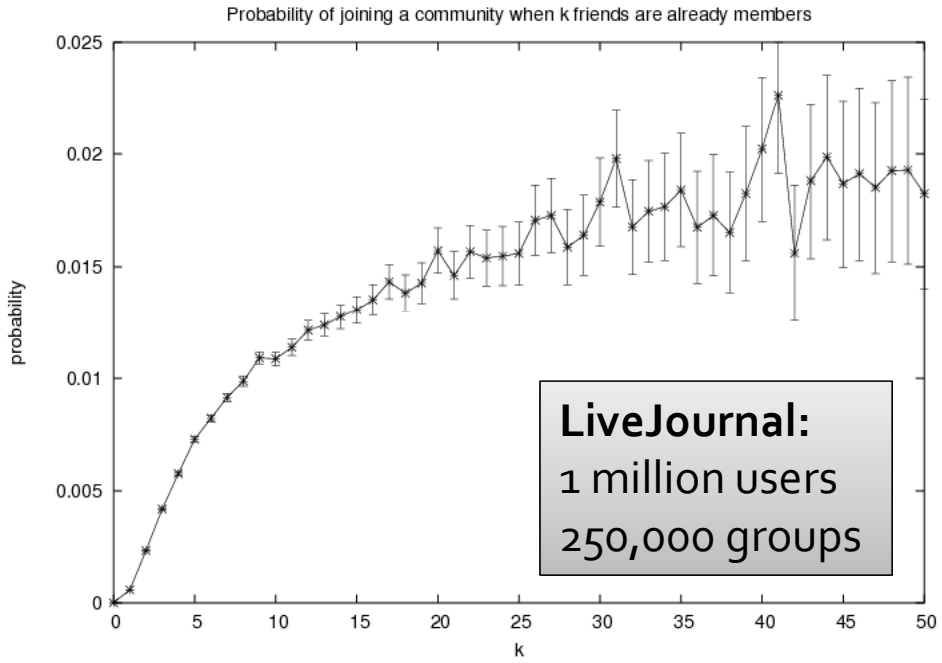
Group growth as diffusion

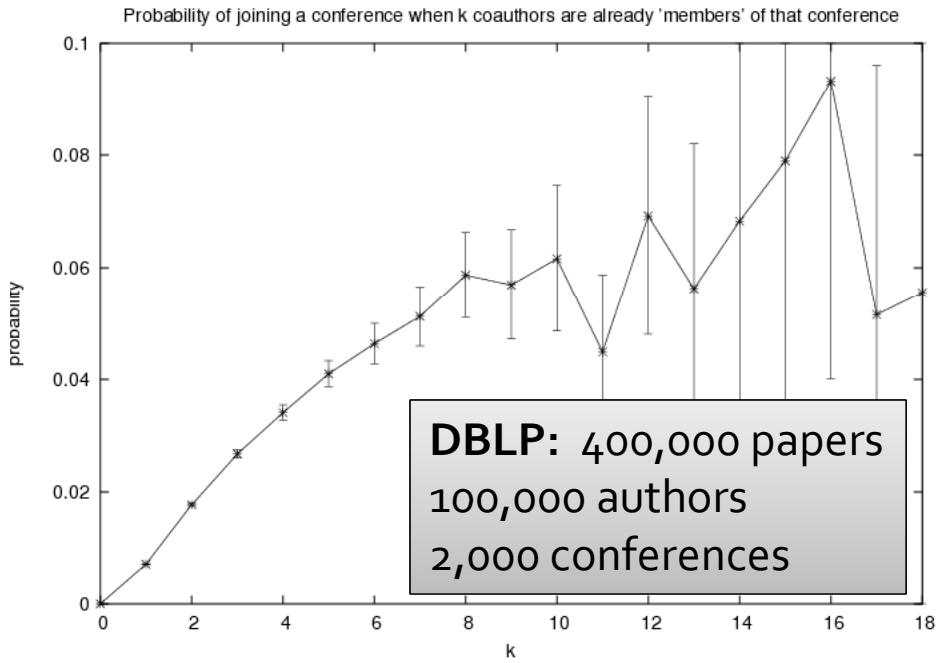
- Analogous to diffusion
  Group memberships spread over the network:
  - Red circles represent existing group members
  - Yellow squares may join
- Question:
  - How does prob. of joining a group depend on the number of friends already in the group?



[Backstrom et al. KDD '06]

P(join) vs. # friends in the group





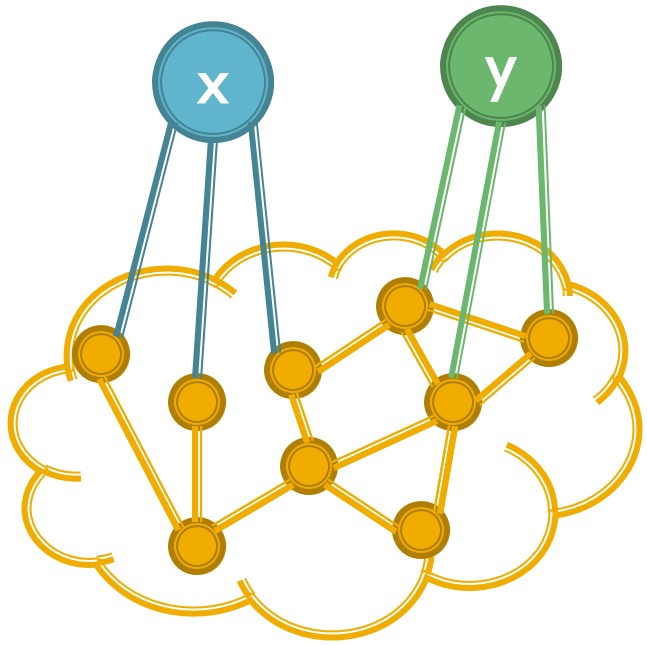
- Diminishing returns:
  - Probability of joining increases with the number of friends in the group
  - But increases get smaller and smaller

[Backstrom et al. KDD '06]

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

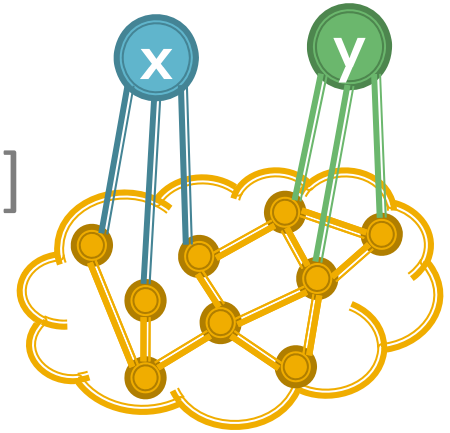
Who is more likely to join?



[Backstrom et al. KDD '06]

Connectedness of Friends

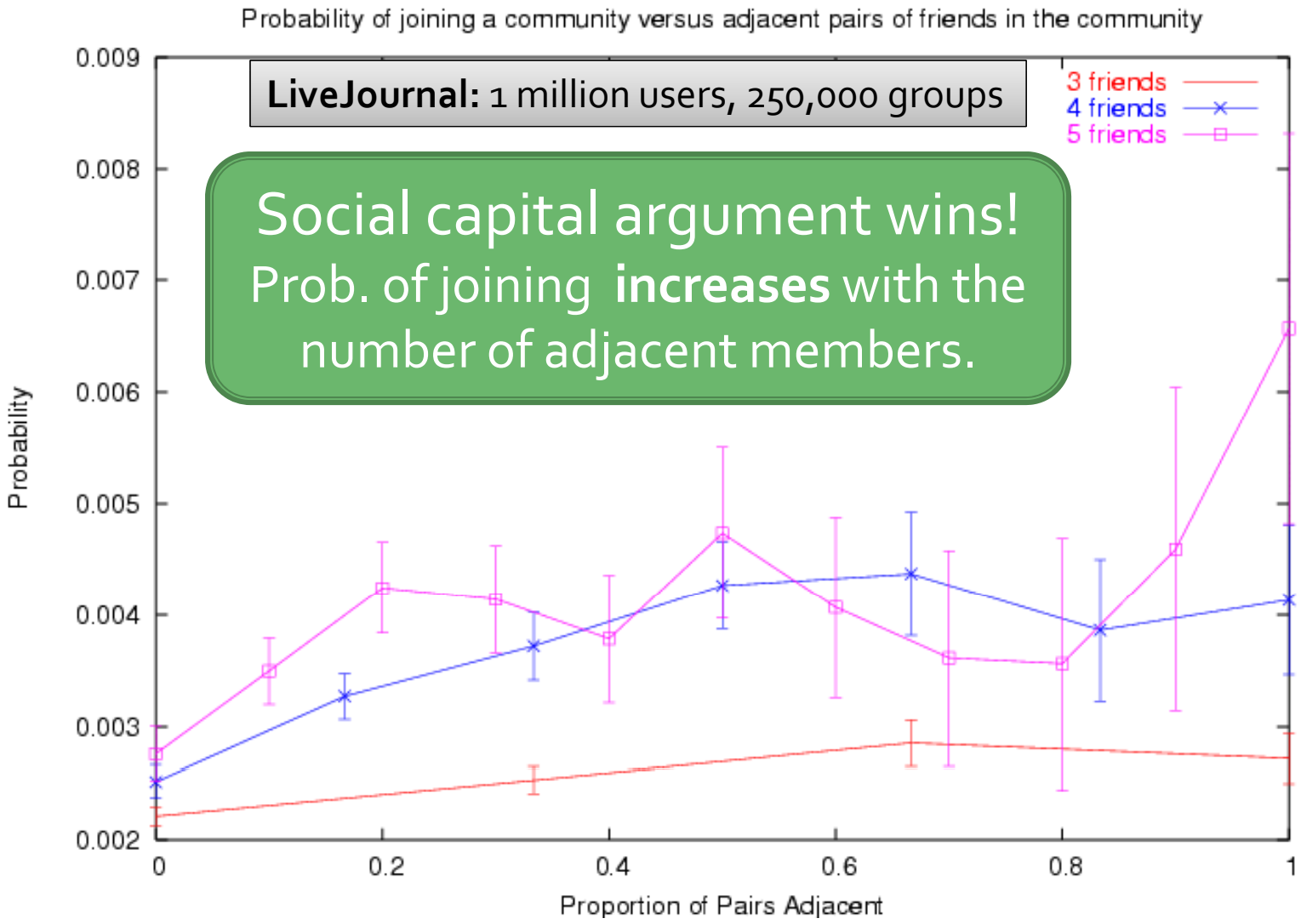
- Competing sociological theories:
  - Information argument [Granovetter '73]
  - Social capital argument [Coleman '88]



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

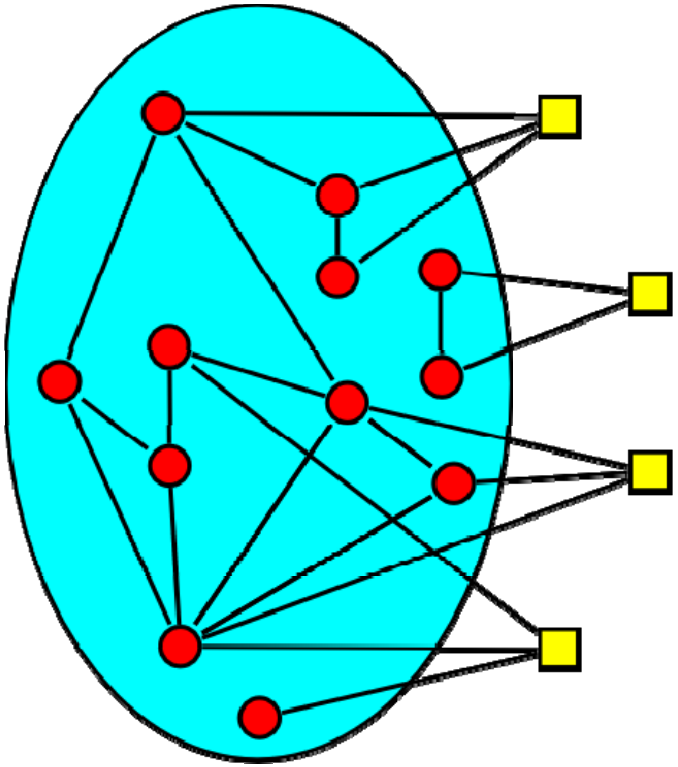
[Backstrom et al. KDD '06]

Connectedness of friends

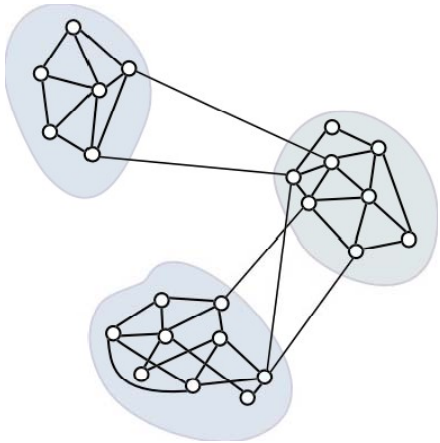


So, this means that

- 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



Clustering and Community Finding



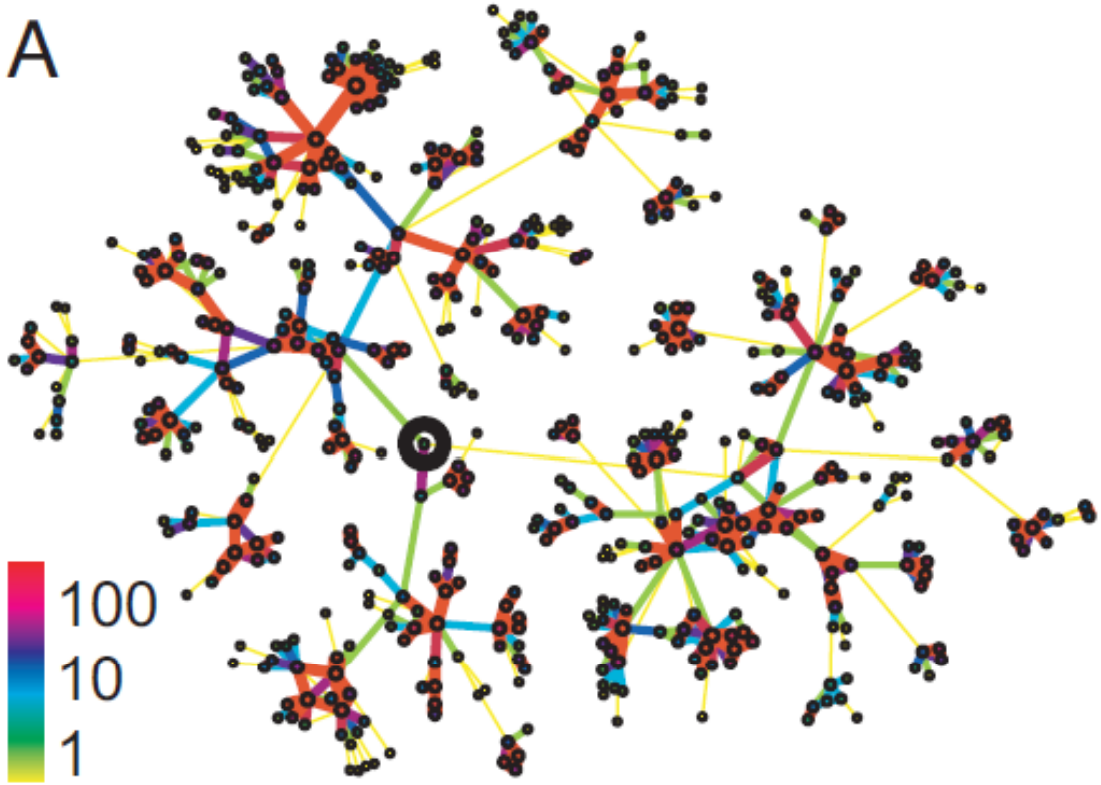
How to extract groups?

Many methods:

- Linear (low-rank) methods:
  - If Gaussian, then low-rank space is good
- Kernel (non-linear) methods:
  - If low-dimensional manifold, then kernels are good
- Hierarchical methods:
  - Top-down and bottom-up – common in social sciences
- Graph partitioning methods:
  - Define "edge counting" metric – conductance, expansion, modularity, etc. – and optimize!

[Onnela et al. '07]

Method 1: Strength of weak ties



- Real edge strengths in mobile call graph

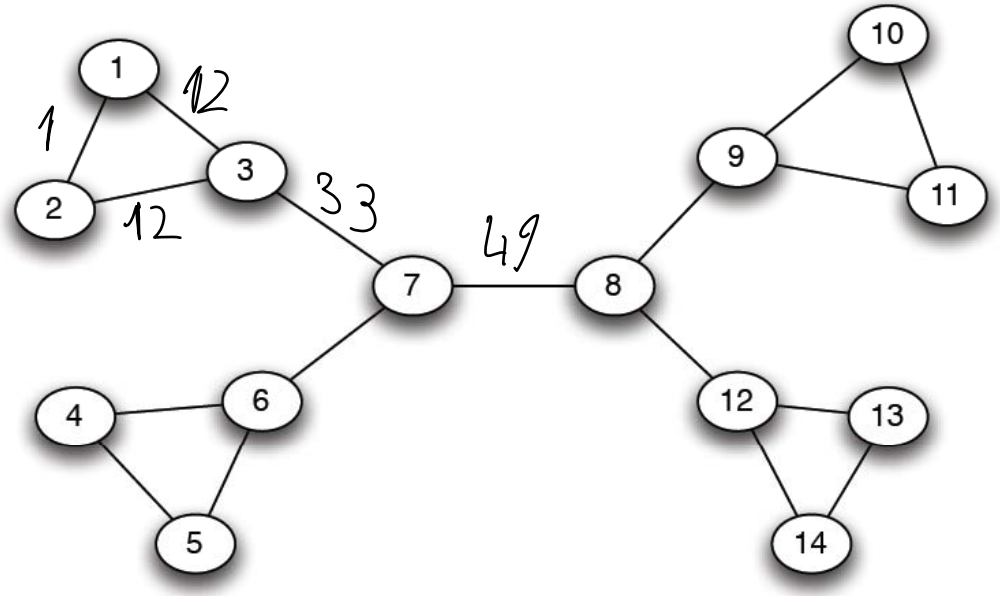
[Girvan-Newman PNAS '02]

Method 1: Girvan-Newman

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

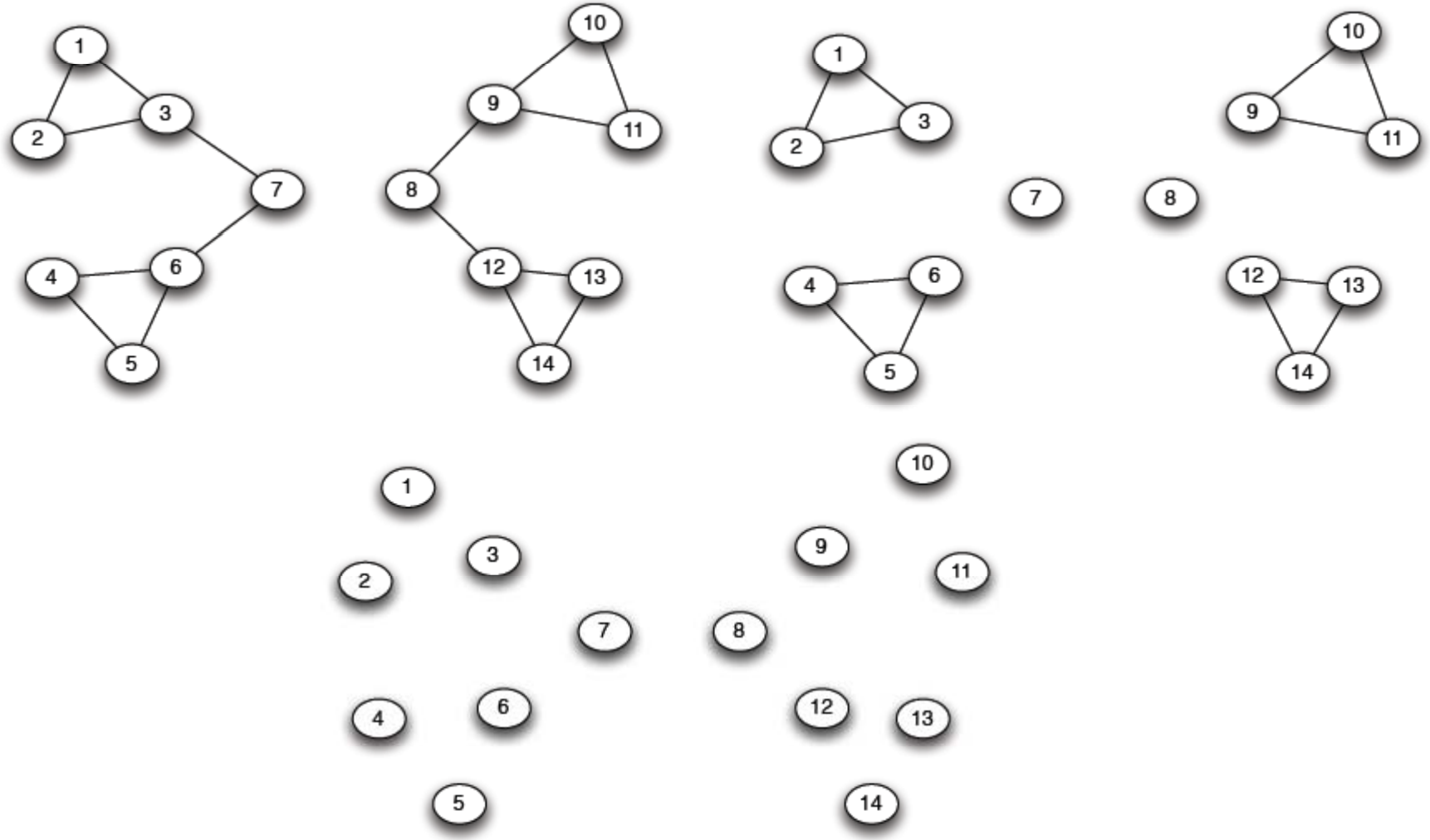
  Number of shortest paths passing through the edge

- Remove edges in decreasing betweenness
- Example:



[Girvan-Newman PNAS '02]

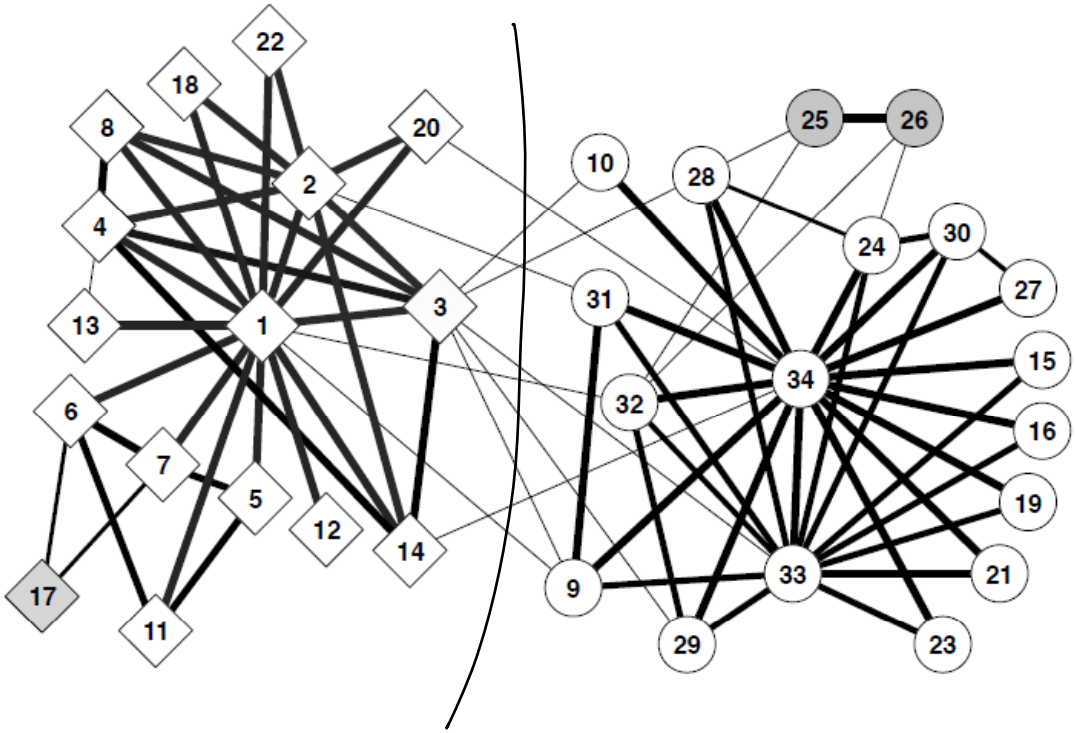
Algorithm of Girvan-Newman

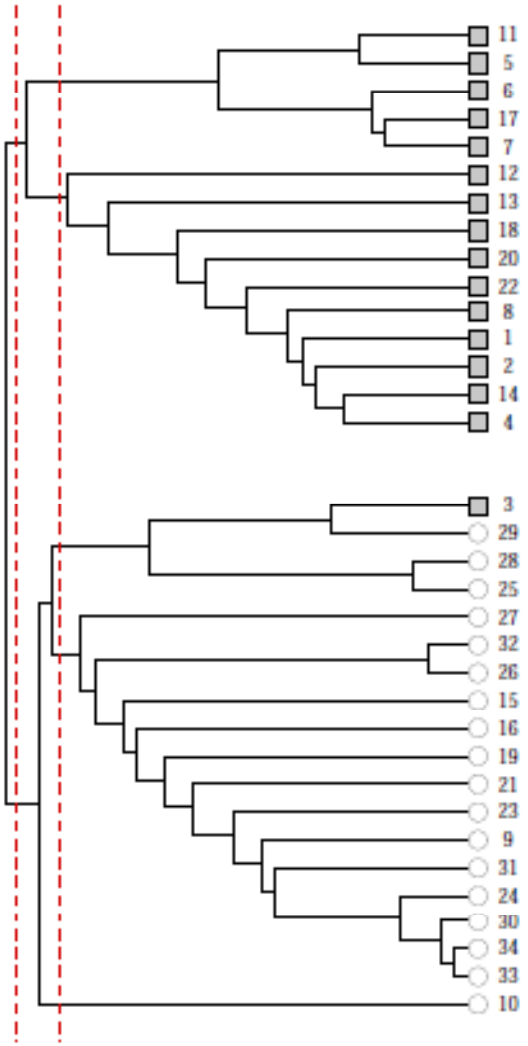


[Newman-Girvan PhysRevE '03]

Girvan-Newman: Results

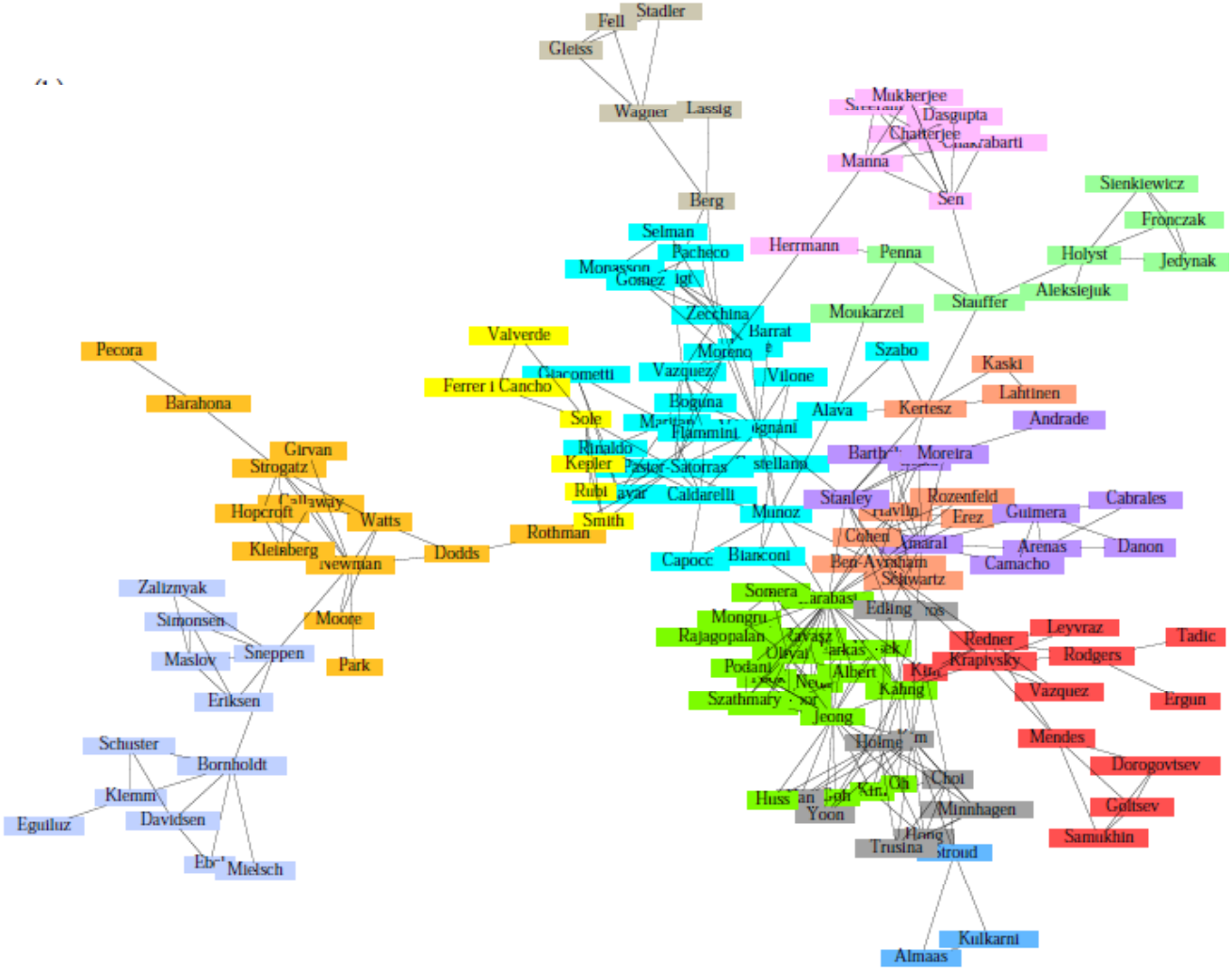
- Zachary's Karate club:
  hierarchical decomposition





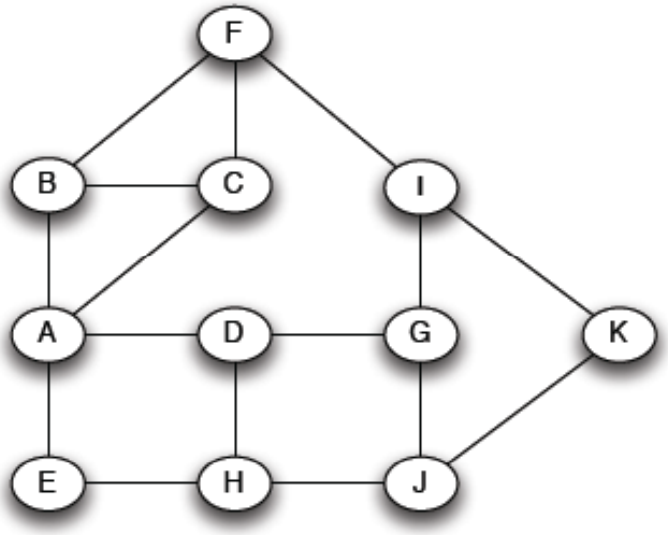
[Newman-Girvan PhysRevE '03]

Girvan-Newman: Results



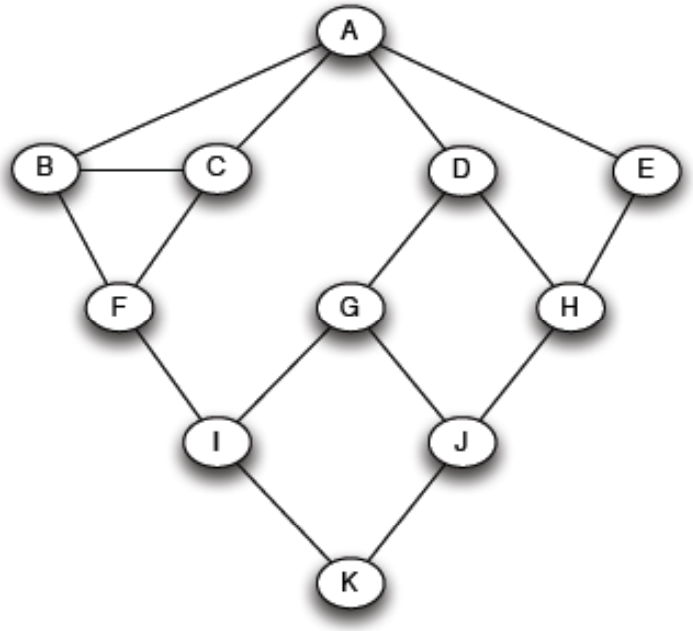
Communities in physics collaborations

How to compute betweenness (1)



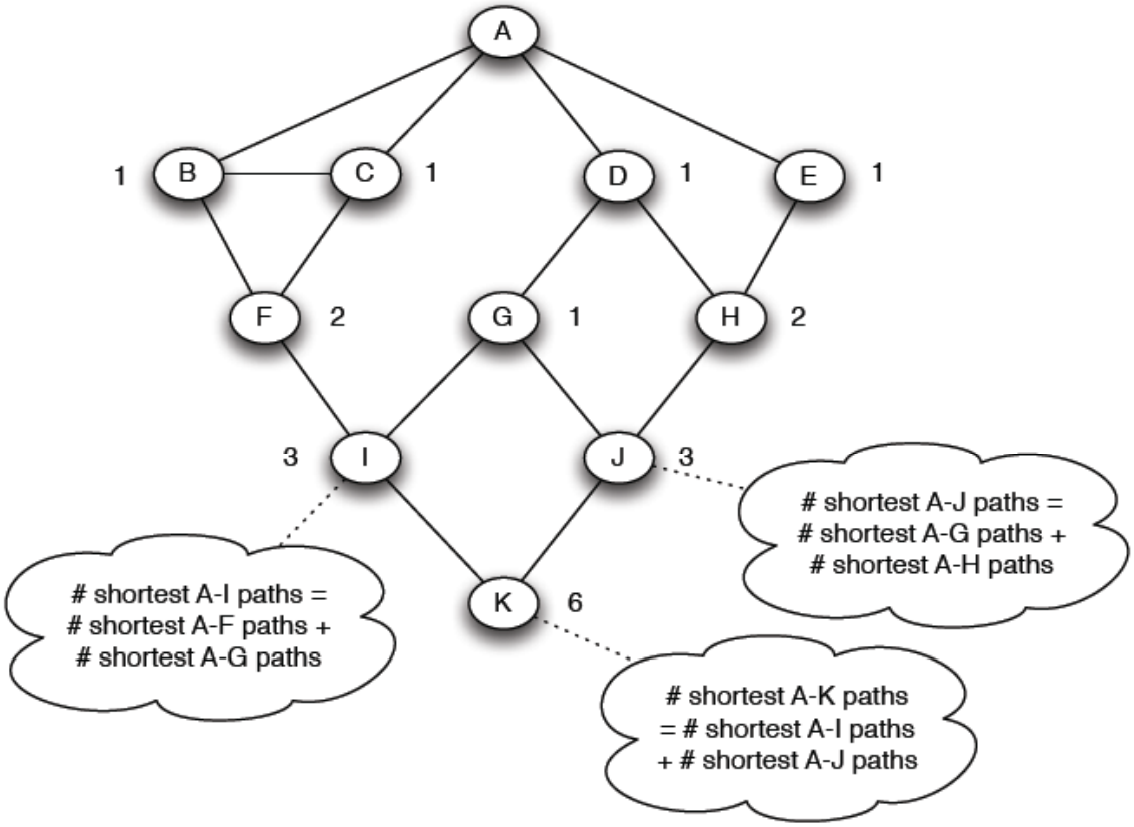
- Want to compute betweenness of paths starting at node A

- Breath first search starting from A:



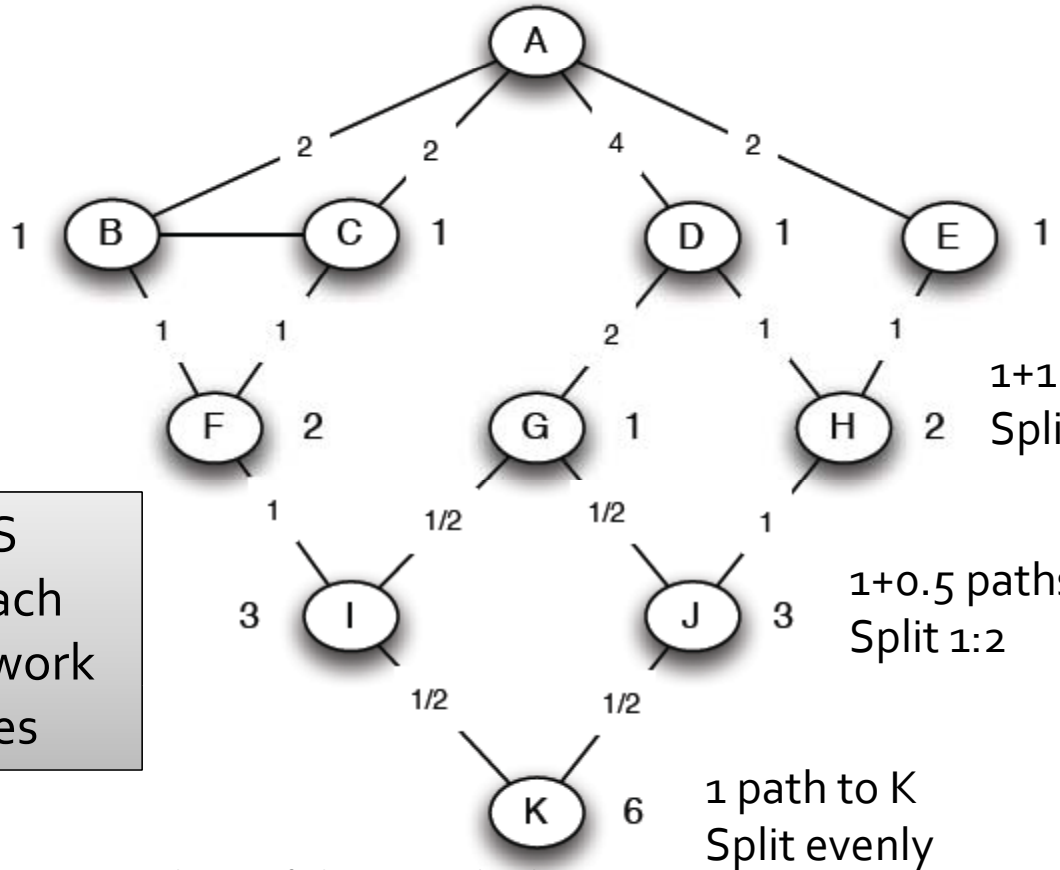
How to compute betweenness (2)

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



How to compute betweenness (3)

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



1+1 paths to H
Split evenly

1+0.5 paths to J
Split 1:2

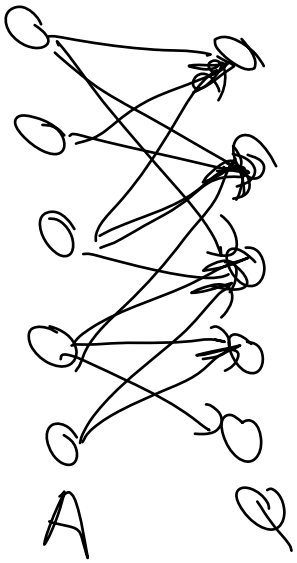
1 path to K
Split evenly

- Repeat the BFS procedure for each node of the network
- Add edge scores

[Kumar et al. '99]

Searching for small communities

- Searching for small communities in a web graph
- (1) The signature of a community/discussion



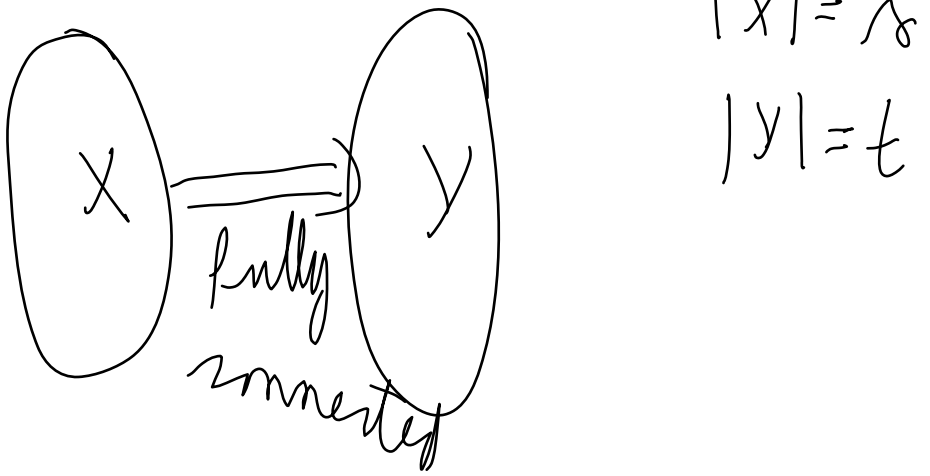
A dense 2-layer graph

Intuition: a bunch of people all talking about the same things

Searching for small communities

- (2) A more well-defined problem:

  enumerate all complete bipartite subgraphs $K_{s,t}$= s nodes each links to the same t other nodes



The Plan: (1), (2) and (3)

- A) From (2) get back to (1):
  - Via: any dense enough graph as in (1) contains a smaller $K_{s,t}$ as a subgraph
- B) How do we solve (2) in a giant graph?
  - What similar problems have been solved on a giant non-graph datsets?
- (3) Frequent itemset enumeration [Agrawal-Srikant '99]

[Agrawal-Srikant '94]

Frequent itemset enumeration

- Example: What items are bought together in a store?
- Setting:
  - Universe U of n items
  - m subsets of U: $S_1, S_2, \ldots, S_m \subseteq U$
    (S_i is a set of items one person bought)
  - Frequency threshold f
- Goal:
  - Find all subsets T s.t. $T \subseteq S_i$ of $\geq f$ sets S_i
    (items in T were bought together $\geq f$ times)

[Agrawal-Srikant '94]

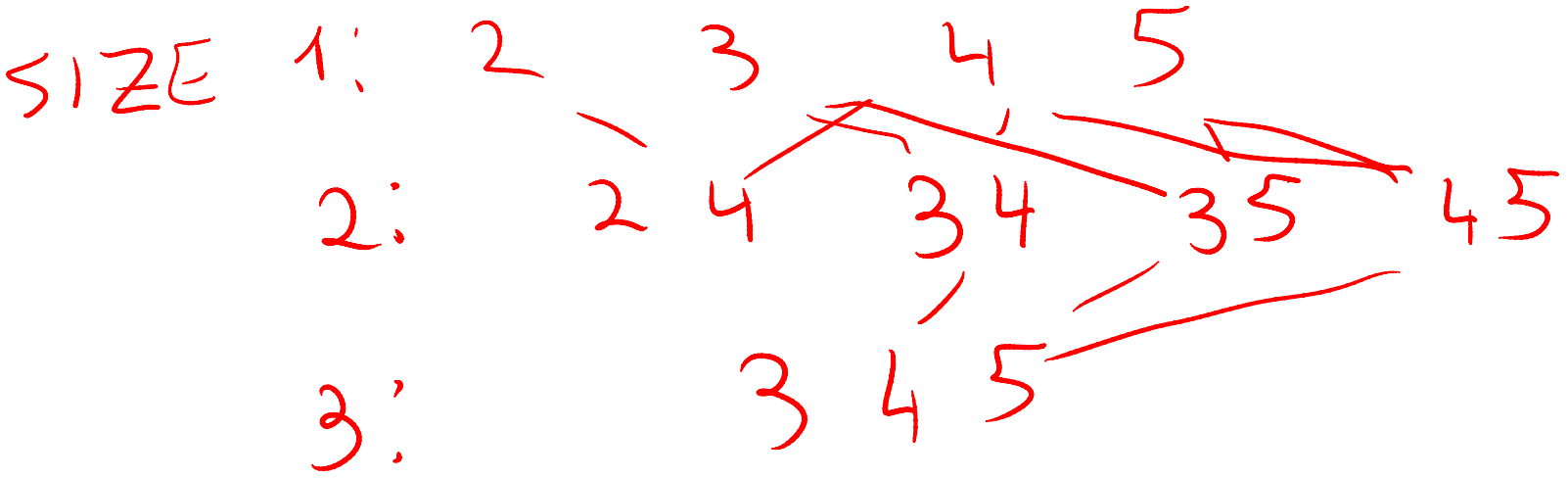
Example (1)

- Example:
  - U={1,2,3,4,5}
  - S_1={1,3,5}, S_2={2,3,4}, S_3={2,4,5}, S_4={3,4,5}, S_5={1,3,4,5}, S_6={2,3,4,5}
  - f=3
- Algorithm: build up the lists
- Insight:

  for a frequent set of size $k \Rightarrow$ all its subsets are also frequent

Example (2)

- U={1,2,3,4,5}
- S_1={1,3,5}, S_2={2,3,4}, S_3={2,4,5}, S_4={3,4,5}, S_5={1,3,4,5}, S_6={2,3,4,5}
- f=3



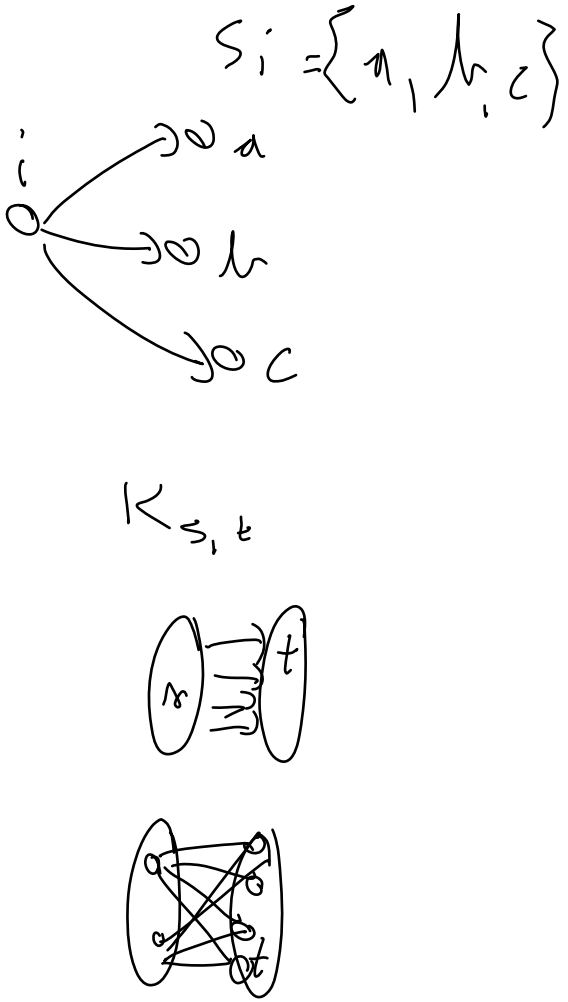
The algorithm

- For i = 1,…,k
  - Find all frequent sets of size i by composing sets of size i-1 that differ in 1 element

- Open question:
  - Efficiently find only maximal frequent sets

Itemsets and $K_{s,t}$

- Claim: (3) (itemsets) solves (2) (bipartite subgraphs)
- How?
  - View each node *i* as a set S_i of nodes *i* points to
  - $K_{s,t}$ = a set *y* of size *t* that occurs in *s* sets S_i
  - Looking for $K_{s,t}$ → set of frequency threshold to *s* and look at layer t – all frequent sets of size *t*.



$K_{s,t}$ and communities

- (2) $\Rightarrow$ (1): Informally, every dense enough bipartite graph *G* contains a $K_{s,t}$ subgraph where *s* and *t* depend on size (# of nodes) and density (avg. degree) of *G* [Kovan-Sos-Turan '53]
- Theorem: Let *G=(X,Y,E)*, $|X|=|Y|=n$ with avg. degree:

$$d = s^{1/t} n^{1-1/t} + t$$

then *G* contains $K_{s,t}$ as a subgraph

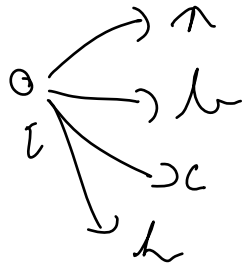
$K_{s,t}$ and communities

- Proof:
  - Recall: $\binom{a}{b} = \frac{a(a-1)\ldots(a-b+1)}{b!}$
  - Let $f(x) = x(x-1)(x-2)\ldots(x-k)$
    Once $x \geq k$, $f(x)$ curves upward (convex)

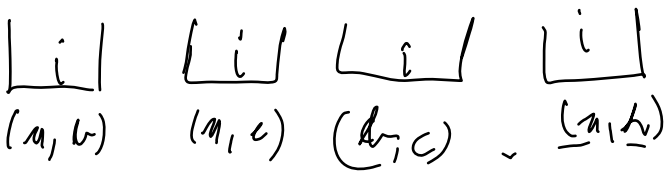
  - Supposed g is convex, want to min $\sum_{i=1}^{n} g(x_i)$ where $\sum_{i=1}^{n} x_i = x$
  - To minimize $\sum_{i=1}^{n} g(x_i)$ make each $x_i = x/n$

Nodes and buckets

- Node *i*, degree d_i:



Potential right-hand sides of $K_{s,t}$ (i.e., all size t subsets of Y)



- Put node *i* in buckets for all size *t* subsets of its neighbors

Nodes and buckets

- As soon as *s* people appear in a bucket we have a $K_{s,t}$
- How many buckets node *i* contributes? $\binom{d_i}{t}$
- What is the total size of all buckets?

$$\sum_{i=1}^{n} \binom{d_i}{t} \geq \sum_{i=1}^{n} \binom{\bar{d}}{t} = n\binom{\bar{d}}{t}$$

by convexity

Nodes and buckets

$$\vec{n}\,(\vec{n}-1)\cdots(\vec{n}-t)$$

- So the total height of all buckets is…

$$n\binom{\vec{n}}{t} \geq n\,\frac{(\vec{n}-t)^t}{t!} = n\,\frac{\delta\, n^{t-1}}{t!} = \frac{n^t \delta}{t!}$$

$$\vec{n} = \delta^{\frac{1}{\varepsilon}}\, n^{1-\frac{1}{\varepsilon}+t}$$

And we are done...

- How many buckets are there?
- What is the average height of buckets?

$$\binom{n}{t} \leq \frac{n^t}{t!}$$

$$\frac{n^t s}{t!} \cdot \frac{t!}{n^t} = s \leq$$

□

- So by pigeonhole principle, there must be a bucket with more than *s* nodes in it.

Summary

- Girvan-Newman:
  - based on strength of weak ties
  - Remove edge of highest betweenness
- Extracting complete bipartite subgraphs:
  - Frequent itemsets and dynamic programming
  - Theorem that complete bipartite subgraphs are embedded in bigger graphs