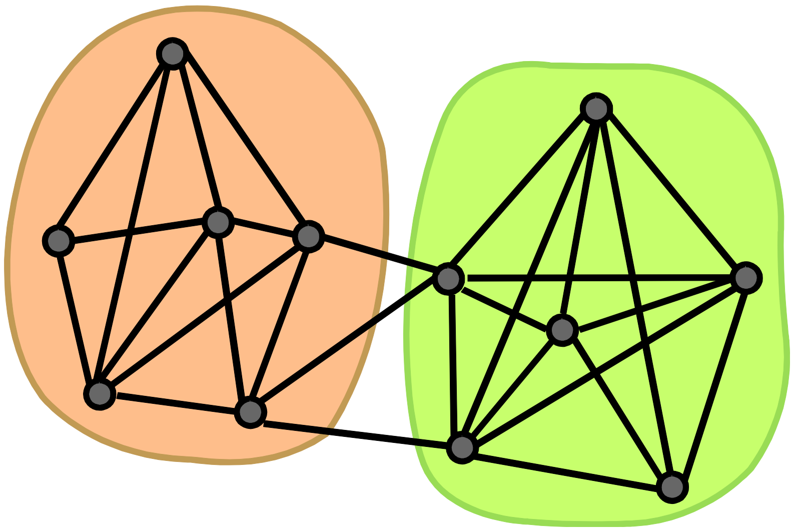
Community Detection: Overlapping Communities

Thanks to Jure Leskovec, Stanford and Panayiotis Tsaparas, Univ. of Ioannina for slides

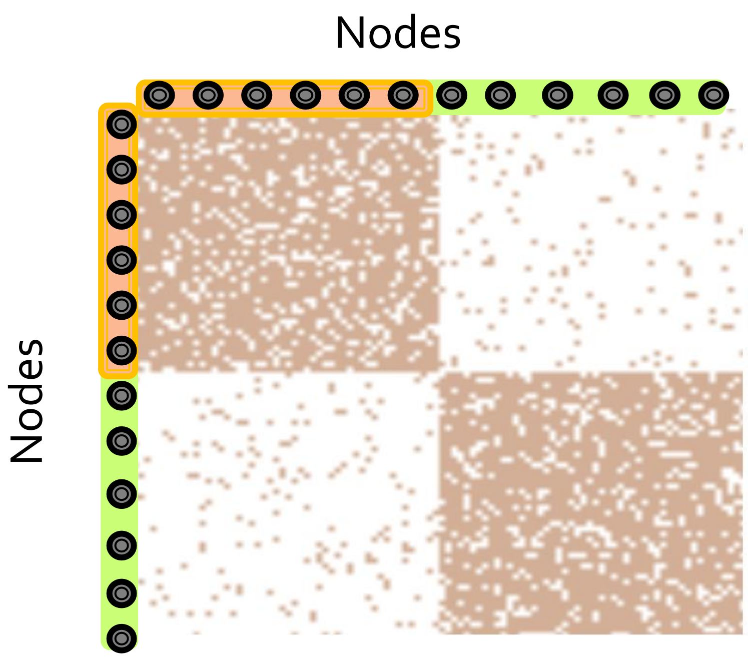
Agenda

- Overlapping Communities
- Cliques
- Clique Percolation Method (CPM)
- Modeling Networks with Communities
  - Community-Affiliation Graph Model (AGM)

Non-overlapping Communities

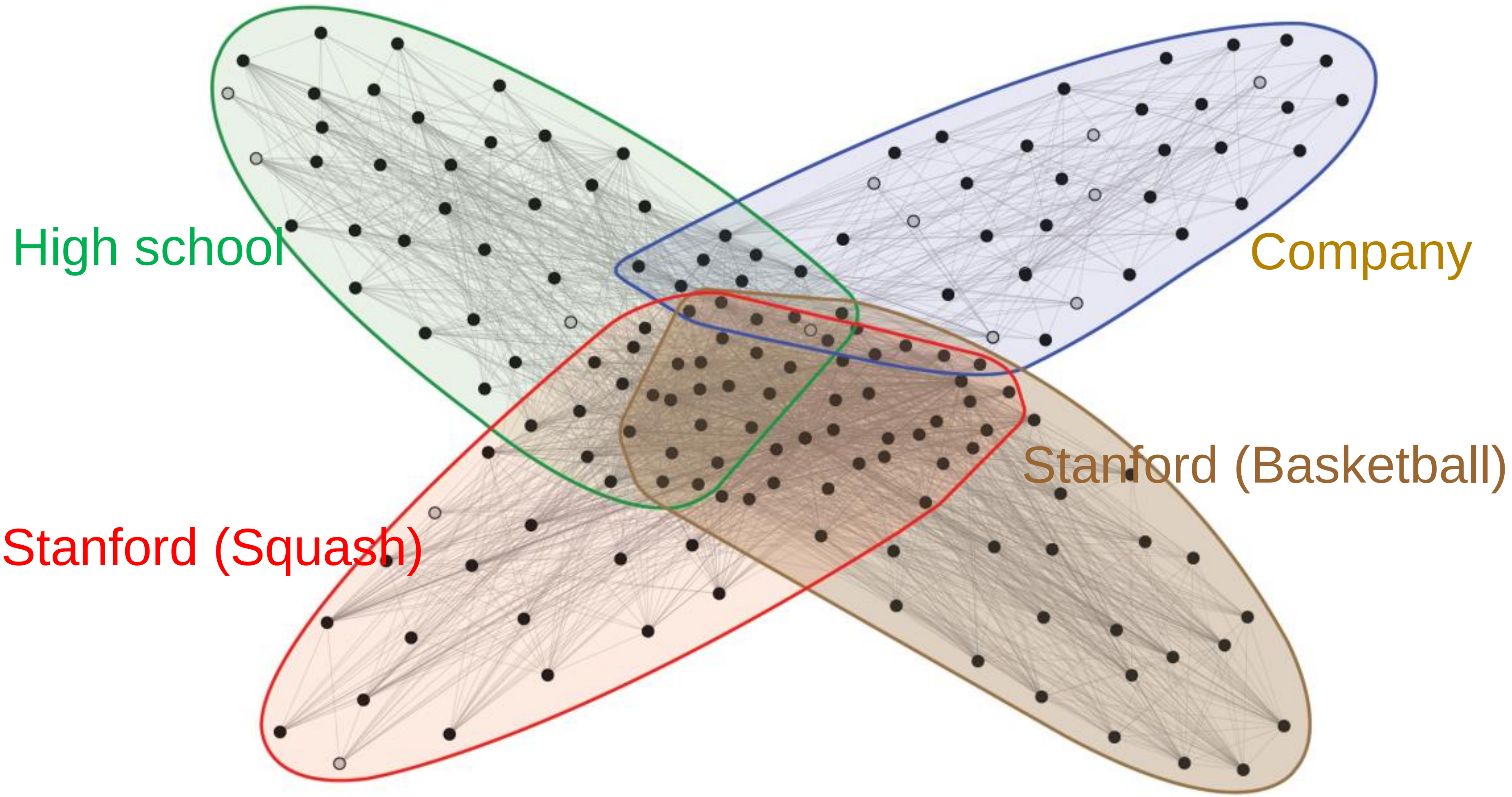


Network



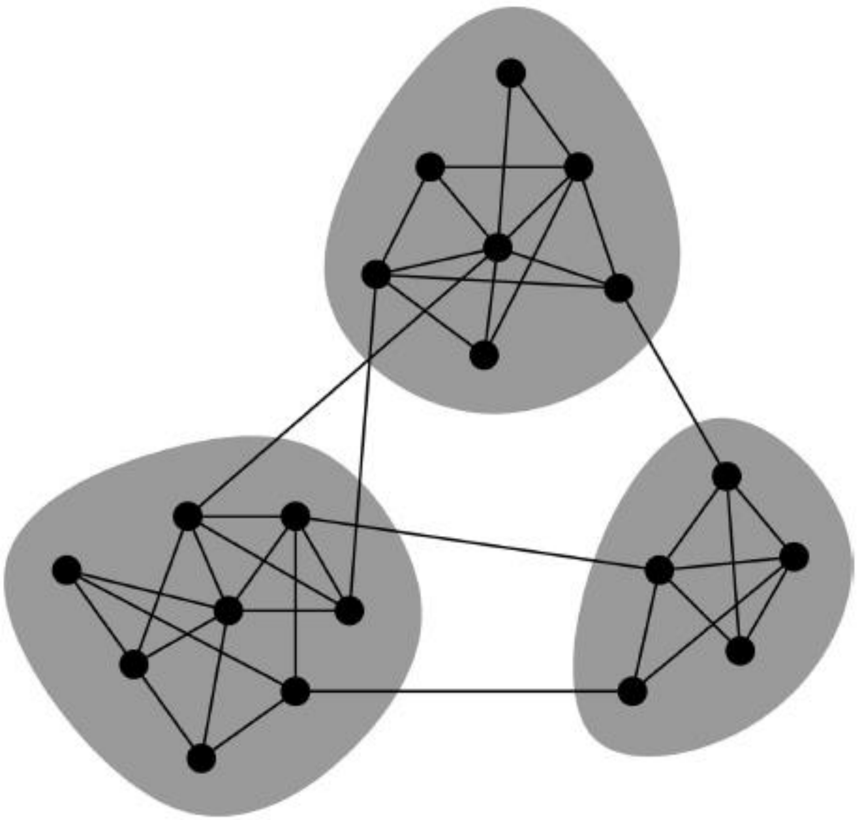
Adjacency matrix

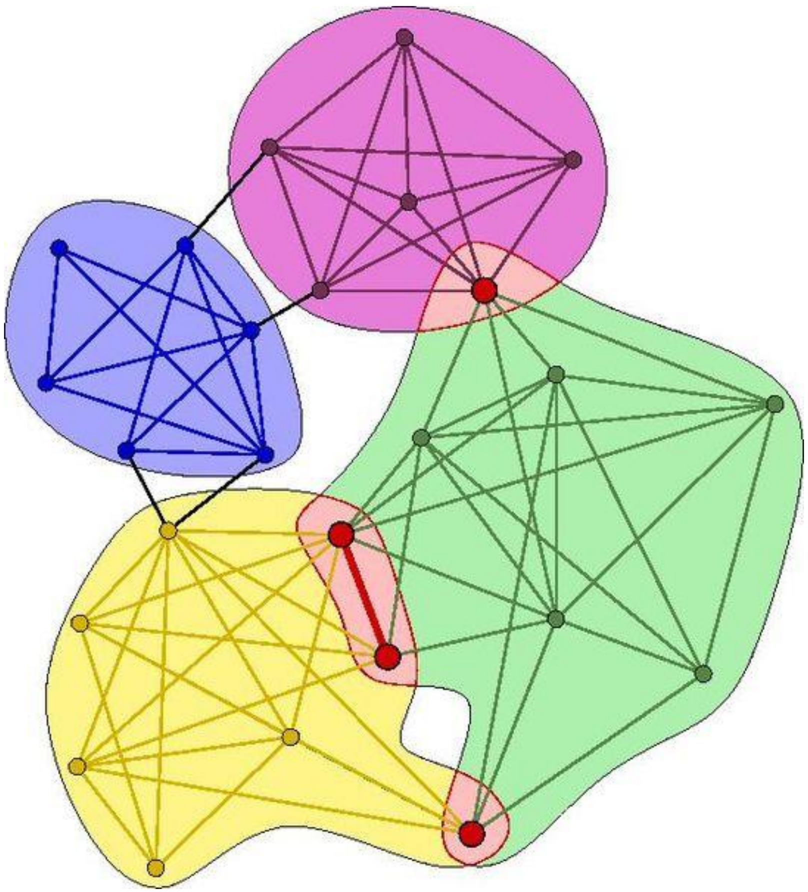
What if communities overlap?



Overlapping Communities

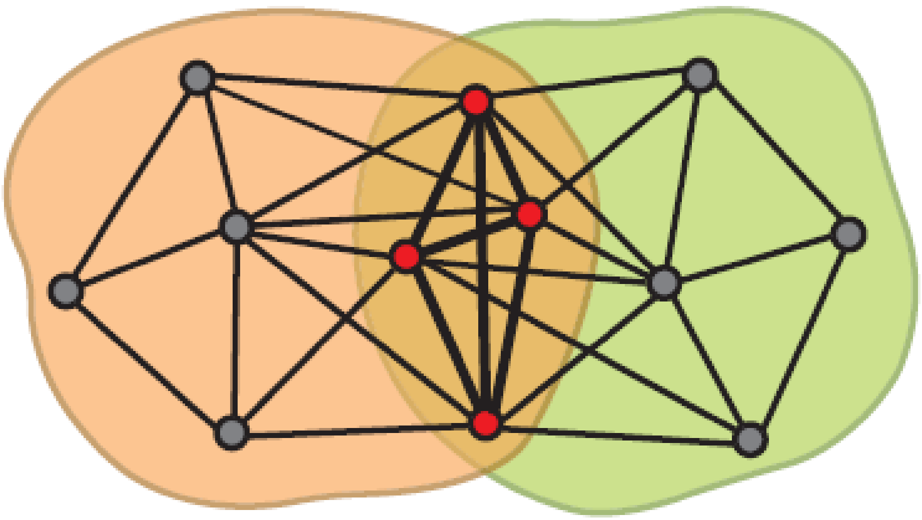
- **Non-overlapping vs. overlapping communities**

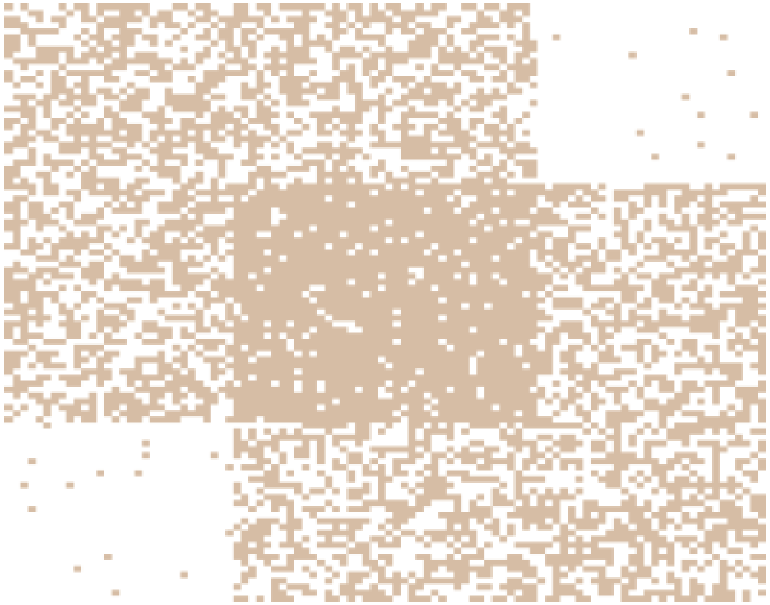




Overlapping Communities

What is the structure of community overlaps:
Edge density in the overlaps is higher!



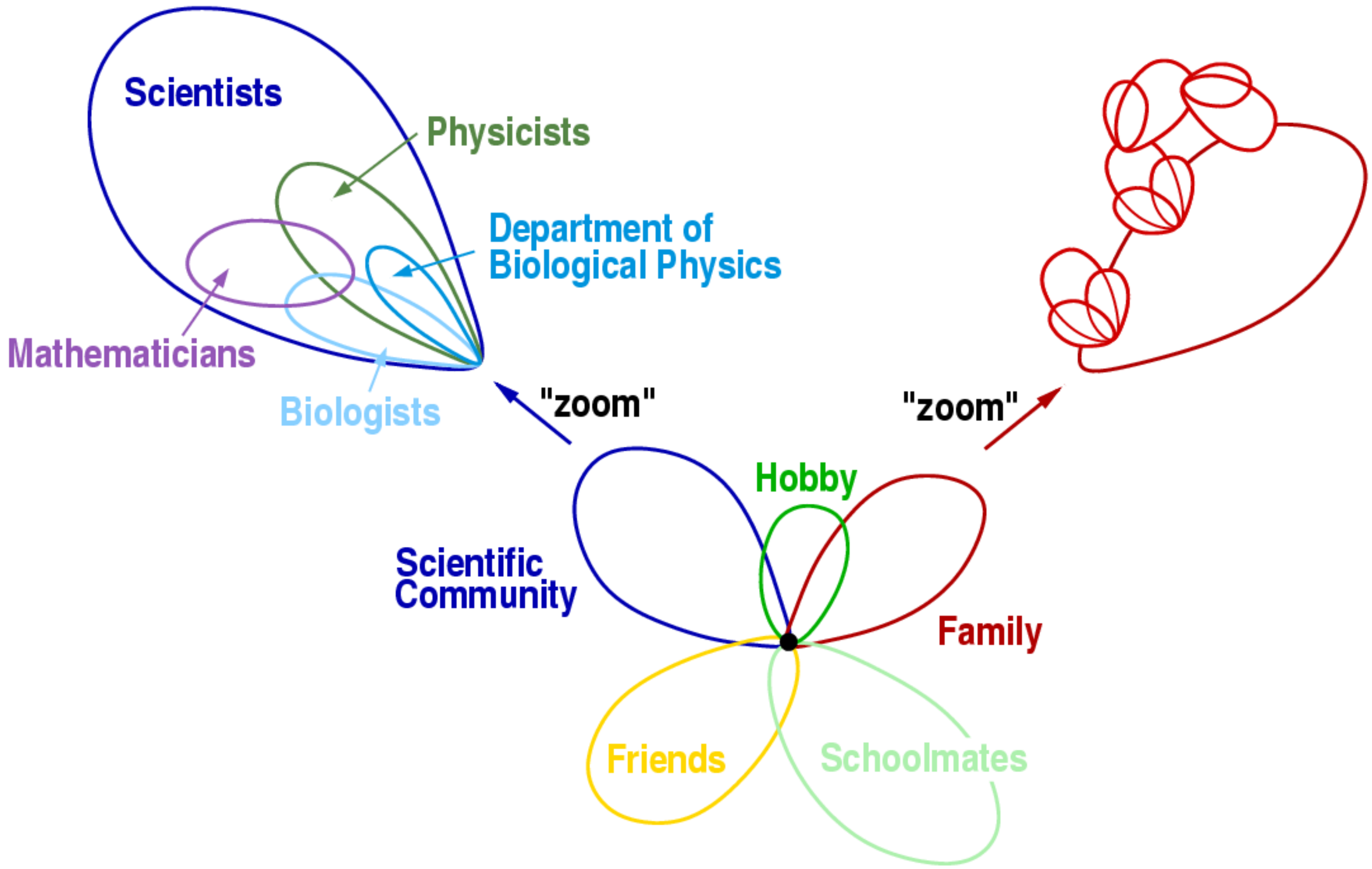


Communities as "tiles"

[Palla et al., '05]

Overlaps of Social Circles

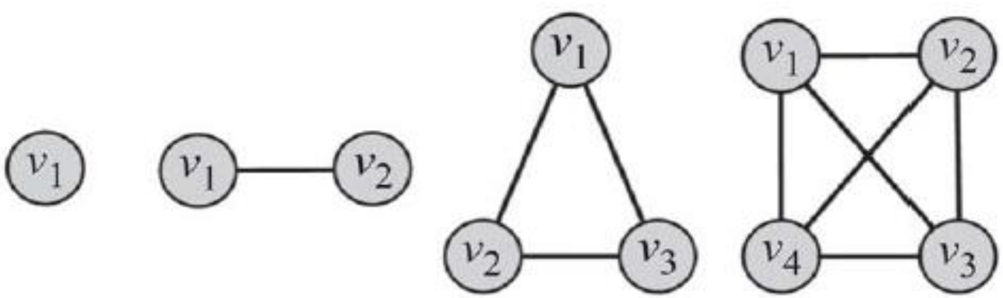
- **A node can belong to many social "circles"**



Cliques

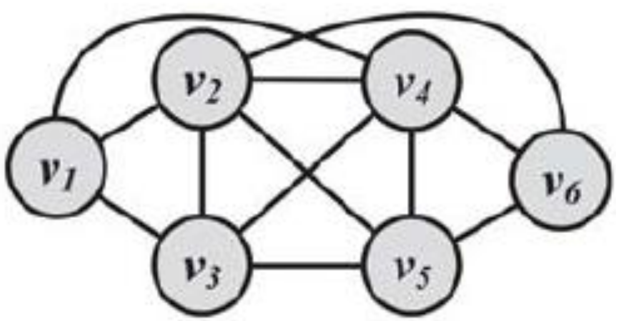
Cliques

- ***Clique***: a maximum ***complete subgraph*** in which all pairs of vertices are connected by an edge
- ***k-Clique***: A ***clique of size k*** is a subgraph of ***k*** vertices where the degree of all vertices in the induced subgraph is ***k-1***



Maximum Clique & Maximal Cliques

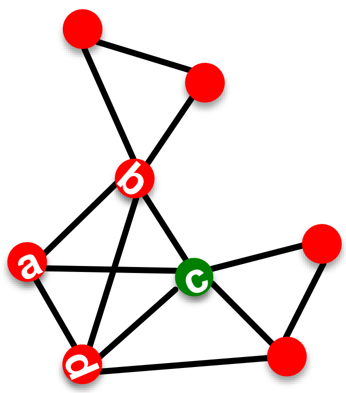
- Two problems
  - Find the ***maximum clique*** (the one with the largest number of vertices) or
  - Find all ***maximal cliques*** (cliques that are not subgraphs of a larger clique; i.e., cannot be expanded further).



- Both problems are ***NP-hard***

How to Find Maximal Cliques?

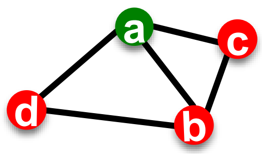
- **No nice way, hard combinatorial problem**
- **Maximal clique:** Clique that can't be extended
  - $\{a, b, c\}$ is a clique but not maximal clique
  - $\{a, b, c, d\}$ is maximal clique



- **Algorithm:** Sketch
  - Start with a seed node
  - Expand the clique around the seed
  - Once the clique cannot be further expanded we found the maximal clique
  - **Note:**
    - This will generate the same clique multiple times

How to Find Maximal Cliques?

- Start with a seed vertex $\boldsymbol{a}$
- **Goal:** Find the max clique $\boldsymbol{Q}$ that $\boldsymbol{a}$ belongs to
  - **Observation:**
    - If some $\boldsymbol{x}$ belongs to $\boldsymbol{Q}$ then it is a neighbor of $\boldsymbol{a}$
      - **Why?** If $\boldsymbol{a}, \boldsymbol{x} \in \boldsymbol{Q}$ but edge $(\boldsymbol{a}, \boldsymbol{x})$ does not exist, $\boldsymbol{Q}$ is not a clique!
- **Recursive algorithm:**
  - $\boldsymbol{Q}$ … current clique
  - $\boldsymbol{R}$ … candidate vertices to expand the clique to
- **Example:** Start with $\boldsymbol{a}$ and expand around it



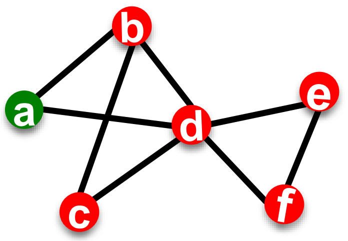
Q=
R=

Steps of the recursive algorithm

Γ(u)…neighbor set of *u*

How to Find Maximal Cliques?

- ***Q*** ... current clique
- ***R*** ... candidate vertices
- **`Expand(R,Q)`**
  - **`while`** `R ≠ {}`
    - `p = vertex in R`
    - $Q_p = Q \cup \{p\}$
    - $R_p = R \cap \Gamma(p)$
    - **if** $R_p \neq \{\}$: $\mathrm{Expand}(R_p, Q_p)$
      else: output Q_p
    - `R = R – {p}`



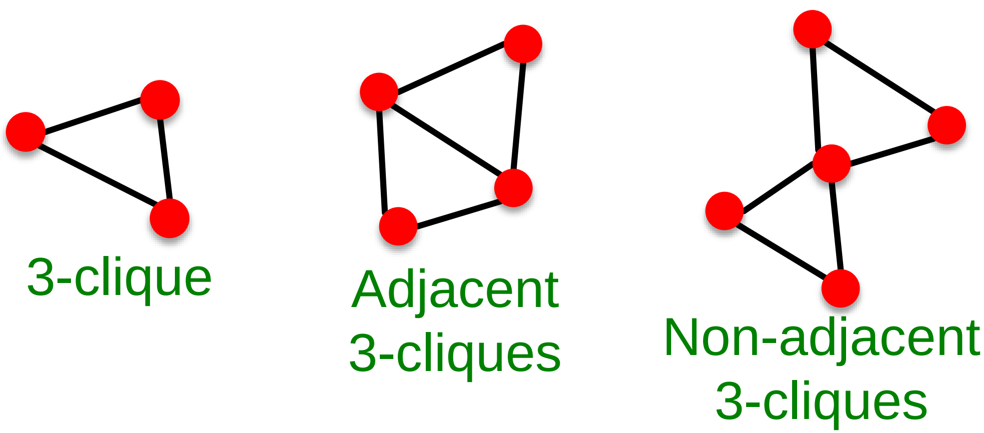
Pruning

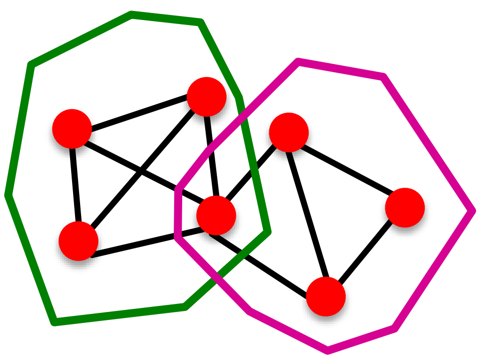
- Prune all vertices (and incident edges) with degrees less than ***k-1***
  - Effective due to the power-law distribution of vertex degrees
- "Exact cliques" are rarely observed in real networks
  - A clique of 1,000 vertices has 499,500 edges
  - A single edge removal results in a subgraph that is no longer a clique (less than 0.0002% of the edges)
- ***Relaxing Cliques***
  - All vertices have a minimum degree but not necessarily ***k-1***

Clique Percolation Method

Clique Percolation Method (CPM)

- **Two nodes belong to the same community if they can be connected through adjacent *k*-cliques:**
  - ***k*-clique:**
    - Fully connected graph on ***k*** nodes
  - **Adjacent *k*-cliques:**
    - overlap in ***k-1*** nodes
- ***k*-clique community**
  - Set of nodes that can be reached through a sequence of adjacent ***k***-cliques

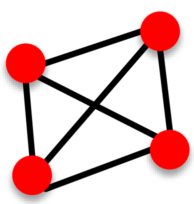




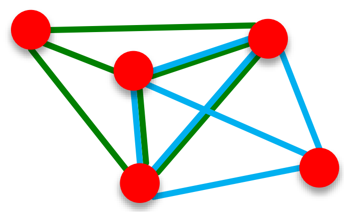
Two overlapping *3*-clique communities

Clique Percolation Method (CPM)

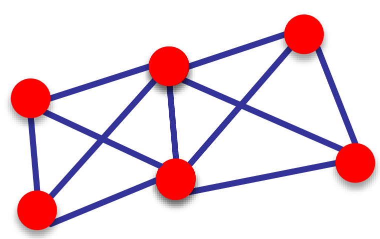
- **Two nodes belong to the same community if they can be connected through adjacent *k*-cliques:**



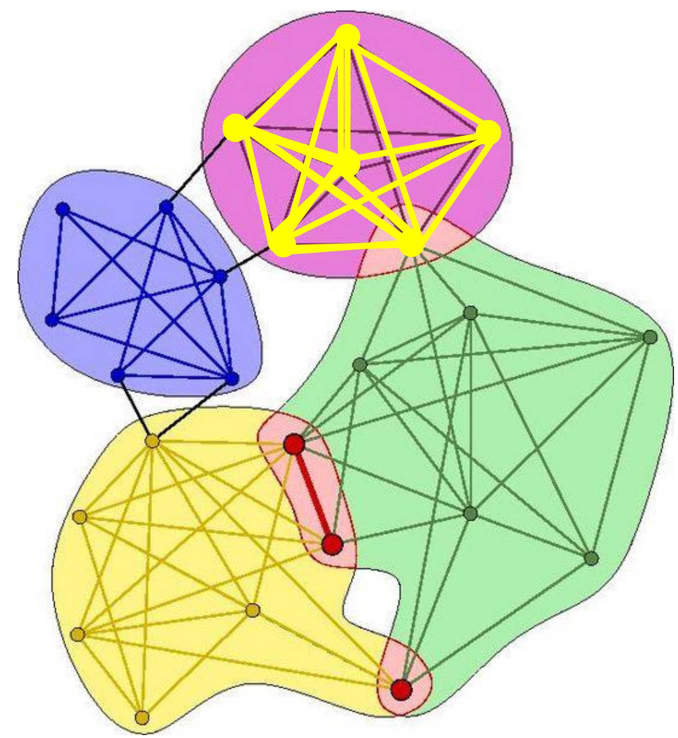
4-clique



Adjacent 4-cliques



Non-adjacent 4-cliques



Communities for k=4

(CPM): Using Cliques as Seeds

- Given k, find all cliques of size k.
- Create graph (clique graph) where all cliques are vertices, and two cliques that share k - 1 vertices are connected via an edge.
- Communities are the connected components of this graph.

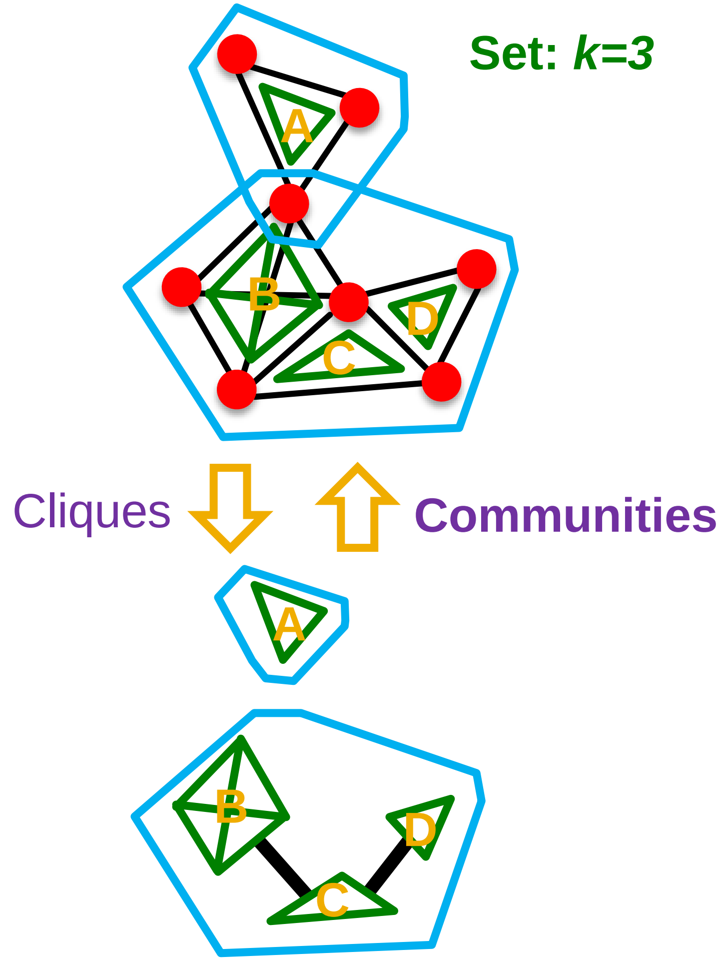
Algorithm 6.2 Clique Percolation Method (CPM)

Require: parameter k

1: **return** Overlapping Communities
2: $Cliques_k$ = find all cliques of size k
3: Construct clique graph $G(V, E)$, where $|V| = |Cliques_k|$
4: $E = \{e_{ij} \mid \text{clique } i \text{ and clique } j \text{ share } k-1 \text{ nodes}\}$
5: Return all connected components of G

CPM: Steps explained

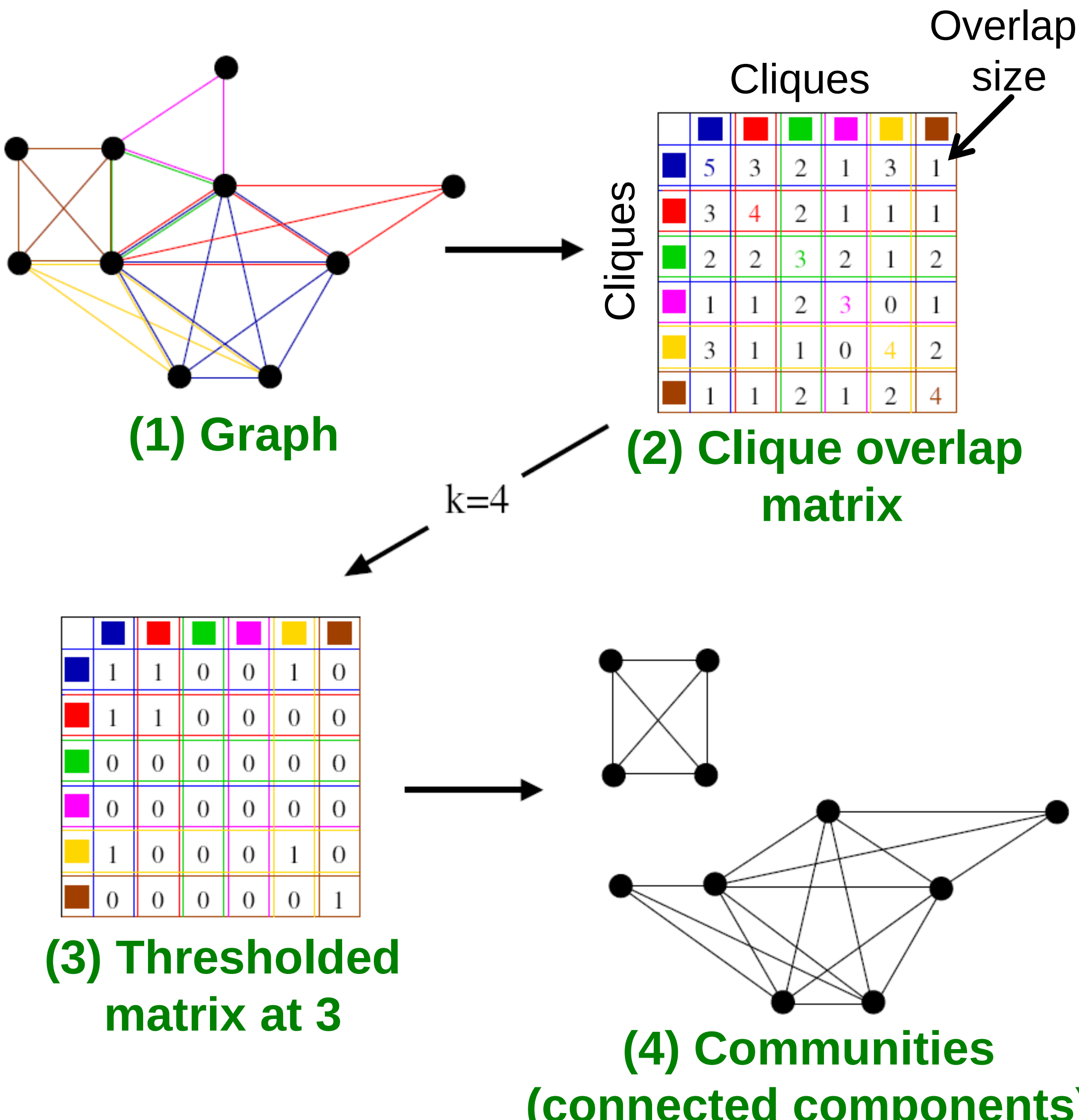
- **Clique Percolation Method:**
  - **Find maximal-cliques**
    - Def: Clique is maximal if no superset is a clique
  - **Clique overlap super-graph:**
    - Each clique is a super-node
    - Connect two cliques if they overlap in at least k-1 nodes
  - **Communities:**
    - Connected components of the clique overlap matrix



- **How to set k?**
  - Set k so that we get the "richest" (most widely distributed cluster sizes) community structure

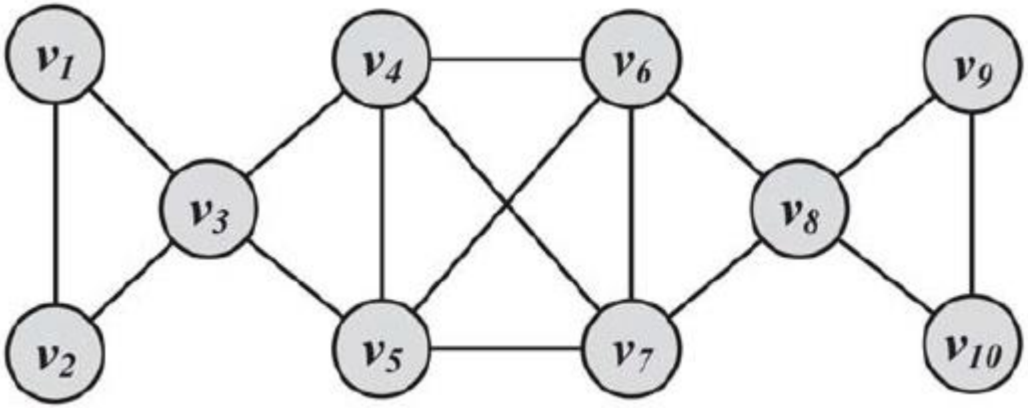
CPM method: Example

- Start with graph
- Find maximal cliques
- Create clique overlap matrix
- Threshold the matrix at value ***k-1***
  - If $a_{ij} < k - 1$ *set* 0
- Communities are the connected components of the thresholded matrix



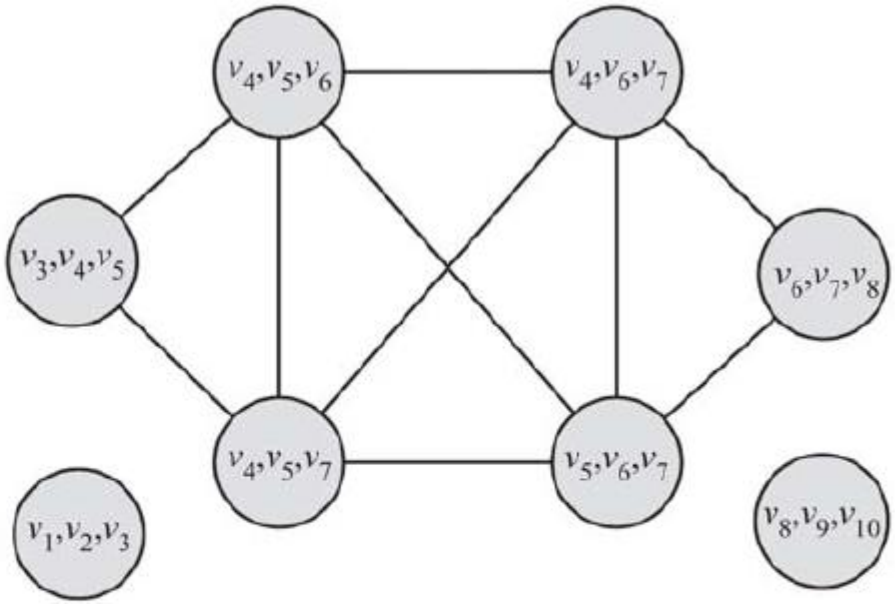
(CPM): Using Cliques as Seeds

- Input graph, let k = 3



(CPM): Using Cliques as Seeds

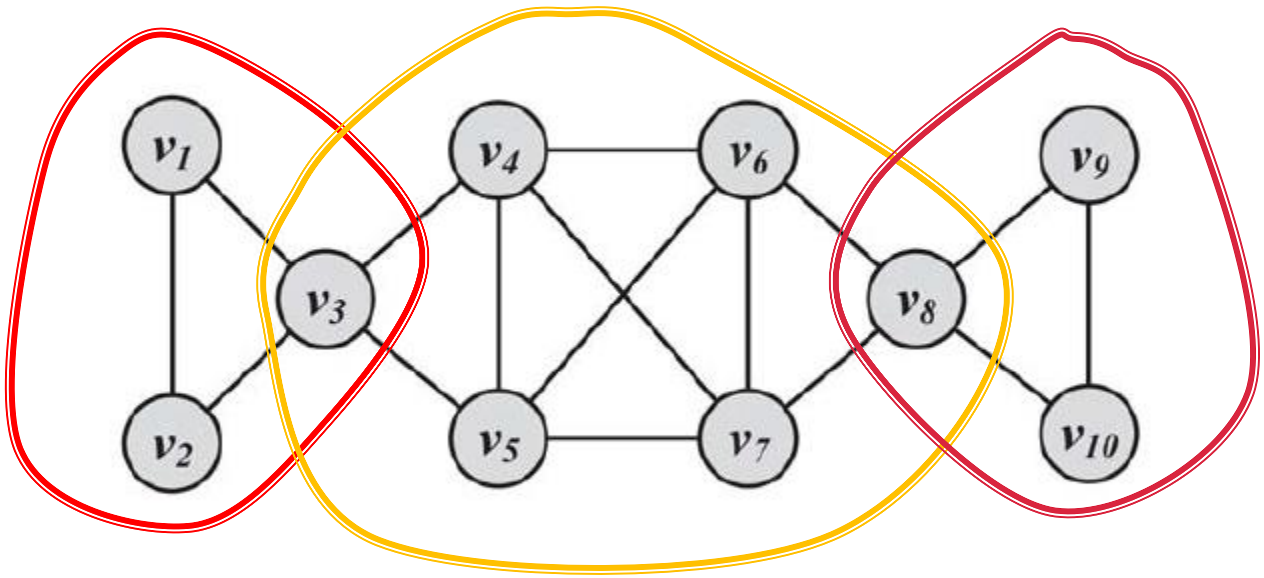
- Clique graph for k = 3



- (v1, v2, ,v3)
- (v8, v9, v10)
- (v3, v4, v5, v6, v7, v8)

(CPM): Using Cliques as Seeds

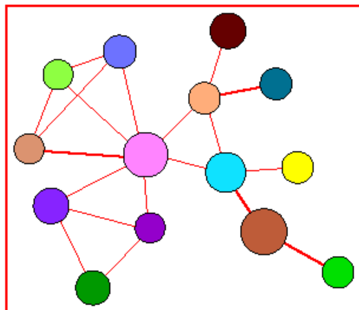
- Result

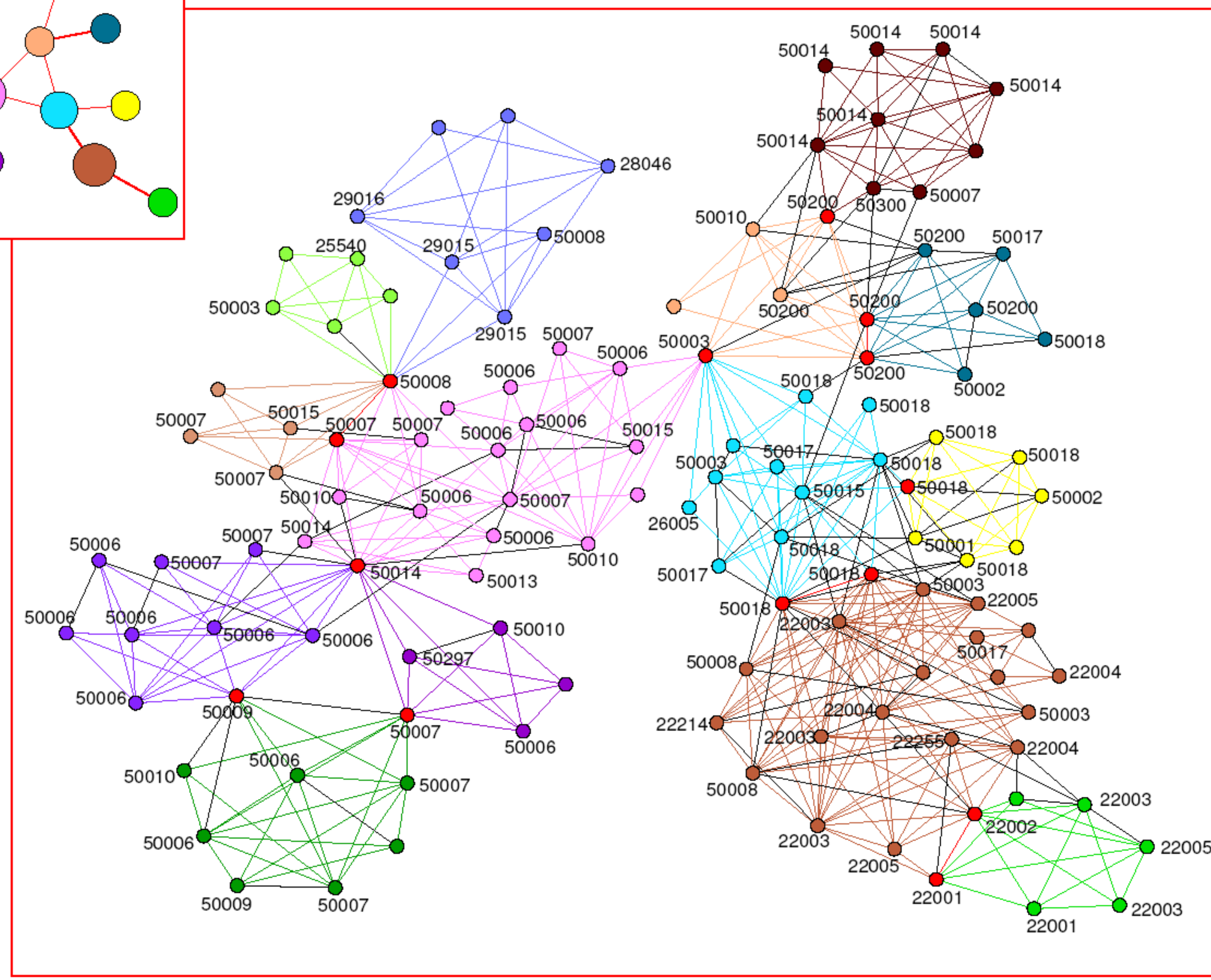


- (v1, v2, v3)
- (v8, v9, v10)
- (v3, v4, v5, v6, v7, v8)

Note: the example protein network was detected using a CPM algorithm

Example: Phone-Call Network



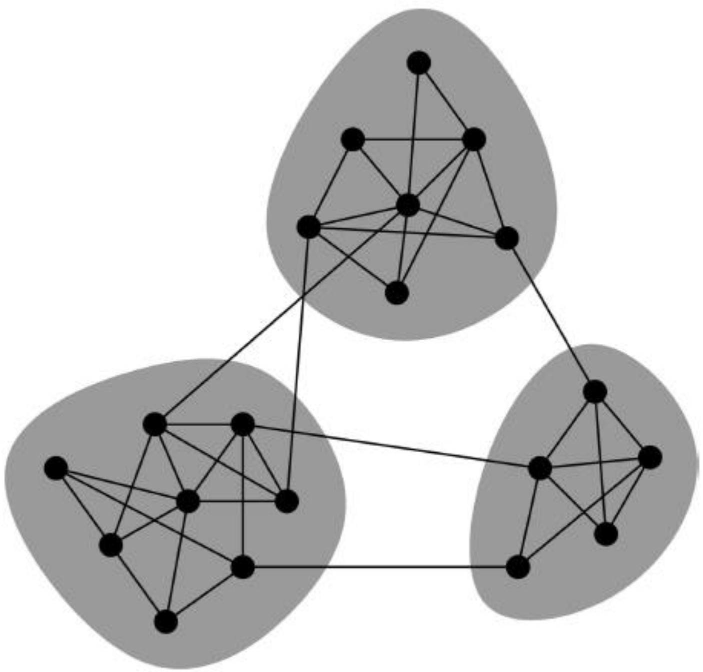


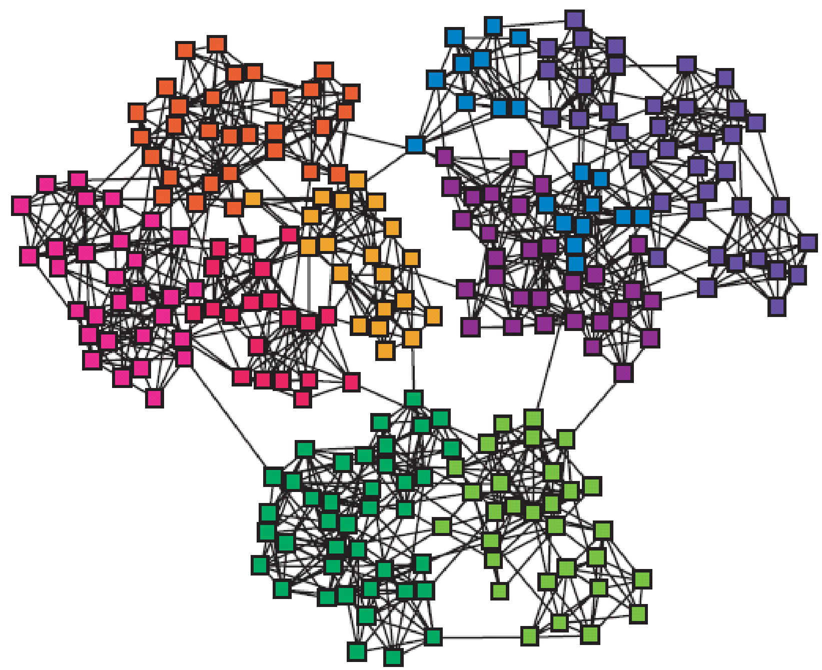
Communities in a "tiny" part of a phone call network of 4 million users [Palla et al., '07]

How to Model Networks with Communities?

Network and Communities

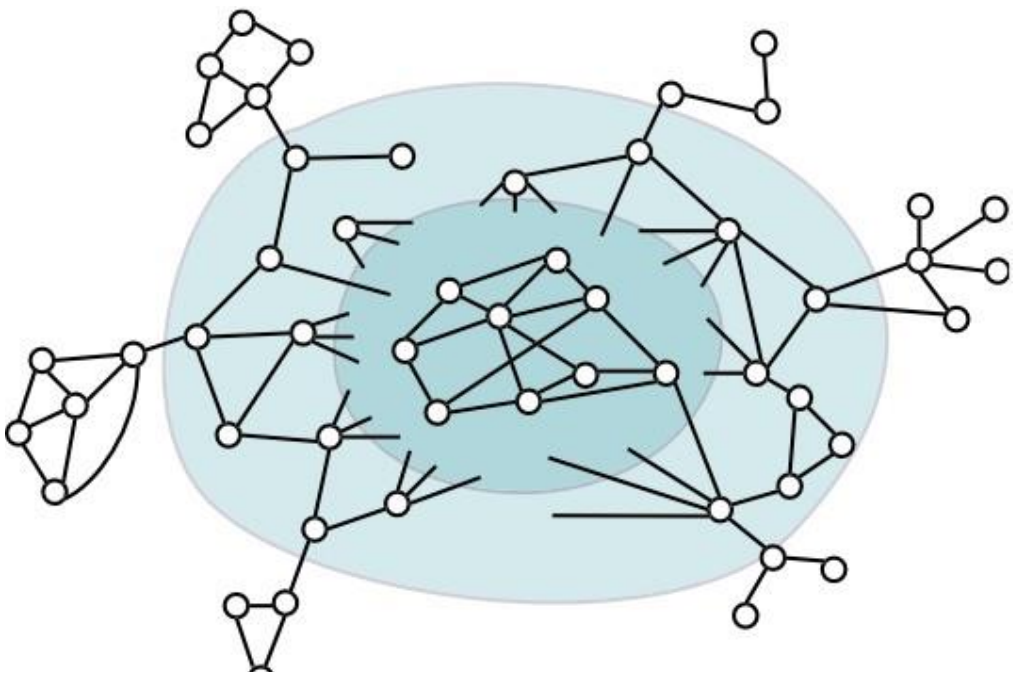
- **How should we think about large scale organization of clusters in networks?**
  - **Finding:** Community Structure





Network and Communities

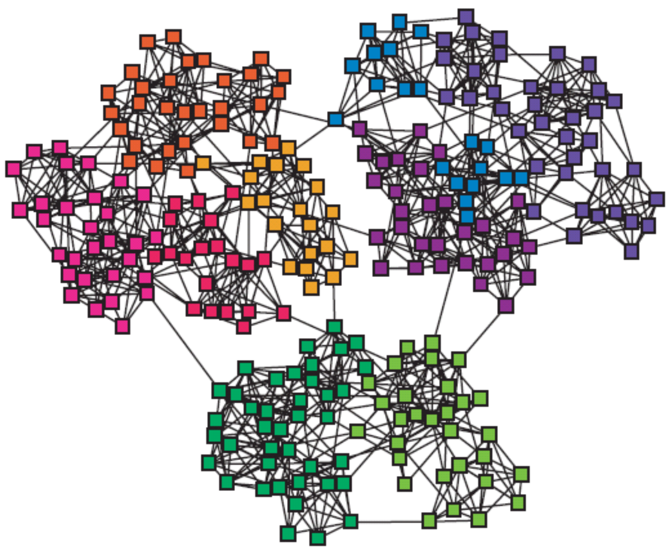
- **How should we think about large scale organization of clusters in networks?**
  - **Finding:** Core-periphery structure



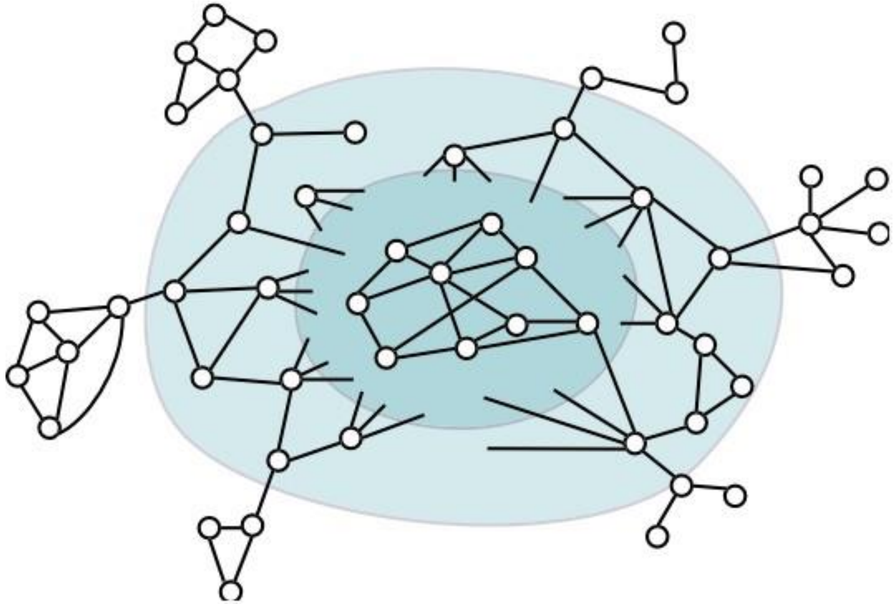
Nested Core-Periphery

Network and Communities

- **How do we reconcile these two views?**
  (and still do community detection)



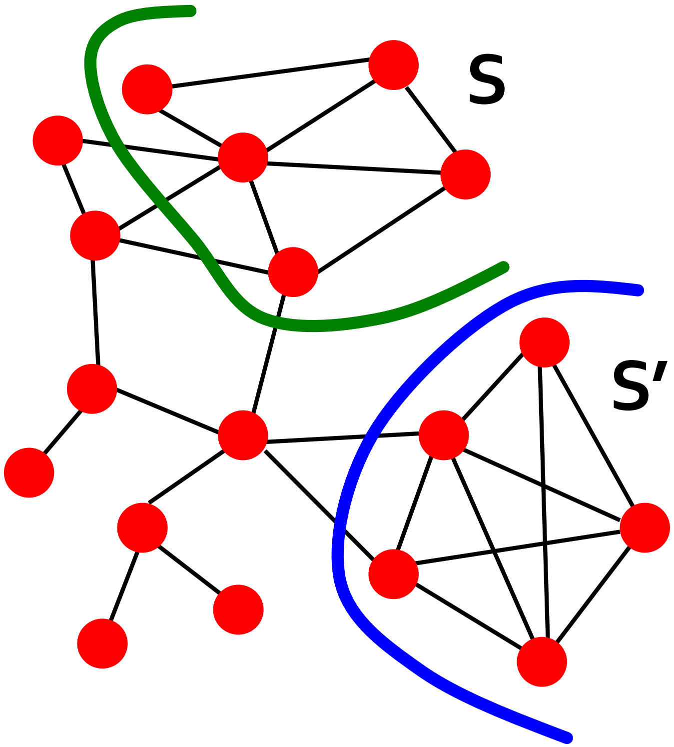
vs.



Community structure **Core-periphery**

Community Score

- **How community-like is a set of nodes?**
- **A good cluster *S* has**
  - Many edges internally
  - Few edges pointing outside
- What's a good metric: **Conductance**



$$\phi(S) = \frac{|\{(i,j) \in E; i \in S, j \notin S\}|}{\sum_{s \in S} d_s}$$

Small conductance corresponds to good clusters

(Note |S| < |V|/2)

Network Community Profile Plot

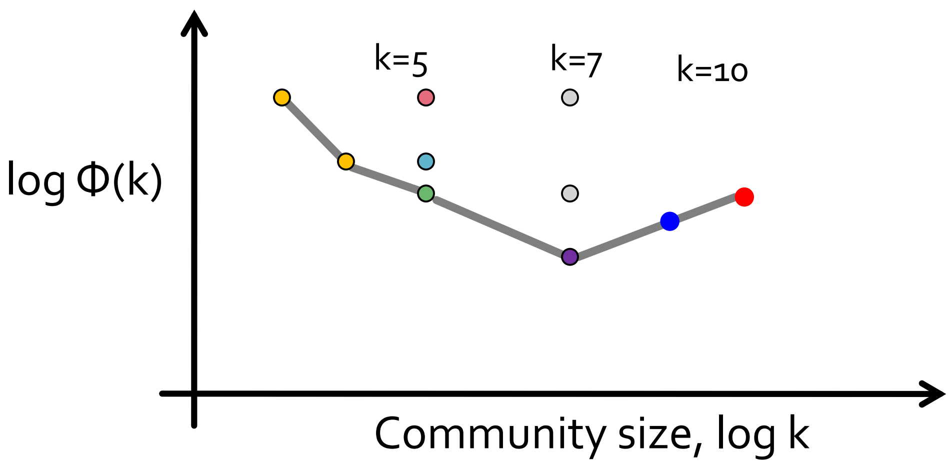
(Note |S| < |V|/2)

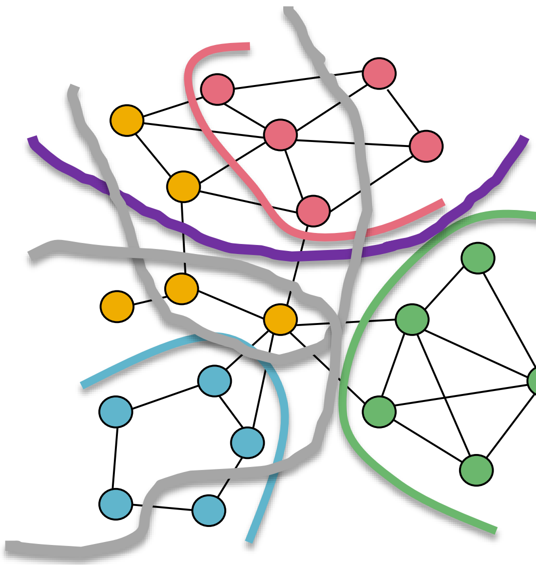
- <u>Define:</u>

  Network community profile (**NCP**) plot

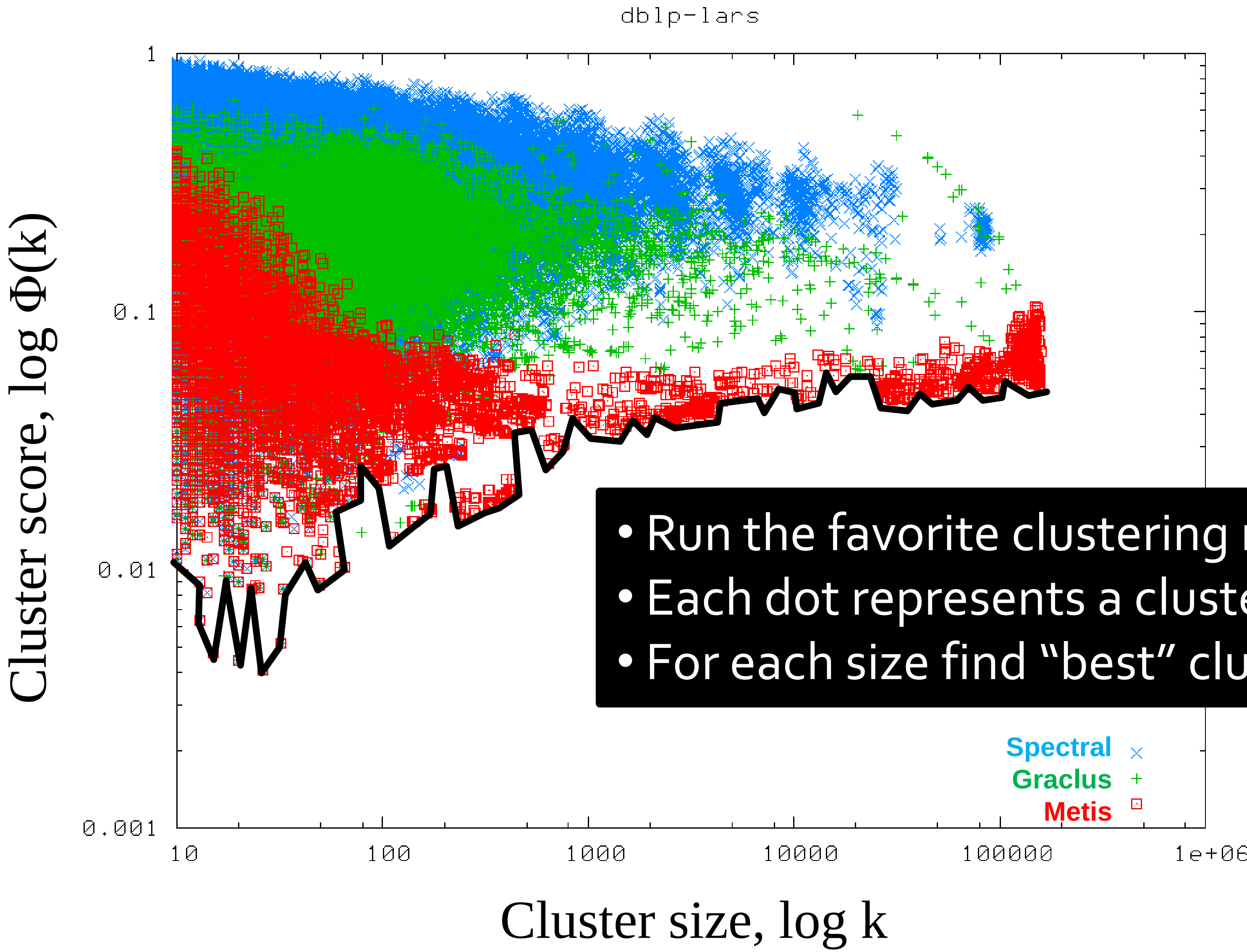
  Plot the score of **best** community of size *k*

$$\Phi(k) = \min_{S \subset V, |S|=k} \phi(S)$$





How to (Really) Compute NCP?

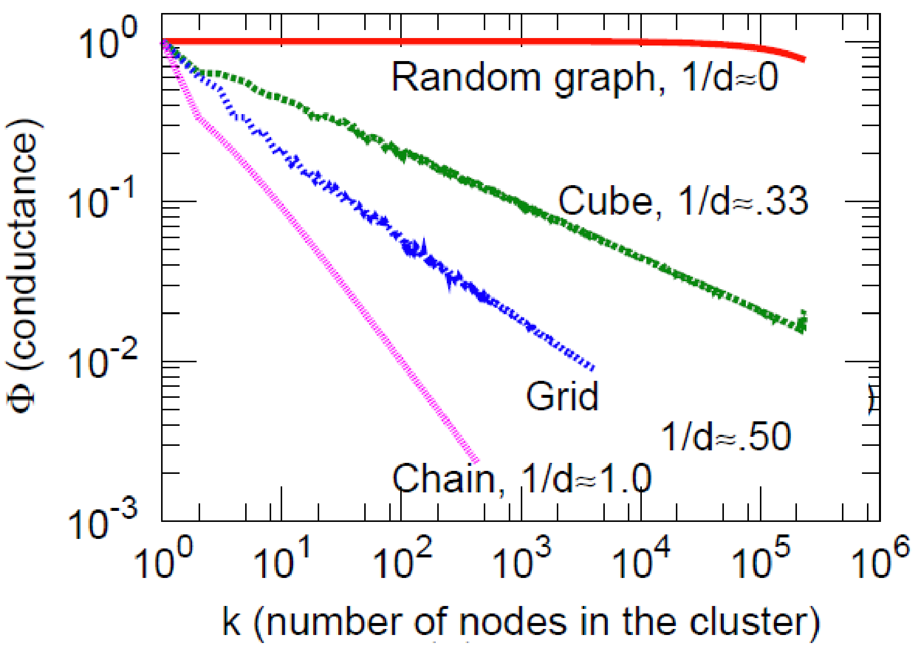


- Run the favorite clustering method
- Each dot represents a cluster
- For each size find “best” cluster

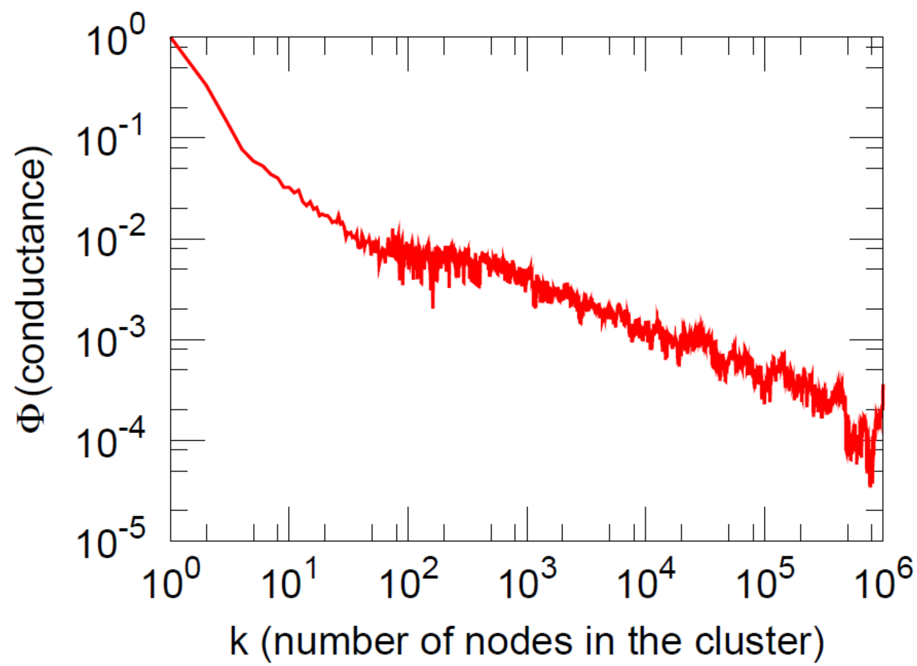
[WWW '08]

NCP Plot: Meshes

- **Meshes, grids, dense random graphs:**



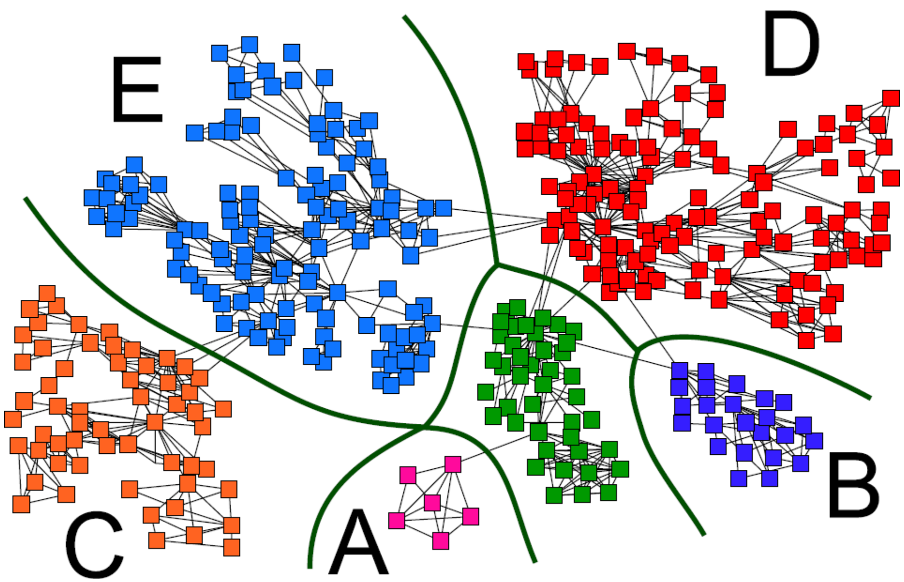
d-dimensional meshes

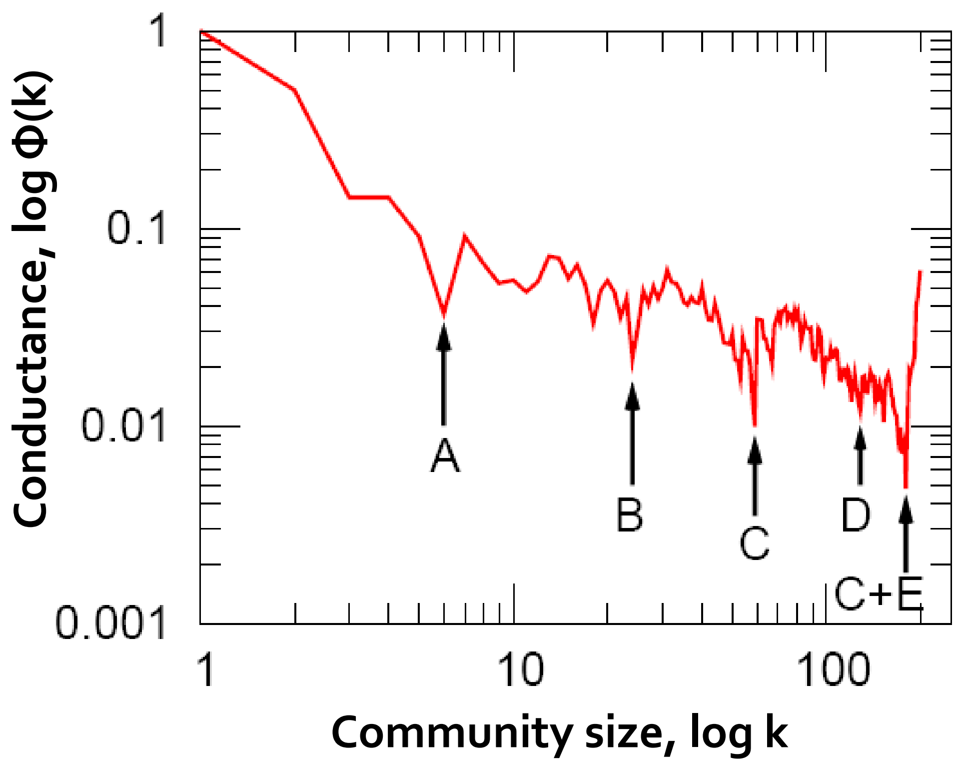


California road network

NCP plot: Network Science

- **Collaborations between scientists in networks**
  [Newman, 2005]



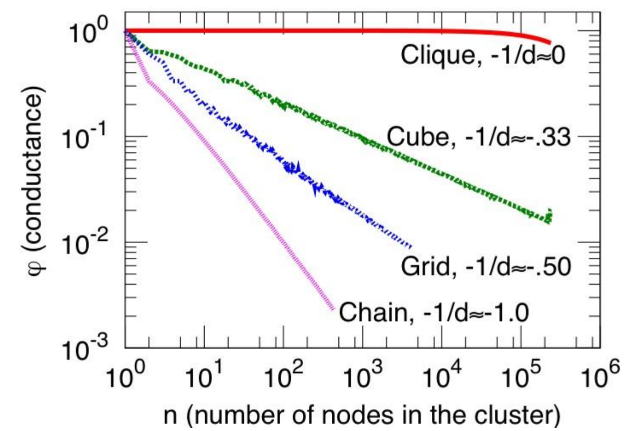


Dips in the conductance graph correspond to the "good" clusters we can visually detect

Natural Hypothesis

Natural hypothesis about NCP:

- NCP of real networks slopes downward
- Slope of the NCP corresponds to the "dimensionality" of the network

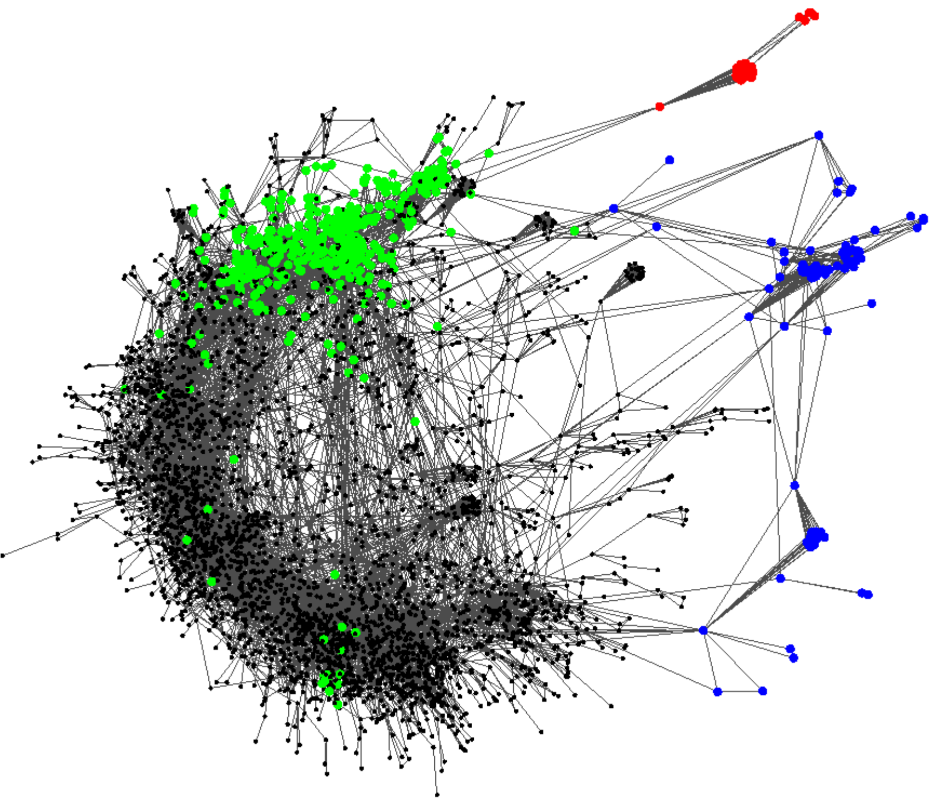


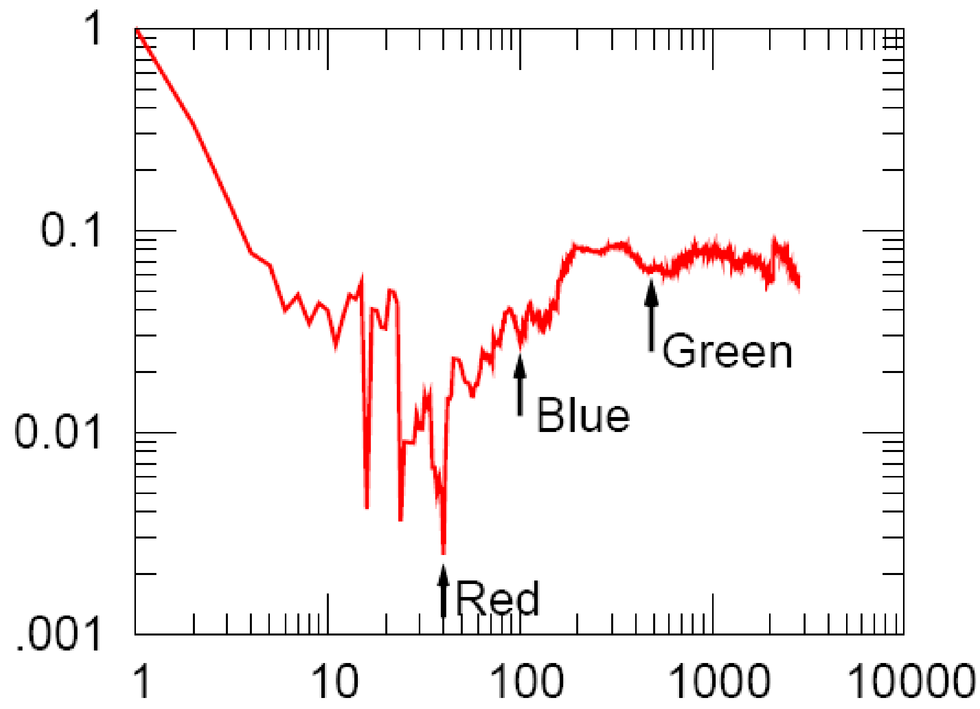
What about large networks?

• Social nets	Nodes	Edges	Description
LiveJournal	4,843,953	42,845,684	Blog friendships [5]
Epinions	75,877	405,739	Trust network [28]
CA-DBLP	317,080	1,049,866	Co-authorship [5]
• Information (citation) networks			
Cit-hep-th	27,400	352,021	Arxiv hep-th [14]
AmazonProd	524,371	1,491,793	Amazon products [8]
• Web graphs			
Web-google	855,802	4,291,352	Google web graph
Web-wt10g	1,458,316	6,225,033	TREC WT10G
• Bipartite affiliation (authors-to-papers) networks			
Atp-DBLP	615,678	944,456	DBLP [21]
AtM-Imdb	2,076,978	5,847,693	Actors-to-movies
• Internet networks			
AsSkitter	1,719,037	12,814,089	Autonom. sys.
Gnutella	62,561	147,878	P2P network [29]

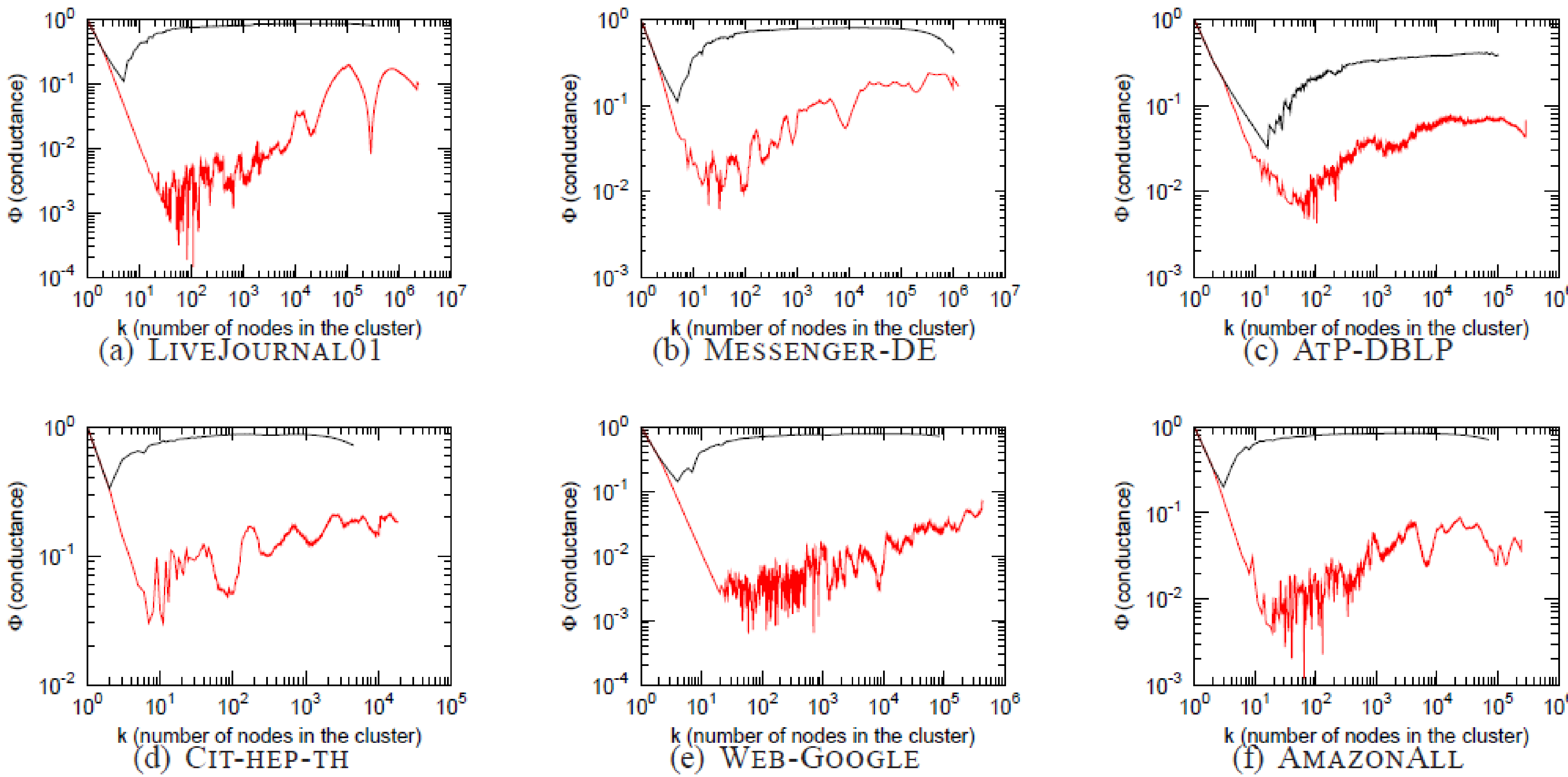
Large Networks: Very Different

Typical example: General Relativity collaborations (n=4,158, m=13,422)





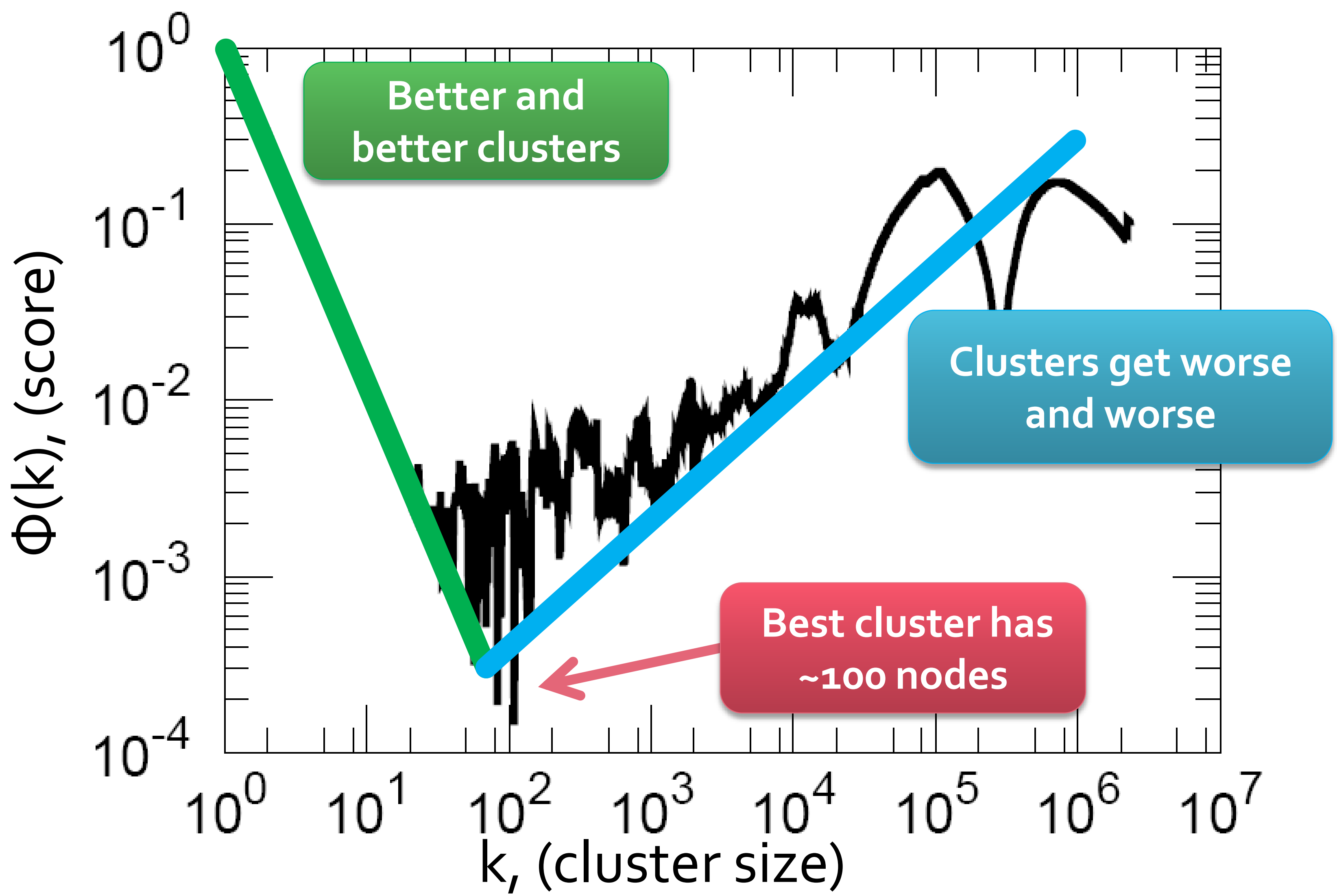
More NCP Plots of Networks



-- Rewired graph

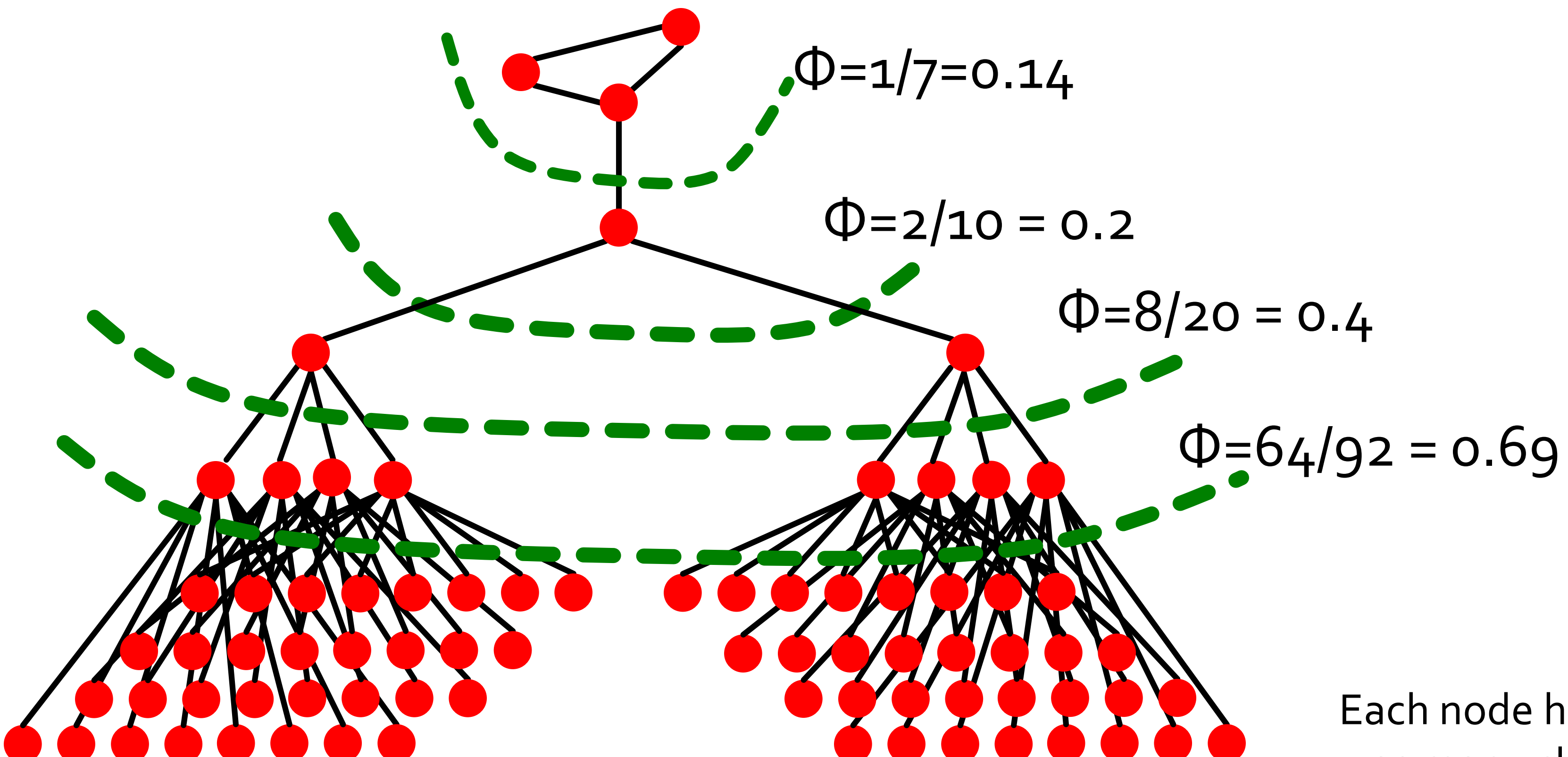
-- Real graph

NCP: LiveJournal (n=5m, m=42m)



Explanation: The Upward Part

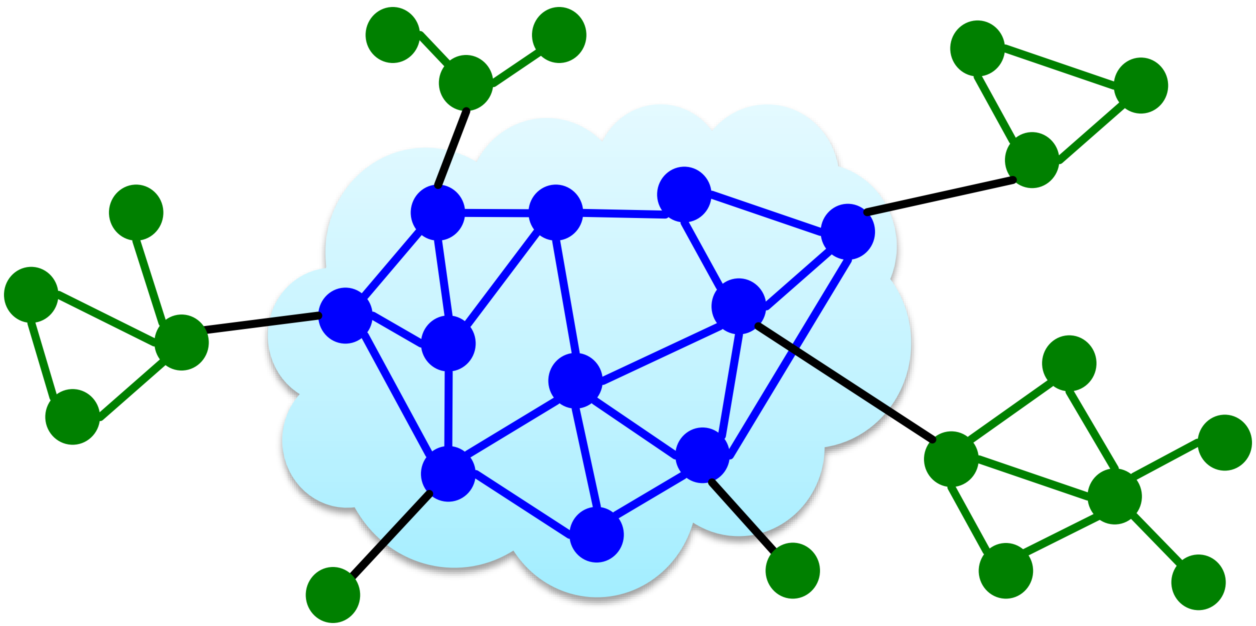
- As clusters grow the number of edges inside grows **slower** that the number crossing

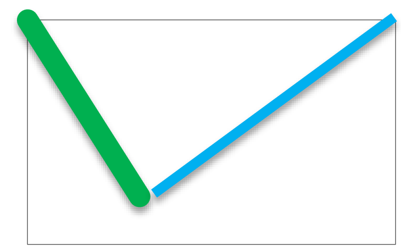


Each node has twice as many children

Explanation: Downward Part

- Empirically we note that **best clusters** (corresponding to **green nodes**) are **barely connected** to the network

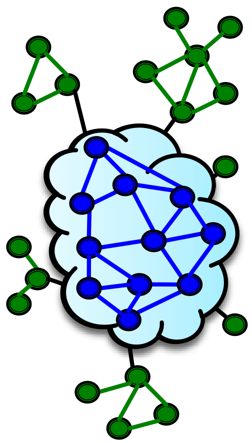


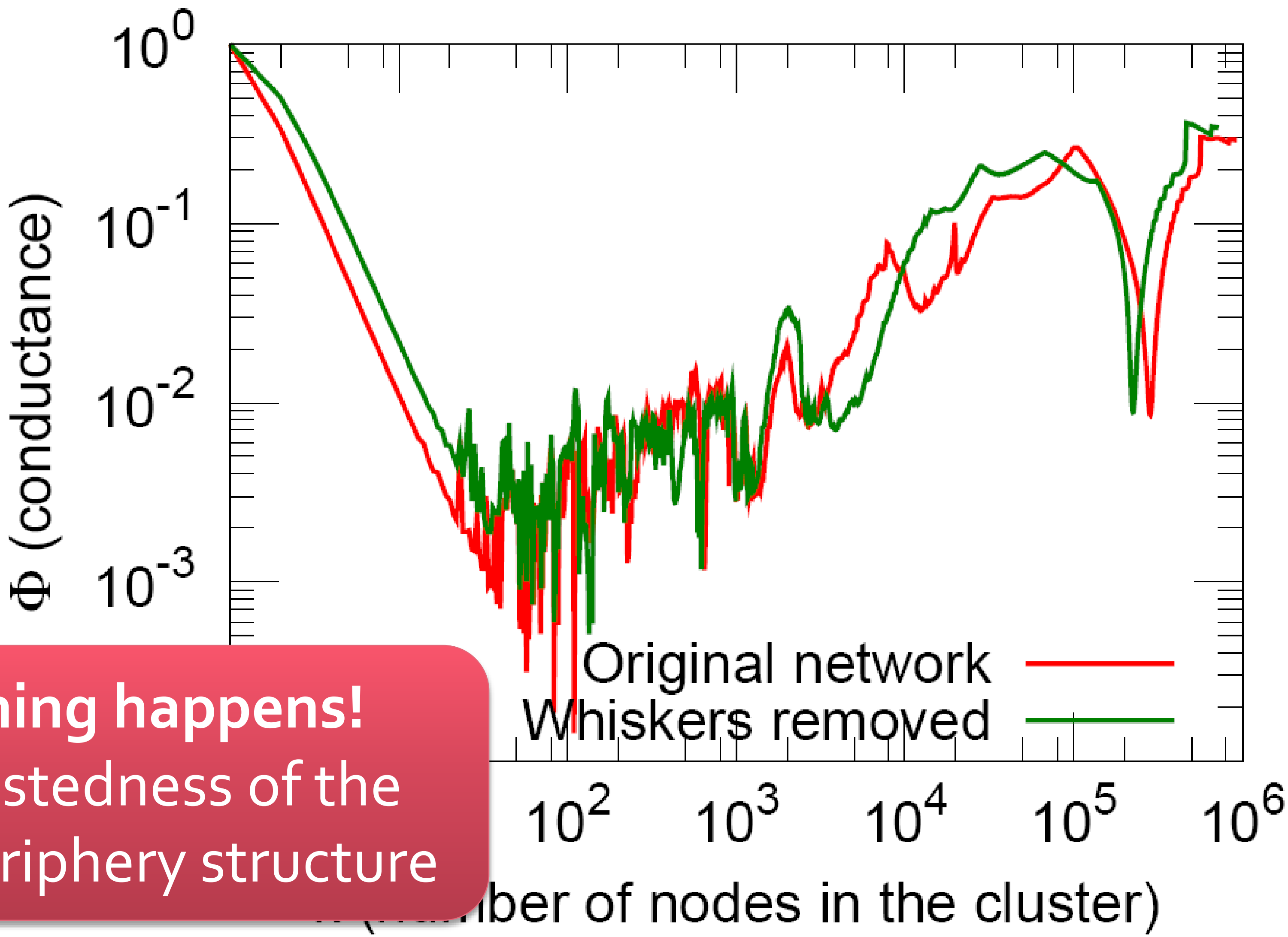


NCP plot

⇒ Core-periphery structure

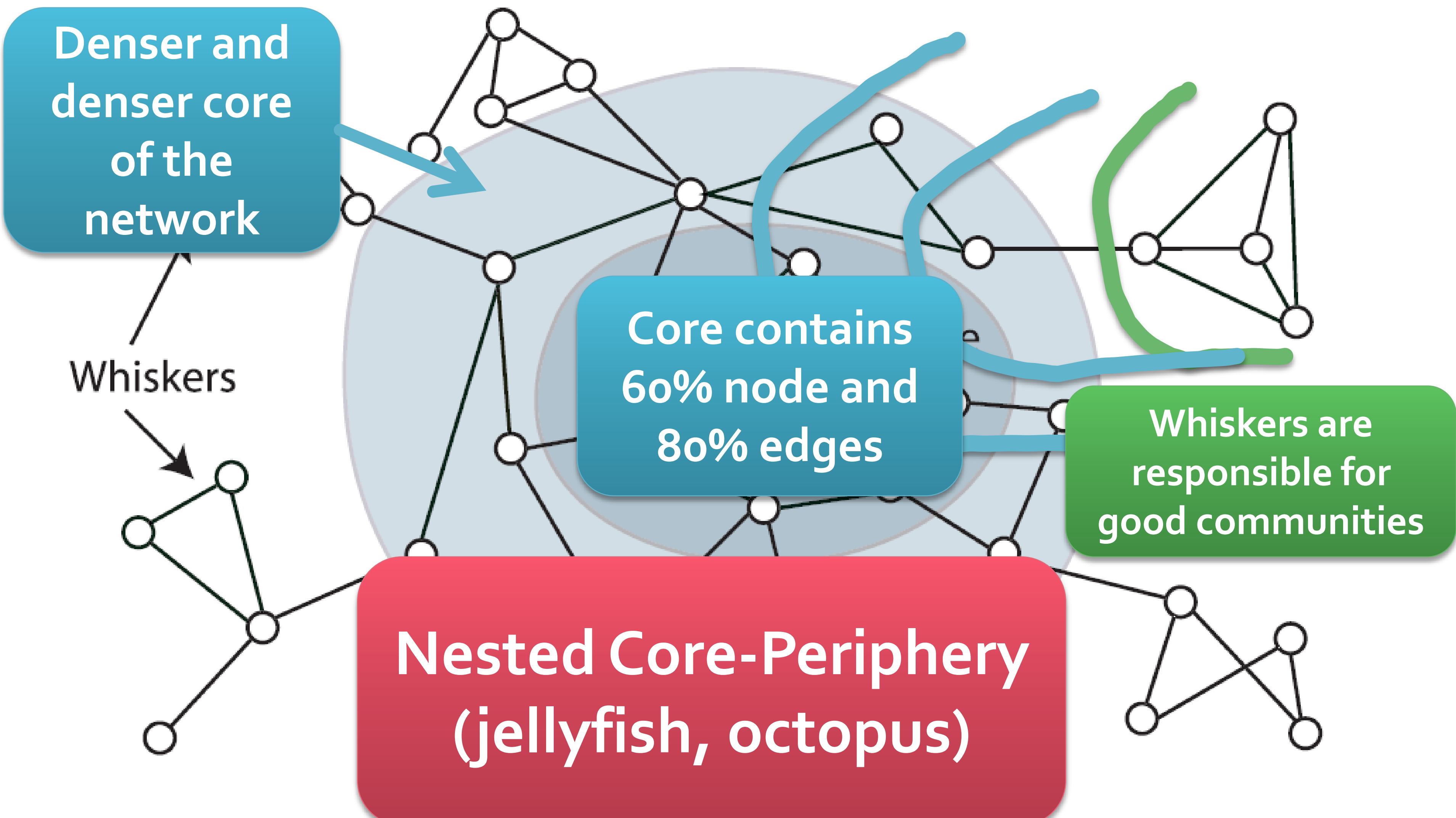
What If We Remove Good Clusters?



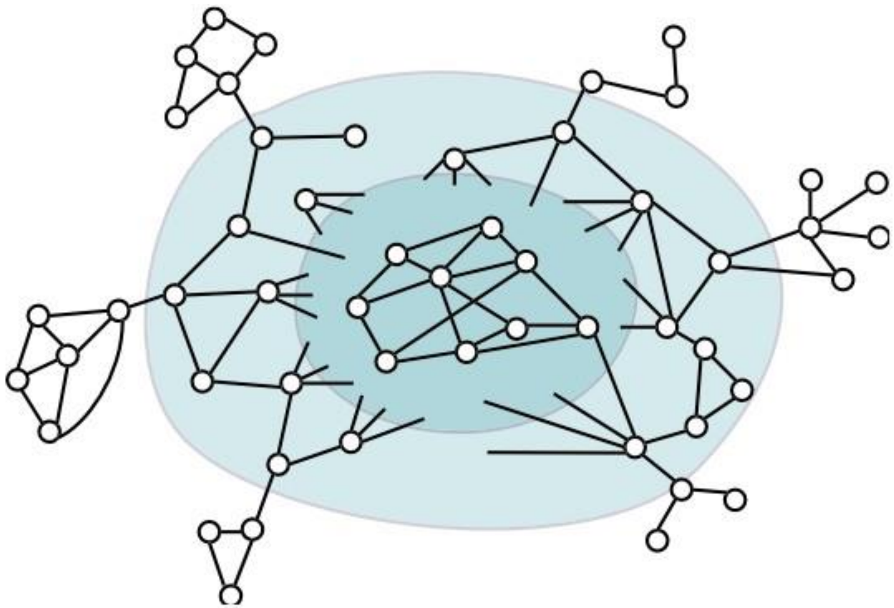


Nothing happens!
$\Rightarrow$ Nestedness of the core-periphery structure

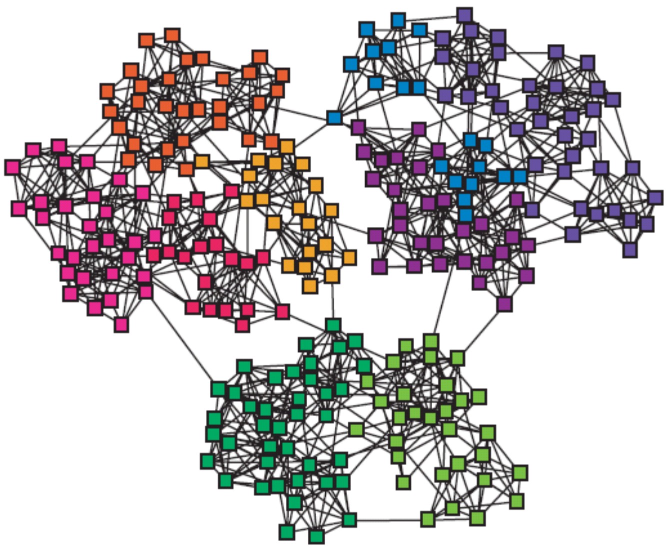
Suggested Network Structure



Part 2: Explanation



vs.



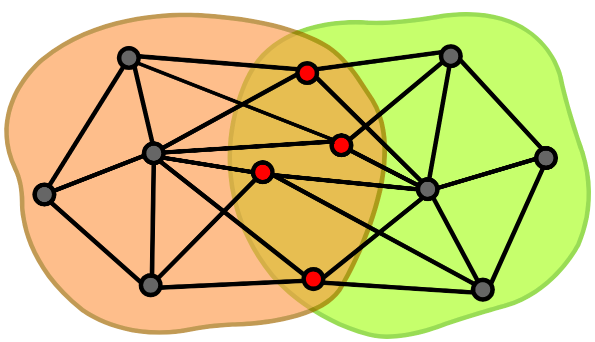
How do we reconcile these two views?

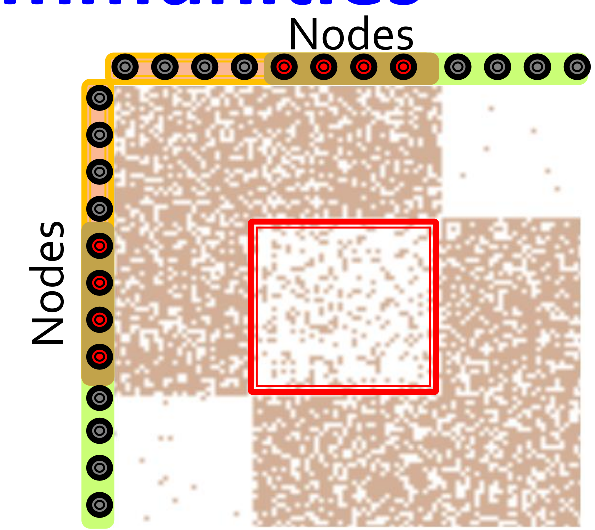
Overlapping Community Detection

- **Many methods for overlapping communities**
  - Clique percolation [Palla et al. '05]
  - Link clustering [Ahn et al. '10] [Evans et al.'09]
  - Clique expansion [Lee et al. '10]
  - Mixed membership stochastic block models [Airoldi et al. '08]
  - Bayesian matrix factorization [Psorakis et al. '11]
- **What do these methods assume about community overlaps?**

Overlapping Communities

- **Many overlapping community detection methods make an implicit assumption:**
  - **Edge probability decreases with the number of shared communities**



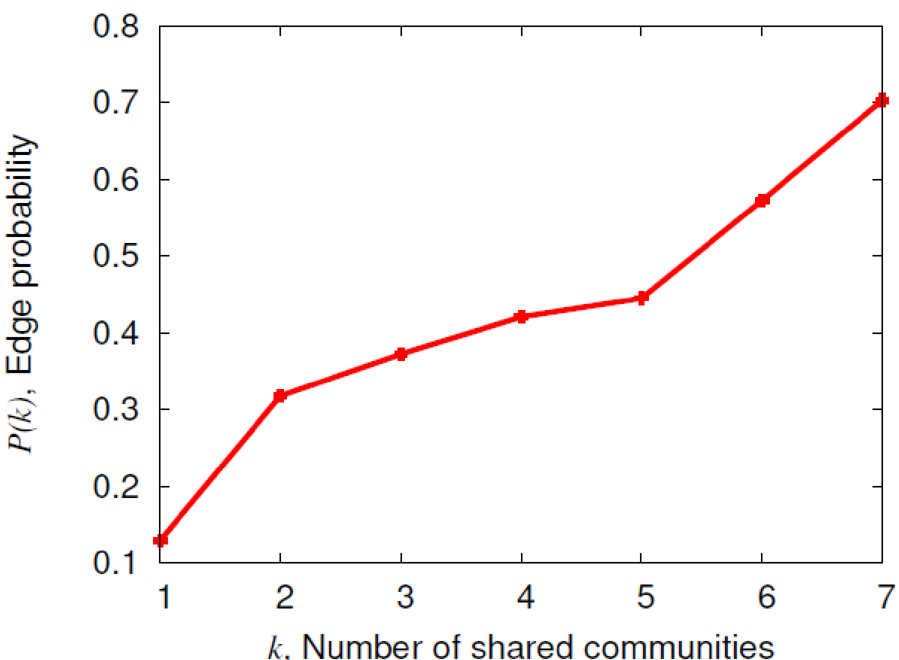


y matrix

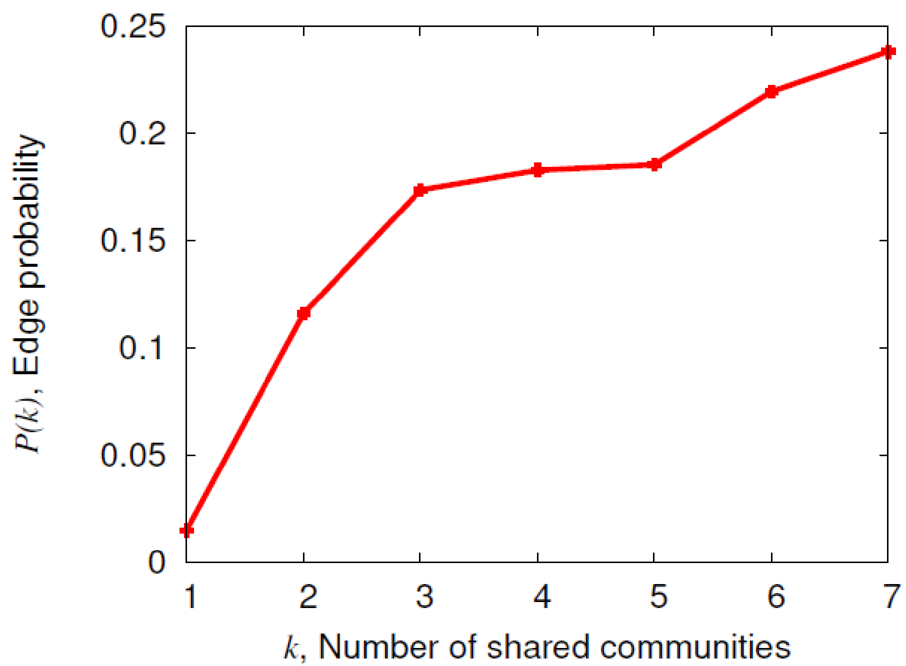
Is this true?

Ground-truth Communities

- **Basic question: nodes u, v share k communities**
- **What's the edge probability?**



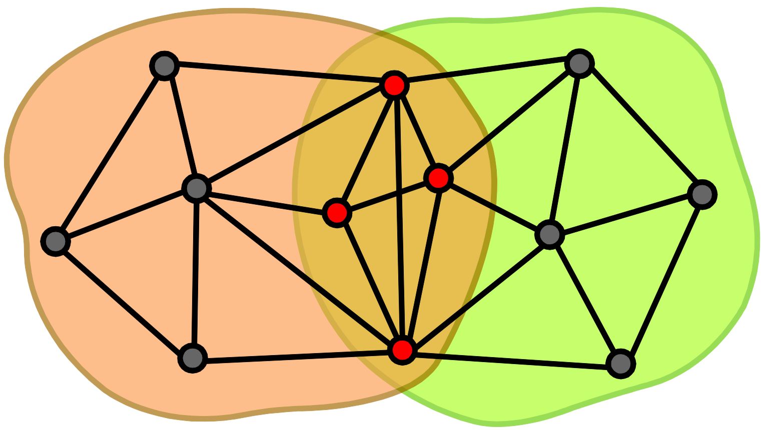
LiveJournal
social network

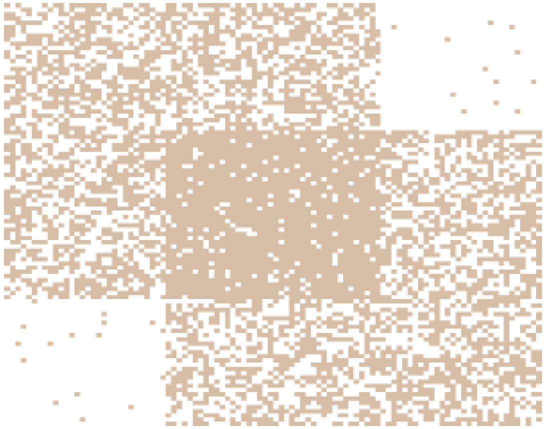


Amazon
product network

Communities as Tiles!

- **Edge density in the overlaps is higher!**

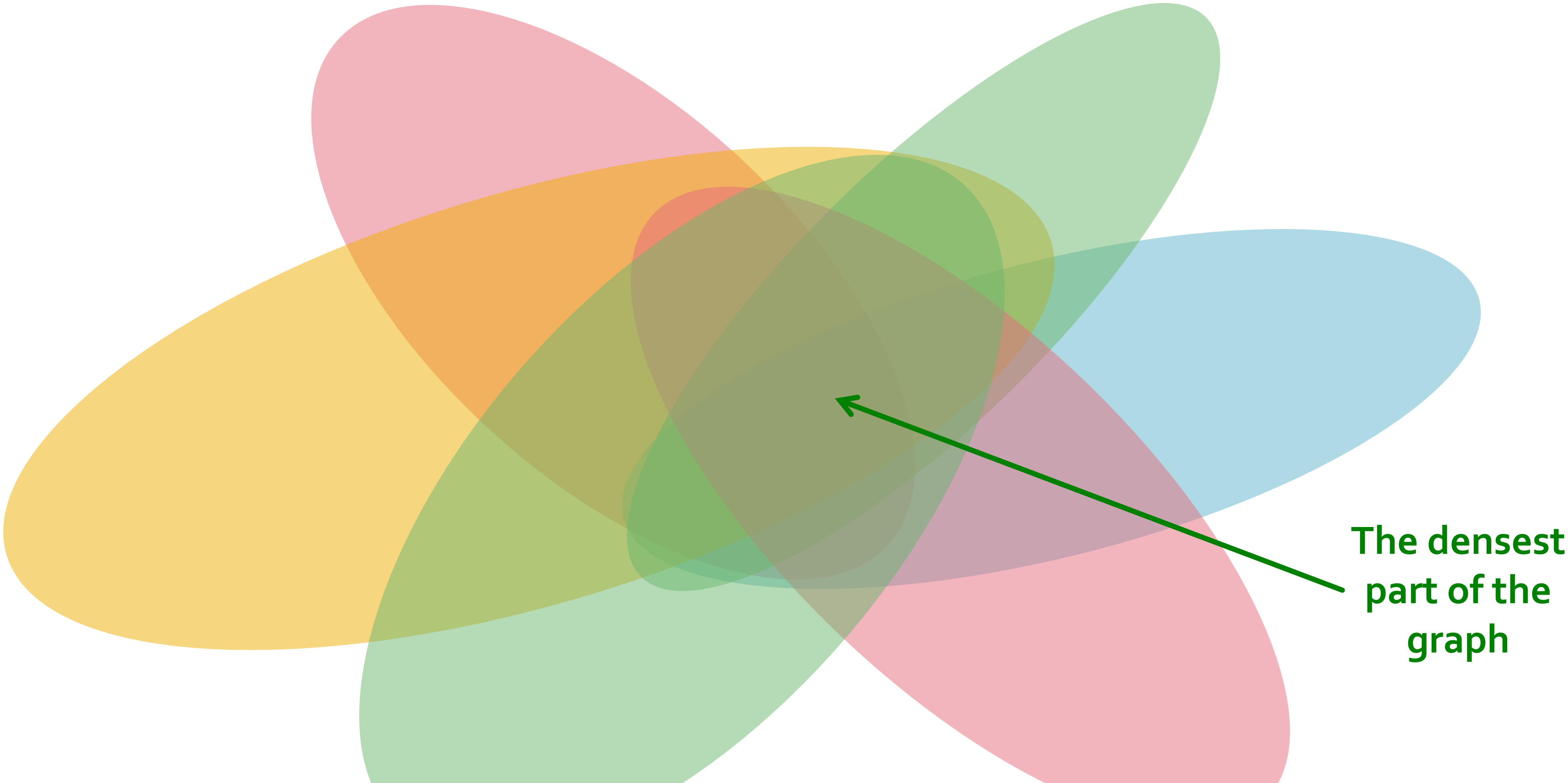




"The more different foci (communities) that two individuals share, the more likely it is that they will be tied" - S. Feld, 1981

Communities as "tiles"

Communities as Tiles/Circles

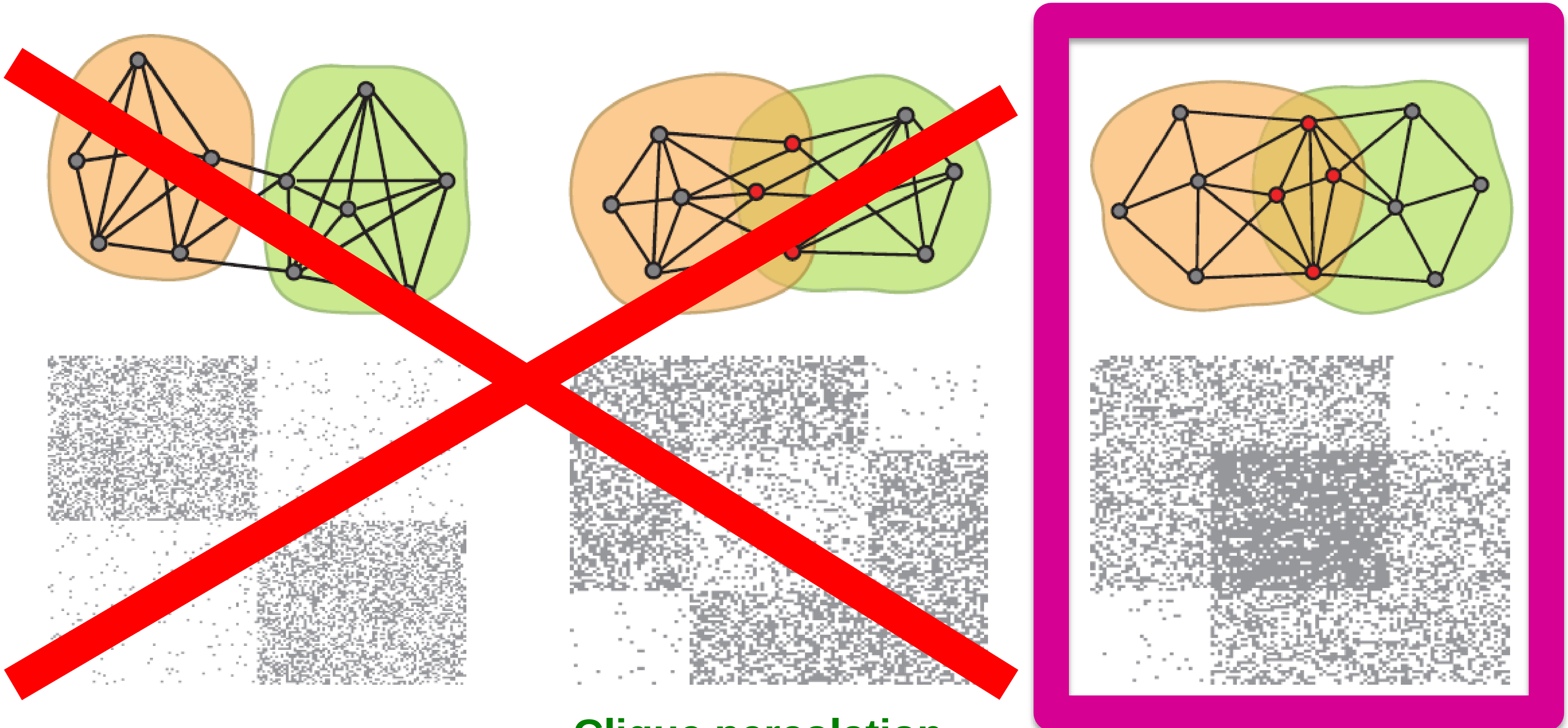


Communities as overlapping tiles

Web of affiliations [Simmel '64]

Communities in Networks

What does this mean?

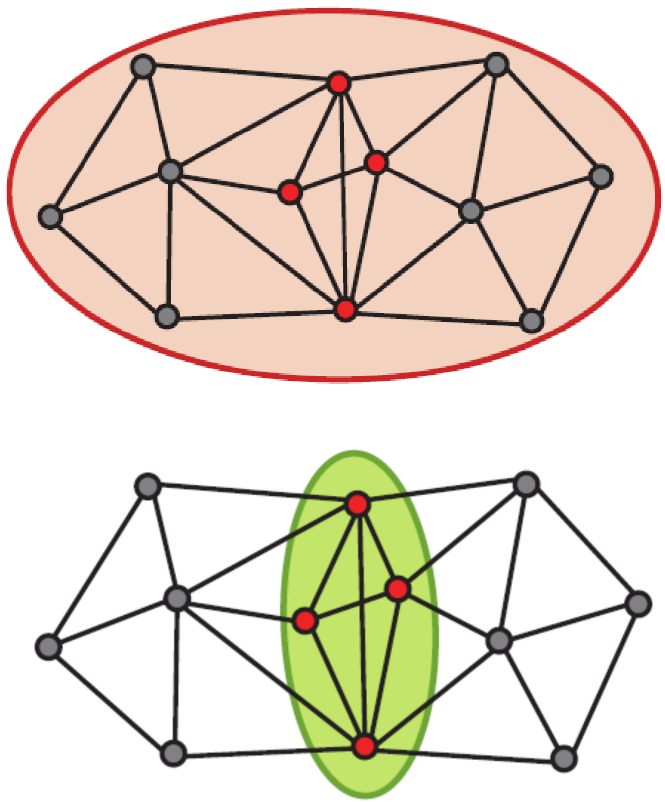


Non-overlapping methods (spectral, modularity optimization)

Clique percolation, and many other overlapping methods as well

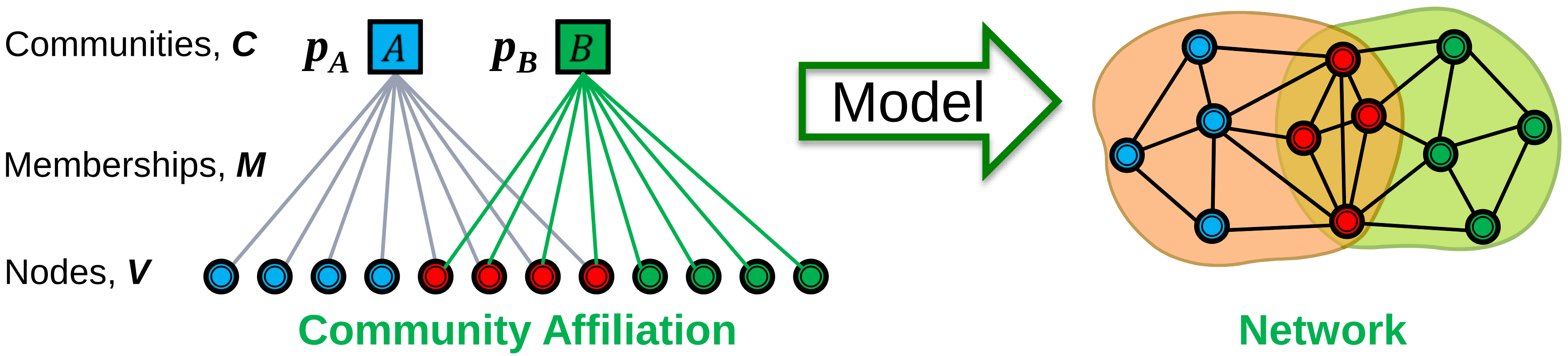
Many Methods Fail

- **Many methods fail to detect dense overlaps:**
  - Clique percolation, …



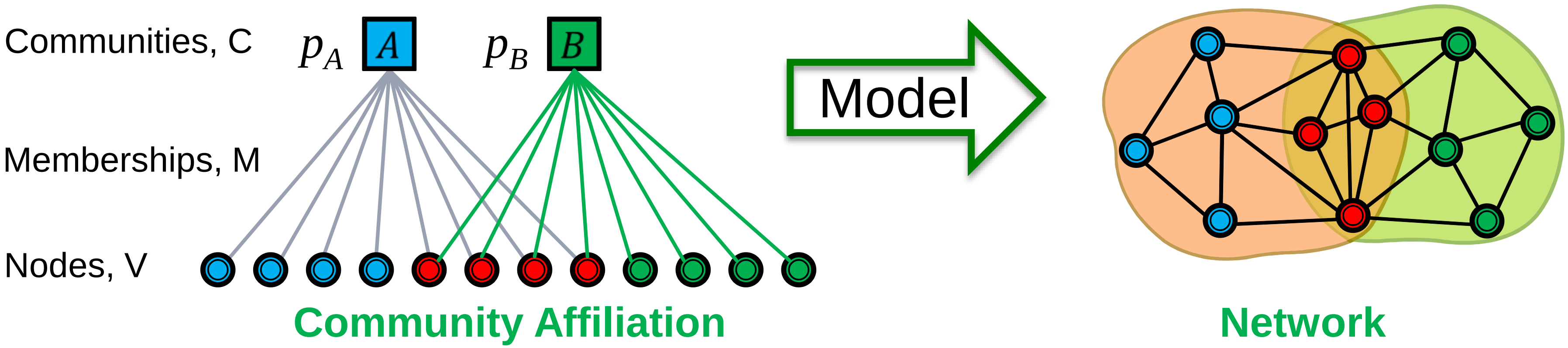
Clique percolation

Community-Affiliation Graph Model (AGM)



- **Generative model: How is a network generated from community affiliations?**
- **Model parameters:**
  - Nodes ***V***, Communities ***C***, Memberships ***M***
  - Each community c has a single probability p_c

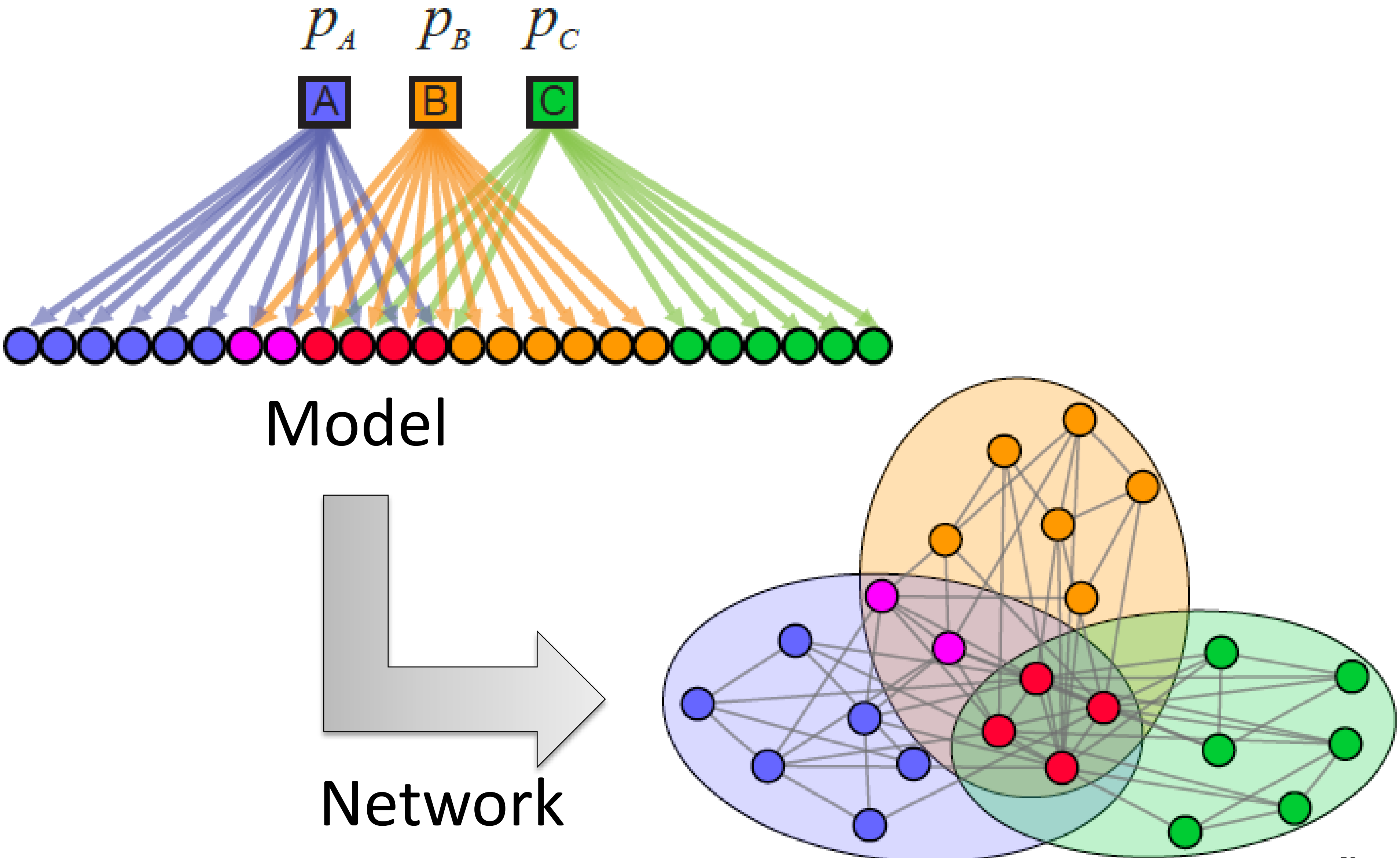
AGM: Generative Process



- **Given parameters (*V, C, M,* $\{p_c\}$)**
  - Nodes in community c connect to each other by flipping a coin with probability p_c
  - **Nodes that belong to multiple communities have multiple coin flips: Dense community overlaps**
    - If they "miss" the first time, they get another chance through the next community"

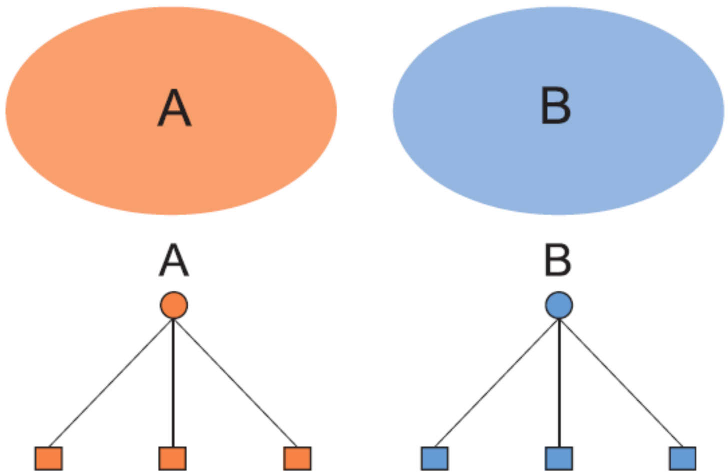
$$p(u,v) = 1 - \prod_{c \in M_u \cap M_v} (1 - p_c)$$

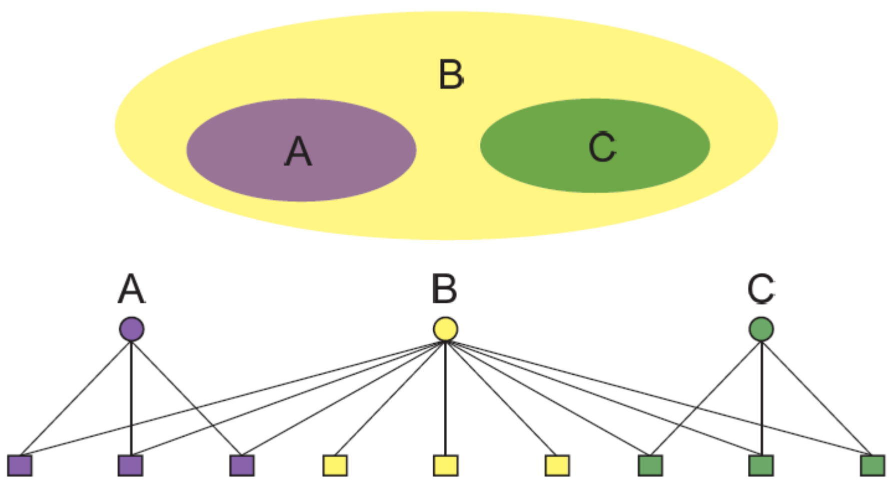
AGM: Dense Overlaps

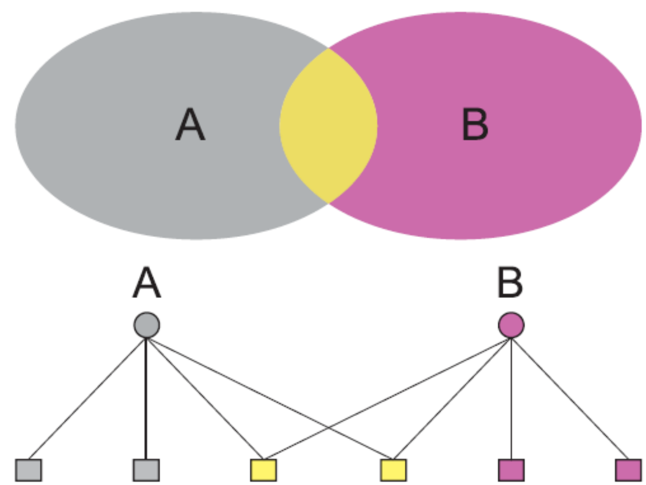


Community-Affiliation Graph Model

- **AGM is flexible and can express variety of network structures:** Non-overlapping, Nested, Overlapping







Community Evaluation: Extras

Community Evaluation

- Without ground truth
- With ground truth

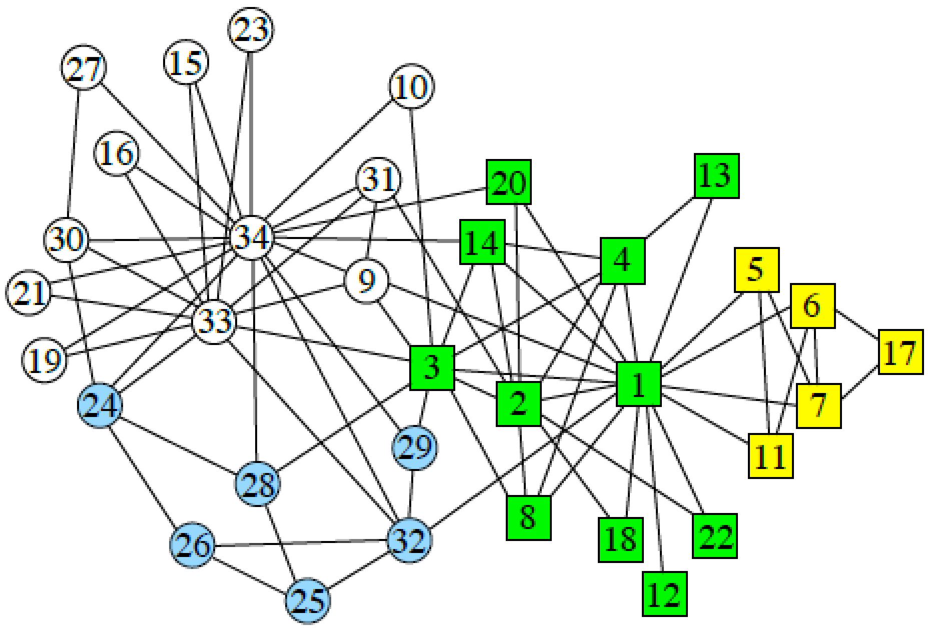
Eval. Without Ground Truth

- Cluster Cohesion: Measures how closely related are objects in a cluster
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters

$$\delta_{int}(C) = \frac{\#\text{ internal edges of } C}{n_c(n_c - 1)/2}$$

$$\delta_{ext}(C) = \frac{\#\text{ inter-cluster edges of } C}{n_c(n - n_c)}$$

Evaluation With Ground Truth



Zachary's Karate Club

Club president (34) (circles) and instructor (1) (rectangles)

Metrics: Purity

the fraction of instances that have labels equal to the label of the community's majority

$$Purity = \frac{1}{N}\sum_{i=1}^{k}\max_{j}|C_i \cap L_j|$$

(5+6+4)/20 = 0.75

Metrics: Pair Counting

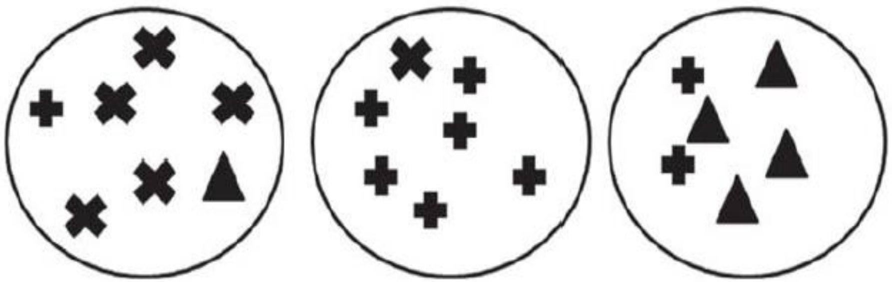
- **Based on pair counting**: the number of pairs of vertices which are classified in the same (or different) clusters
  - **True Positive (TP)**: when **similar** members are assigned to the **same** community (**correct** decision)
  - **True Negative (TN)**: when **dissimilar** members are assigned to **different** communities (**correct** decision)
  - **False Negative (FN)**: when **similar** members are assigned to **different** communities (**incorrect** decision)
  - **False Positive (FP)**: when **dissimilar** members are assigned to the **same** community (**incorrect** decision)

Metrics: Pair Counting

For TP, we need to compute the number of pairs with the **same** label that are in the **same** community

$$TP = \underbrace{\binom{5}{2}}_{Community\ 1} + \underbrace{\binom{6}{2}}_{Community\ 2} + \underbrace{(\binom{4}{2} + \binom{2}{2})}_{Community\ 3} = 32$$

Metrics: Pair Counting



Community 1 *Community 2* *Community 3*

For TN: compute the number of **dissimilar** pairs in **dissimilar** communities

$$
\begin{aligned}
TN &= \underbrace{(\overbrace{5\times 6}^{\times,+} + \overbrace{1\times 1}^{+,\times} + \overbrace{1\times 6}^{\triangle,+} + \overbrace{1\times 1}^{\triangle,\times})}_{\textit{Communities 1 and 2}} \\
&+ \underbrace{(\overbrace{5\times 4}^{\times,\triangle} + \overbrace{5\times 2}^{\times,+} + \overbrace{1\times 4}^{+,\triangle} + \overbrace{1\times 2}^{\triangle,+})}_{\textit{Communities 1 and 3}} \\
&+ \underbrace{(\overbrace{6\times 4}^{+,\triangle} + \overbrace{1\times 2}^{\times,+} + \overbrace{1\times 4}^{\times,\triangle}}_{\textit{Communities 2 and 3}} = 104.
\end{aligned}
$$

Metrics: Pair Counting

For FP, compute **dissimilar** pairs that are in the **same** community

$$FP = \underbrace{(5 \times 1 + 5 \times 1 + 1 \times 1)}_{\textit{Community 1}} + \underbrace{(6 \times 1)}_{\textit{Community 2}} + \underbrace{(4 \times 2)}_{\textit{Community 3}} = 25$$

For FN, compute **similar** members that are in **different** communities

$$FN = \underbrace{(5 \times 1)}_{\times} + \underbrace{(6 \times 1 + 6 \times 2 + 2 \times 1)}_{+} + \underbrace{(4 \times 1)}_{\triangle} = 29$$

Metrics: Pair Counting

- **Precision (P)**: the fraction of pairs that have been correctly assigned to the same community

  TP/(TP+FP)

- **Recall (R)**: the fraction of pairs assigned to the same community of all the pairs that should have been in the same community.

  TP/(TP+FN)

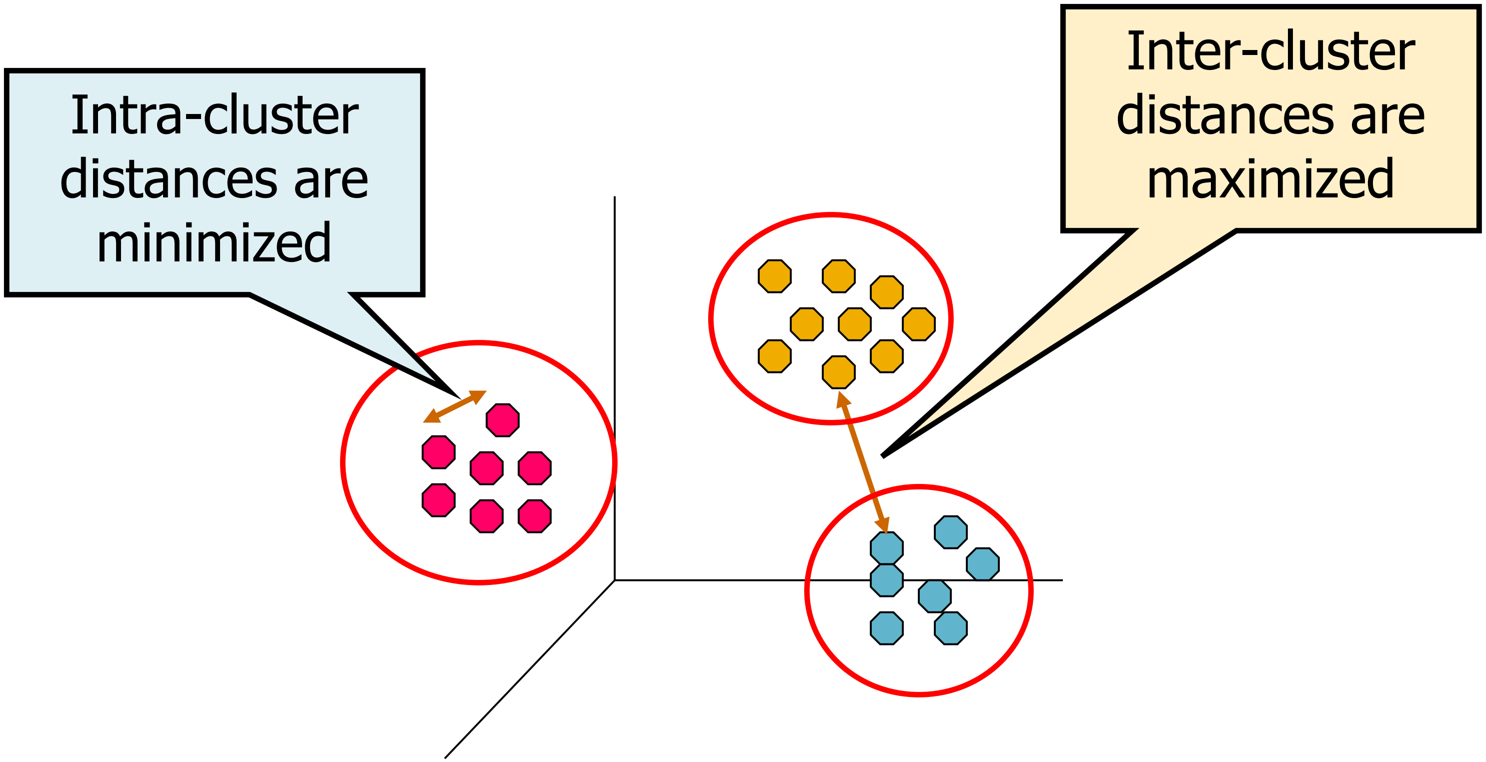
- **F-measure**:

  2PR/(P+R)

Communities:
Issues and Questions

What is Cluster Analysis?

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Clusters Can Be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

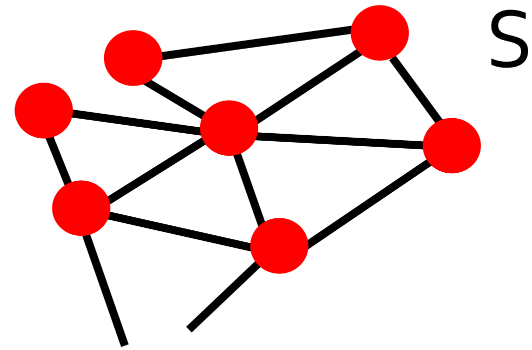
Communities: Issues and Questions

- **Some issues with community detection:**
  - Many different formalizations of clustering objective functions
  - Objectives are **NP-hard** to optimize exactly
  - Methods can find clusters that are systematically "biased"
- **Questions:**
  - **How well do algorithms optimize objectives?**
  - **What clusters do different methods find?**

[WWW '09]

Many Different Objective Functions

- **Single-criterion:**
  - Modularity: $m\text{-}E(m)$
  - Edges cut: c
- **Multi-criterion:**
  - Conductance: $c/(2m+c)$
  - Expansion: c/n
  - Density: $1\text{-}m/n^2$
  - CutRatio: $c/n(N\text{-}n)$
  - Normalized Cut: $c/(2m+c) + c/2(M\text{-}m)+c$
  - Flake-ODF: *frac. of nodes with more than ½ edges pointing outside S*



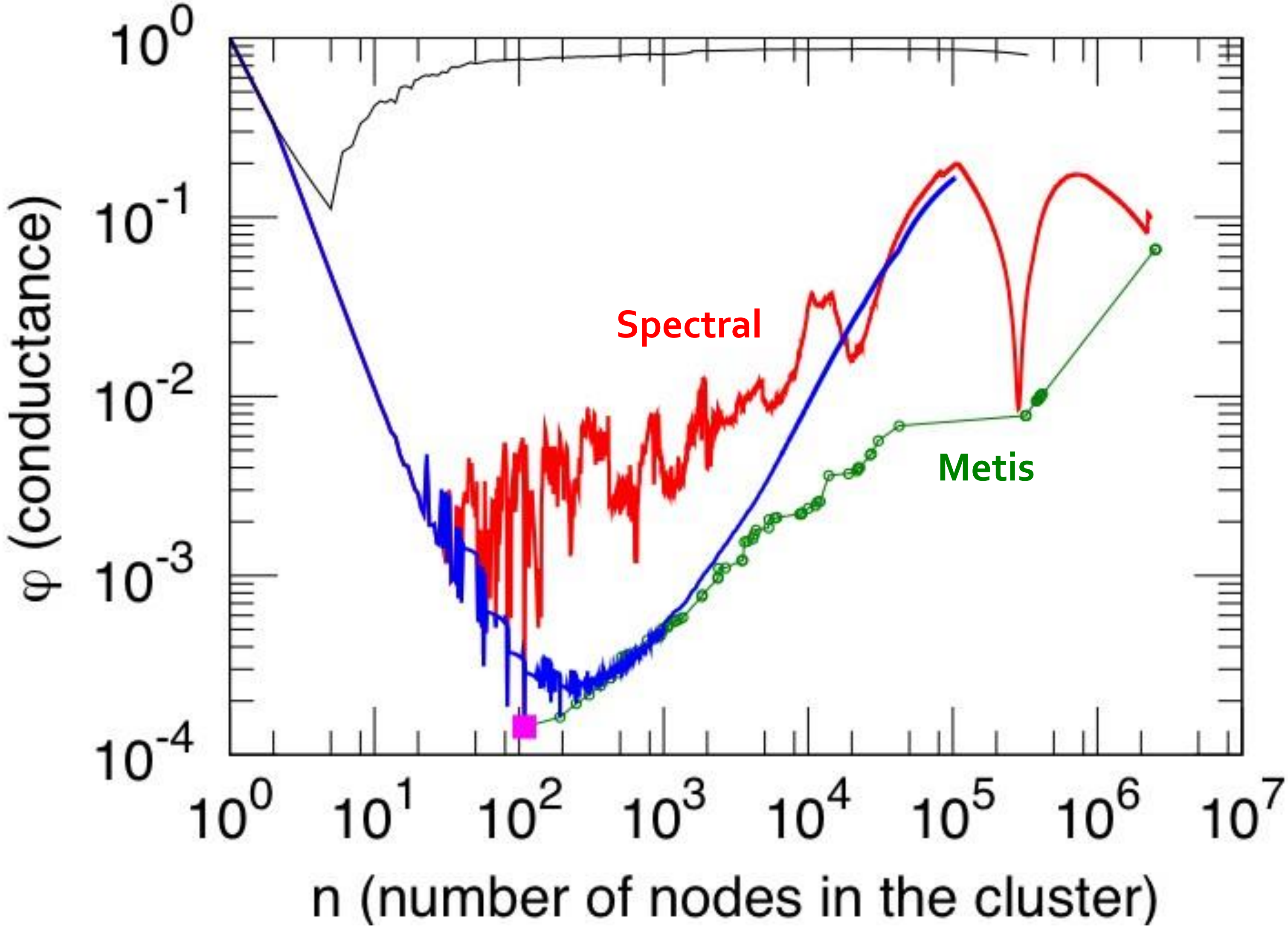
n: nodes in S
m: edges in S
c: edges pointing outside S

Many Classes of Algorithms

Many algorithms to implicitly or explicitly optimize objectives and extract communities:

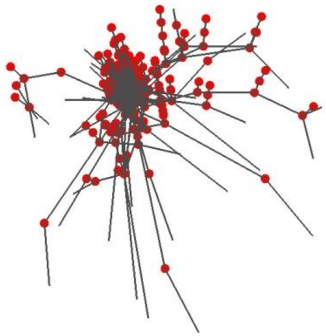
- **Heuristics:**
  - Girvan-Newman, Modularity optimization: popular heuristics
  - Metis: multi-resolution heuristic [Karypis-Kumar '98]
- **Theoretical approximation algorithms:**
  - Spectral partitioning

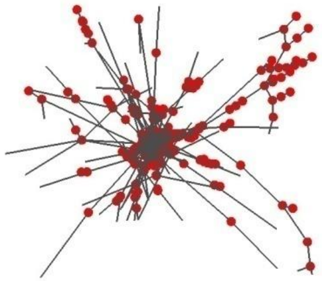
NCP: Live Journal



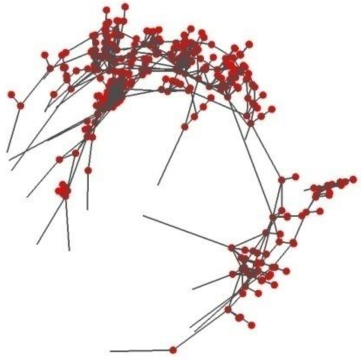
Properties of Clusters (1)

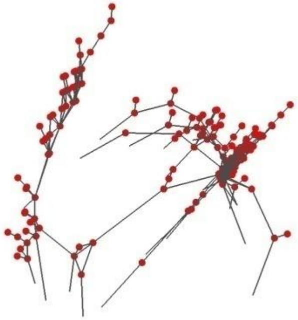
500 node communities from Spectral:





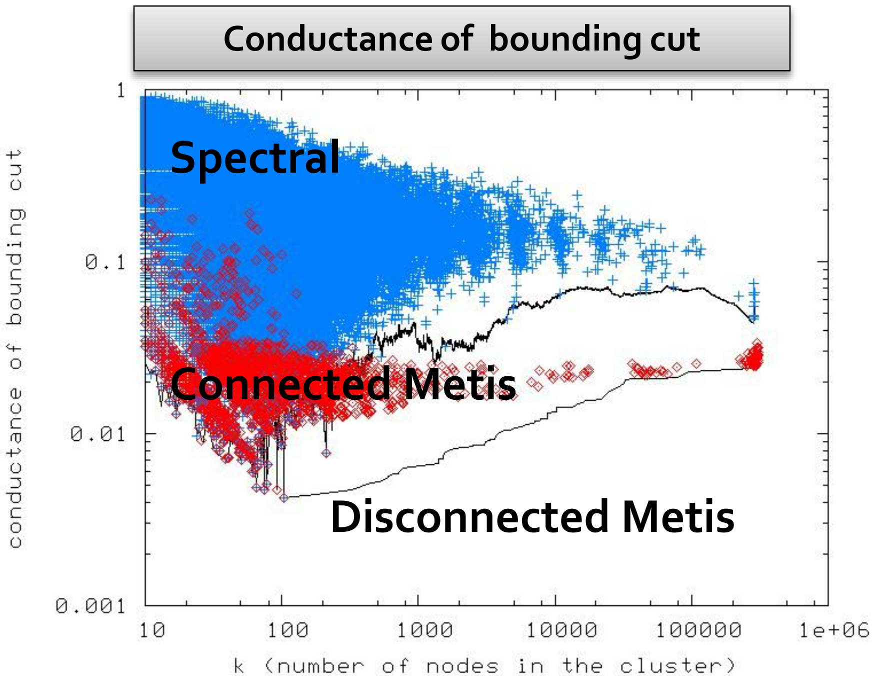
500 node communities from Metis:

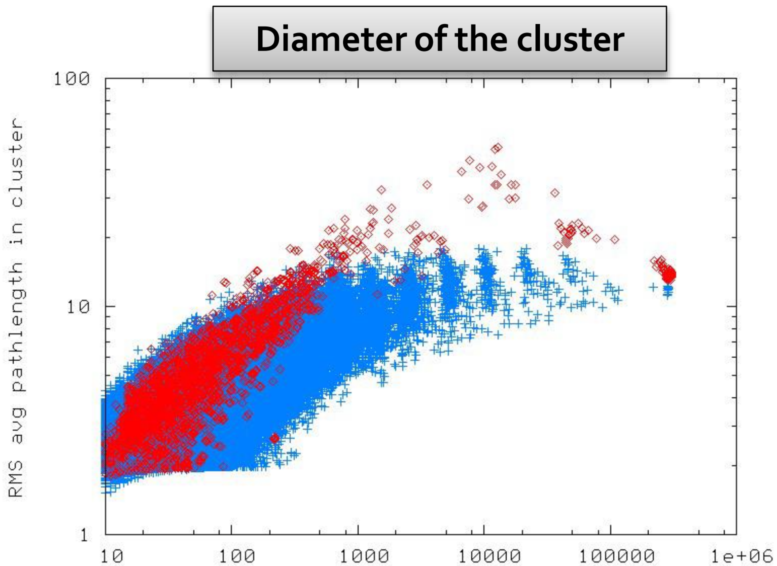


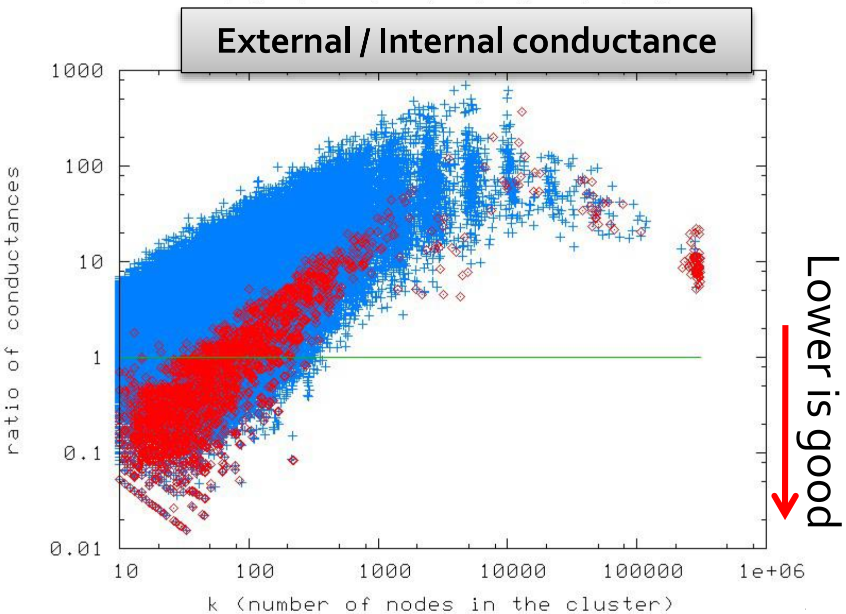


[WWW '09]

Properties of Clusters (2)



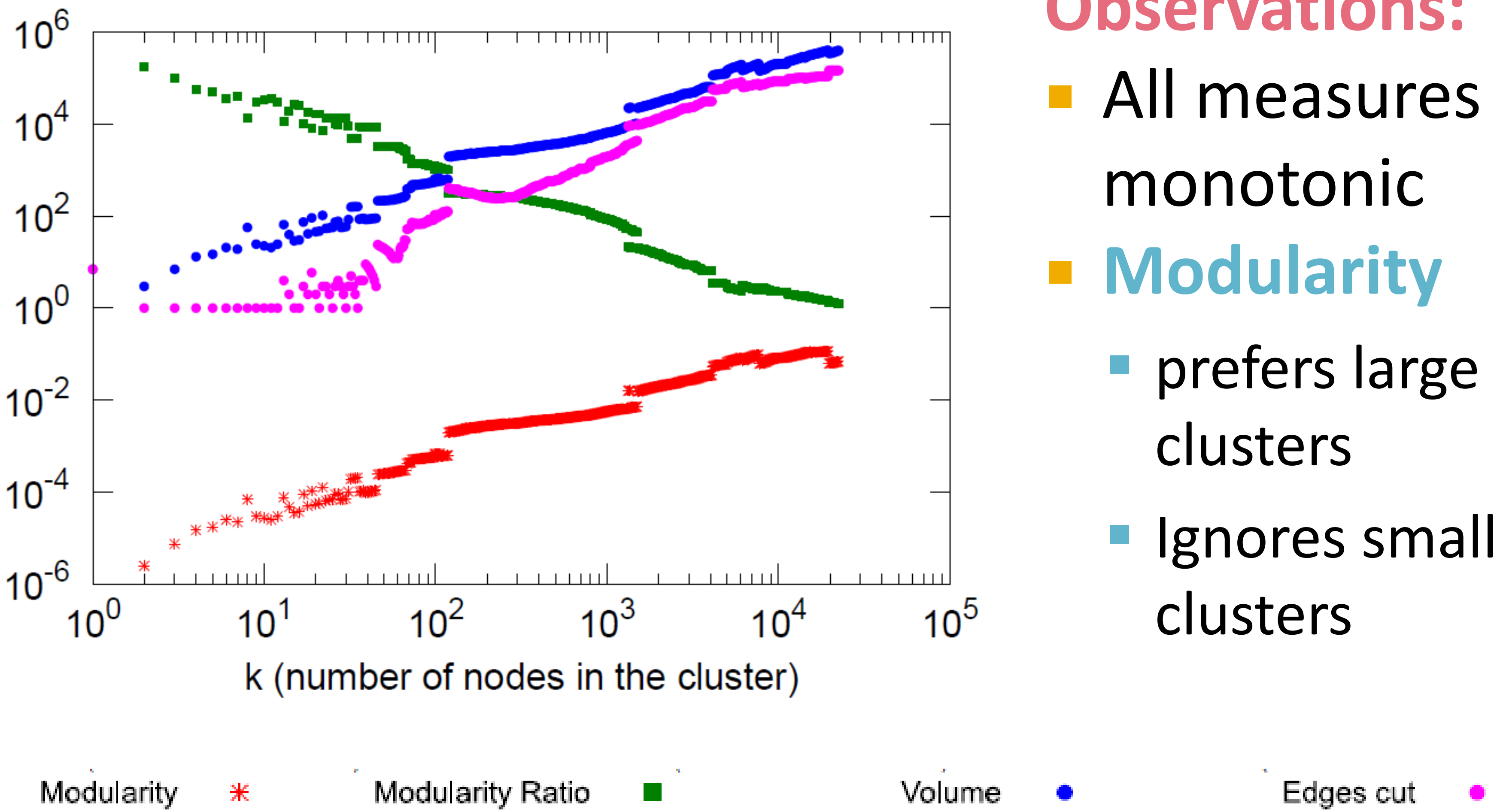




- Metis gives sets with better conductance
- Spectral gives tighter and more well-rounded sets

[WWW '09]

Single-criterion Objectives

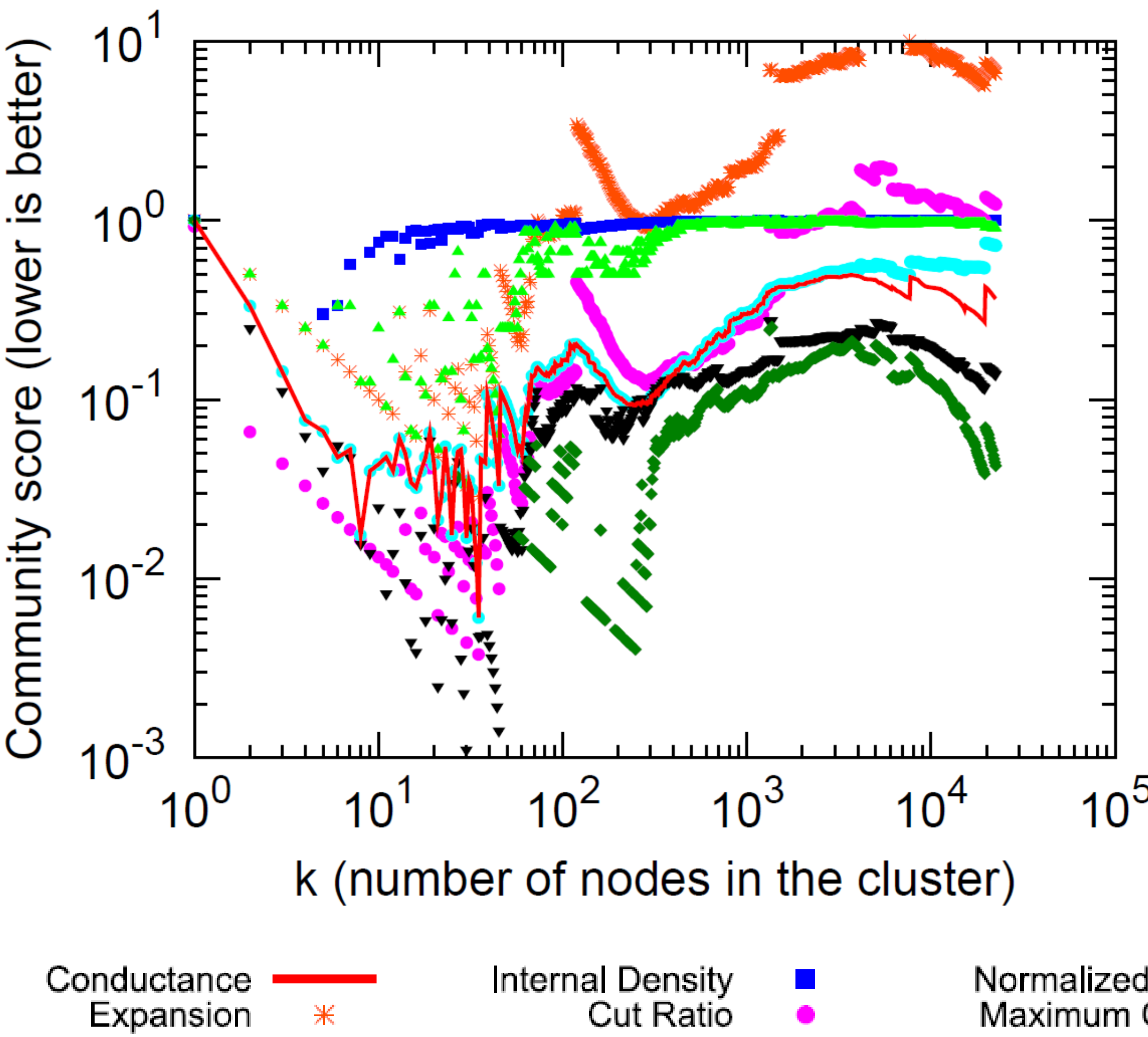


Observations:

- All measures are monotonic
- **Modularity**
  - prefers large clusters
  - Ignores small clusters

[WWW '09]

Multi-criterion Objectives



- **All qualitatively similar**
- **Observations:**
  - Conductance, Expansion, Norm-cut, Cut-ratio are similar
  - Flake-ODF prefers larger clusters
  - Density is bad
  - Cut-ratio has high variance