Overlapping Communities & Models of Community Structure in Large Networks

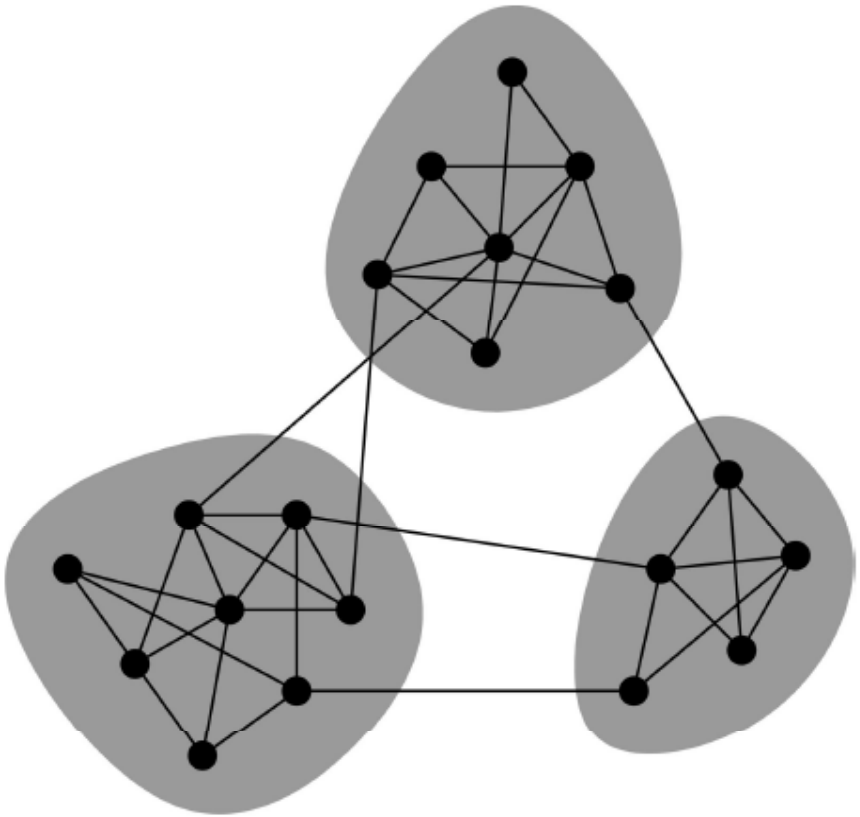
CS224W: Social and Information Network Analysis

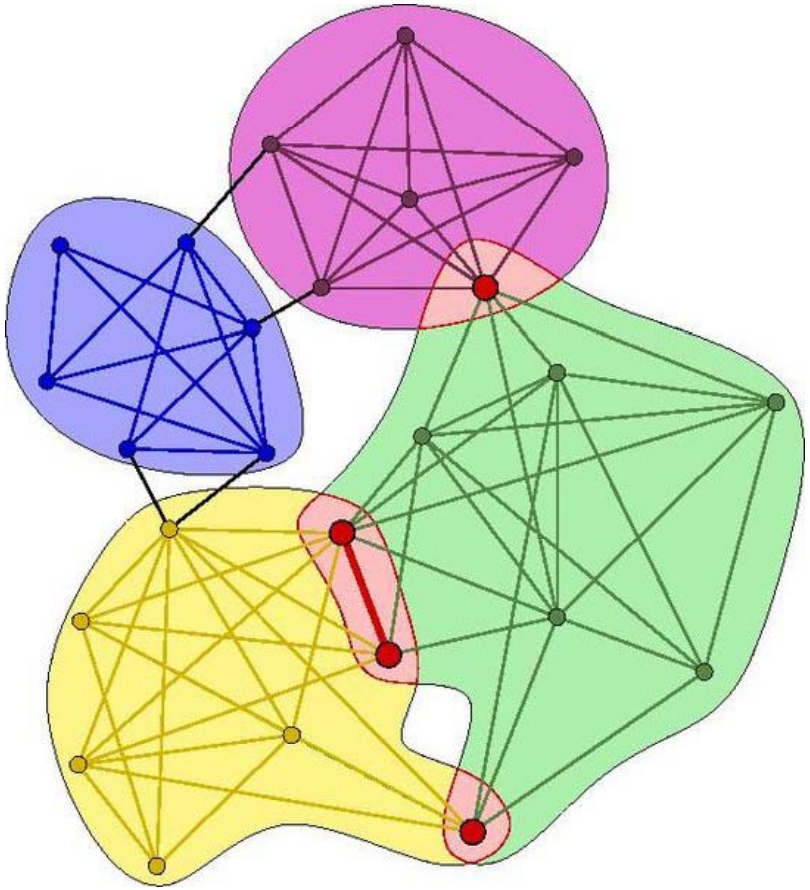
Jure Leskovec, Stanford University

http://cs224w.stanford.edu

Overlapping communities

- Non-overlapping vs. overlapping communities

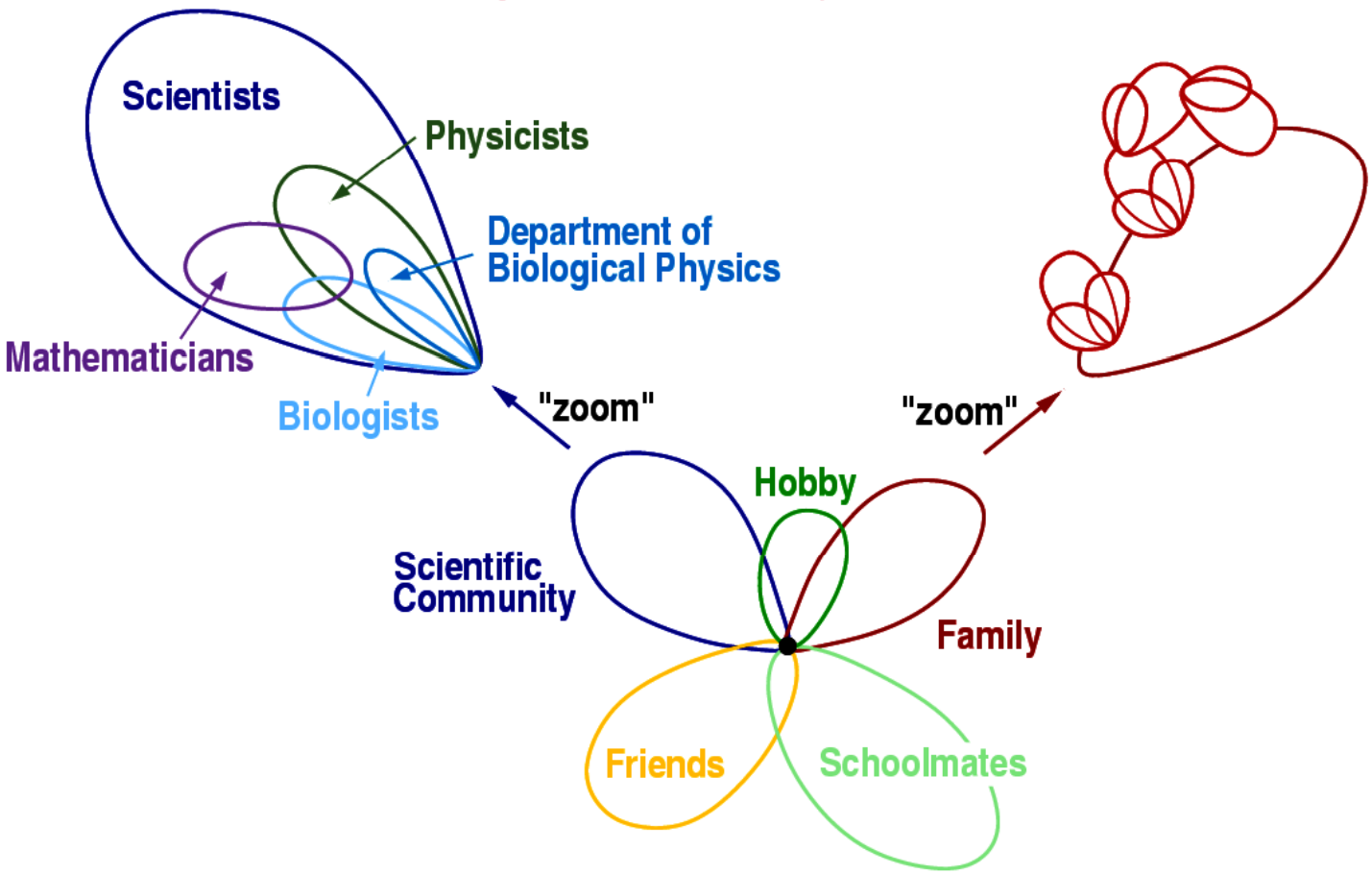




[Palla et al., '05]

Overlaps of social circles

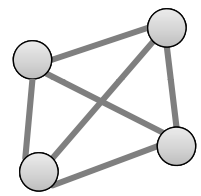
- A node belongs to many social circles



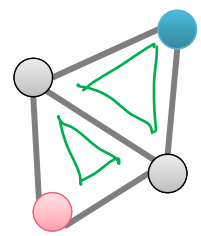
[Palla et al., '05]

Clique Percolation Method (CPM)

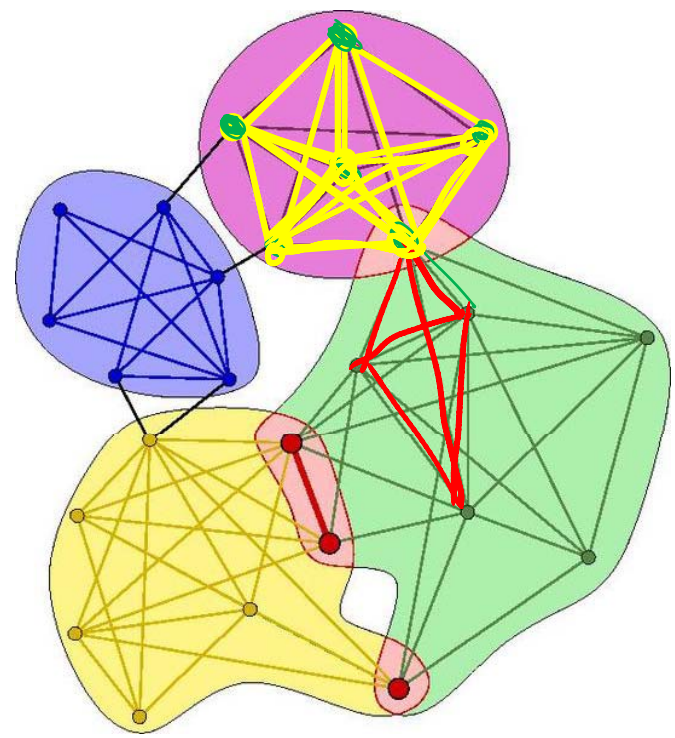
- Two nodes belong to the same community if they can be connected through adjacent k-cliques:
  - k-clique:
    - Fully connected graph on k nodes
  - Adjacent k-cliques:
    - overlap in k-1 nodes
- k-clique community
  - Set of nodes that can be reached through a sequence of adjacent k-cliques



4-clique



adjacent 3-cliques

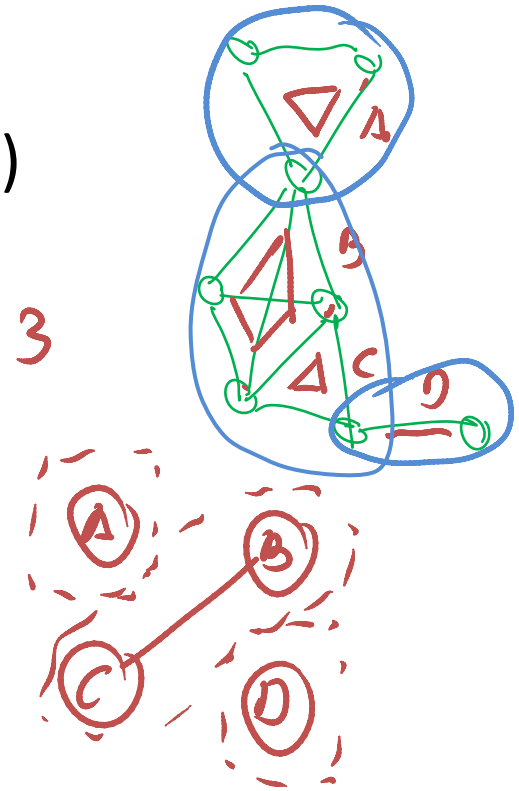


[Palla et al., '05]

Method5: CPM: Steps

- Clique Percolation Method:
  - Find maximal-cliques (not k-cliques!)
  - Clique overlap matrix:
    - Each clique is a node
    - Connect two cliques if they overlap in at least k-1 nodes
  - Communities:
    - Connected components of the clique overlap matrix

$k = 3$



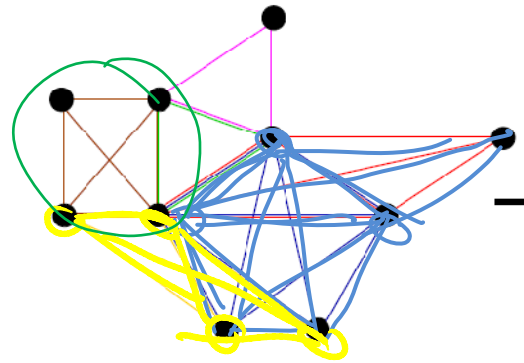
In practice
$k = 3, 4, 5$

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

[Palla et al., '05]

CPM method: Example

- Start with graph and 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



(1) Graph

5	3	2	1	3	1
3	4	2	1	1	1
2	2	3	2	1	2
1	1	2	3	0	1
3	1	1	0	4	2
1	1	2	1	2	4

(2) Clique overlap matrix

k=4

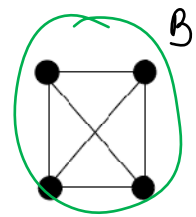
B→R→Y

BR

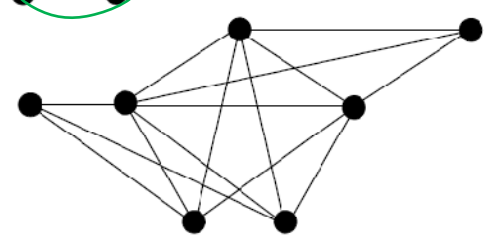
1	1	0	0	1	0
1	1	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	0	0	0	1	0
0	0	0	0	0	1

(3) Thresholded matrix at k=4

BROWN



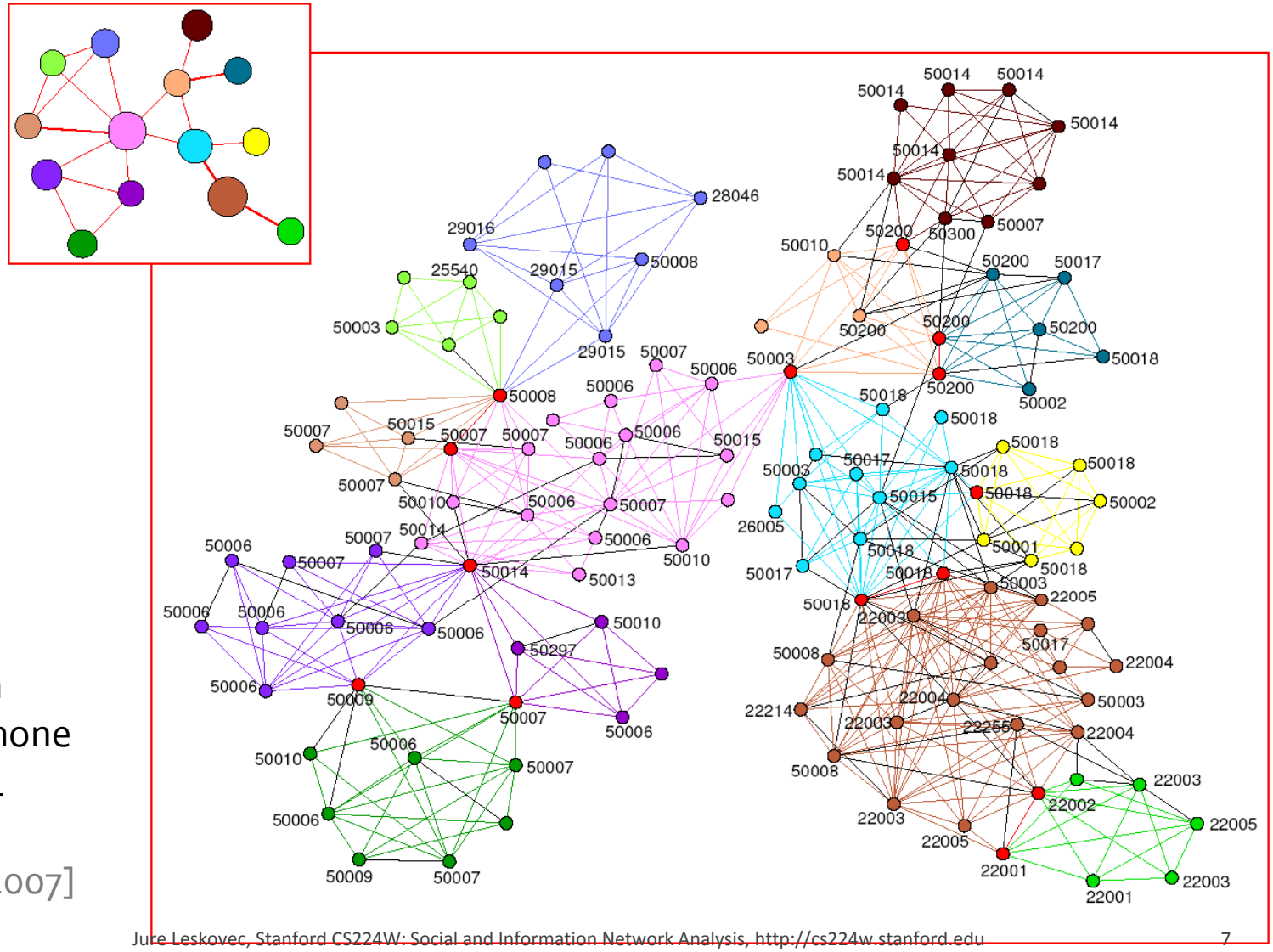
B+R+Y



(4) Communities (connected components)

[Palla et al., '07]

Example: Phone call network

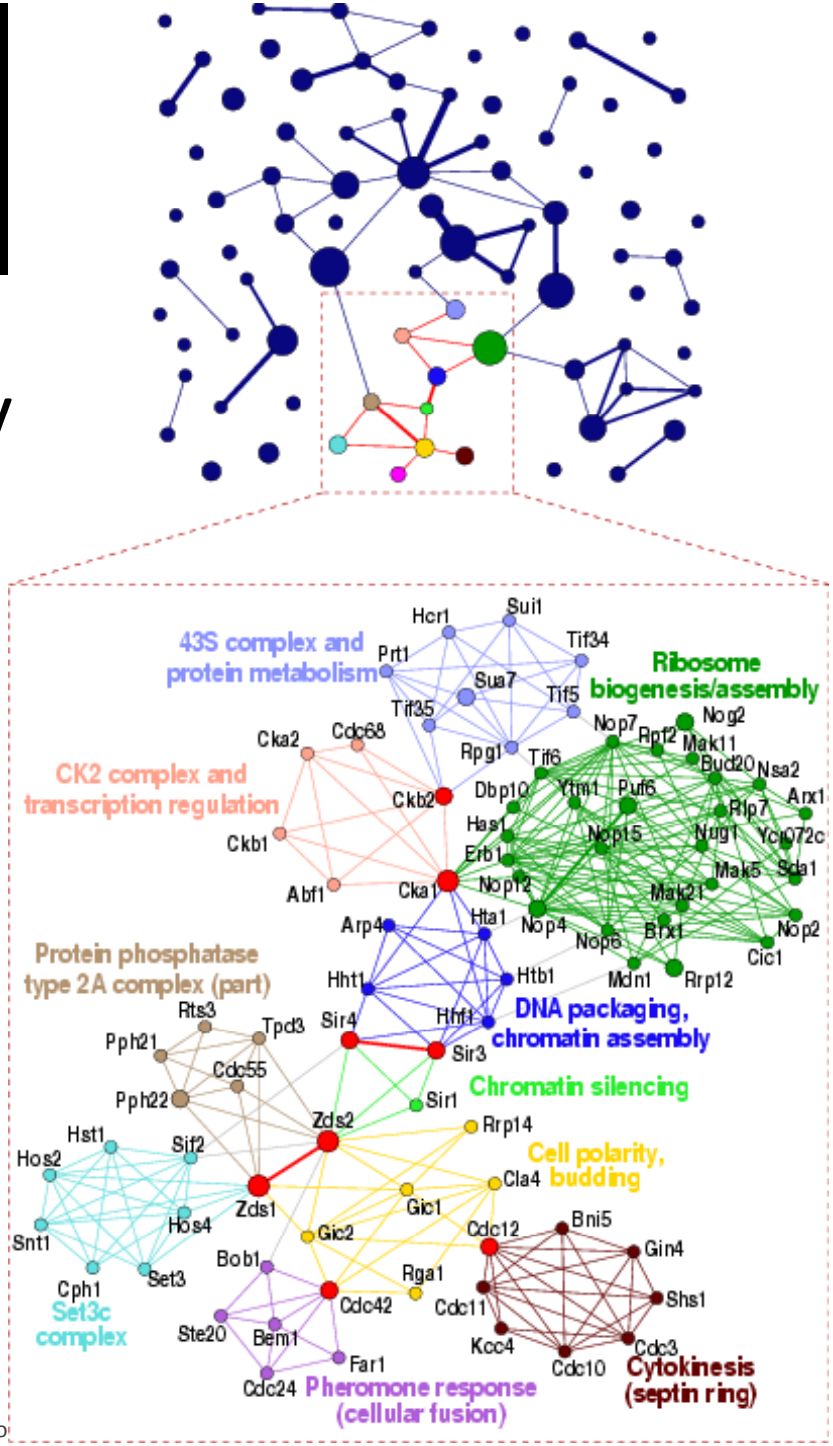


Communities in a "tiny" part of a phone calls network of 4 million users [Barabasi-Palla, 2007]

Example (2)

- Each node is a community
- Nodes are weighted for community size
- Links are weighted for overlap size

- DIP "core" data base of protein interactions (S. cerevisiase, yeast)



How to find max-cliques?

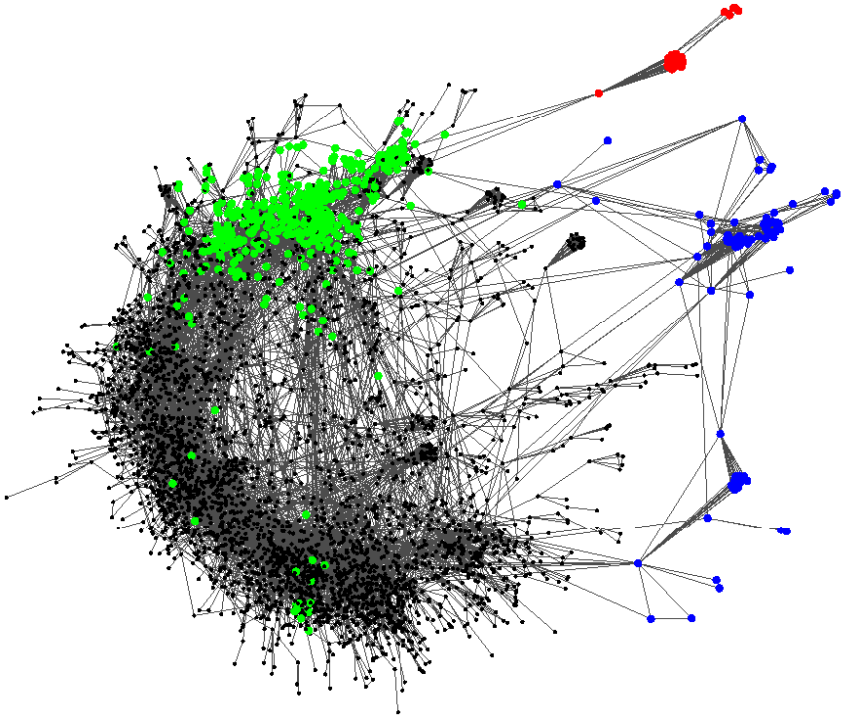
- No nice way, NP-hard combinatorial problem
- Simple Algorithm:
  - Start with max-clique size s
  - Choose node u, extract cliques of size s node u is member of
  - Delete u and its edges
  - When graph is empty, $s=s-1$, restart on original graph

[Palla et al., '05]

How to find max-cliques?

- Finding cliques around u of size s:
  - 2 sets A and B:
    - Each node in B links to all nodes in A
    - Set A grows by moving nodes from B to it
  - Start with $A=\{u\}$, $B=\{v: (u,v) \in E\}$
  - Recursively move each possible $v \in B$ to A and prune B
    - If B runs out of nodes before A reaches size s,
      - backtrack the recursion and try a different v

Reflections: Finding communities



- Let's rethink what we are doing…
  - Given a network
  - Want to find clusters!
- Need to:
  - Formalize the notion of a cluster
  - Need to design an algorithm that will find sets of nodes that are "good" clusters
- More generally:
  - How to think about clusters in large networks?

Clustering Objective functions

What is a good cluster?

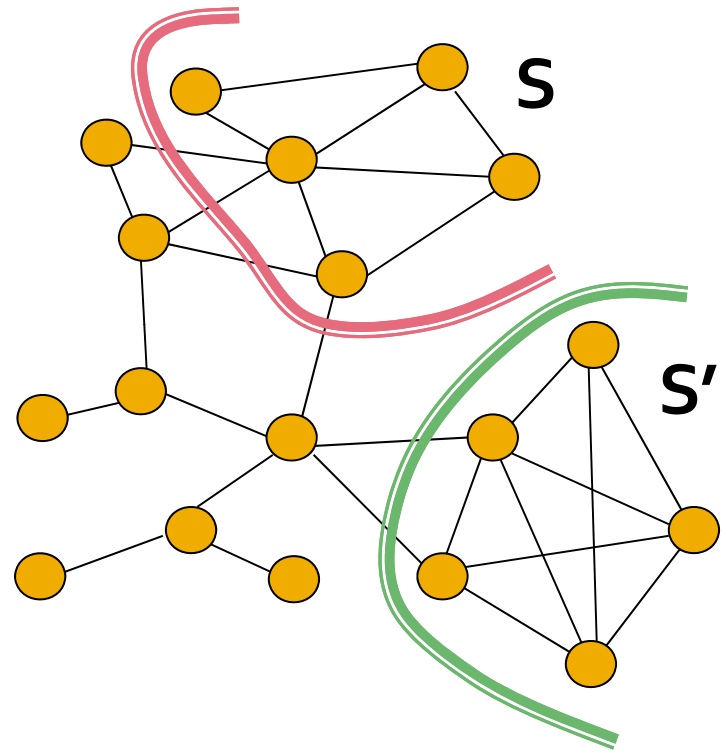
- Many edges internally
- Few pointing outside

Formally, **conductance:**

$$\phi(S) = \frac{\sum_{i \in S, j \notin S} A_{ij}}{\min\{A(S), A(\overline{S})\}}$$

Where: A(S)....volume $A(S) = \sum_{i \in S} \sum_{j \in V} A_{ij}$

Small Φ(S) corresponds to good clusters



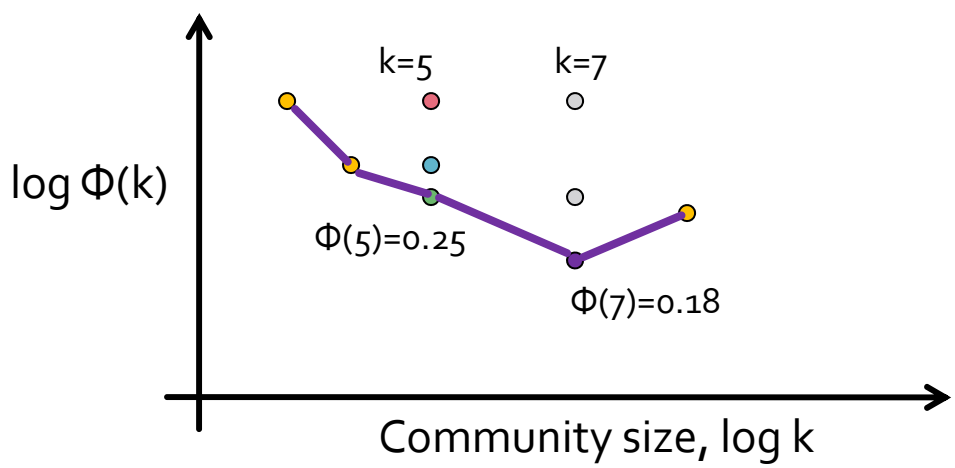
[WWW '08]

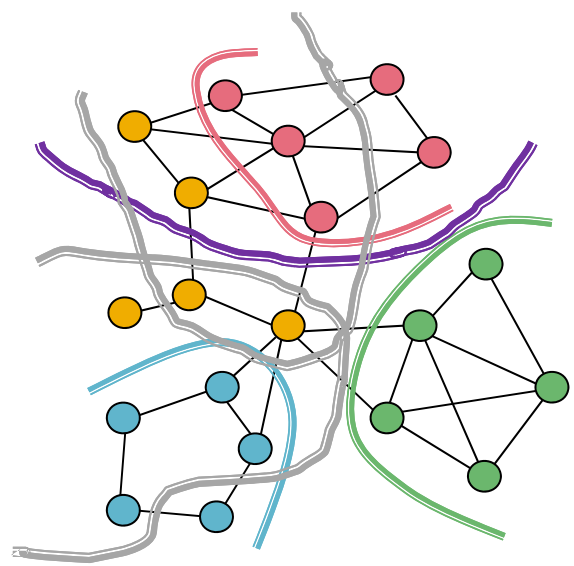
Network Community Profile Plot

- <u>Define:</u>
  Network community profile (**NCP**) plot

  Plot the score of best community of size *k*

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

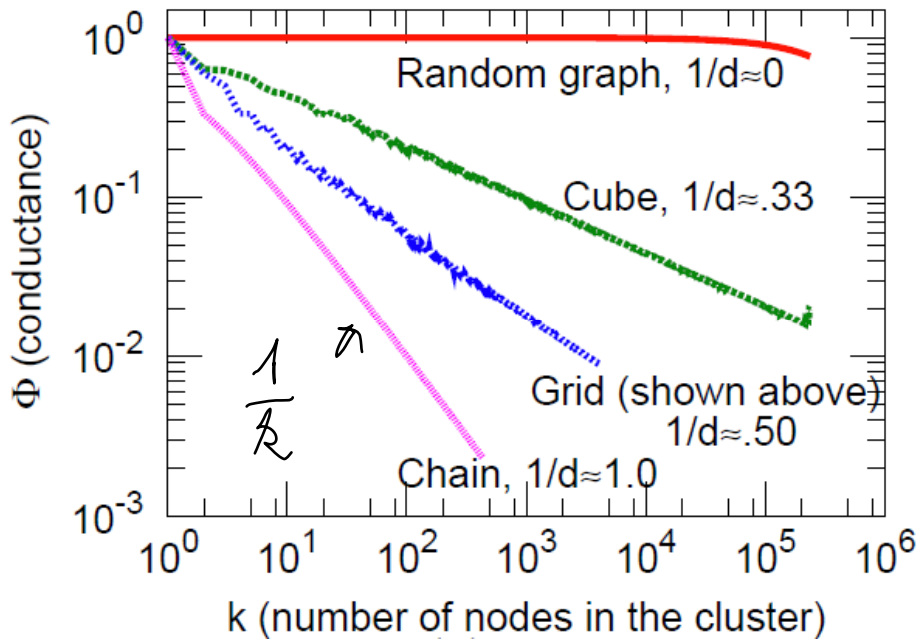


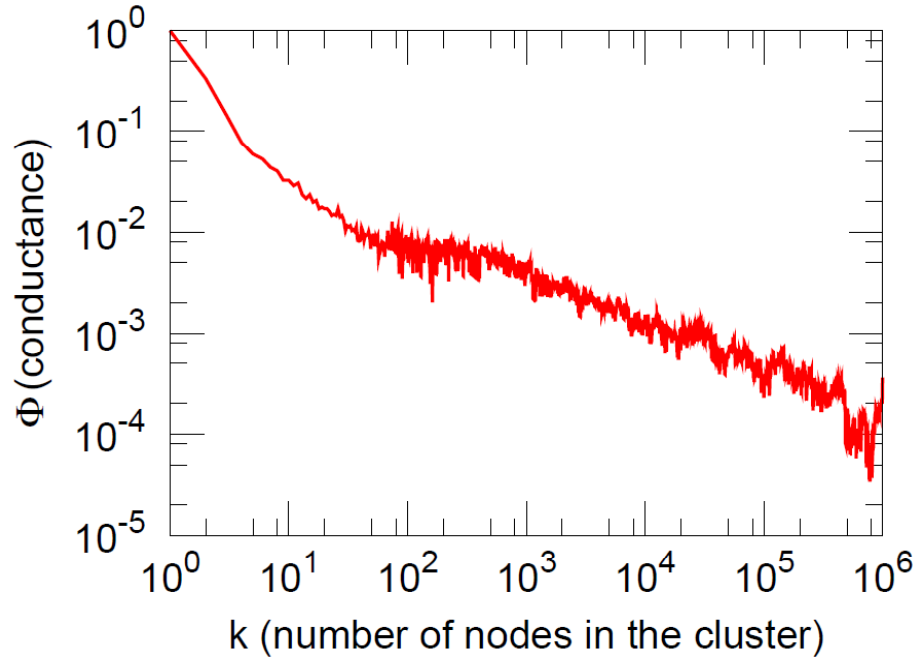


[WWW '08]

NCP plot: Meshes

- Meshes, grids, dense random graphs:

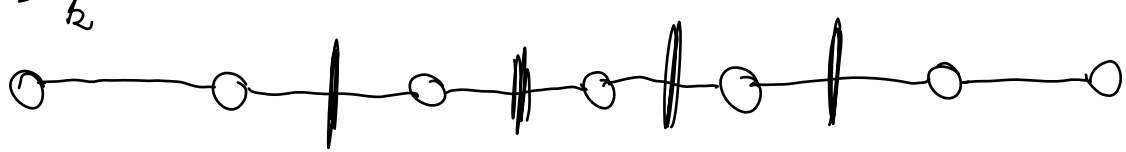




$\phi(k) = \frac{1}{k}$

d-dimensional meshes

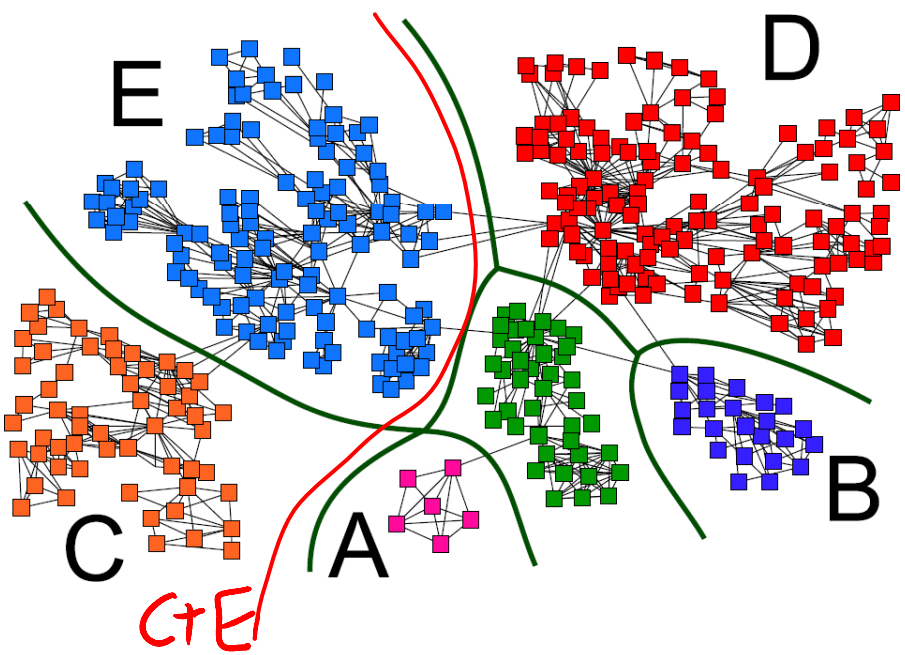
California road network

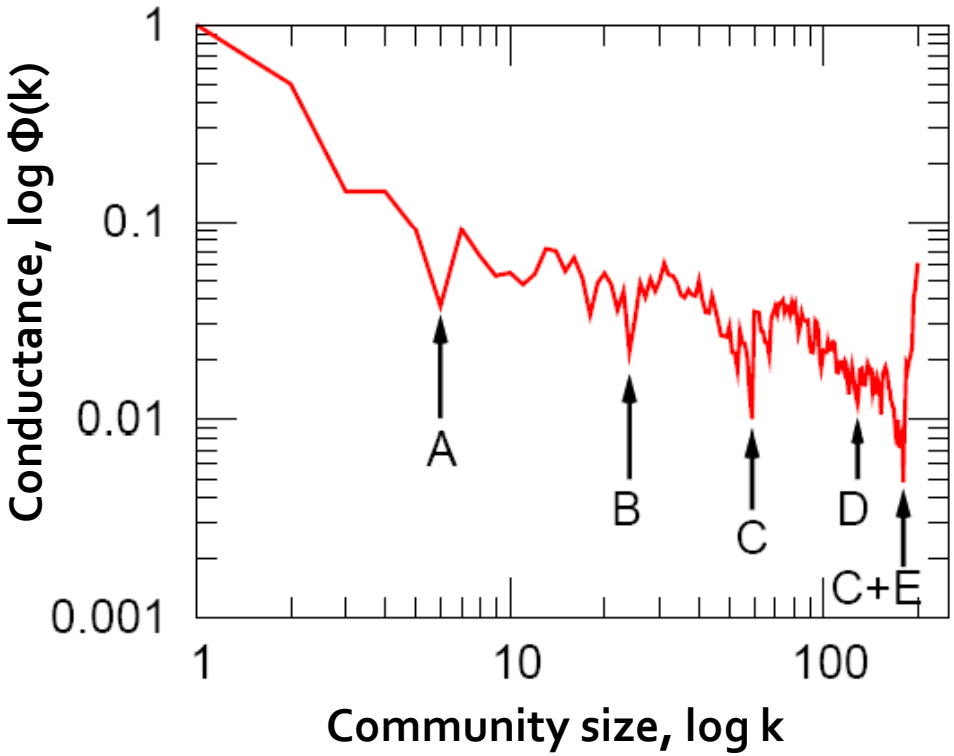


[WWW '08]

NCP plot: Network Science

- Collaborations between scientists in networks
  [Newman, 2005]



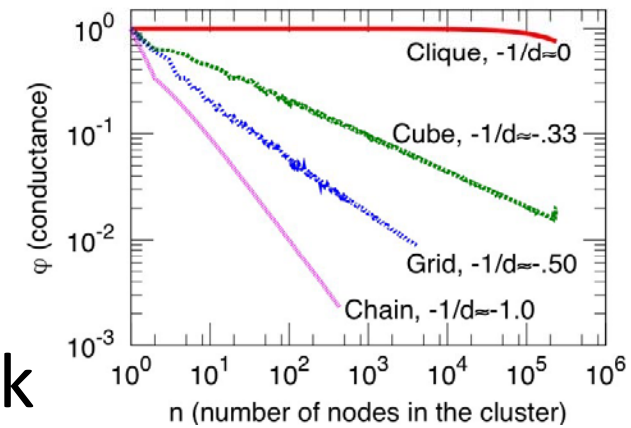


[Internet Mathematics '09]

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?

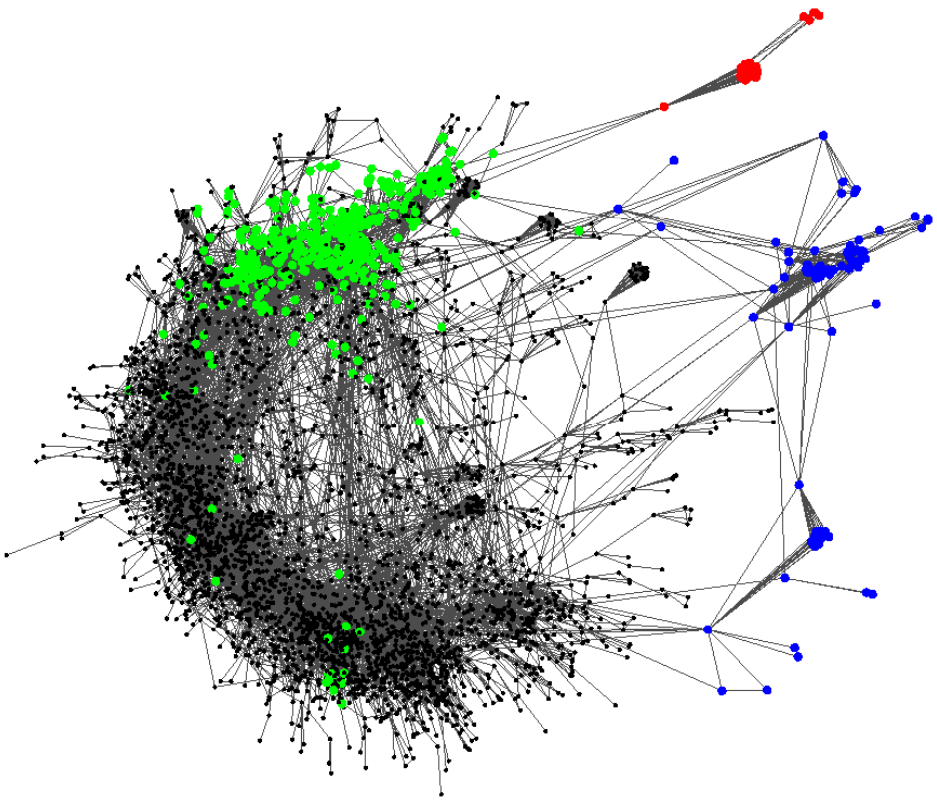
• Social nets	Nodes	Edges	Description
LIVEJOURNAL	4,843,953	42,845,684	Blog friendships [5]
EPINIONS	75,877	405,739	Trust network [28]
CA-DBLP	317,080	1,049,866	Co-authorship [5]
• Information (citation) networks			
CIT-HEP-TH	27,400	352,021	Arxiv hep-th [14]
AMAZONPROD	524,371	1,491,793	Amazon products [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	2,076,9		
• Internet networks			
ASSKITTER	1,719,0		
GNUTELLA	62,5		

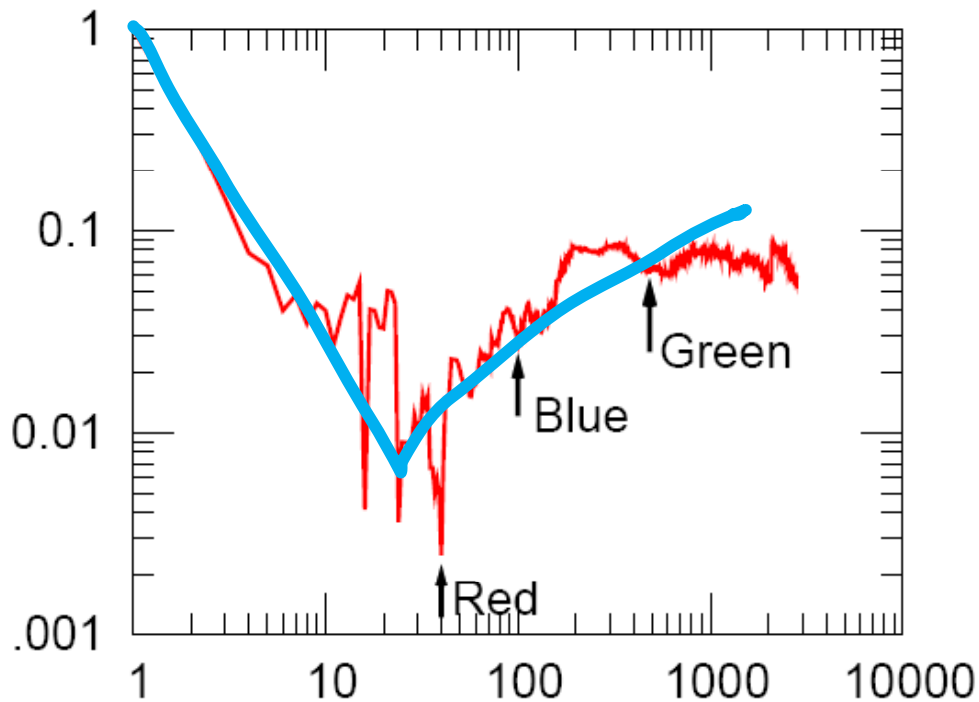
Examine more than **100 large networks**

[Internet Mathematics '09]

Large networks: Very different

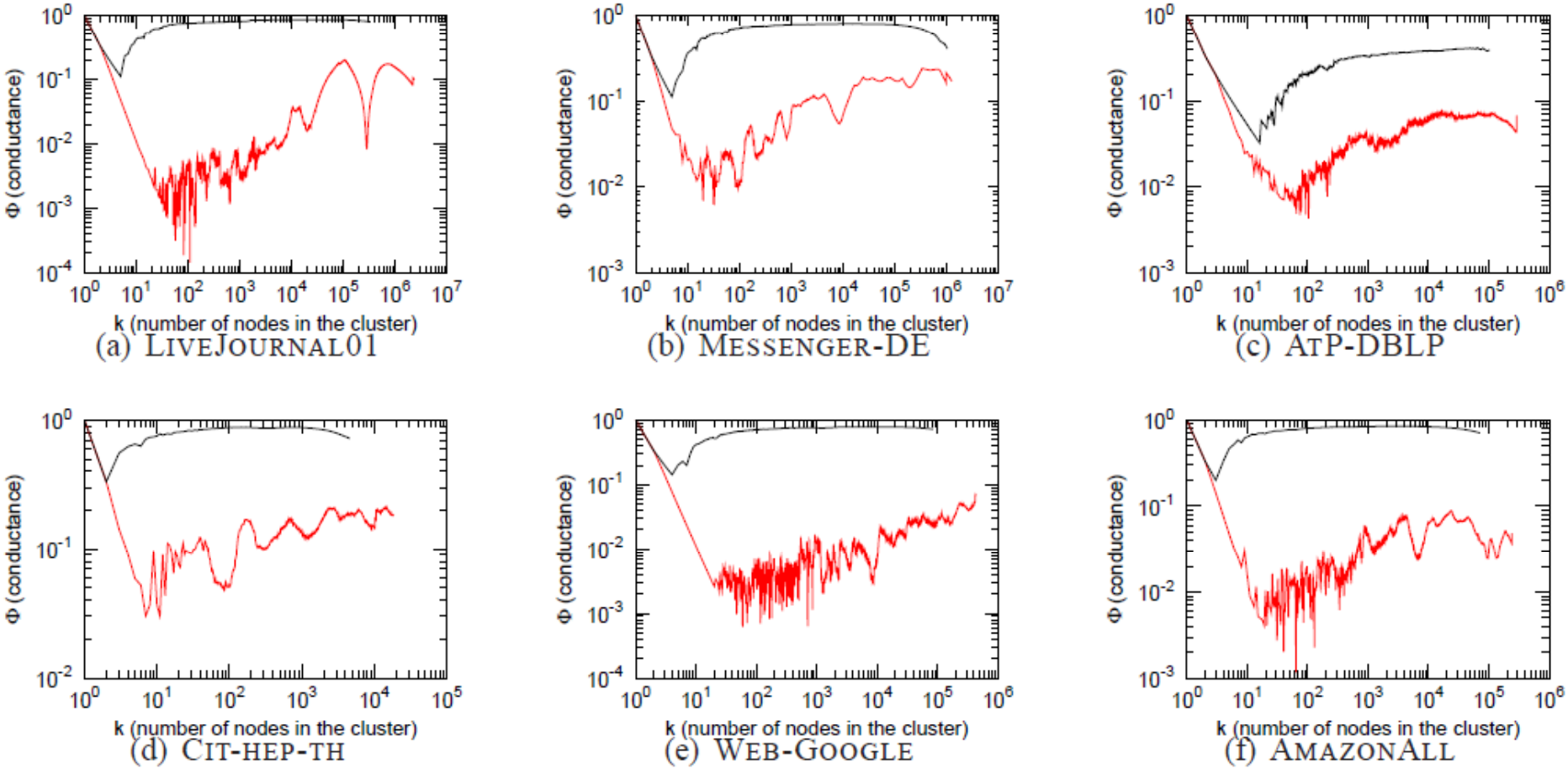
Typical example: General Relativity collaborations (n=4,158, m=13,422)





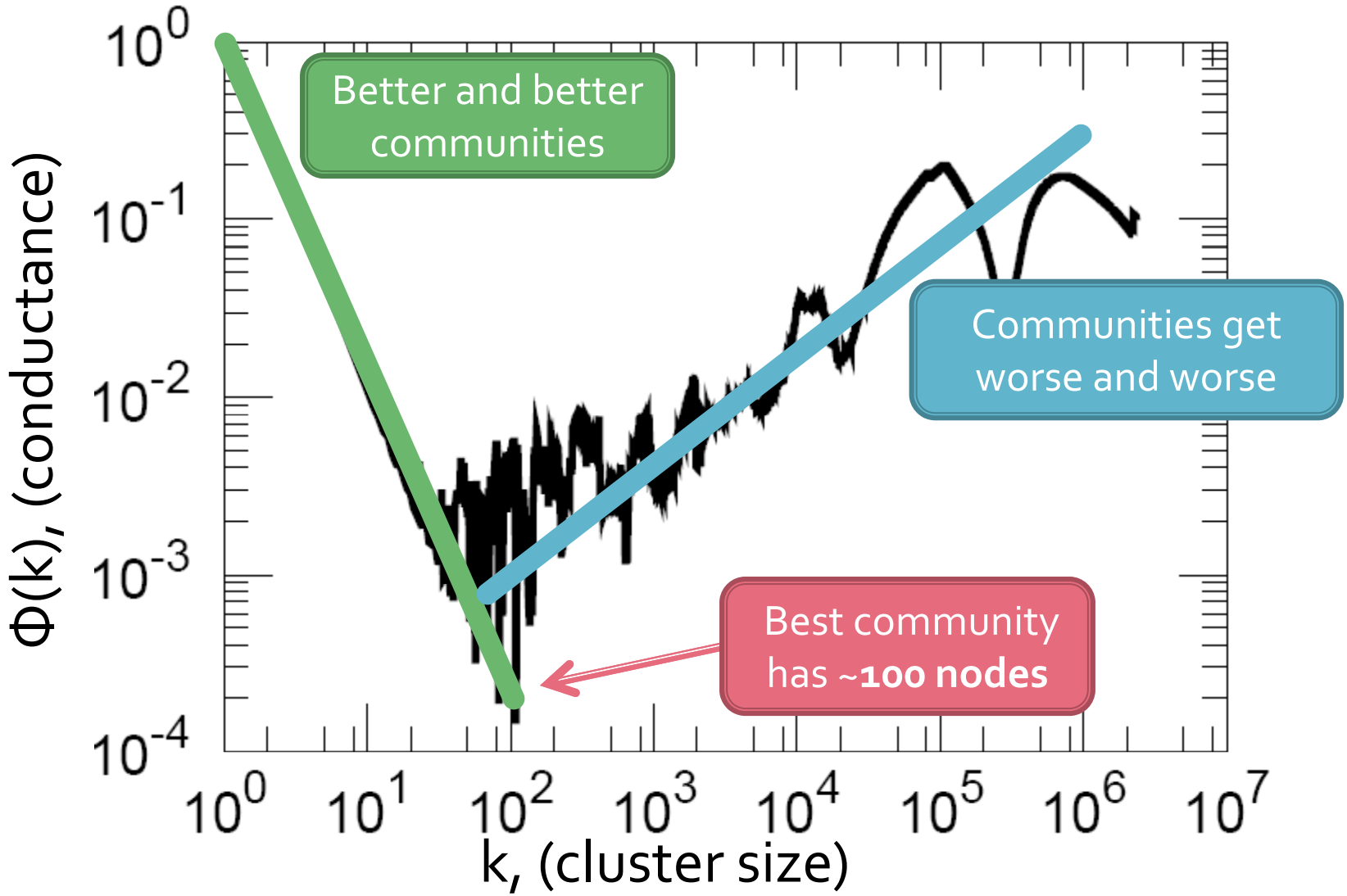
[Internet Mathematics '09]

More NCP plots of networks



[Internet Mathematics '09]

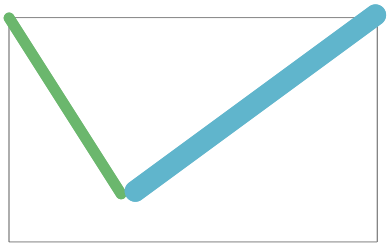
NCP: LiveJournal (n=5m, m=42m)



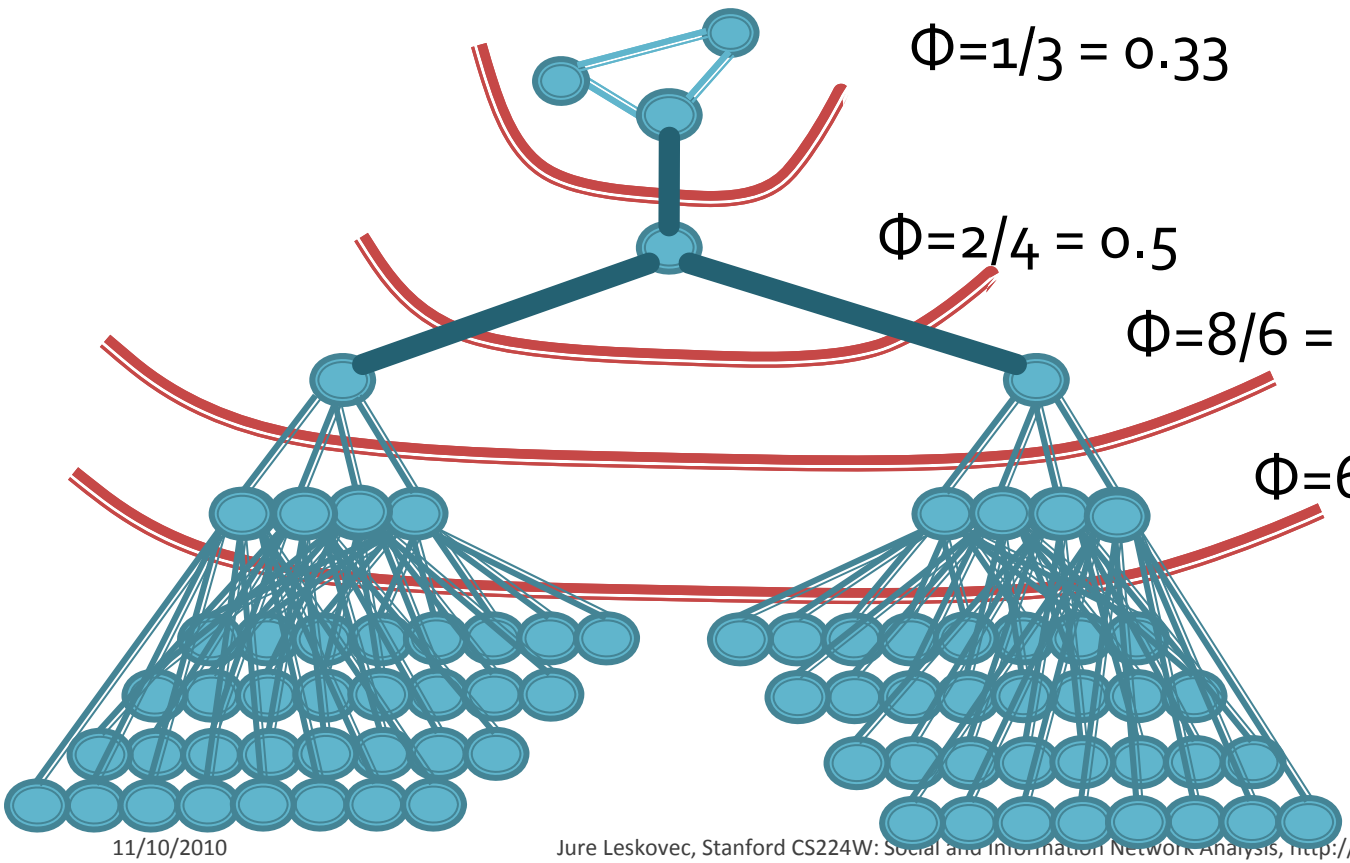
[Internet Mathematics '09]

Explanation: Upward part

- Each successive edge inside the community costs **more** cut-edges



NCP plot

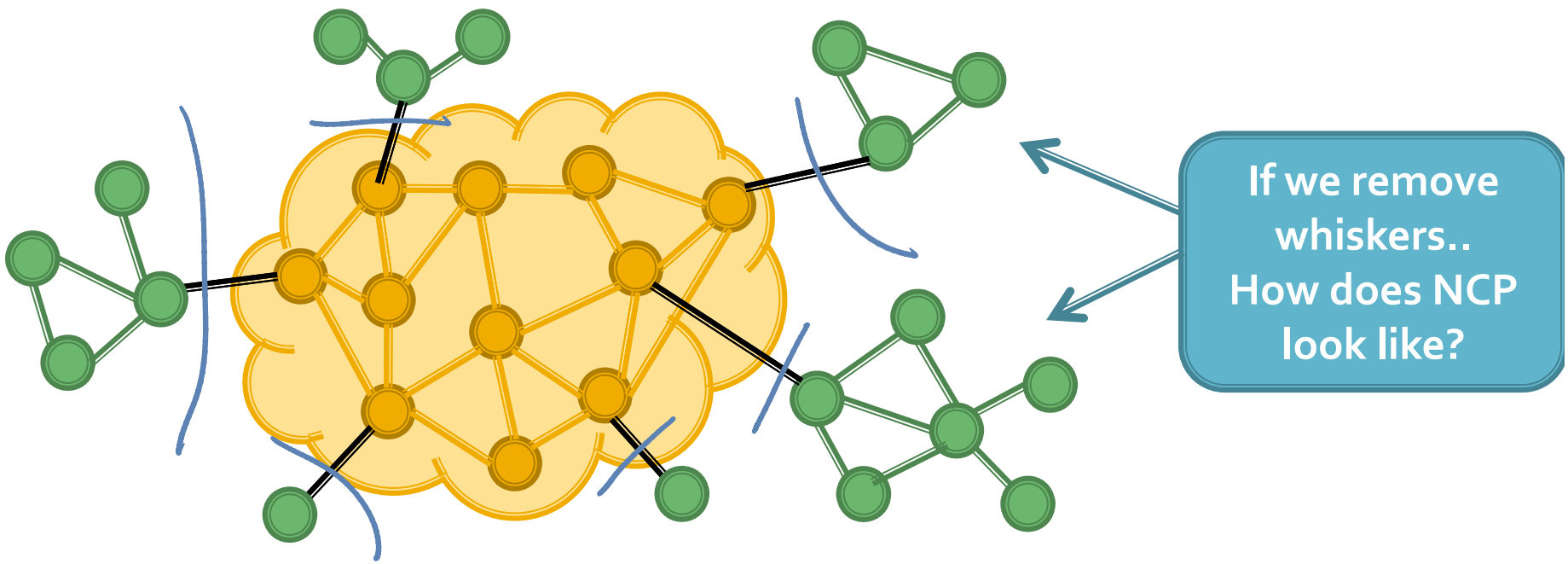


Each node has twice as many children

[Internet Mathematics '09]

Explanation: Downward part

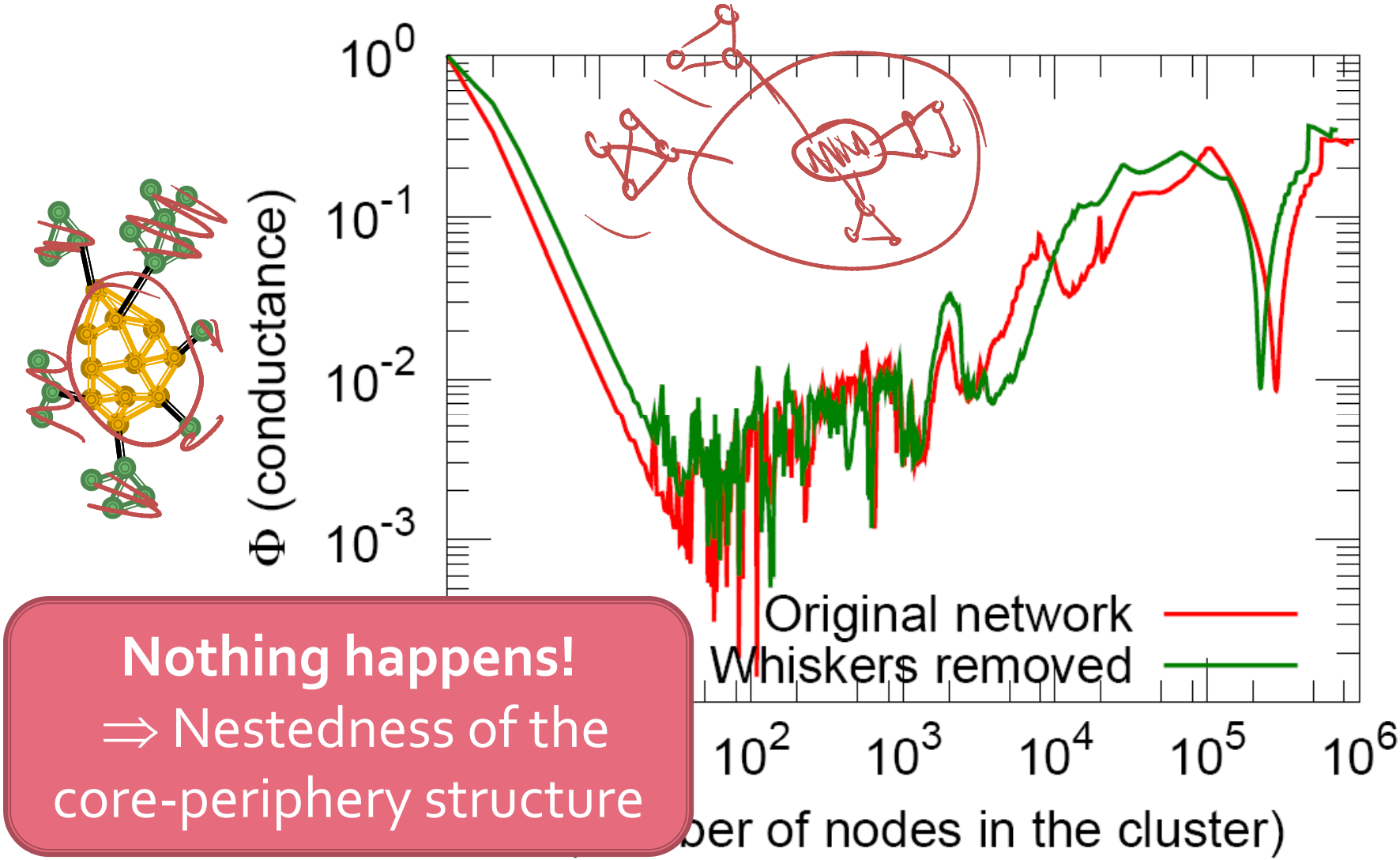
- Empirically we note that best clusters (call them **whiskers**) are barely connected to the network



$\Rightarrow$ Core-periphery structure

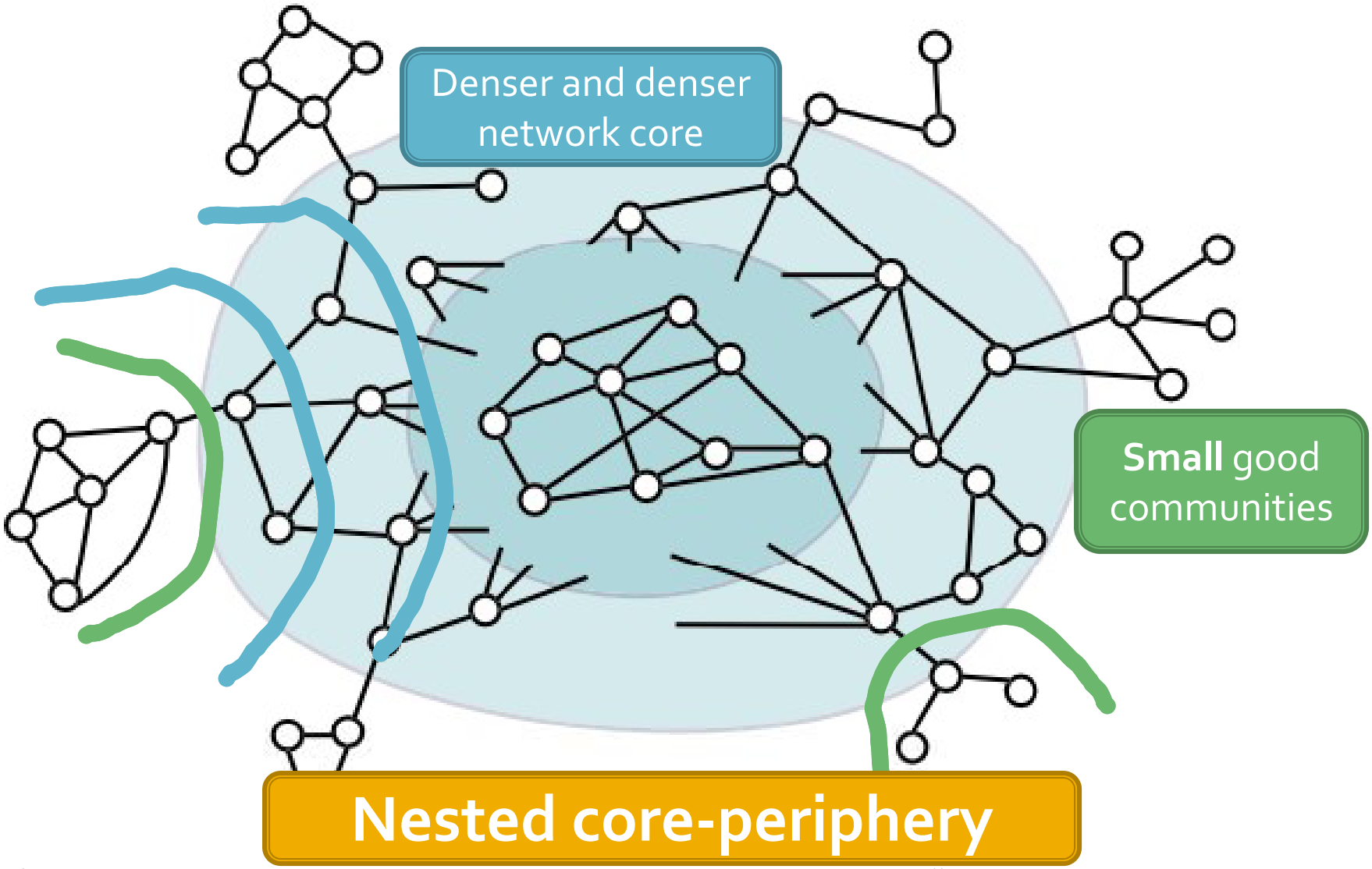
[Internet Mathematics '09]

What if we remove whiskers?



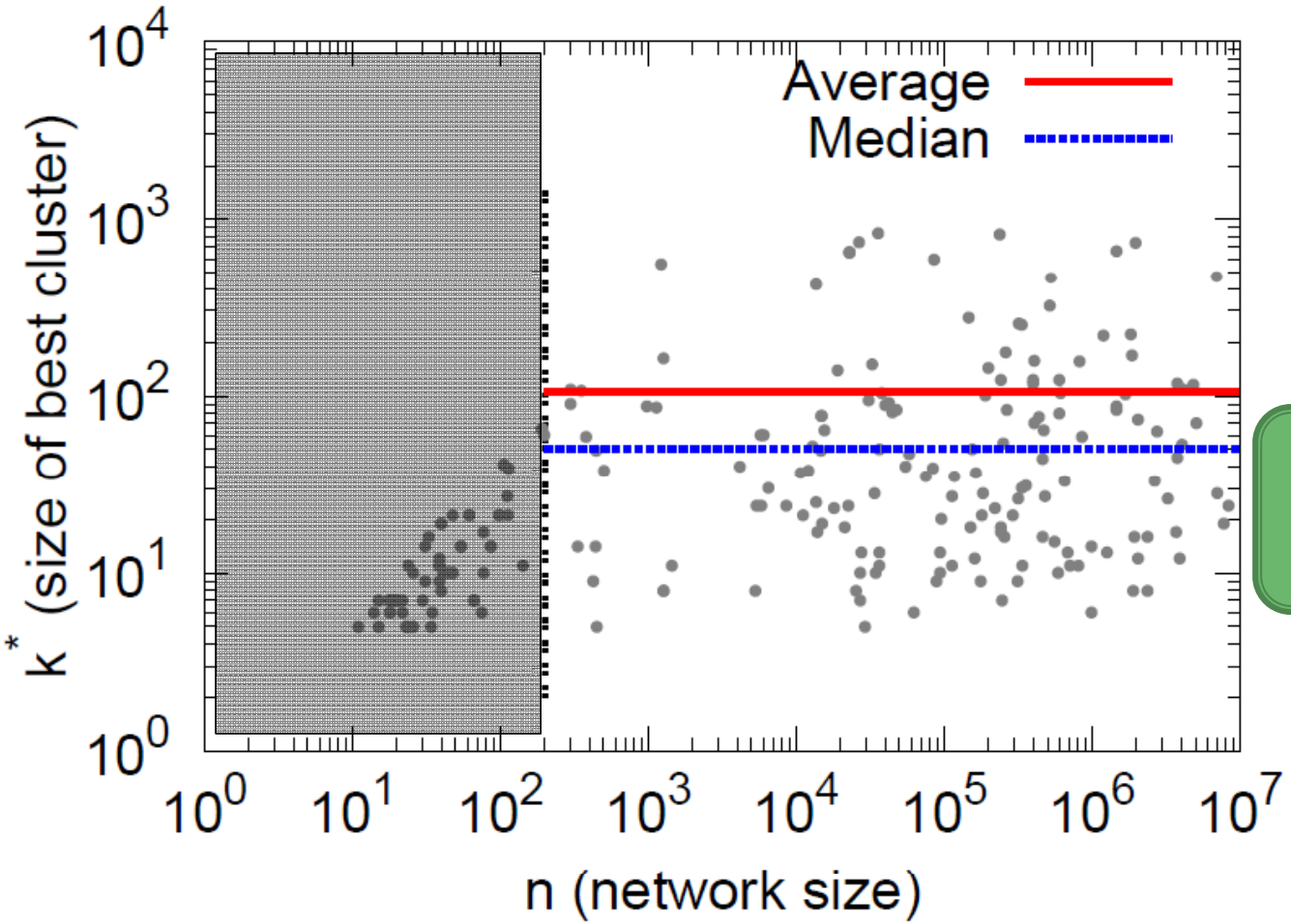
Nothing happens!
⇒ Nestedness of the core-periphery structure

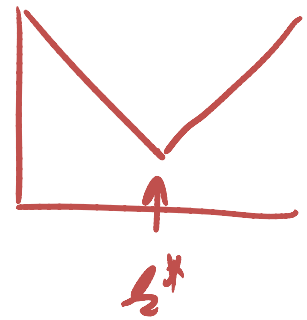
Nested core-periphery



Cluster size is independent of the network size

[Internet Mathematics '09]



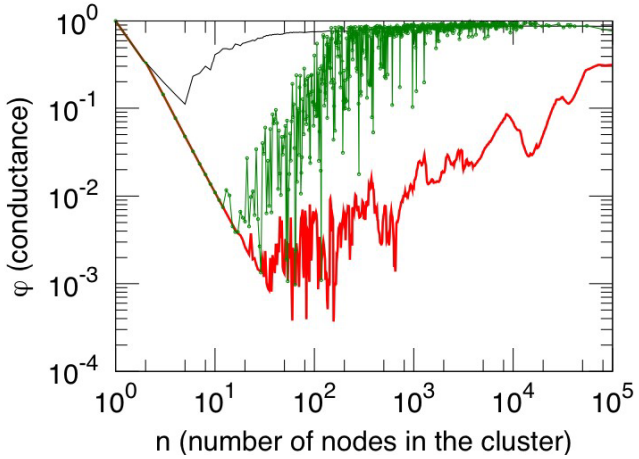


Practically **constant!**

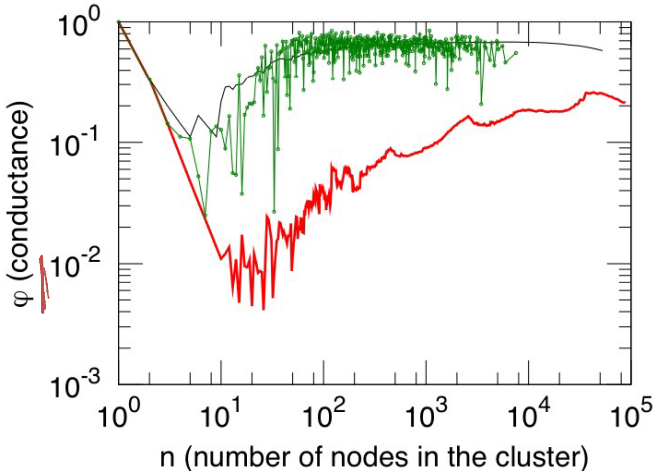
- Each dot is a different network

[Internet Mathematics '09]

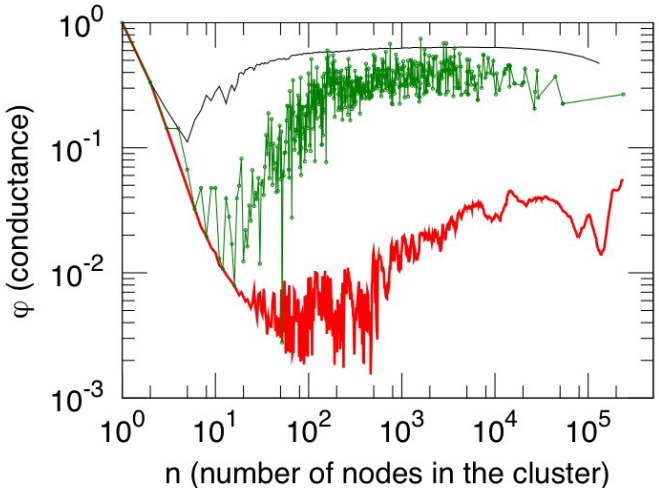
Comparison with "Ground truth"



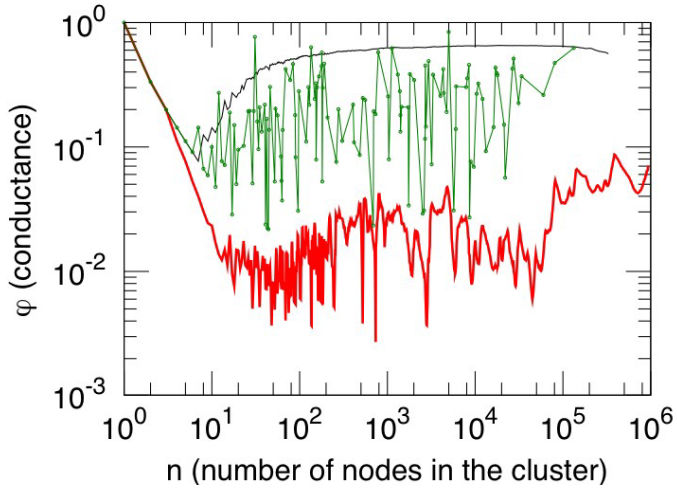
LiveJournal



DBLP



Amazon



IMDB

Rewired

Network

Ground truth

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"
  - Methods can perform well/poorly on some kinds of graphs
- Questions:
  - How well do algorithms optimize objectives?
  - What clusters do different methods find?

[WWW '09]

Many classes of algorithms

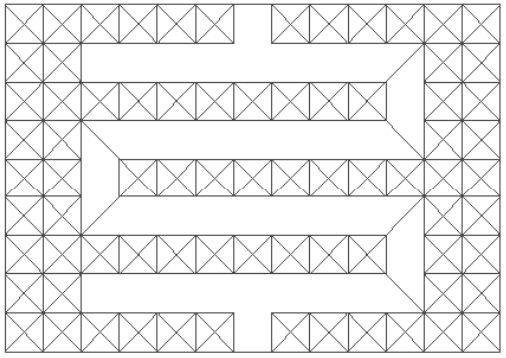
Many algorithms to extract clusters:

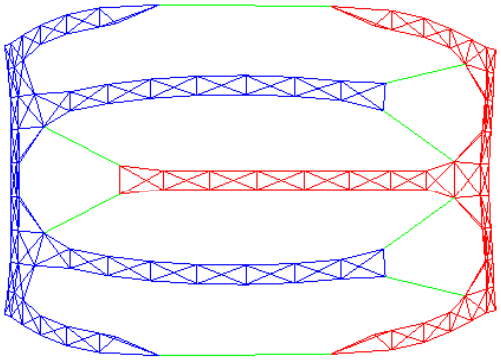
- Heuristics:
  - Girvan-Newman, Modularity optimization: popular heuristics
  - Metis (multi-resolution heuristic): common in practice [Karypis-Kumar '98]
- Theoretical approximation algorithms:
  - Spectral partitioning: most practical but confuses "long paths" with "deep cuts"
    - Local Spectral [Andersen-Chung '07]
  - Leighton-Rao: based on multi-commodity flow

[WWW '09]

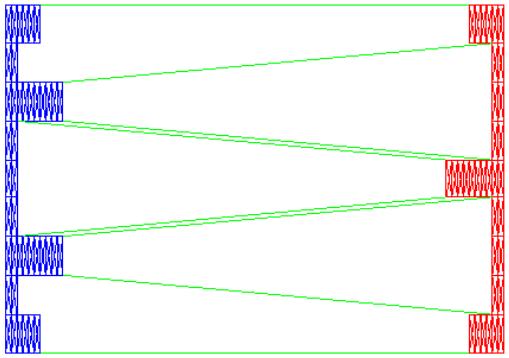
Spectral vs. Flow methods

- Spectral embeddings stretch along directions in which the random-walk mixes slowly
  - Resulting hyperplane cuts have "good" conductance cuts, but may not yield the optimal cuts





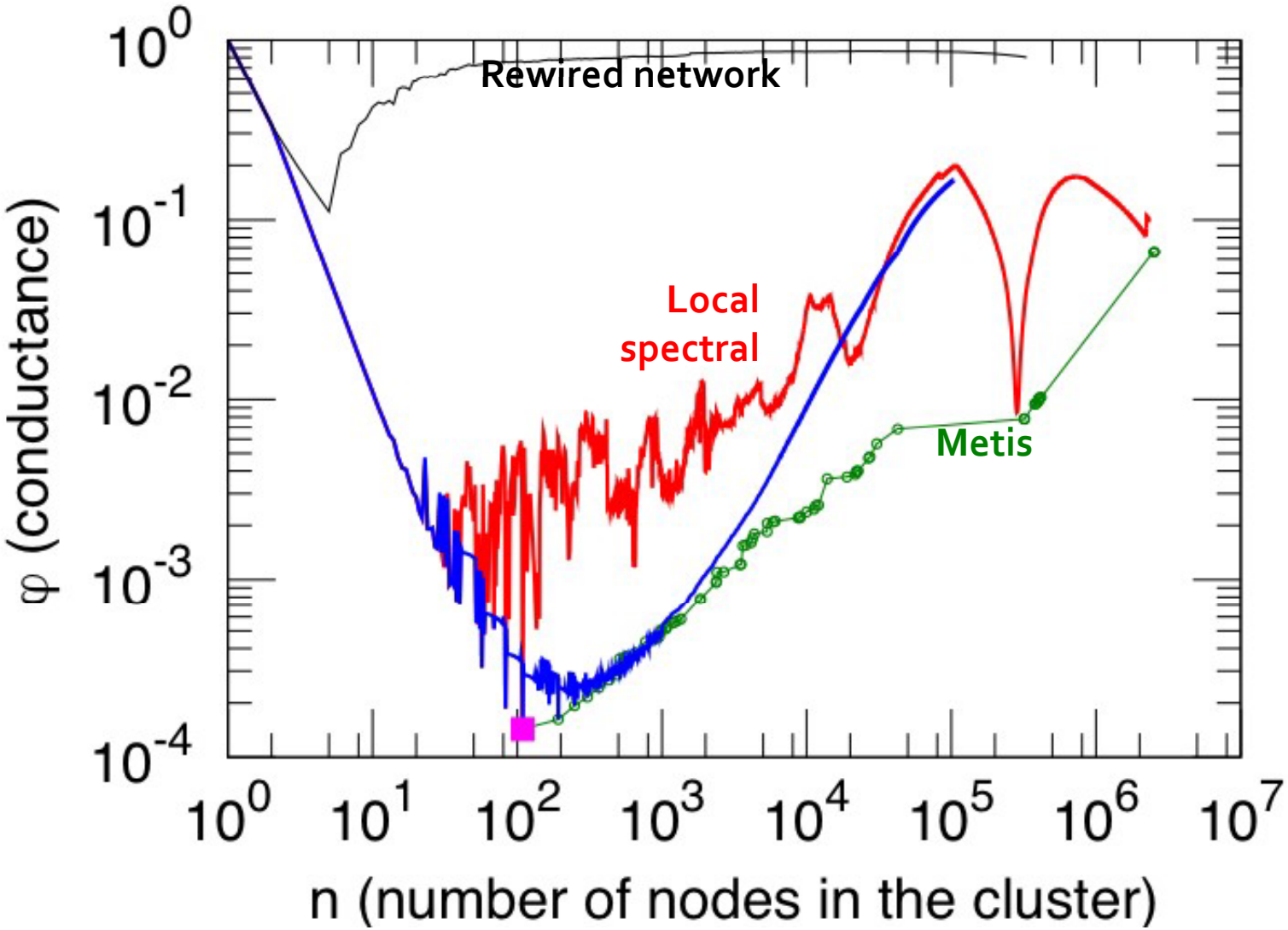
spectral embedding



flow based embedding

[WWW '09]

NCP: Live Journal



[WWW '09]

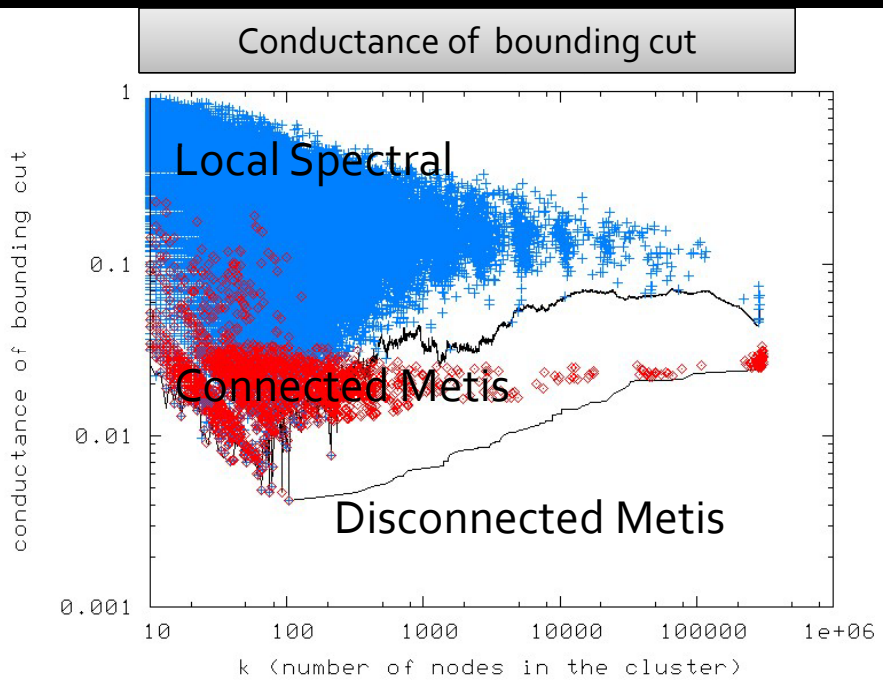
Properties of clusters (1)

500 node communities from Local Spectral:

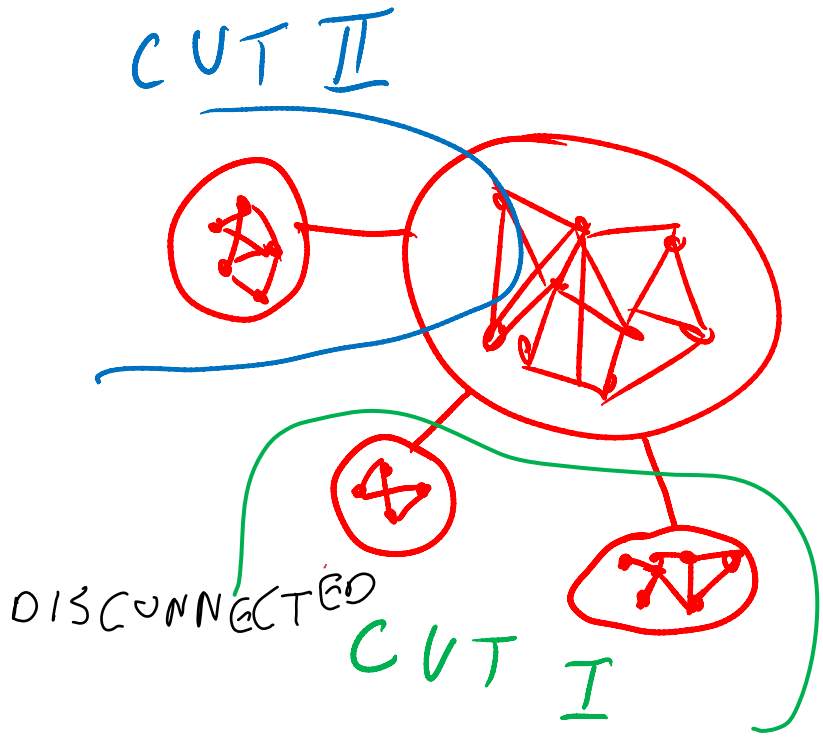
500 node communities from Metis:

[WWW '09]

Properties of clusters (2)



- Metis (red) gives sets with better conductance
- Local Spectral (blue) gives tighter and more well-rounded sets

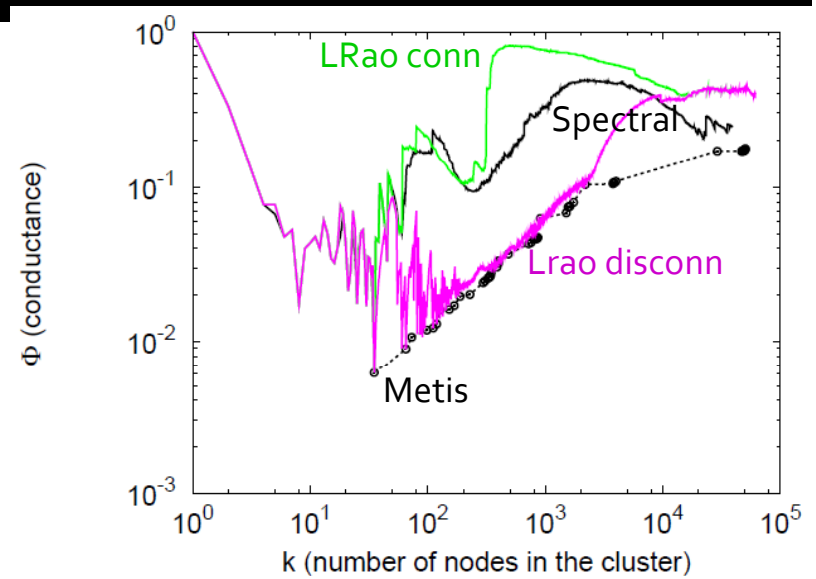


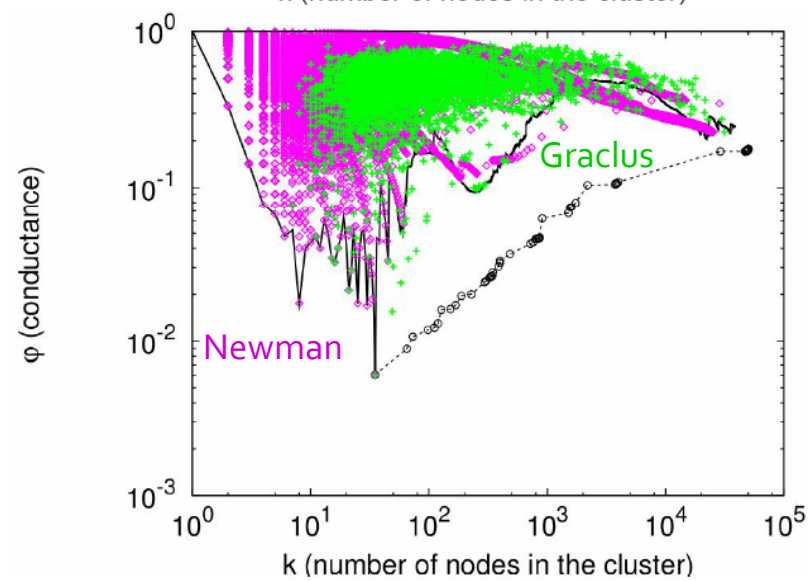
CUT I is better than CUT II

[WWW '09]

Other clustering methods

- LeightonRao: based on multi-commodity flow
  - Disconnected clusters vs. Connected clusters
- Graclus prefers larger clusters
- Newman's modularity optimization similar to Local Spectral

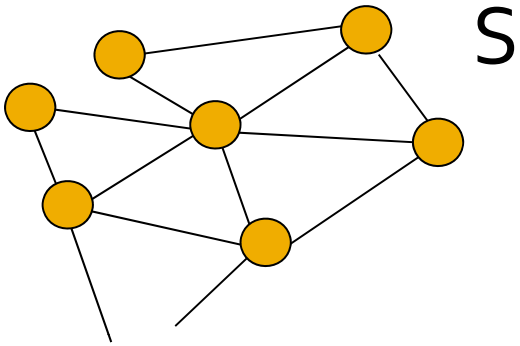




[WWW '09]

Many different objective functions

- Single-criterion:
  - Modularity: $m\text{-}E(m)$
  - Modularity Ratio: $m\text{-}E(m)$
  - Volume: $\sum_u d(u)=2m+c$
  - Edges cut: c
- Multi-criterion:
  - <u>Conductance</u>: $c/(2m+c)$
  - Expansion: c/n
  - Density: $1\text{-}m/n^2$
  - CutRatio: $c/n(N\text{-}n)$
  - Normalized Cut: $c/(2m+c) + c/2(M\text{-}m)+c$
  - Max ODF: *max frac. of edges of a node pointing outside S*
  - Average-ODF: *avg. frac. of edges of a node pointing outside*
  - Flake-ODF: *frac. of nodes with mode than ½ edges inside*



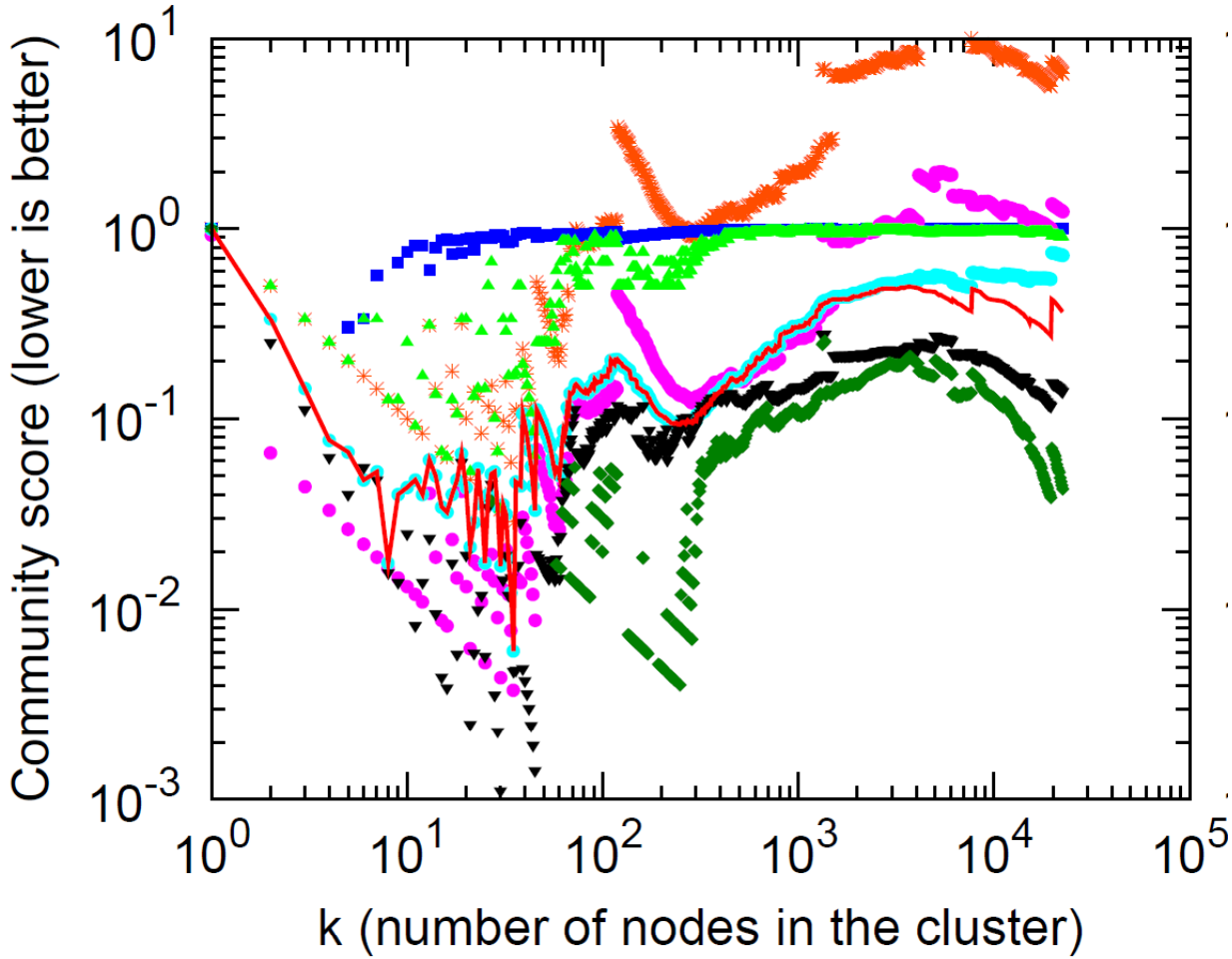
n: nodes in S

m: edges in S

c: edges pointing outside S

[WWW '09]

Multi-criterion objectives

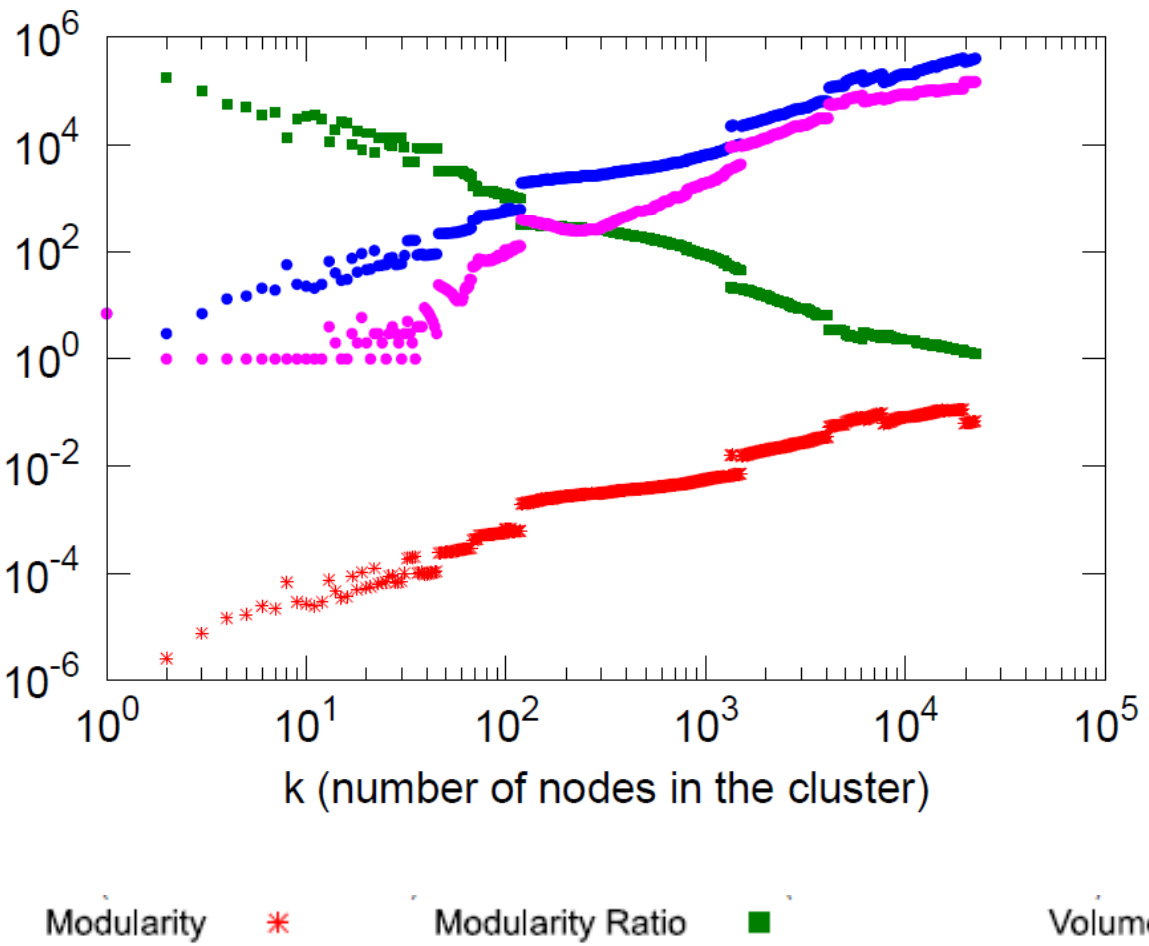


- Qualitatively similar
- Observations:
  - Conductance, Expansion, Norm-cut, Cut-ratio and Avg-ODF are similar
  - Max-ODF prefers smaller clusters
  - Flake-ODF prefers larger clusters
  - Internal density is bad
  - Cut-ratio has high variance

Conductance — Expansion — Internal Density — Cut Ratio — Normalized Cut — Maximum ODF — Avg ODF — Flake ODF

[WWW '09]

Single-criterion objectives



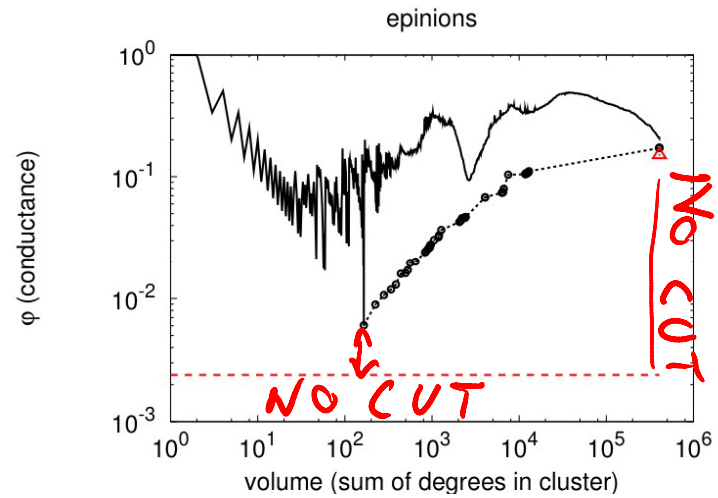
Observations:

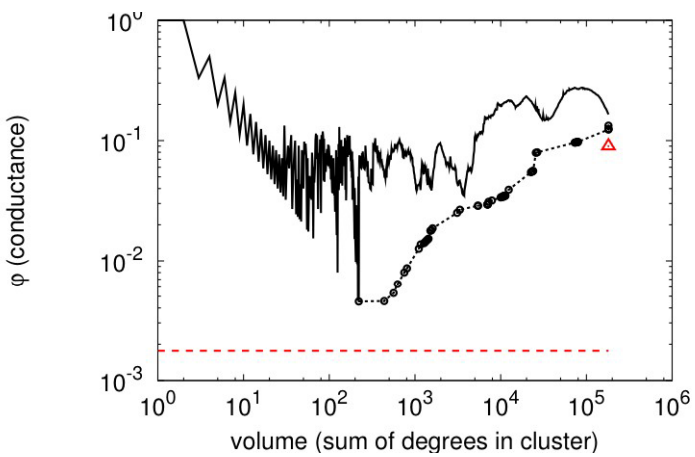
- All measures are monotonic
- Modularity
  - prefers large clusters
  - Ignores small clusters

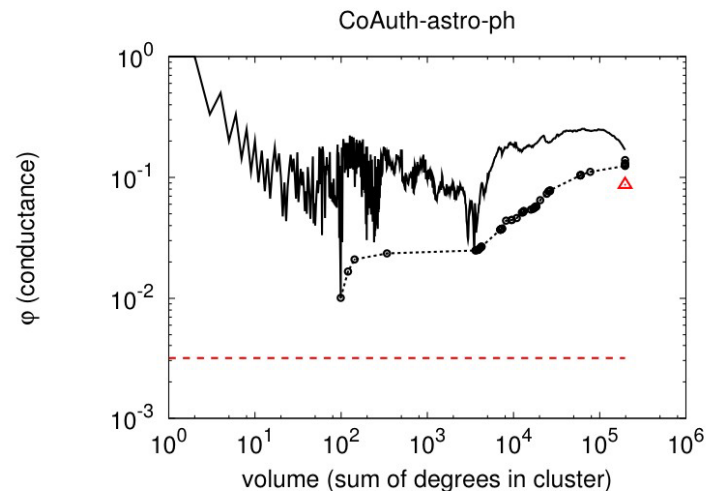
[WWW '09]

Lower and upper bounds

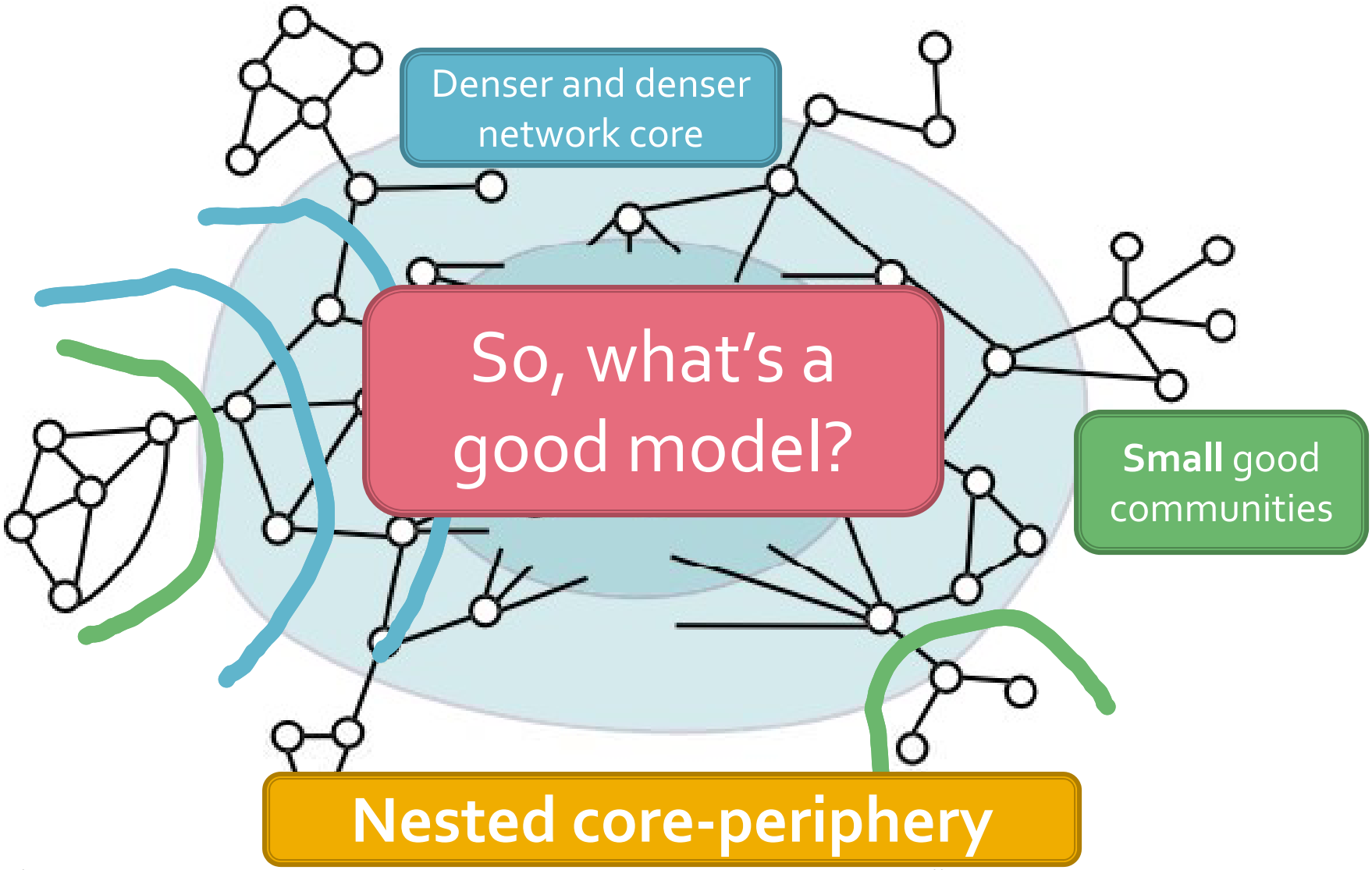
- Lower bounds on conductance can be computed from:
  - Spectral embedding LINE (independent of balance)
  - SDP-based methods (for volume-balanced partitions) △
- Algorithms find clusters close to theoretical lower bounds





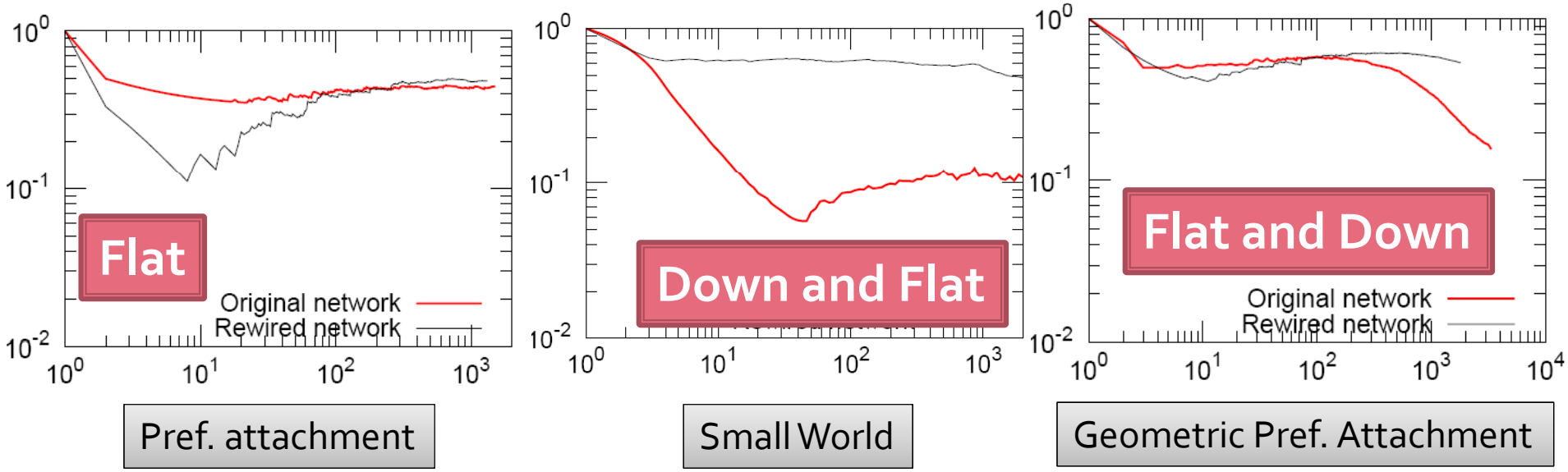


Nested core-periphery



[Internet Mathematics '09]

Modeling nested core-periphery

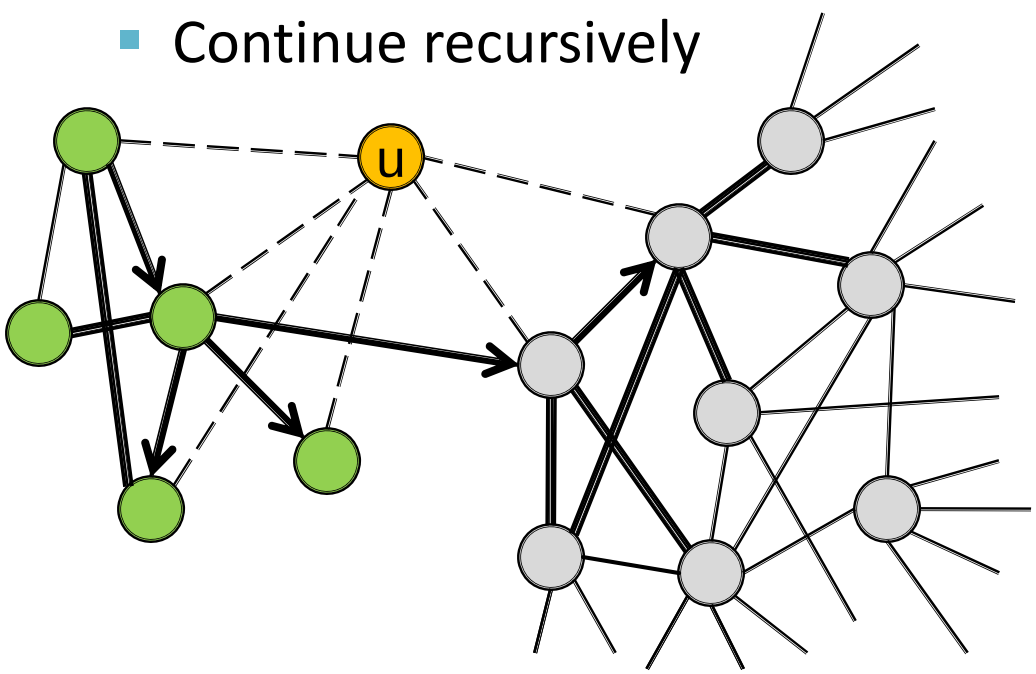


- None of the common models works!

[Internet Mathematics '09]

Forest Fire model works!

- **Forest Fire:**
  connections spread like a fire
  - New node joins the network
  - Selects a seed node
  - Connects to some of its neighbors
  - Continue recursively



Model ingredients:

- Preferential attachment: second neighbor is not uniform at random
- Copying flavor: - since we burn seed's neighbors
- Hierarchical flavor:- burn around the seed node
- "Local" flavor:- burn "near" the seed node

As cluster **grows** it **blends** into the **core** of the **network**

[Internet Mathematics '09]

Forest Fire NCP plot