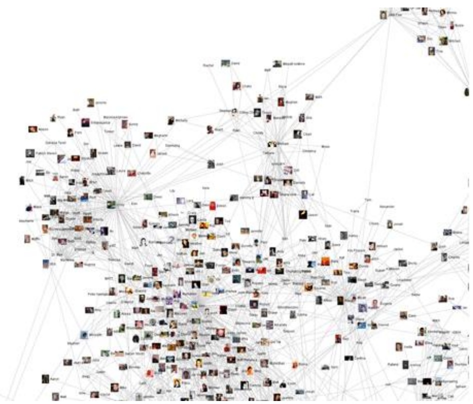
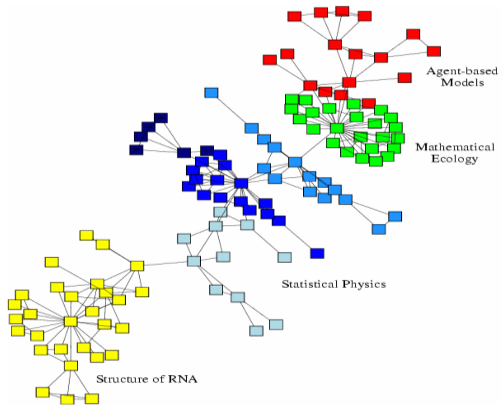
Networks, Communities and the Ground-Truth

Jure Leskovec
Stanford University &
Inštitut Jožef Stefan

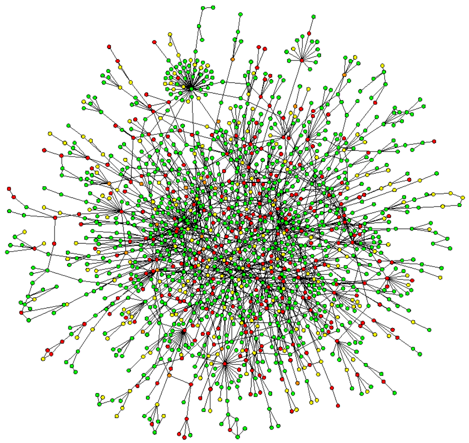
Many Data ARE Networks



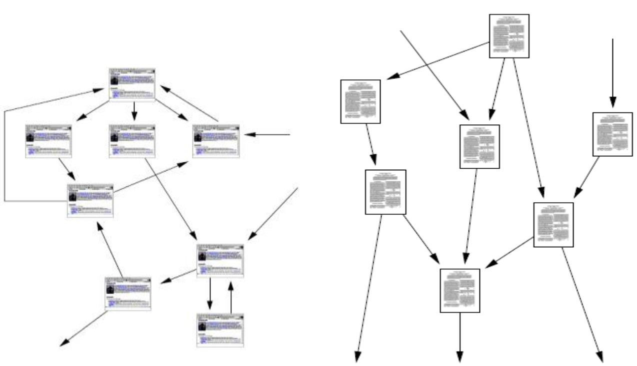
Online social networks



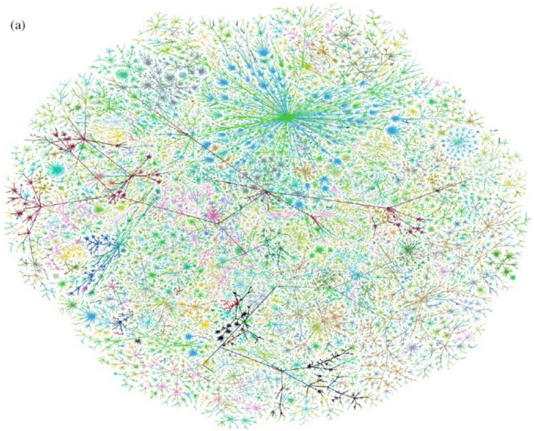
Collaboration networks



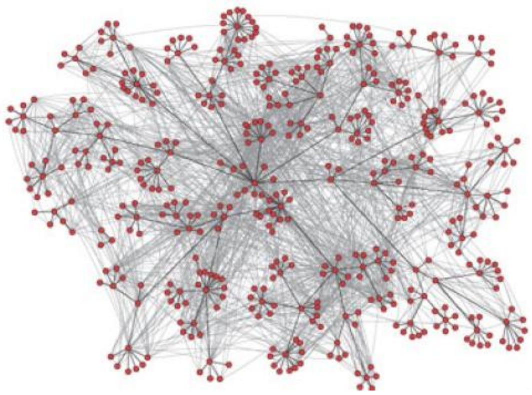
Systems biology networks



Web graphs & citation networks



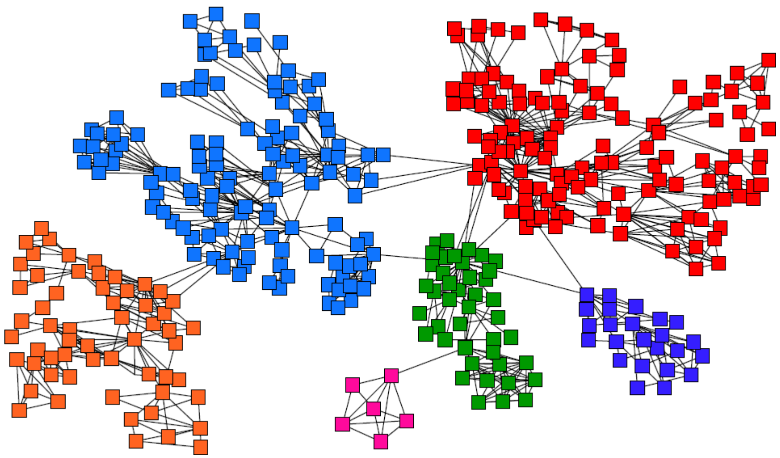
Computer & Internet networks



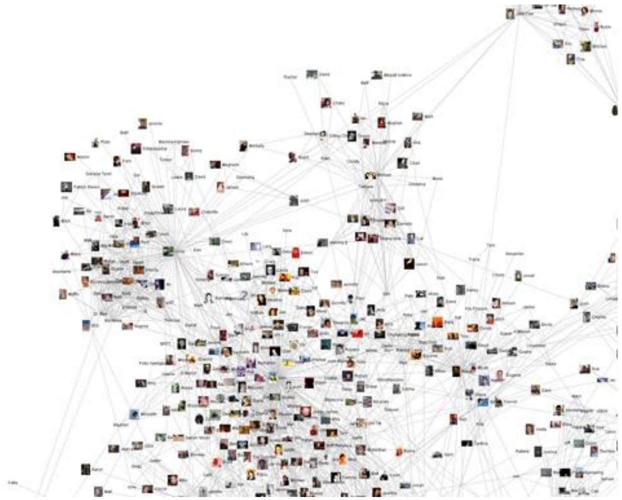
Communication networks

Organization of Networks

How are networks organized?



Collaborations in Network Science

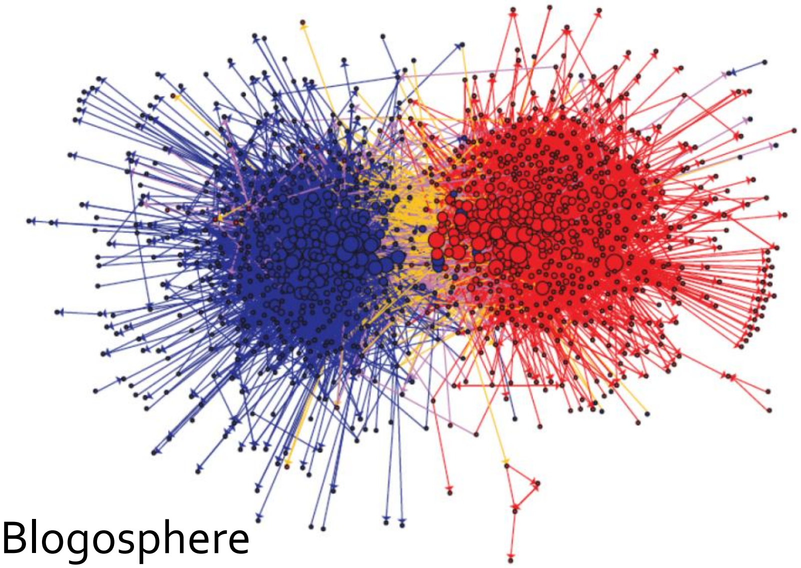


Tiny part of a large social network

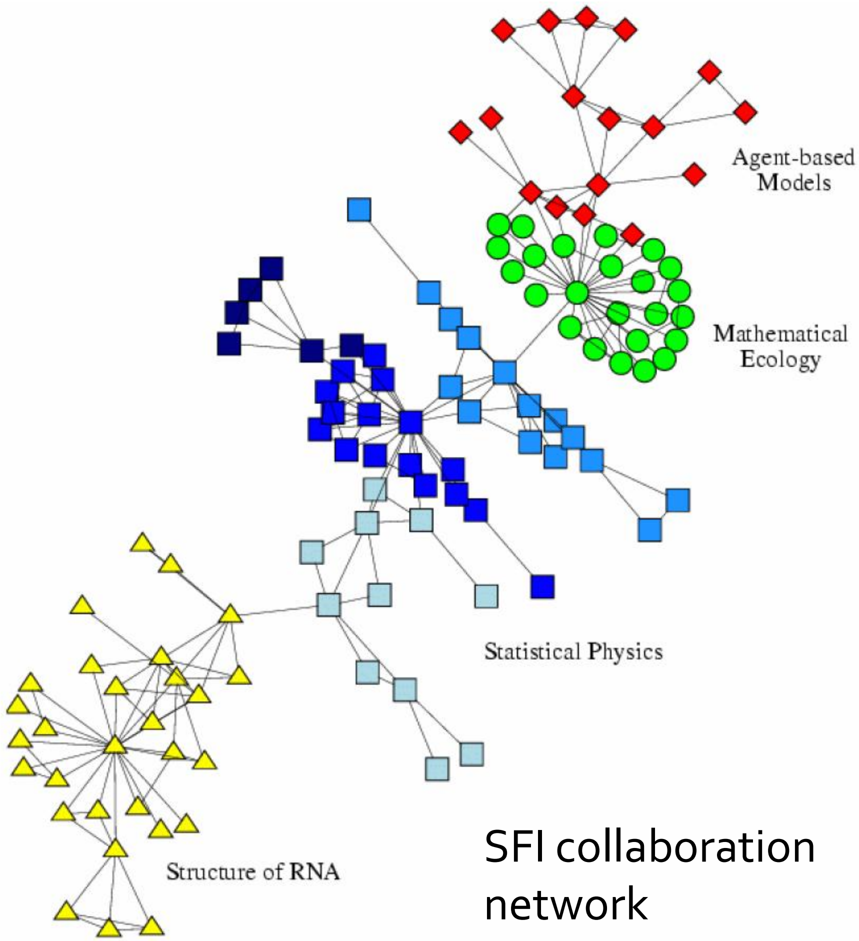
What is the structure of the network?
How should we think about it?

Network Communities

- **Communities are of interest in…**
  - World Wide Web
  - Citation networks
  - Social networks
  - Metabolic networks



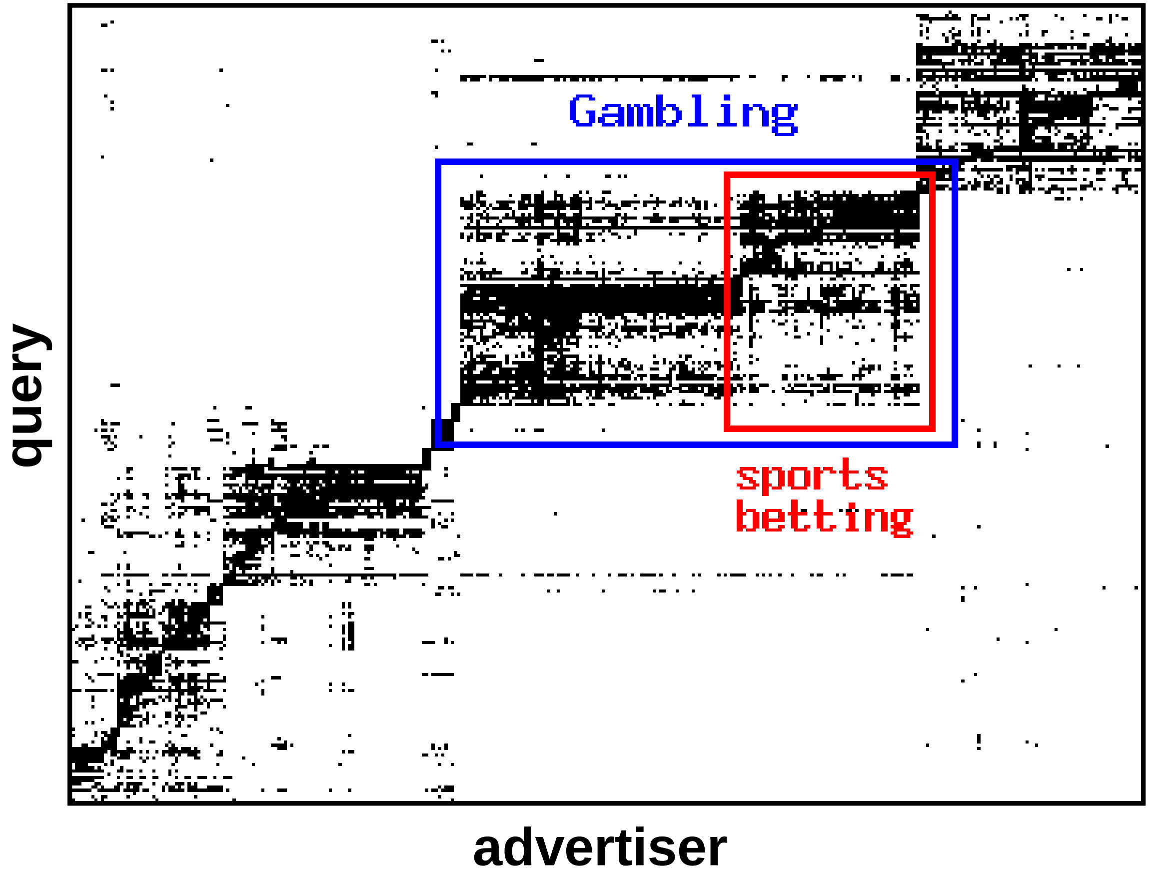
Blogosphere



SFI collaboration network

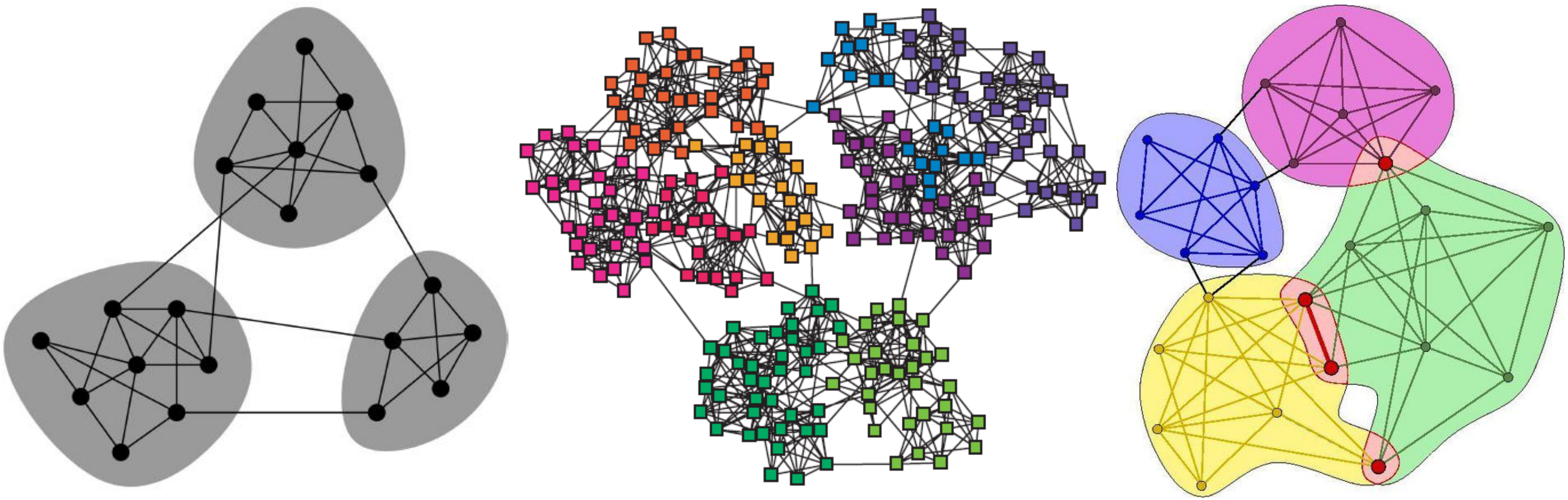
Communities as Micro-markets

- Communities as Micro-markets in "query - advertiser" graph

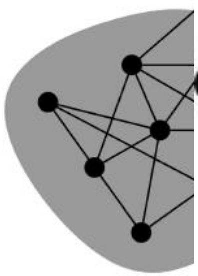


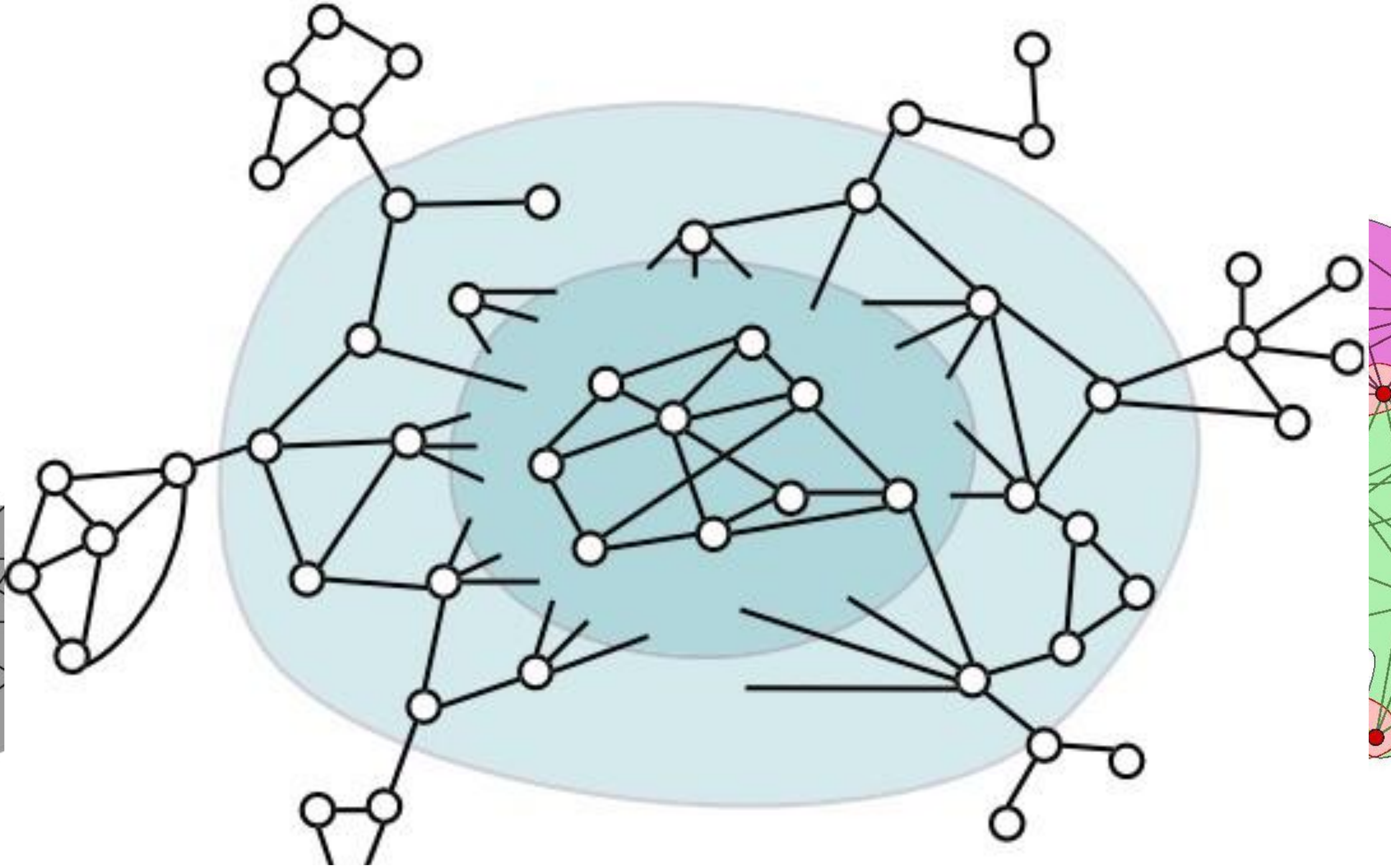
This Talk: Networks & Communities

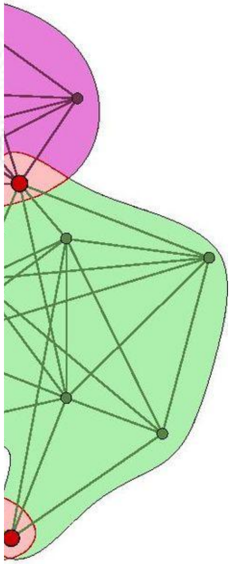
Network Communities



This Talk: Networks & Communities



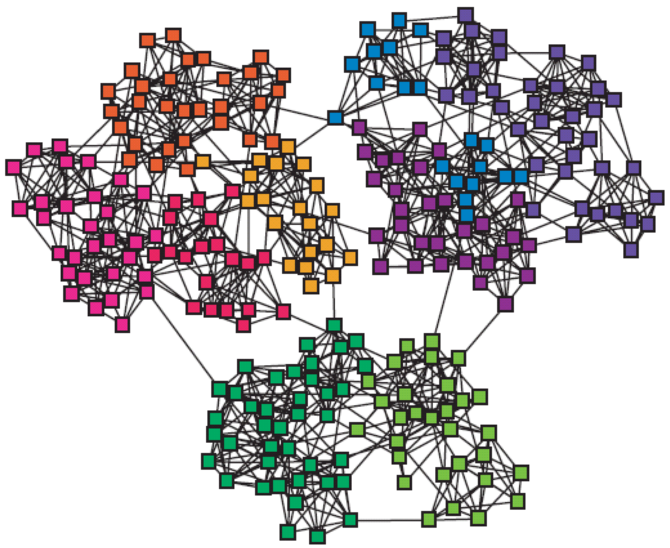




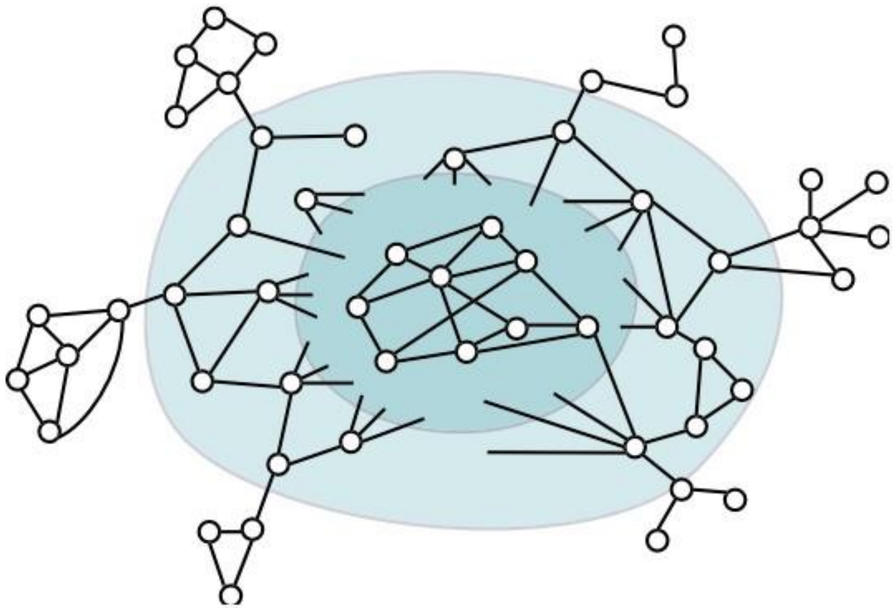
Nested Core-Periphery

(Leskovec et al., Internet Mathematics, 2009)

This Talk: Networks & Communities



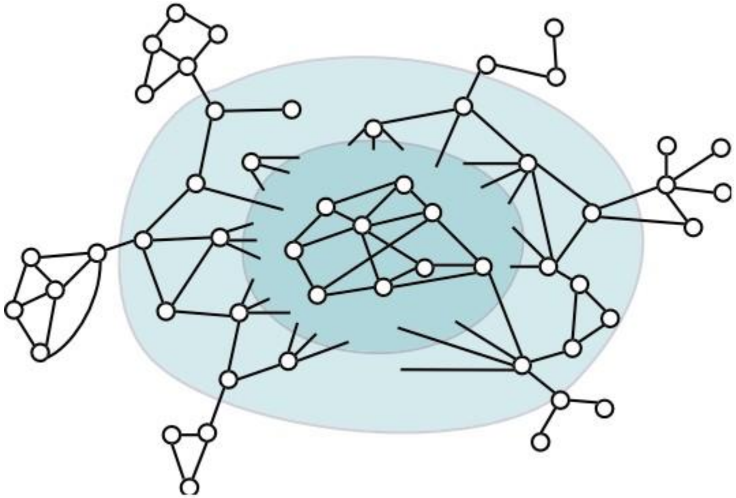
vs.



How do we reconcile these two views?

Part 1: Core-Periphery

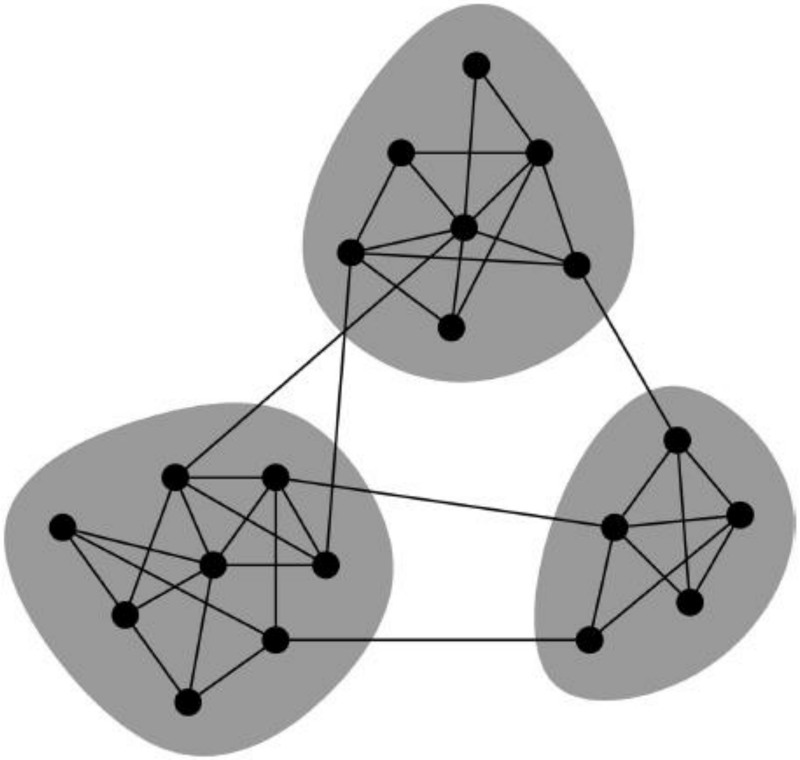
- **How network organize into clusters?**
  - **What computational experiment should we design reveal network organization?**



  - **Idea:** Use **approximation algorithms** for the NP-hard graph partitioning problems as **experimental probes** of the network structure

Network Communities

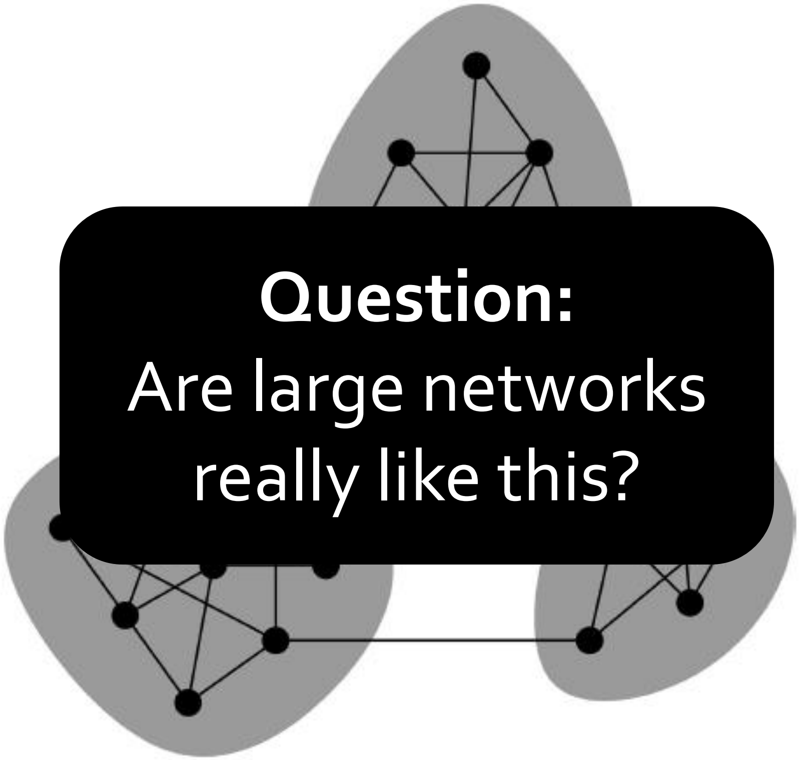
- **Communities:**
  - Working definition: Sets of nodes with **lots** of links **inside** the set and **few** to the **outside** (the rest of the network)
- **Industry:**
  - Develop methods that extract such "community-like" sets of nodes



Communities, clusters, groups, modules

Network Communities

- **Communities:**
  - Working definition: Sets of nodes with **lots** of links **inside** the set and **few** to the **outside** (the rest of the network)
- **Industry:**
  - Develop methods that extract such "community-like" sets of nodes

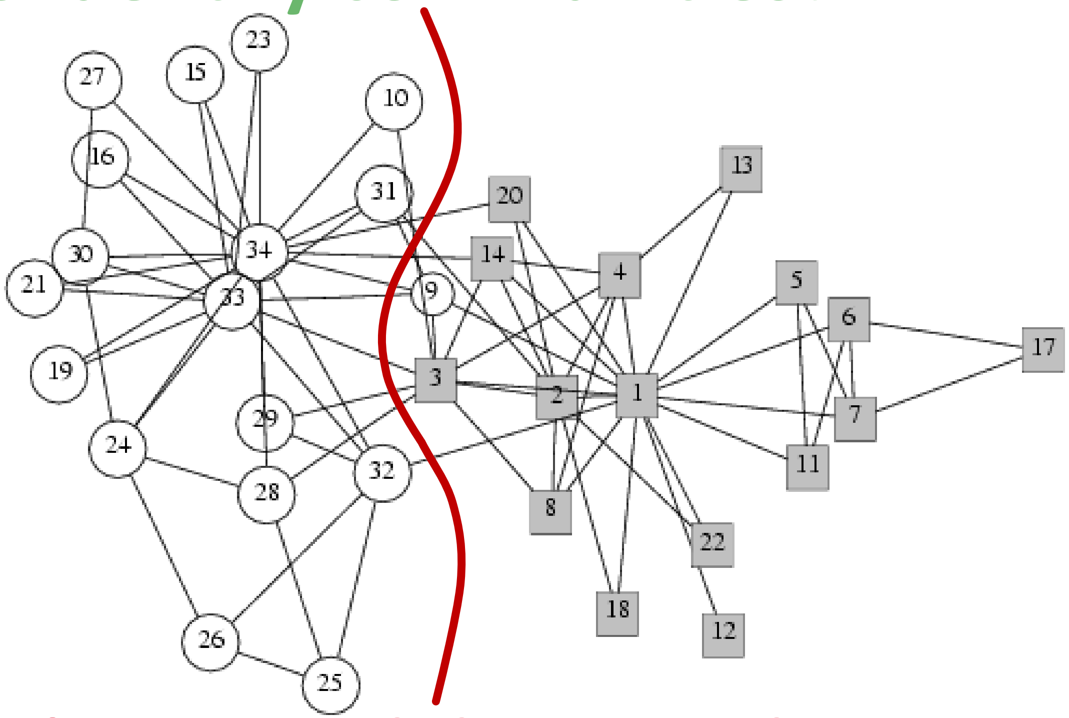


Question:
Are large networks really like this?

Communities, clusters, groups, modules

Communities: Social Networks

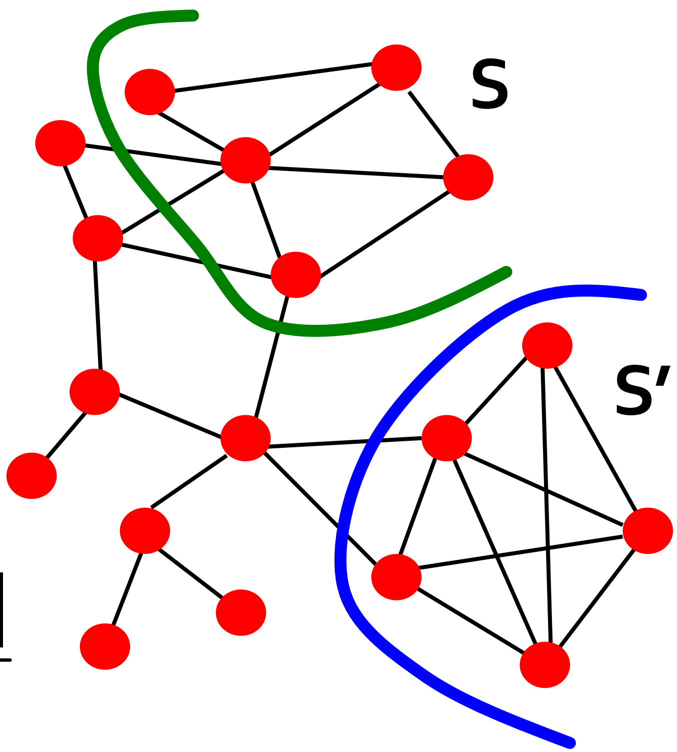
- How to identify communities?



- Zachary's Karate club network
  - Ties in a karate club, conflicts led the group to split
  - Split could be explained by a minimum cut

Community Score

- How "community-like" is a set of nodes?
- A good cluster *S* has
  - Many edges internally
  - Few edges pointing outside
- Simplest objective function: **Conductance**



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

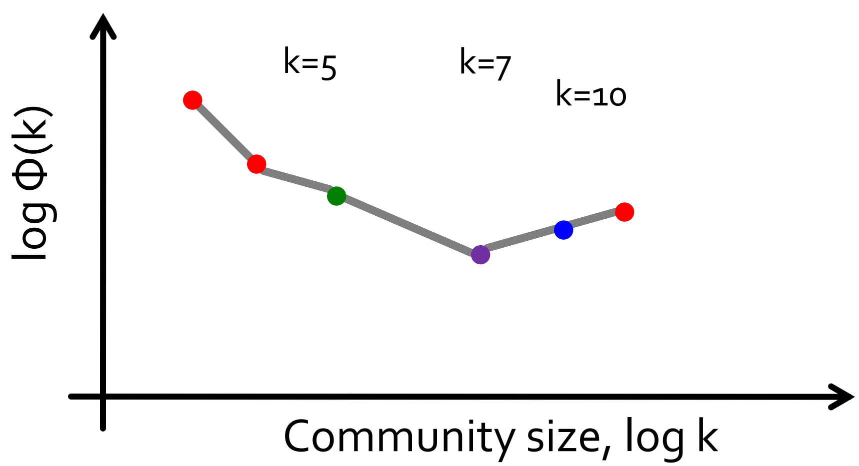
Small conductance corresponds to good clusters

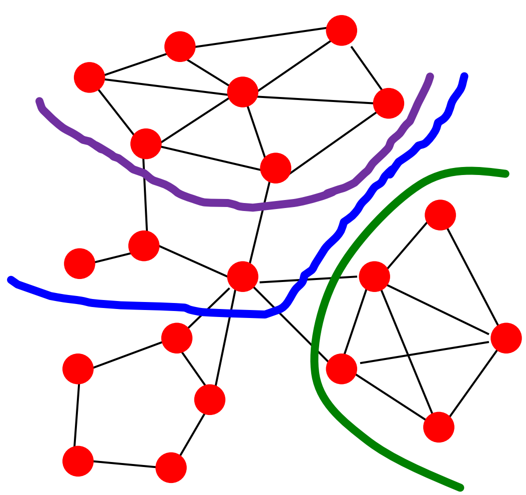
Network Community Profile

- Define:
  Network Community Profile (NCP) plot

  Plot the score of best community of size *k*

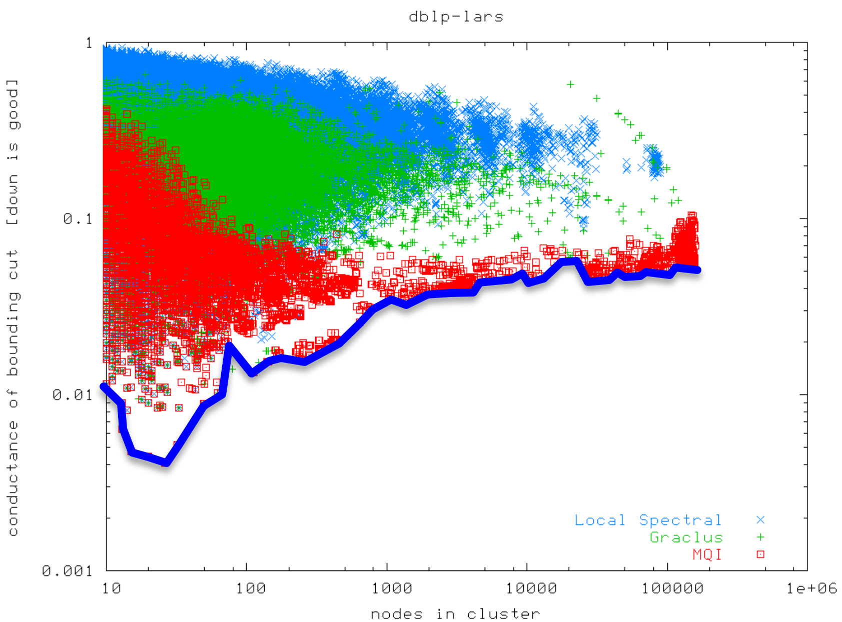
$$\Phi(k) = \min_{S \subset V, |S|=k} \phi(S)$$





Network Community Profile

- **Computing $\Phi(k) = \min_{S \subset V, |S|=k} \phi(S)$ is intractable!**
  - **Use approx. algorithms to graph partitioning**
    - Spectral (quadratic approx.):
      - confuses "long paths" with "deep cuts"
    - Multi-commodity flow (*log(n)* approx.):
      - difficulty with expanders
    - SDP (√*log(n)*):
      - best in theory
    - Metis (heuristic):
      - common in practice
  - **Bottom line: they all reveal similar NCP**

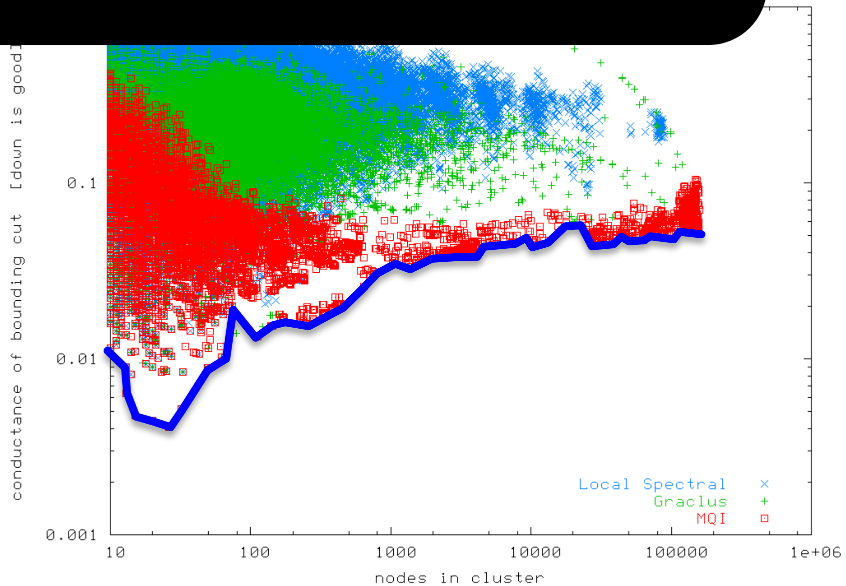


Network Community Profile

- Computing $\Phi(k) = \min_{S \subset V, |S|=k} \phi(S)$ is intractable!
  - Use approx. algorithms to graph partitioning

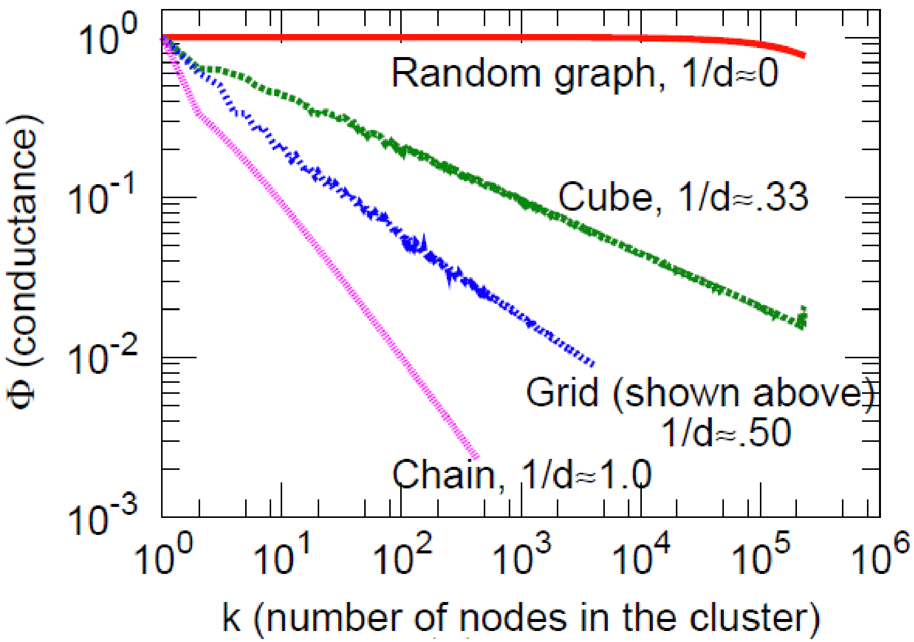
What does NCP tell us about the organization of networks?

- SDP (√*log(n)*):
  - best in theory
- Metis (heuristic):
  - common in practice
- **Bottom line: they all reveal similar NCP**

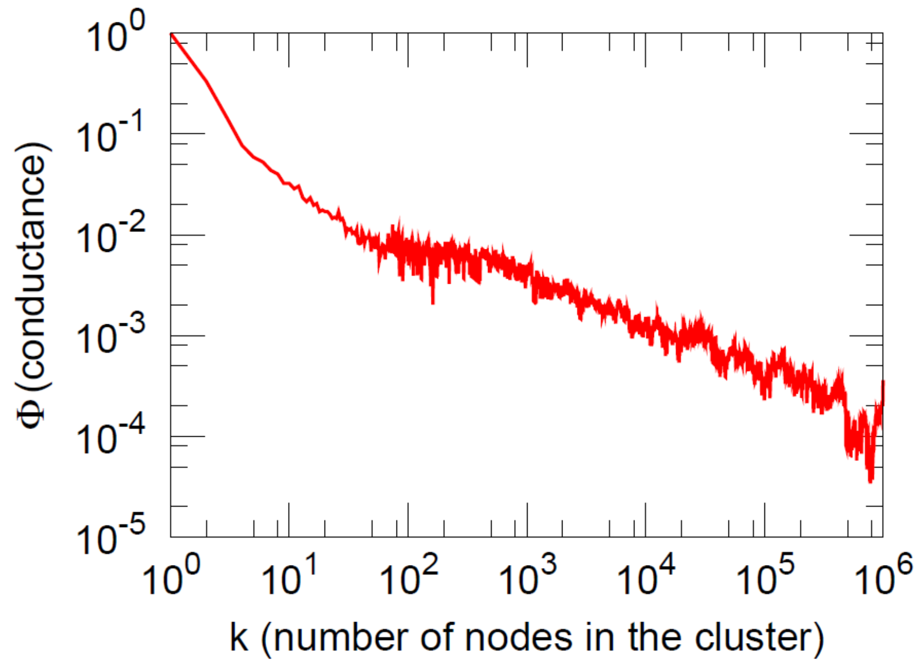


NCP Plot: Lattices

- **Lattices and Dense Random Graphs:**



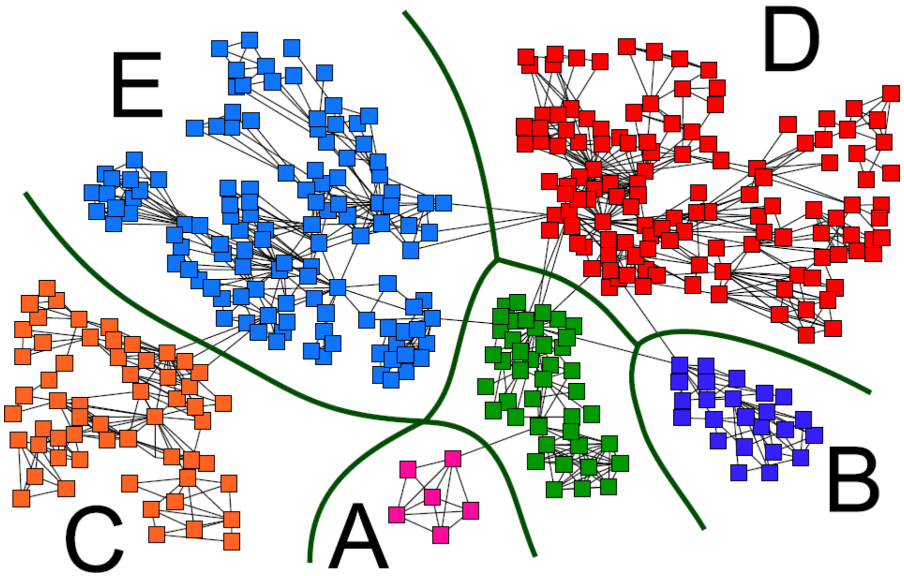
d-dimensional lattices

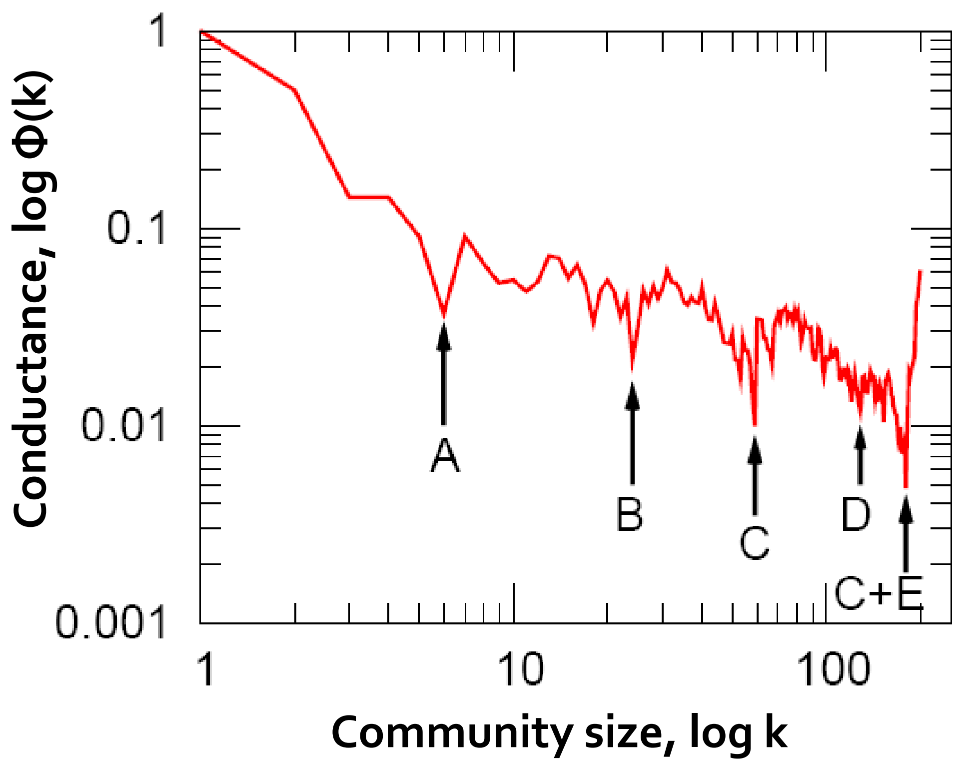


California road network

NCP Plot: Network Science

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

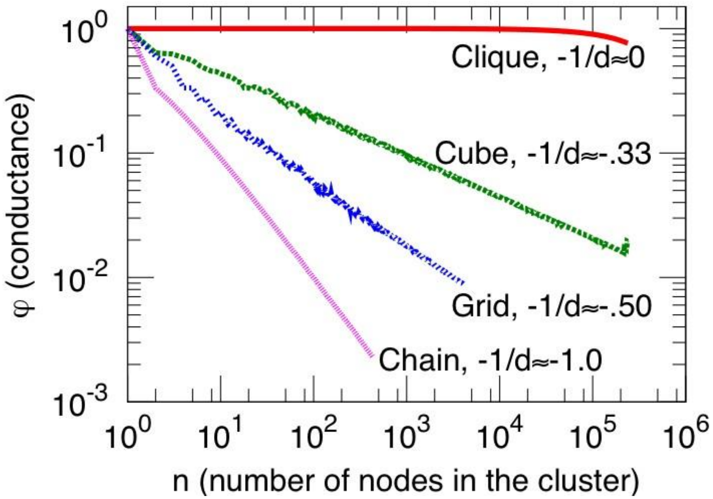




Natural Hypothesis

Natural hypothesis about NCP:

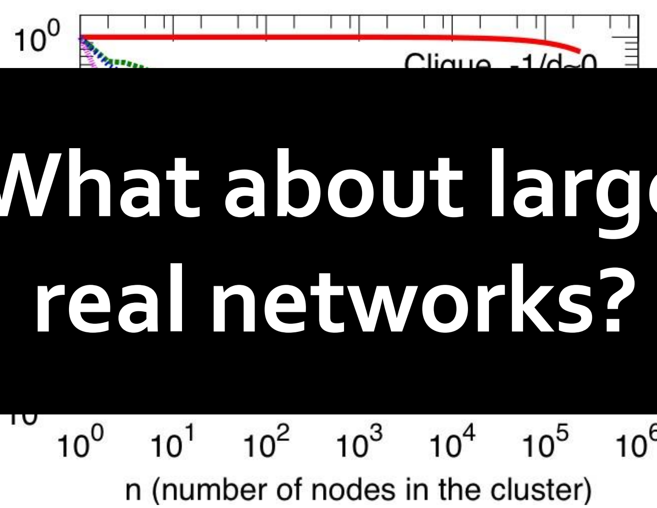
- NCP of real networks slopes **downward**
- Slope of the NCP corresponds to the "**dimensionality**" of the network



Natural Hypothesis

Natural hypothesis about NCP:

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



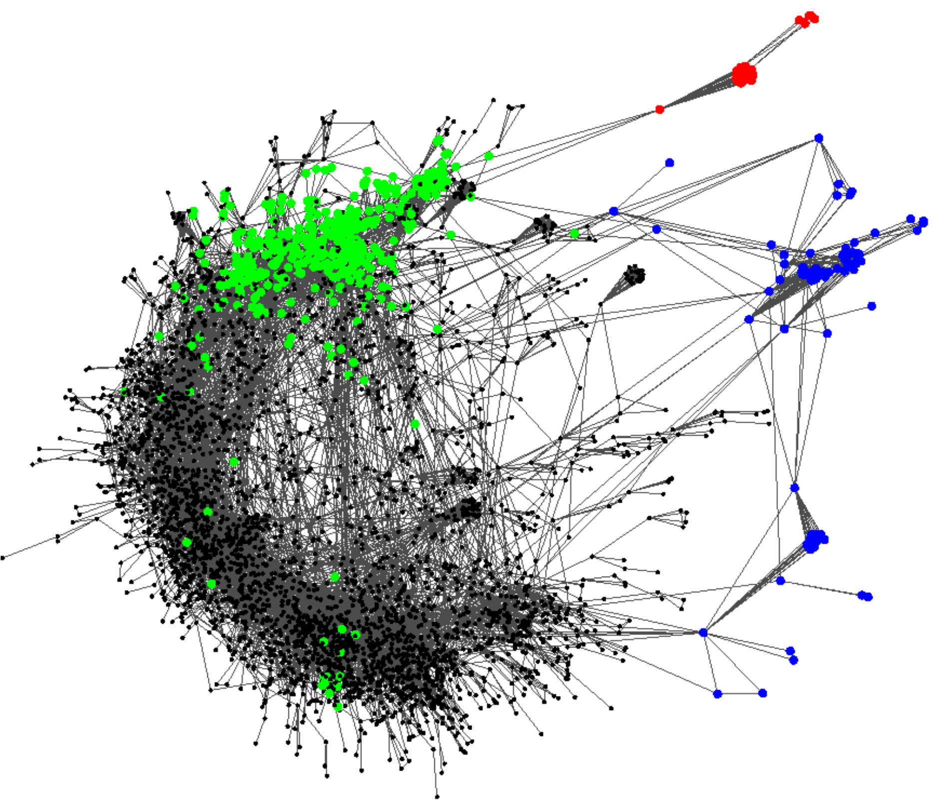
What about large real networks?

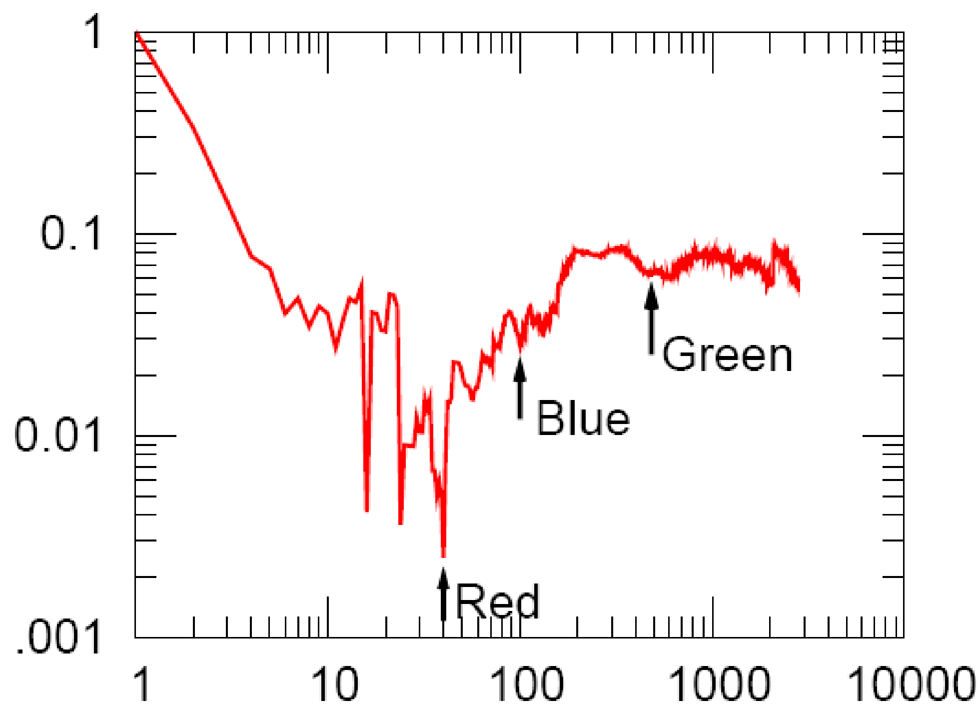
• Social nets	Nodes	Edges	Description
LiveJournal	4,843,953	42,845,684	Blog friendships [5]
Epinions	75,877	405,739	Trust network [28]
CA-DBLP	317,080	1,049,866	Co-authorship [5]
• Information (citation) networks			
Cit-hep-th	27,400	352,021	Arxiv hep-th [14]
AmazonProd	524,371	1,491,793	Amazon products [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		

We examined more than **200 large networks**

Large Networks: Very Different!

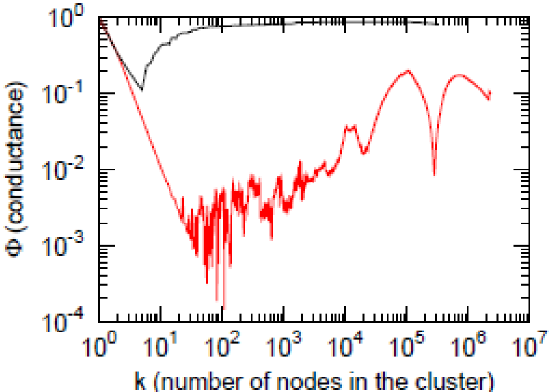
Typical example: General-Relativity Collaborations (n=4,158, m=13,422)



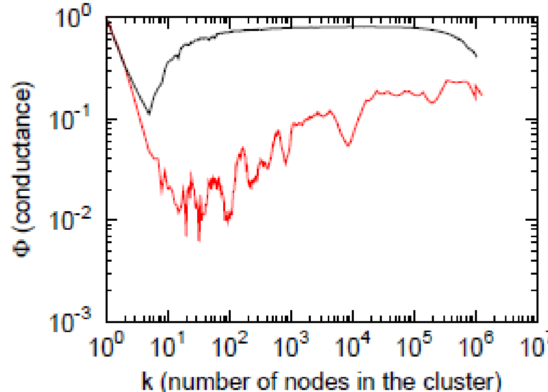


[Leskovec et al. WWW '08]

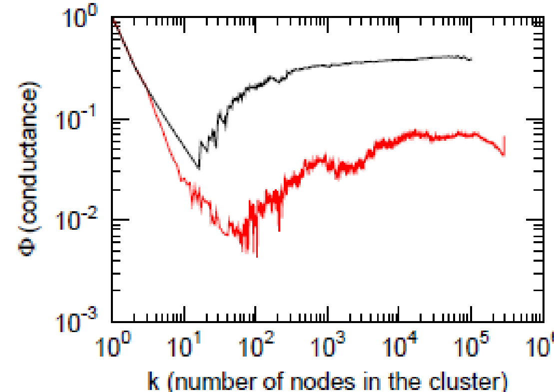
More NCP Plots



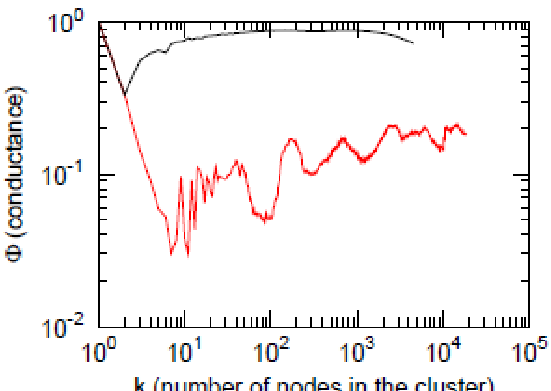
(a) LiveJournal01



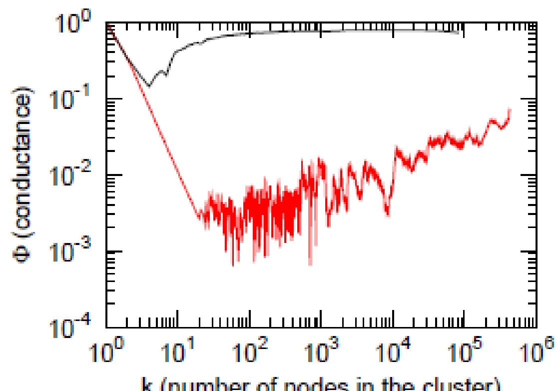
(b) Messenger-DE



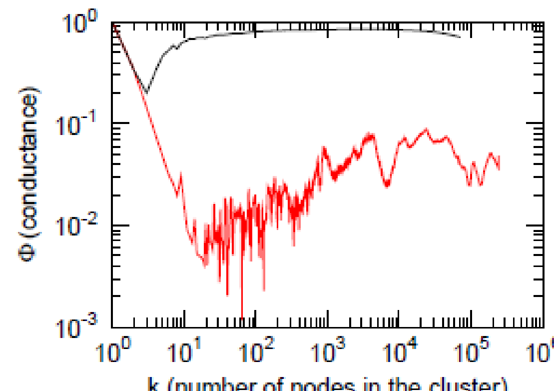
(c) AtP-DBLP



(d) Cit-hep-th

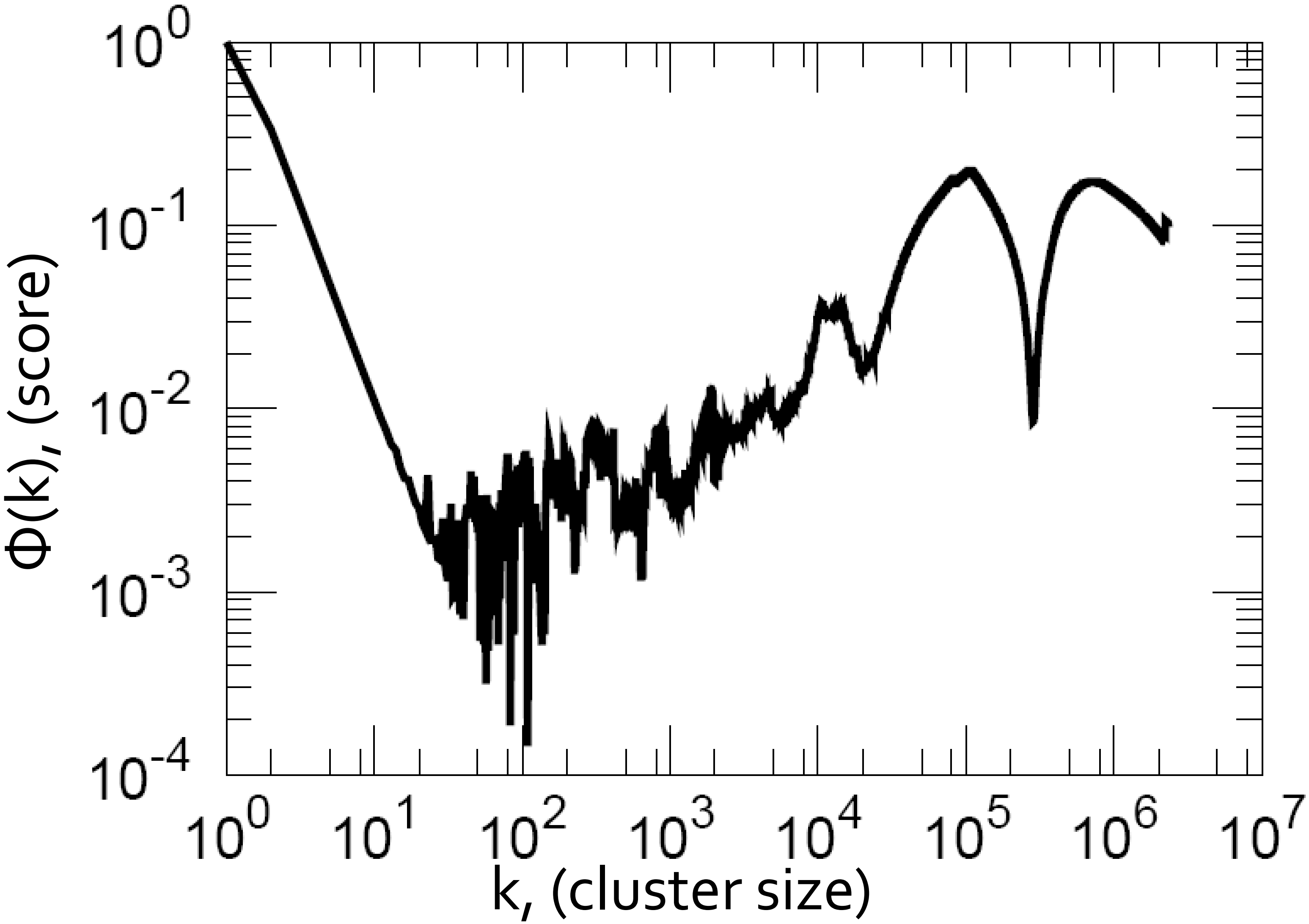


(e) Web-Google

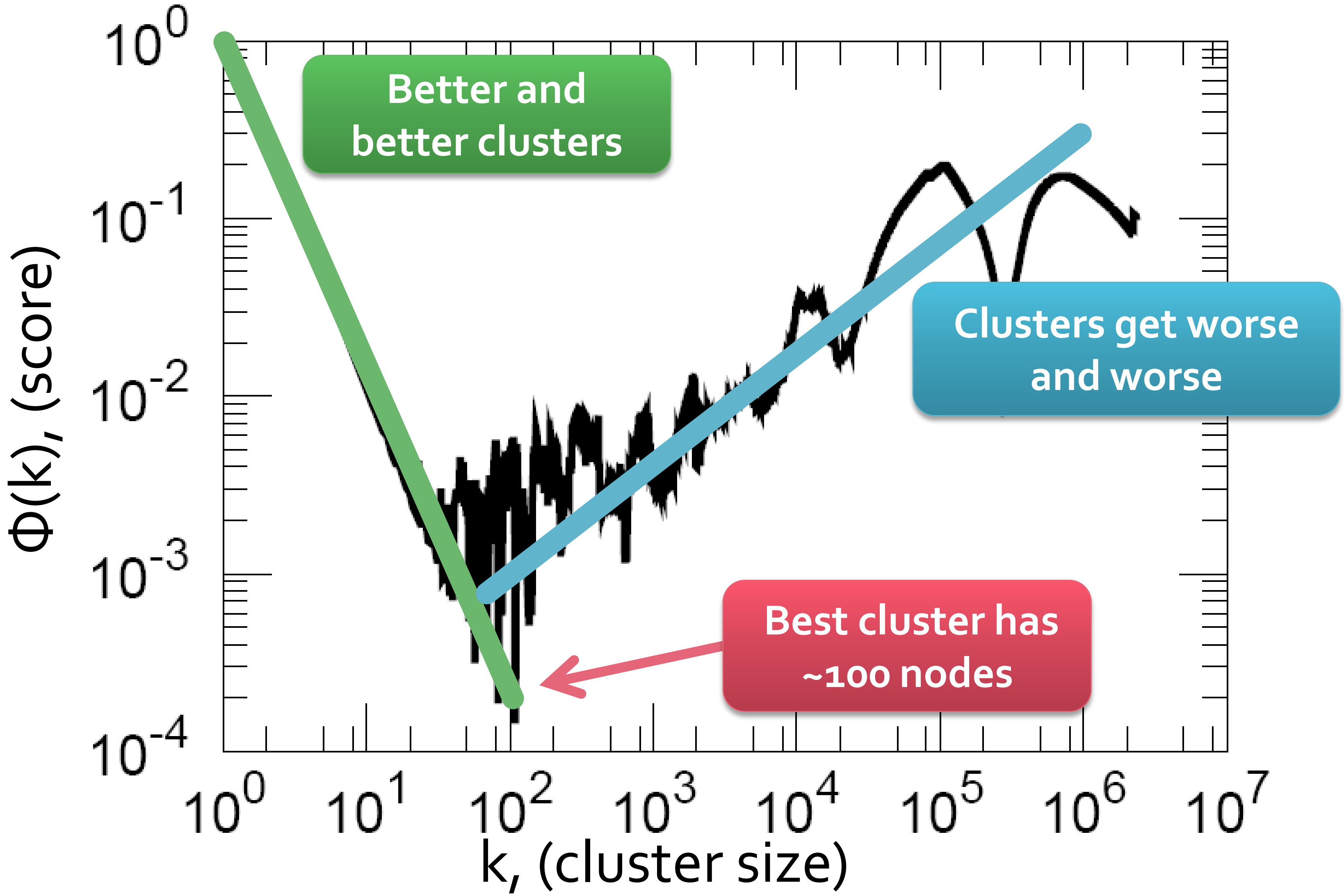


(f) AmazonAll

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

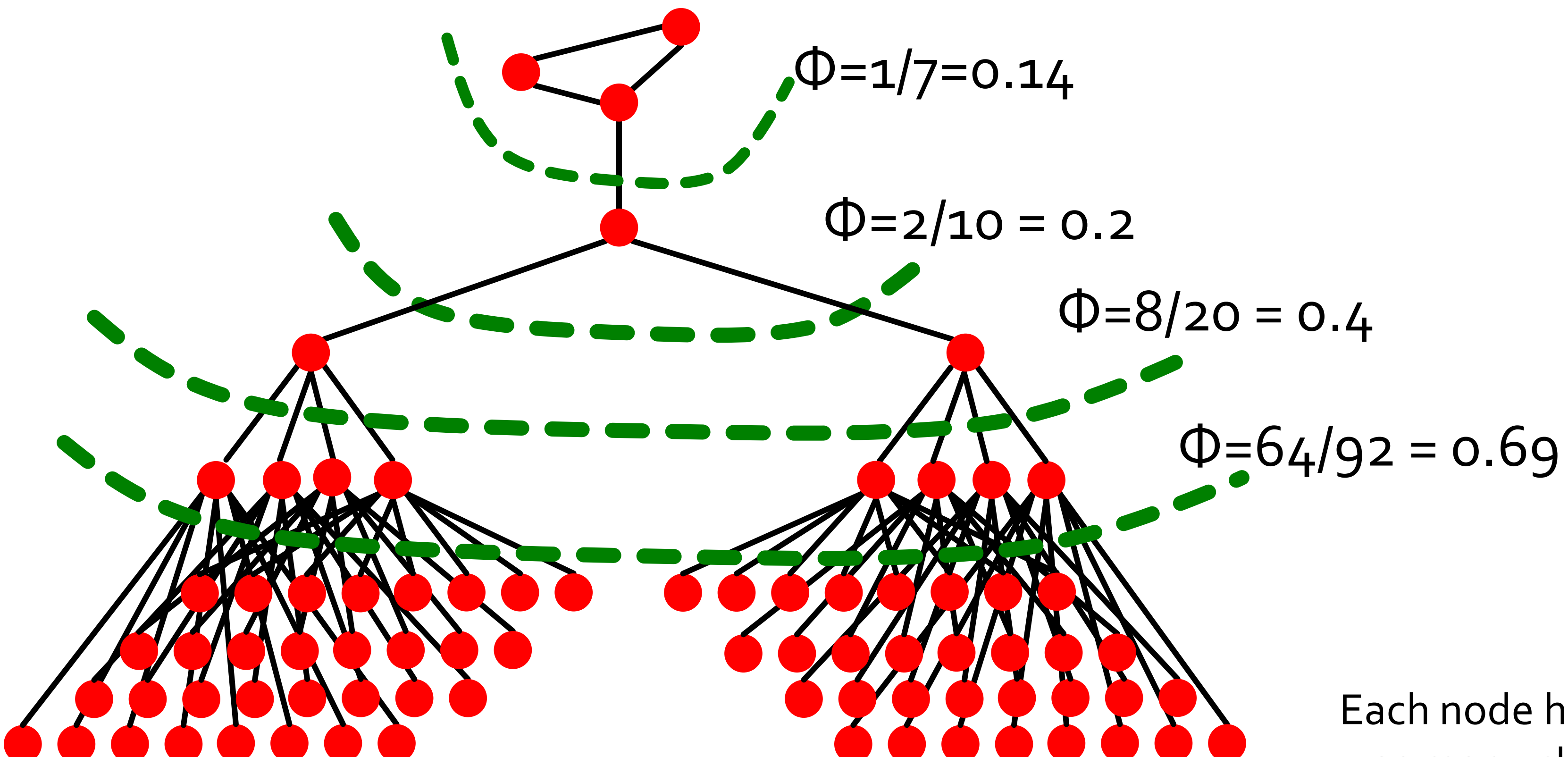


NCP: LiveJournal (n=5m, m=42m)



Explanation: The Upward Part

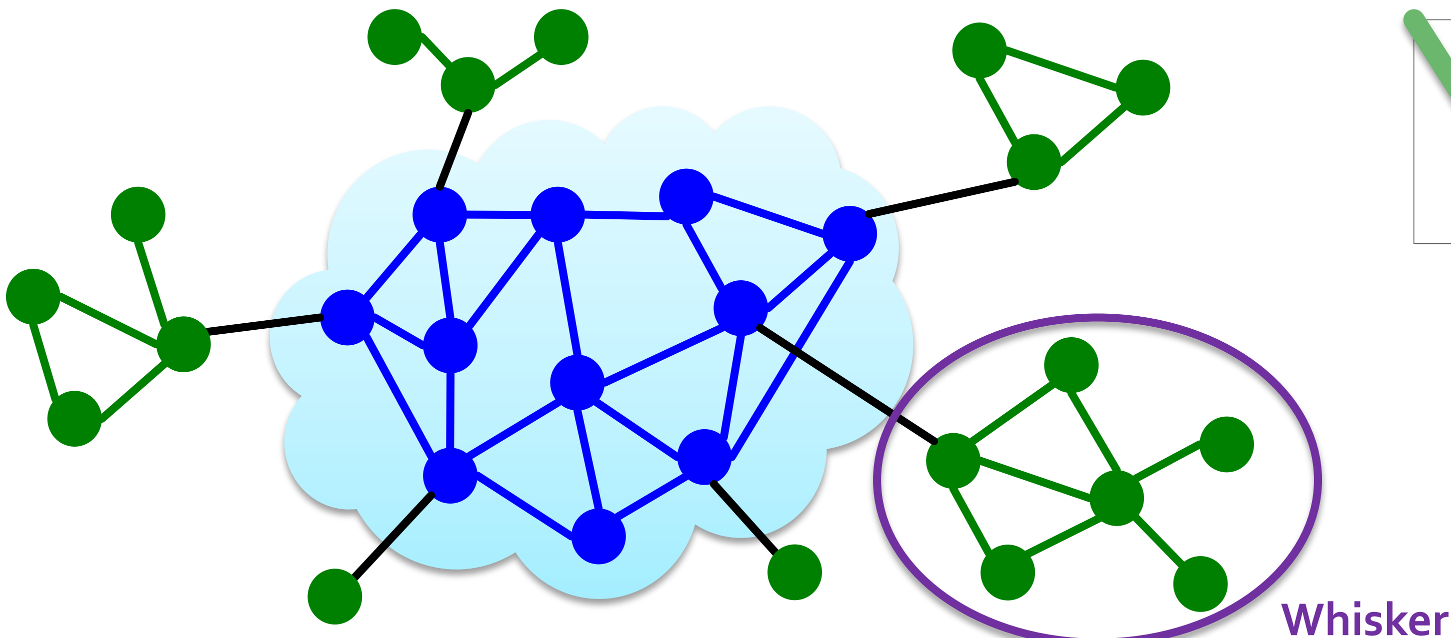
- As clusters grow the number of edges inside grows slower that the number crossing

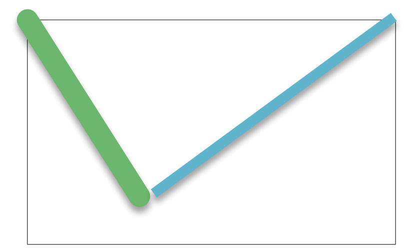


Each node has twice as many children

Explanation: Downward part

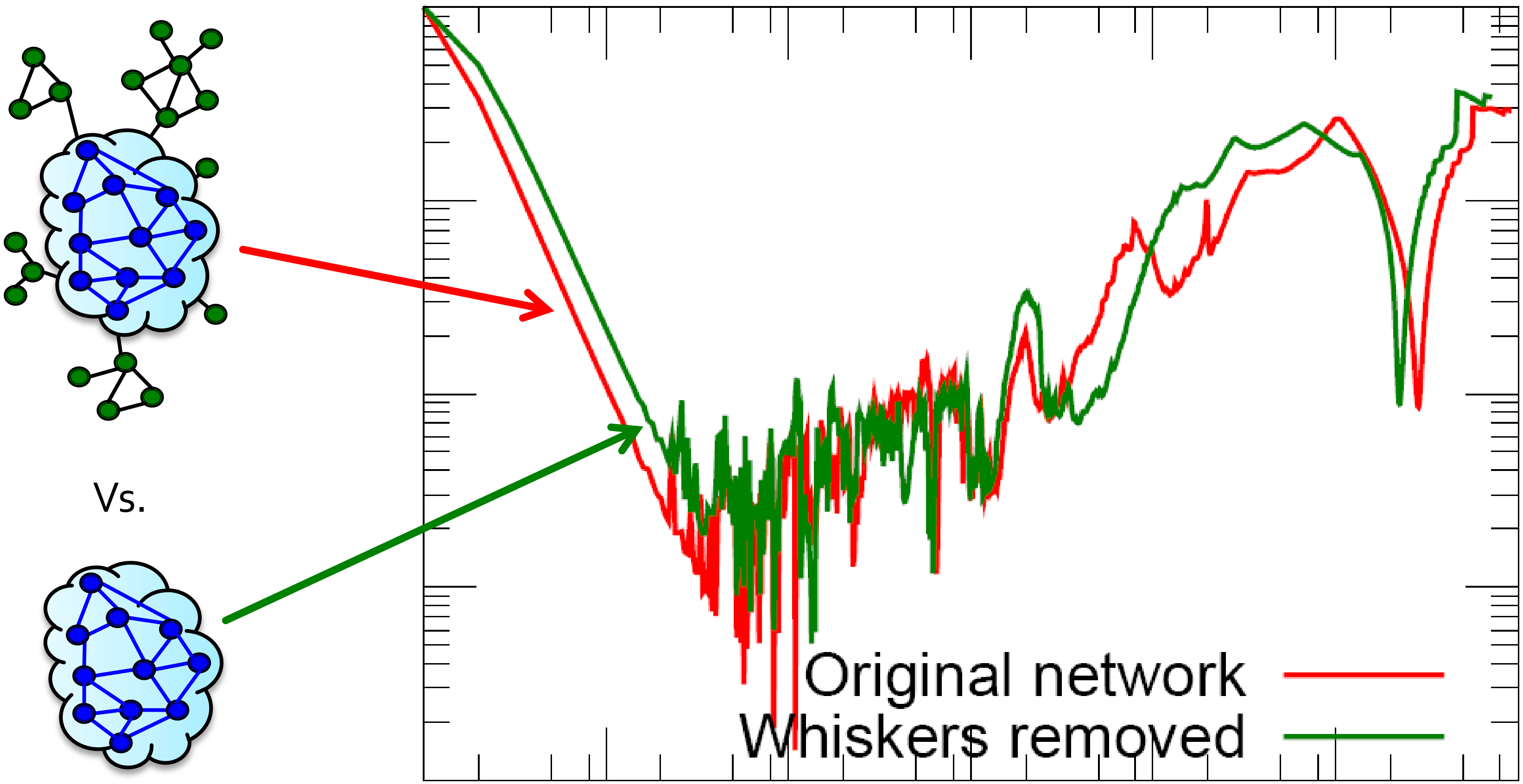
- Empirically we note that best clusters are barely connected to the network





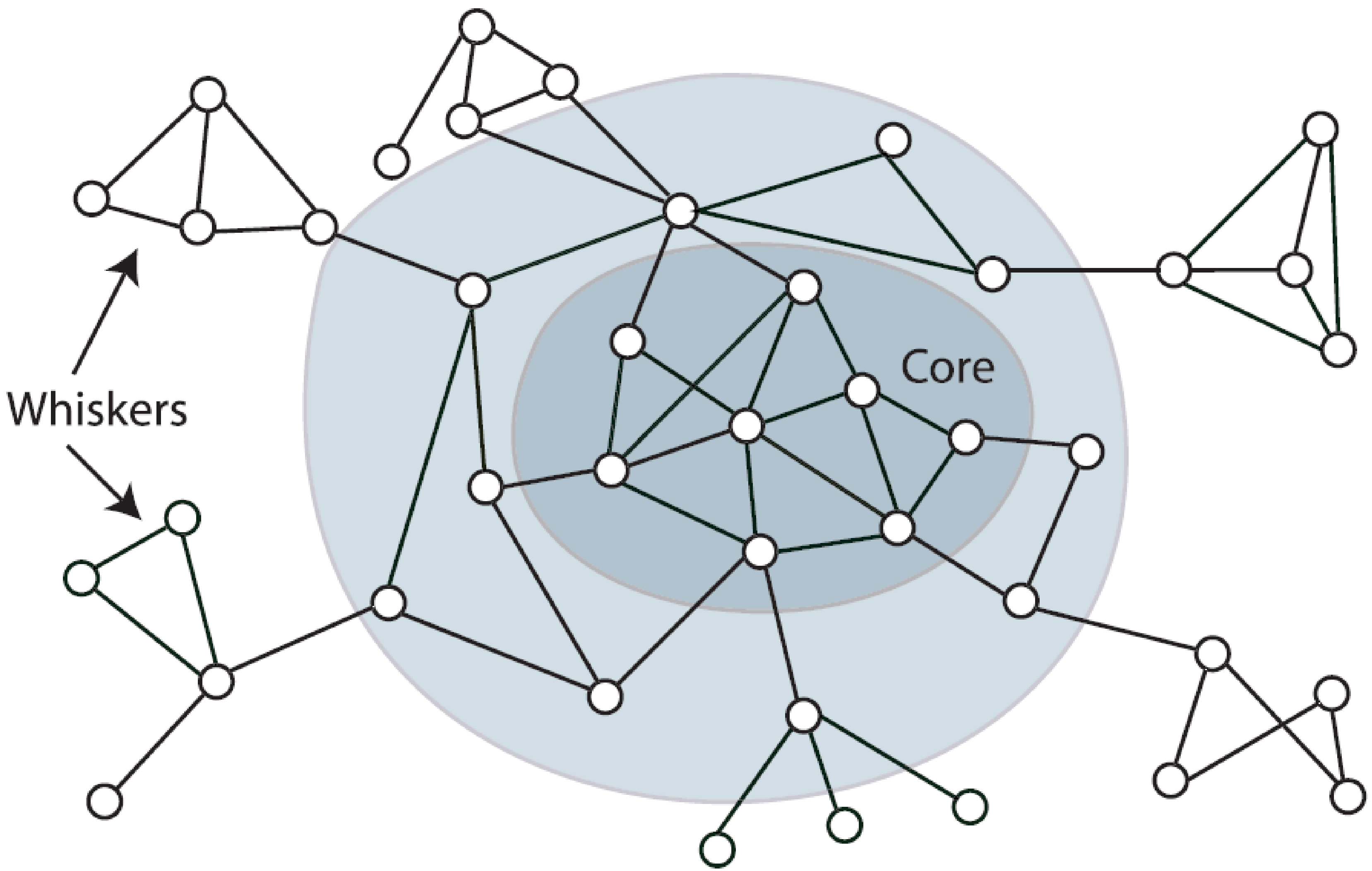
⇒ Core-periphery structure

What If We Remove Good Clusters?

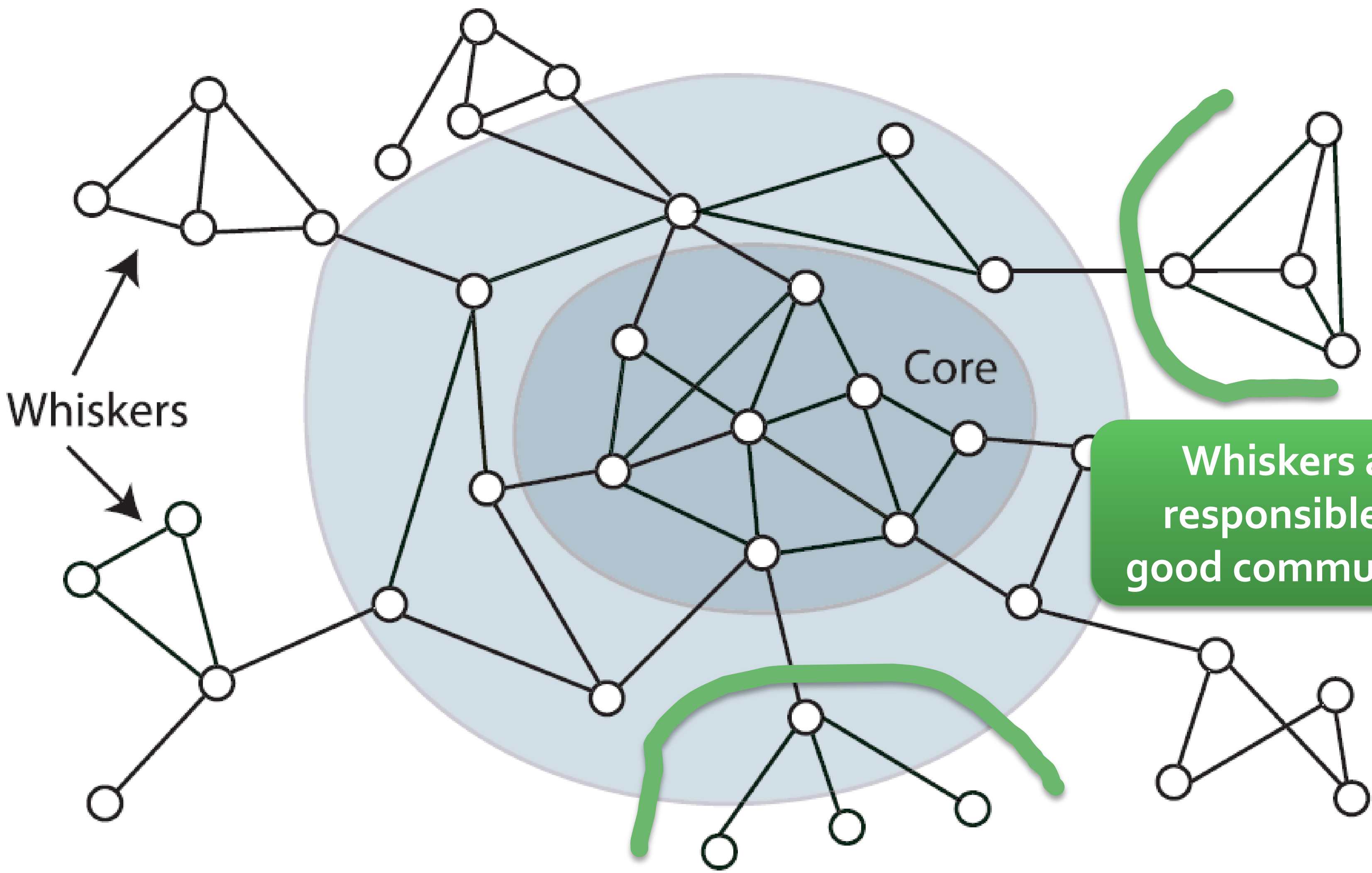


Nothing happens! $\Rightarrow$ Nestedness of the core-periphery structure

Suggested Network Structure

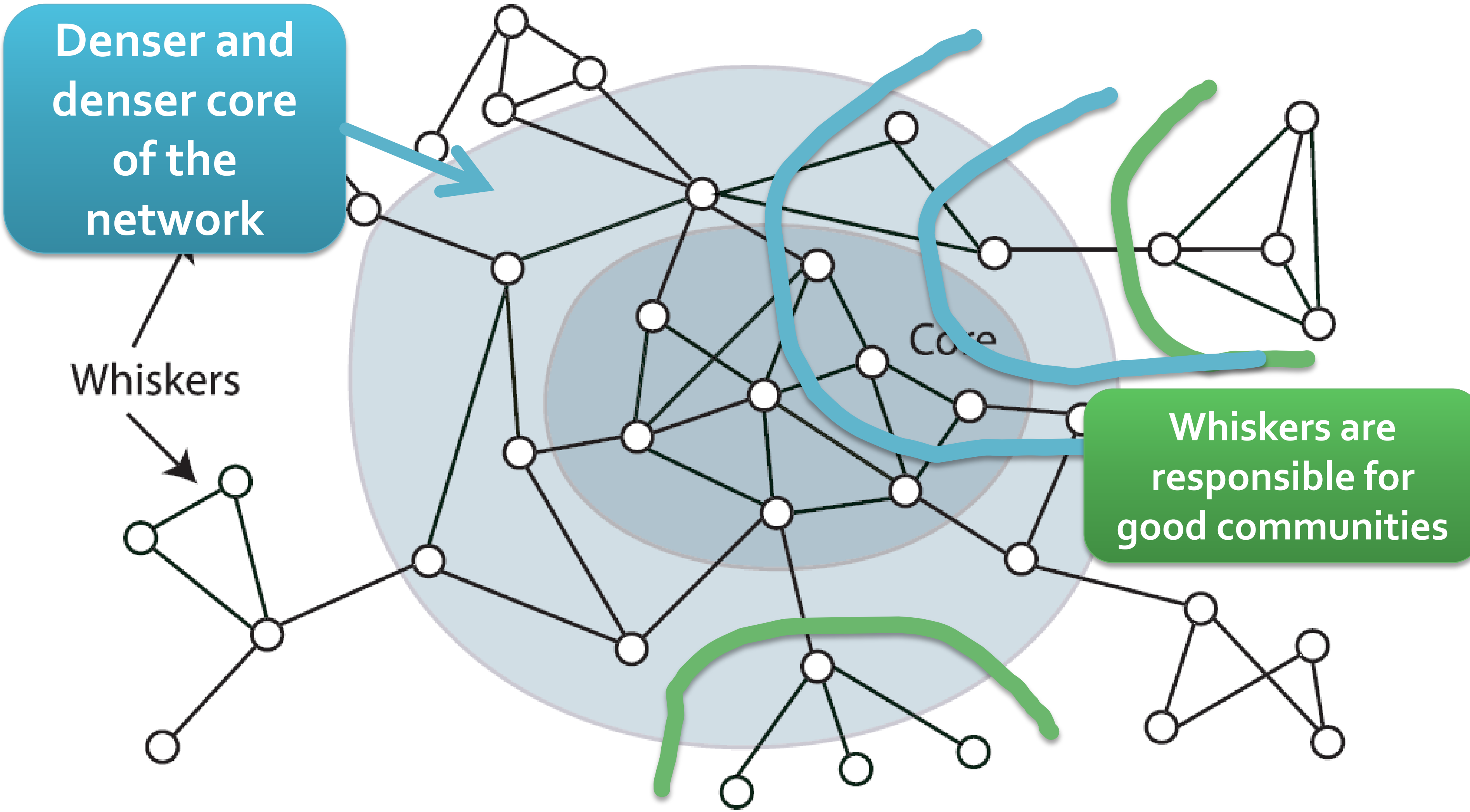


Suggested Network Structure

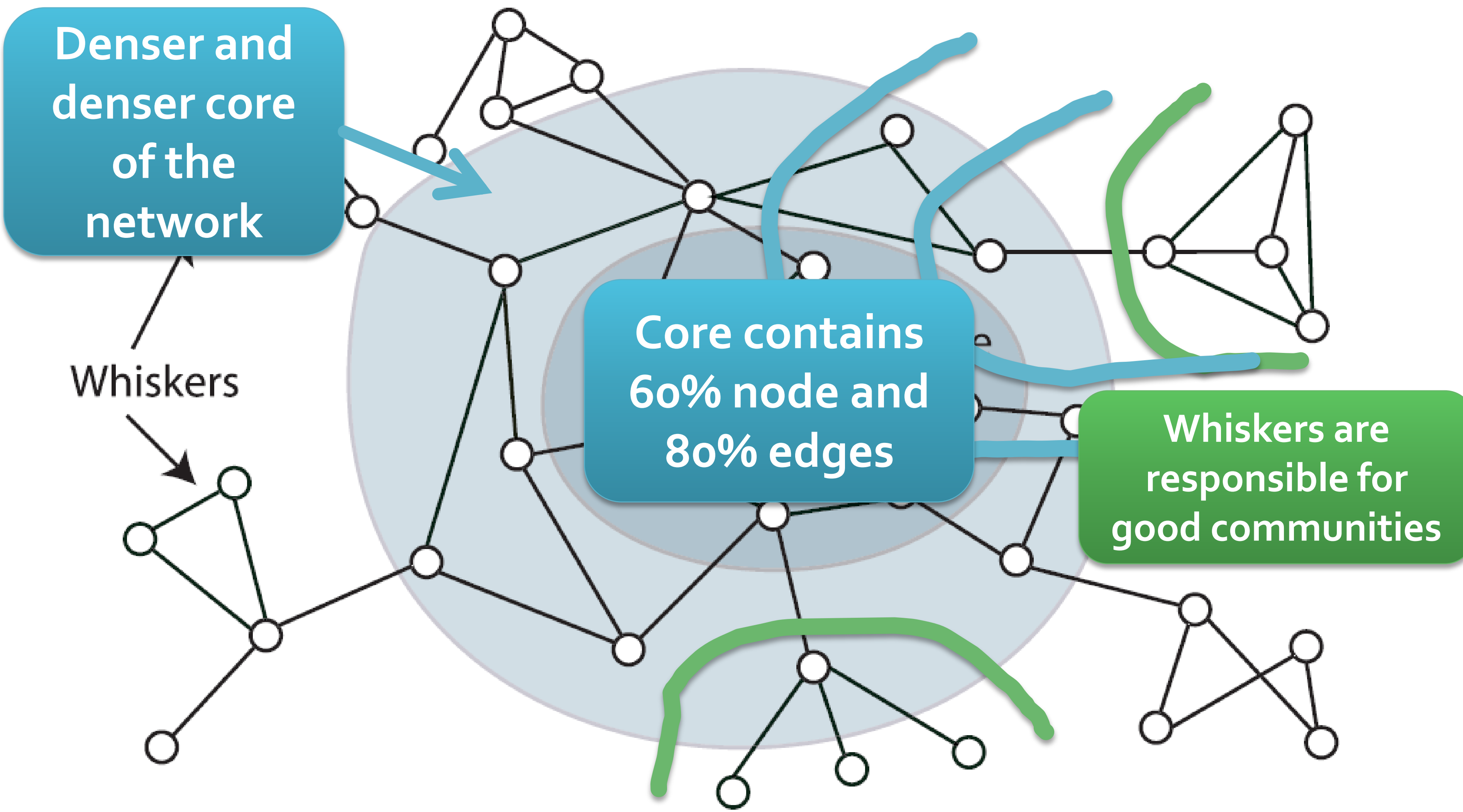


Whiskers are responsible for good communities

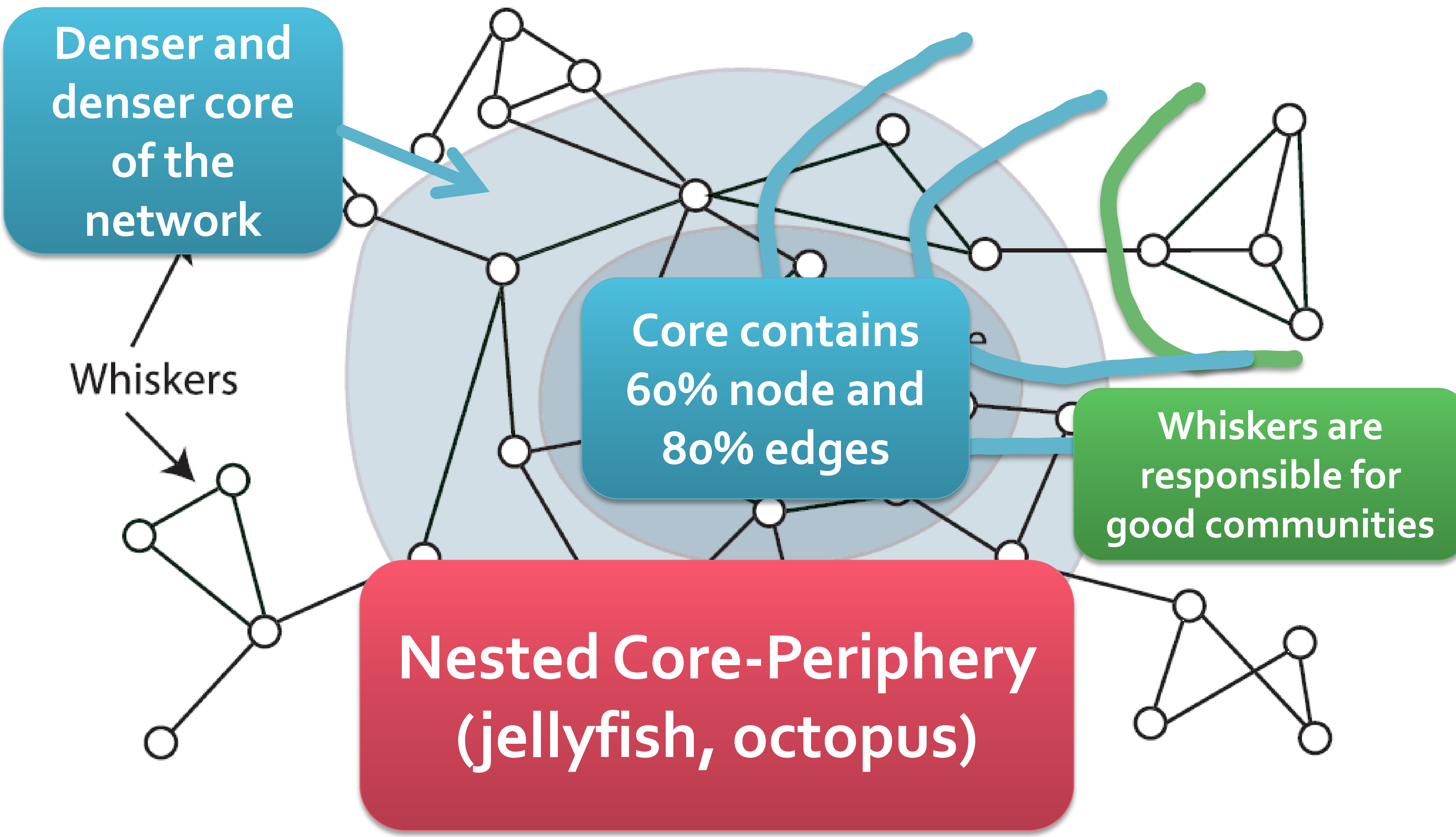
Suggested Network Structure



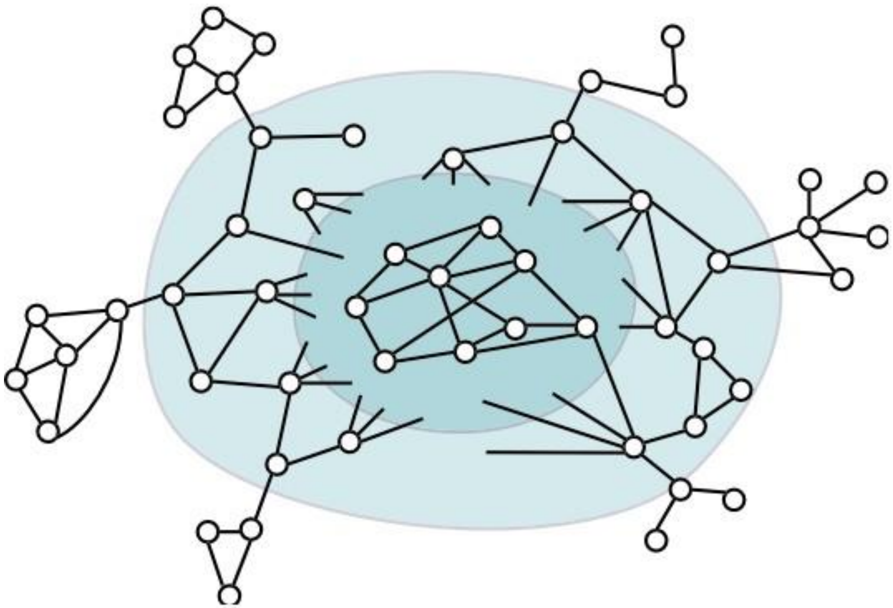
Suggested Network Structure



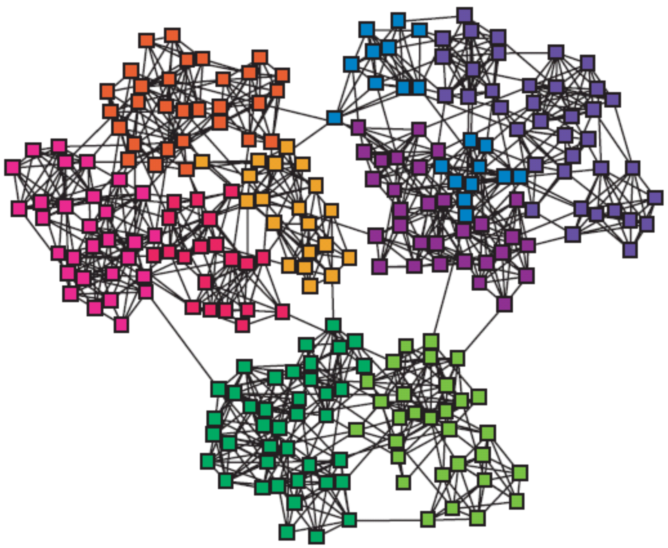
Suggested Network Structure



Part 2: Networks & Communities



vs.

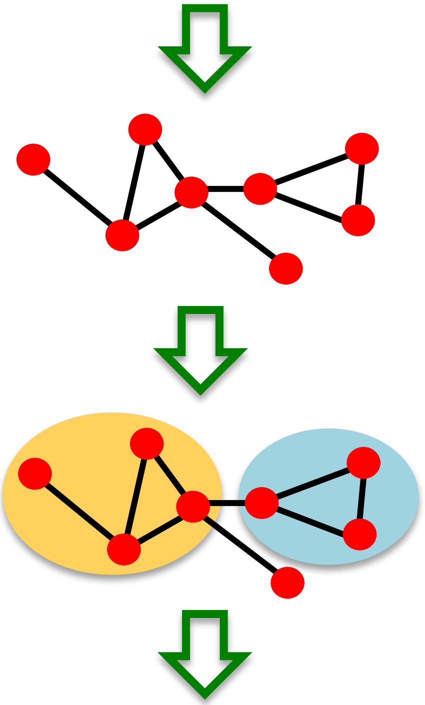


How do we reconcile these two views?

Step Back: Community Detection

(1) Take a dataset

dblp.uni-trier.de
Computer Science Bibliography



(2) Represent it as a graph

(3) Identify communities

(really, clusters)

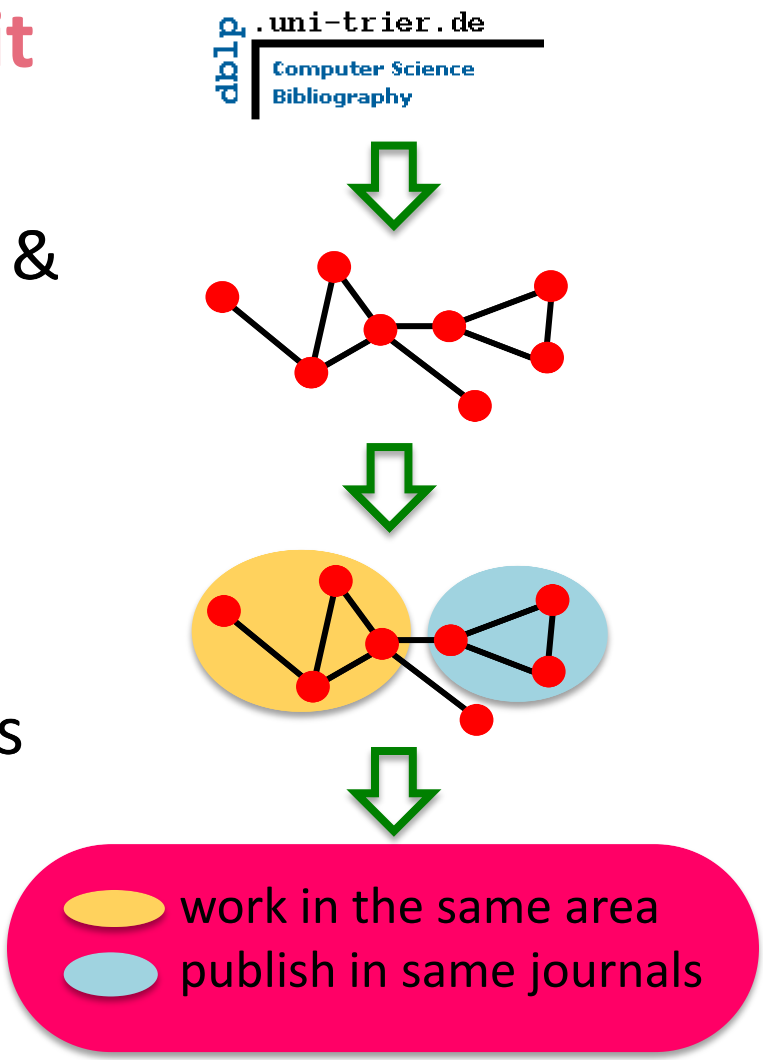
(4) Interpret clusters as "real" communities

work in the same area

publish in same journals

Ground-Truth

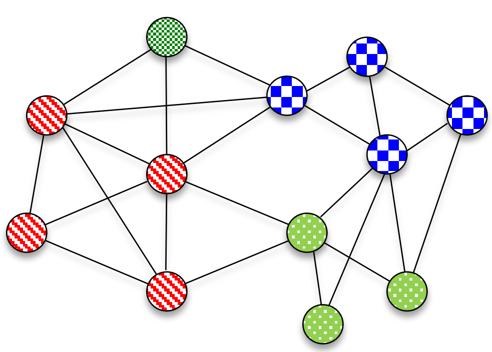
- Networks with a an explicit notion of Ground-Truth:
  - Collaborations: Conferences & Journals as proxies for areas
  - Social Networks: People join to groups, create lists
  - Information Networks: Users create topic based groups



Networks with Ground-Truth

Dataset	N	E	C	S	A
LiveJournal	4,036,538	34,916,684	311,782	40.06	3.09
Friendster	117,751,379	2,586,147,869	1,449,666	26.72	0.33
Orkut	3,072,441	117,185,083	8,455,253	34.86	95.93
Flickr	1,727,127	15,555,041	103,631	82.46	4.95
Youtube	1,138,873	2,990,443	30,087	9.75	0.26
DBLP	425,957	1,348,244	2,547	429.79	2.57
Amazon	334,863	925,872	49,732	99.86	14.83

- N ... # of nodes
- E ... # of edges
- C ... # of ground-truth communities
- S ... average community size
- A ... memberships per node



Youtube social network

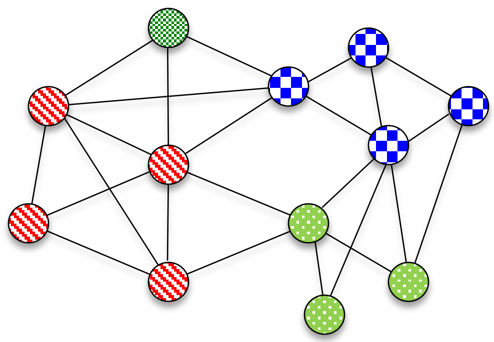
For example:

... fans of Real Madrid

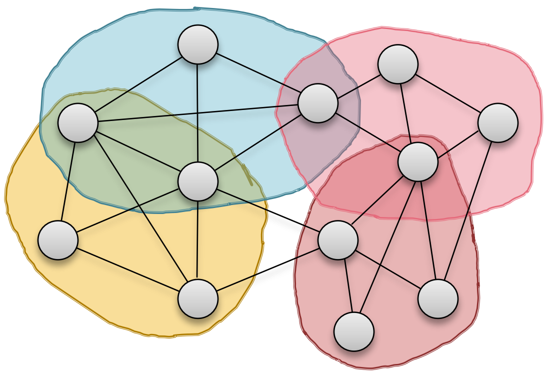
... subscribe to Lady Gaga videos

... follow Volvo Ocean Race

Ground-Truth: Consequences



Ground-truth groups

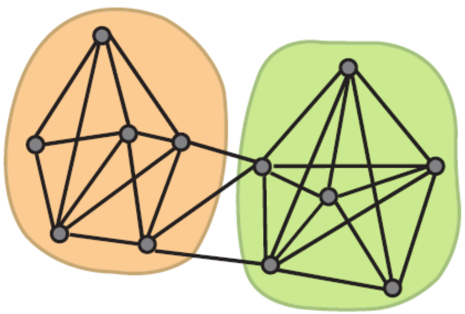


Inferred communities

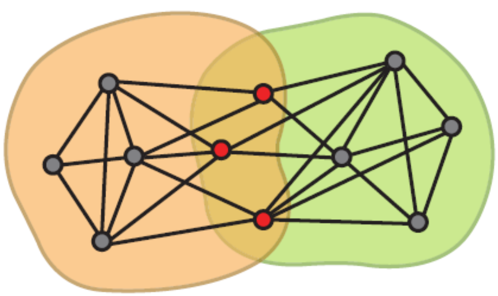
- **How groups map on the network?**
  - ⇒ **Insights for Better Algorithms**
- **How to evaluate and interpret?**
  - ⇒ **“Accuracy” of Algorithms**

Groups and Networks

- Nodes *u* and *v* share *k* groups
- **What is edge prob. P(edge | *k*) as a func. of *k*?**
- **Today's wisdom:**



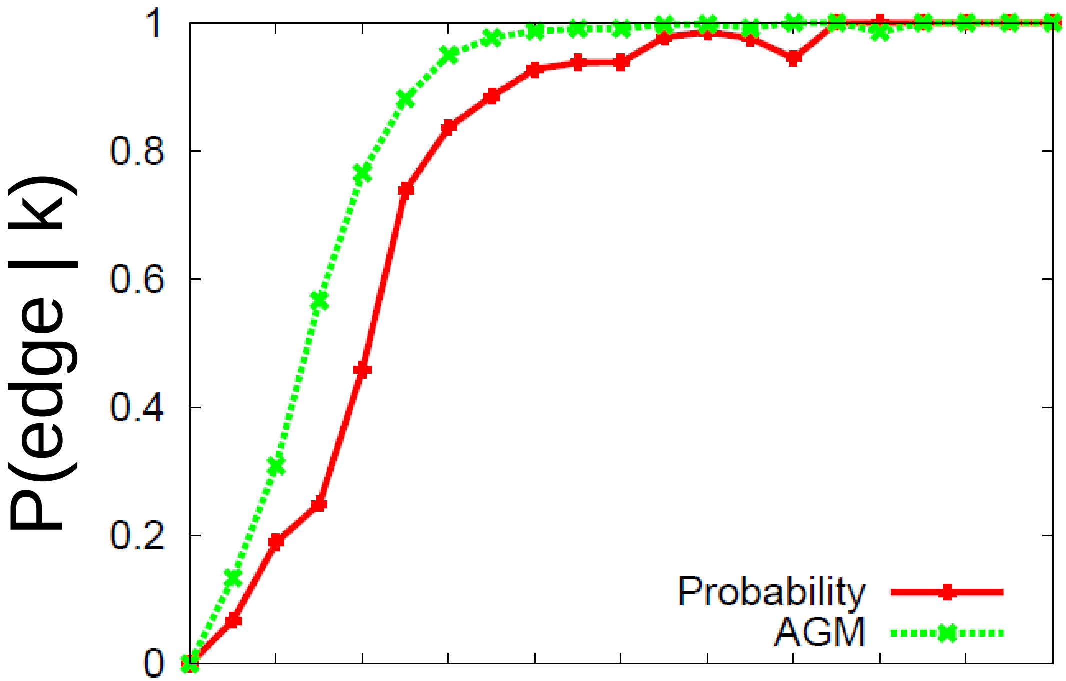
P(edge | k) = N/A



P(edge | k) = decreasing

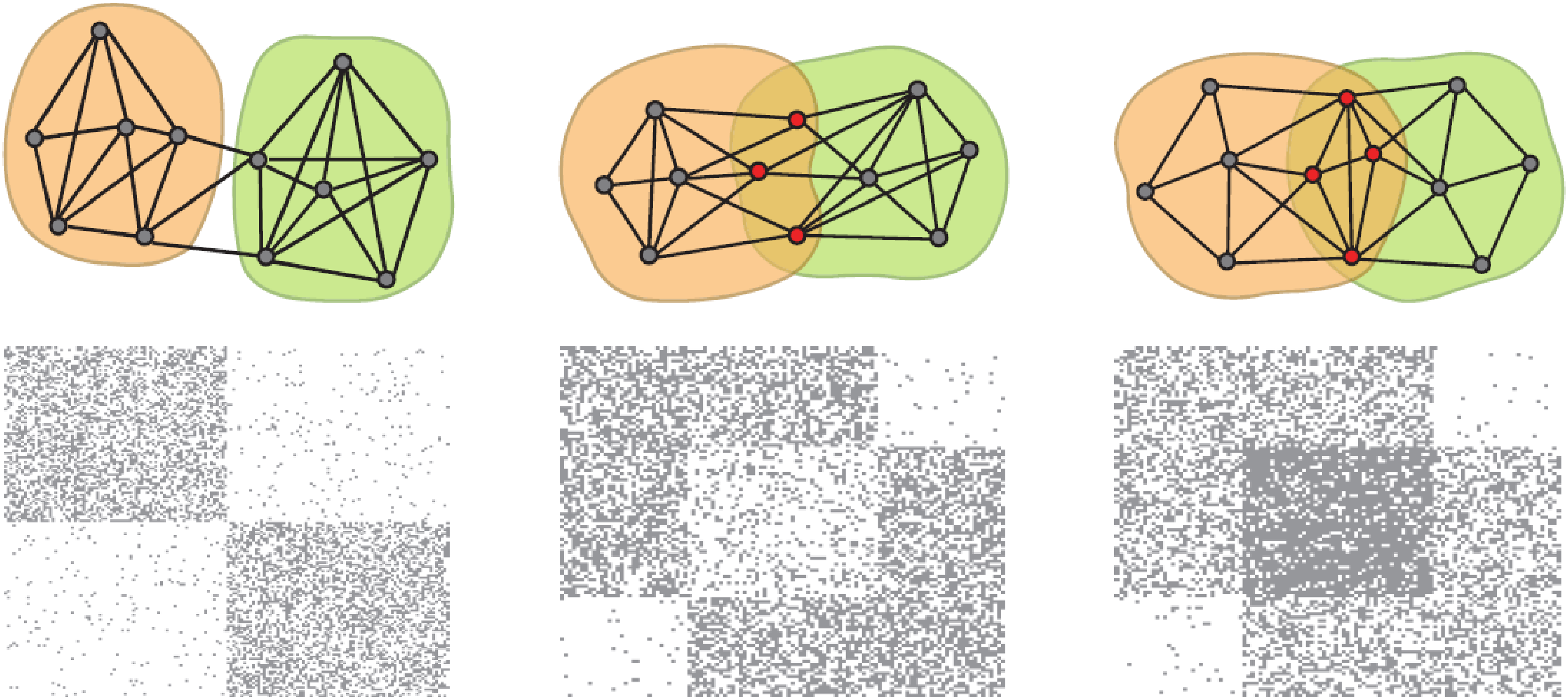
Edge Probability

- Nodes *u* and *v* share *k* groups
- **What is edge prob. P(edge | *k*) as a func. of *k*?**



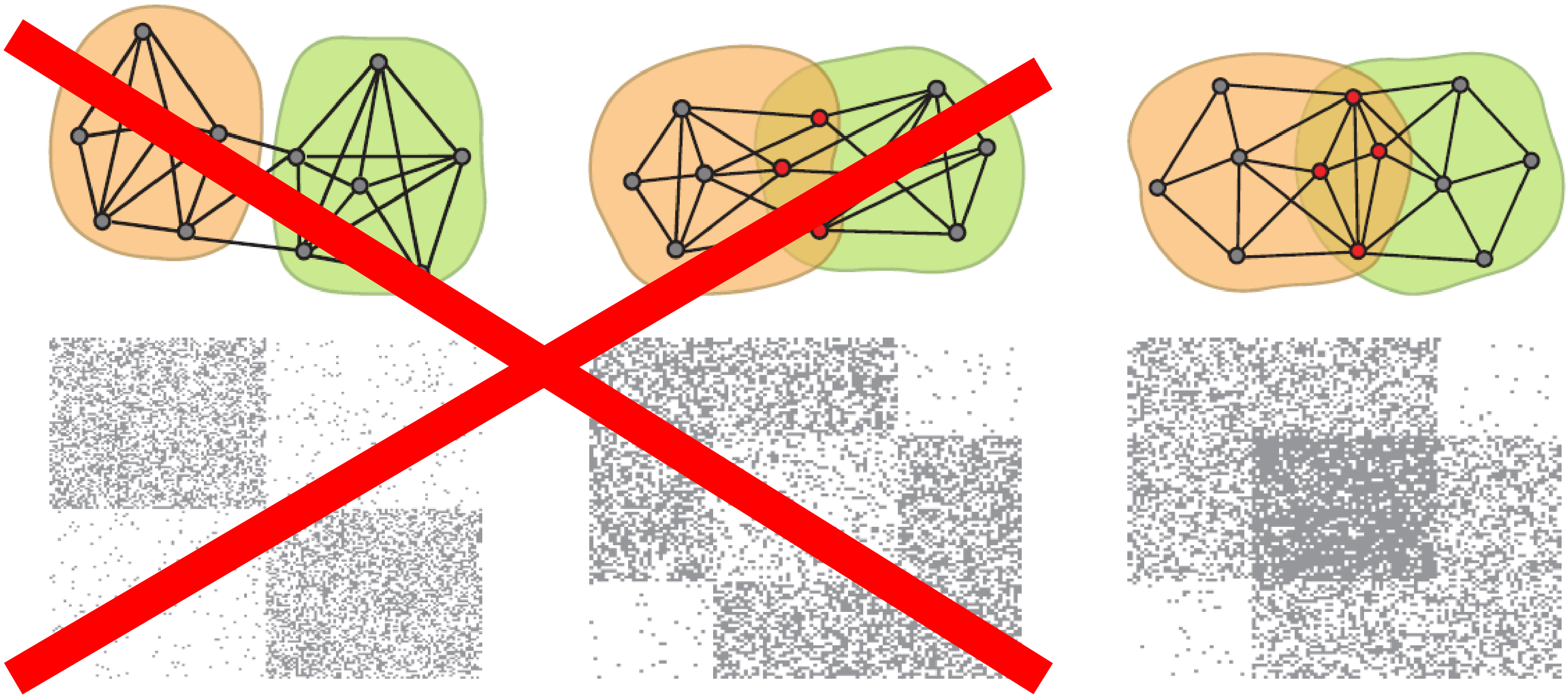
Overlaps are DENSER!

Communities in Networks



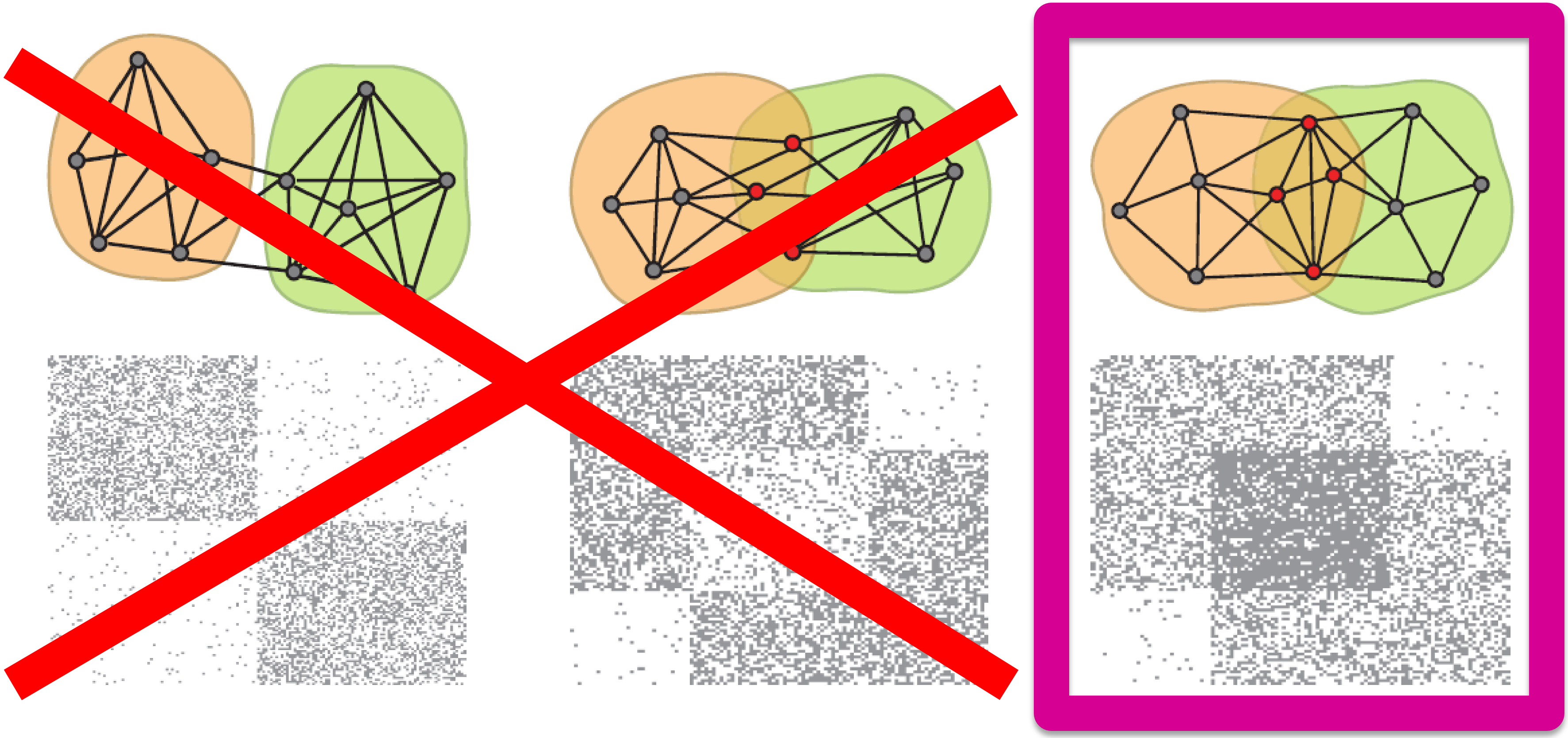
Overlaps are DENSER!

Communities in Networks



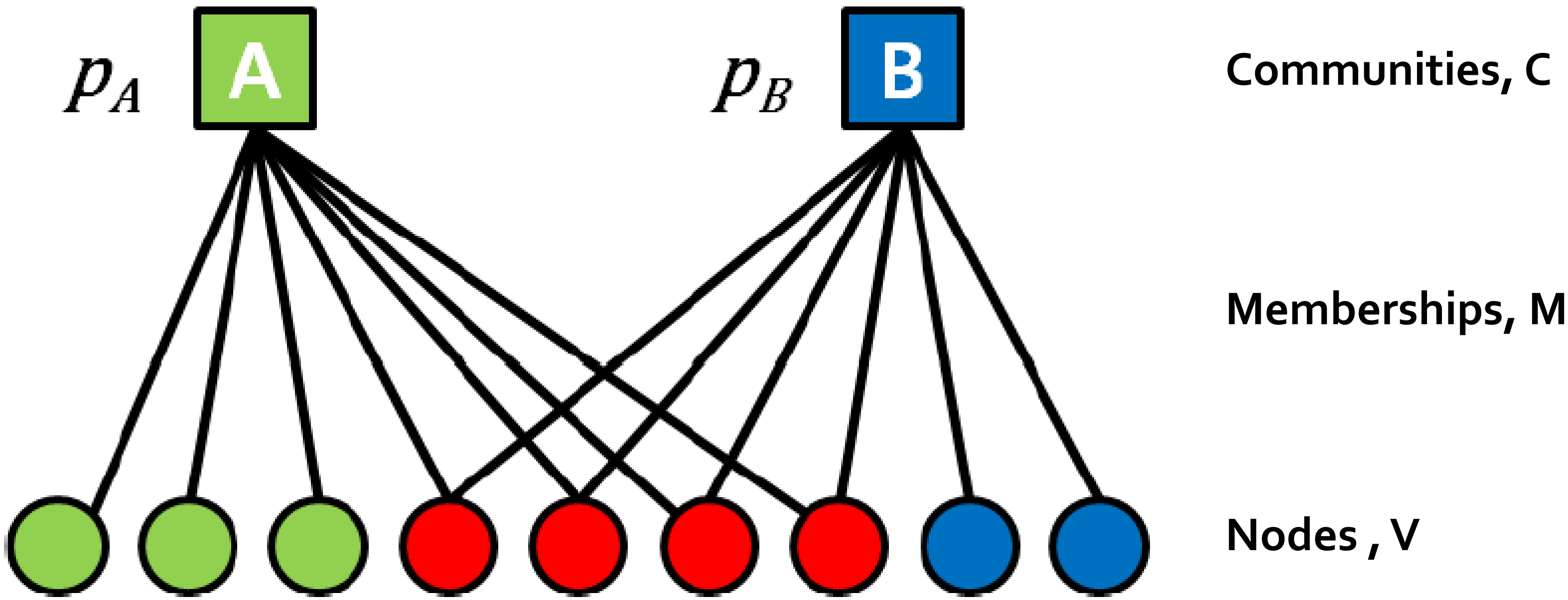
Overlaps are DENSER!

Communities in Networks



Overlaps are DENSER!

Natural Model



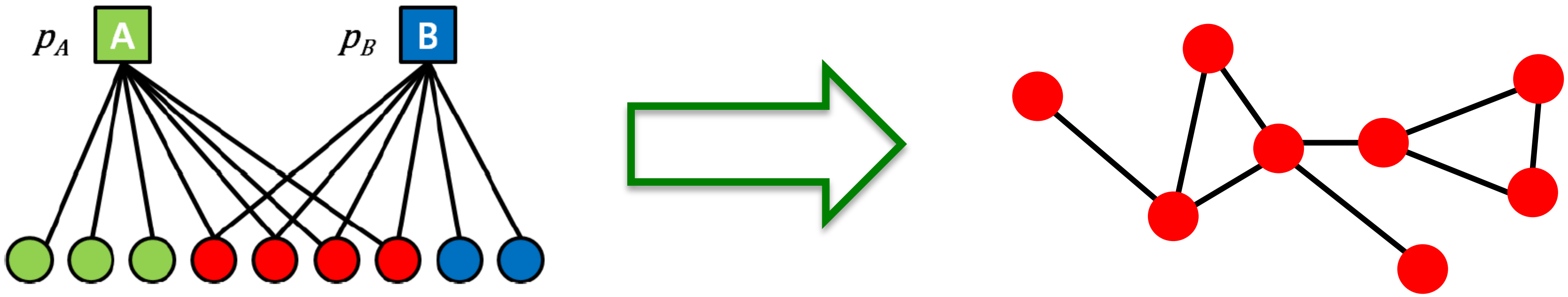
Community-Affiliation Graph Model

B(V,C,M,{p_c})

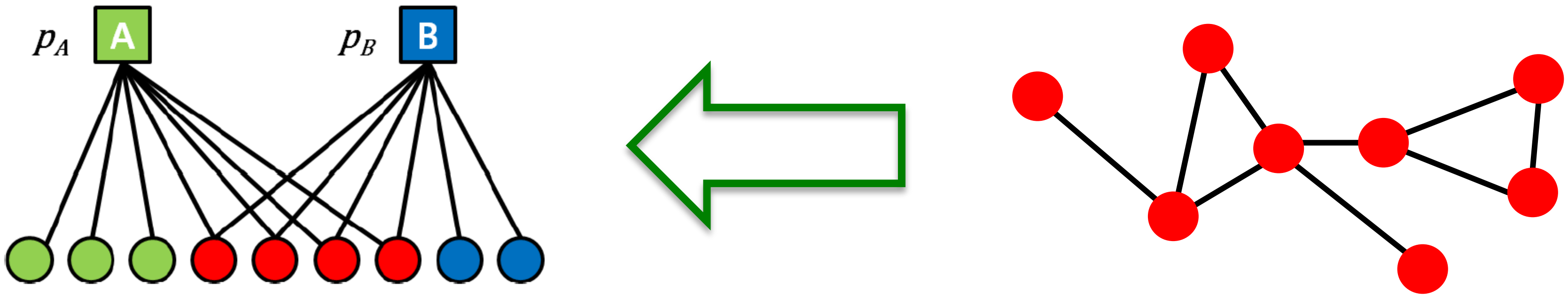
$$P(i,j) = 1 - \prod_{c \in M_i \cap M_j} (1 - p_c)$$

Provably generates power-law degree distributions and other patterns real-world networks exhibit [Lattanzi, Sivakumar, STOC '09]

Model-based Community Detection



Model-based Community Detection



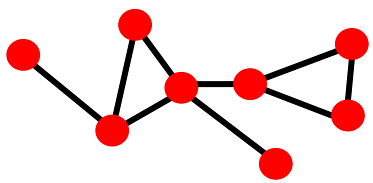
Given a Graph, find the Model

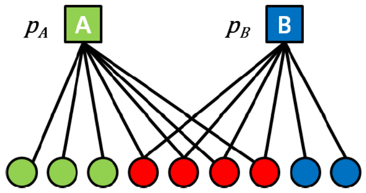
1) Affiliation graph B
2) Number of communities
3) Parameter p_i

Yes, we can!

MAG Model Fitting

- **Task:**





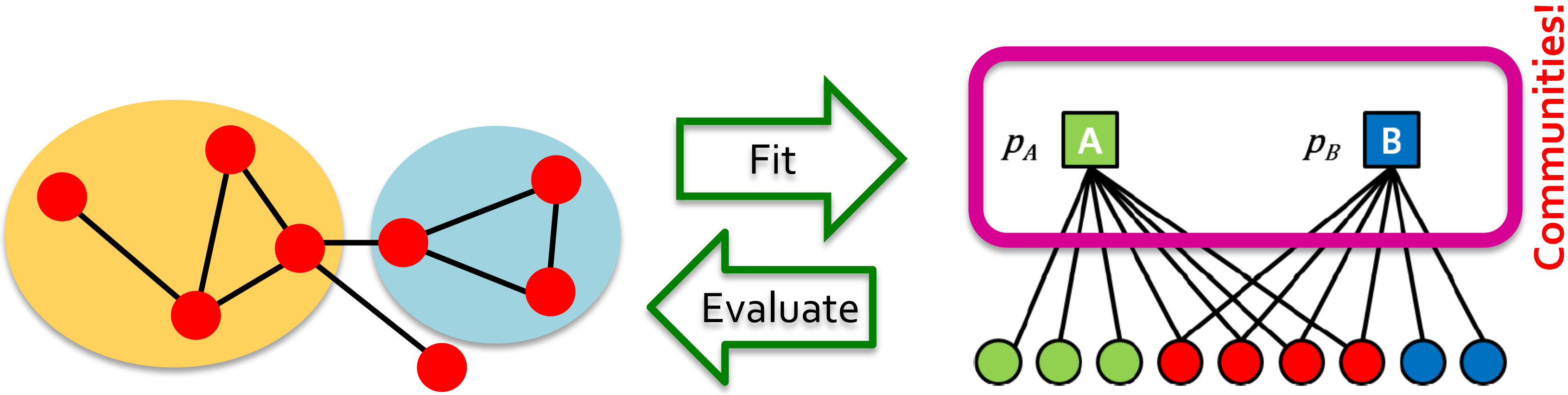
  - Given network *G(V,E)*. Fit *B(V, C, M, {p_c})*
- **Optimization problem (MLE):**

$$\arg\max_B P(G \mid B) = \prod_{(i,j)\in E} P(i,j) \prod_{(i,j)\notin E} (1 - P(i,j))$$

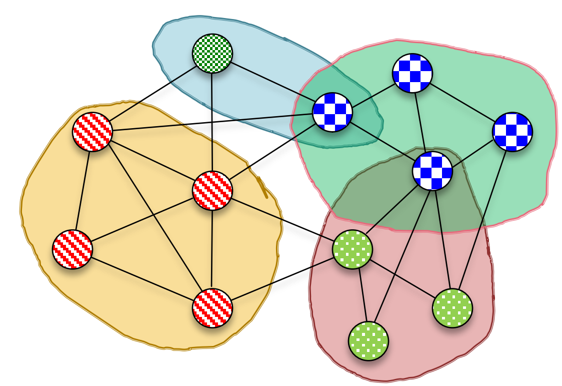
$$P(i,j) = 1 - \prod_{c \in M_i \cap M_j} (1 - p_c)$$

- **How to solve?**
  - Approach: **Coordinate ascent**
    - **(1)** Stochastic search over *B*, while keeping *{p_c}* fixed
    - **(2)** Optimize *{p_c}*, while keeping *B* fixed
  - **Works well in practice!**

Experimental Setup



- **Evaluation:** How well do inferred group memberships correspond to the ground-truth?
  - F-score: Precision, Recall
  - Mutual Information
  - Ω - index
- **Algorithms for comparison:**
  - Clique Percolation [Palla et al., Nature '05]
  - Link Clustering [Ahn et al., Nature '10]



Experiments: Vs. Link Clustering

AGM vs. Link Clustering

Score	LiveJ	Frster	Orkut	DBLP	Amzn	YouTube	Flickr	Improvement
F	0.75	0.78	0.84	0.36	0.54	0.63	0.91	**0.69**
Ω	0.24	0.11	0.20	0.28	1.48	0.02	0.19	**0.36**
MI	0.01	0.22	0.21	-0.42	-0.23	-0.39	-0.11	-0.10
$\|C^*-C\|/C$	80.96	107.27	95.00	40.04	6.32	54.13	129.3	**73.30**

Relative improvement of our method over Link Clustering

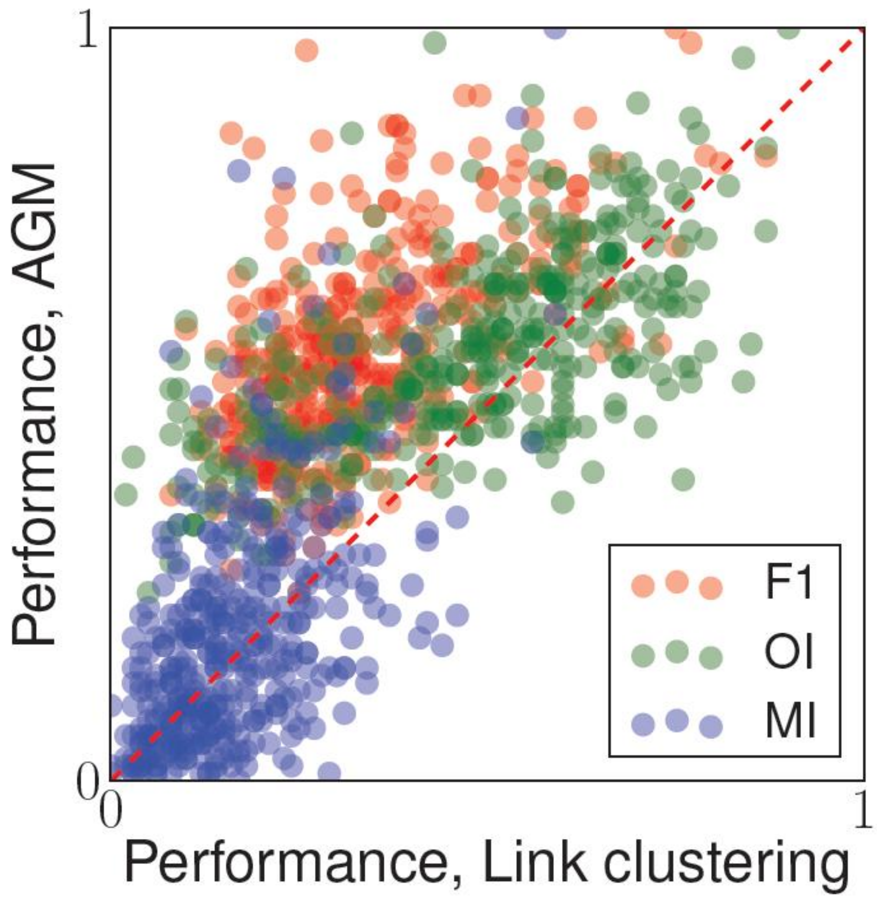
Experiments: Vs. CPM

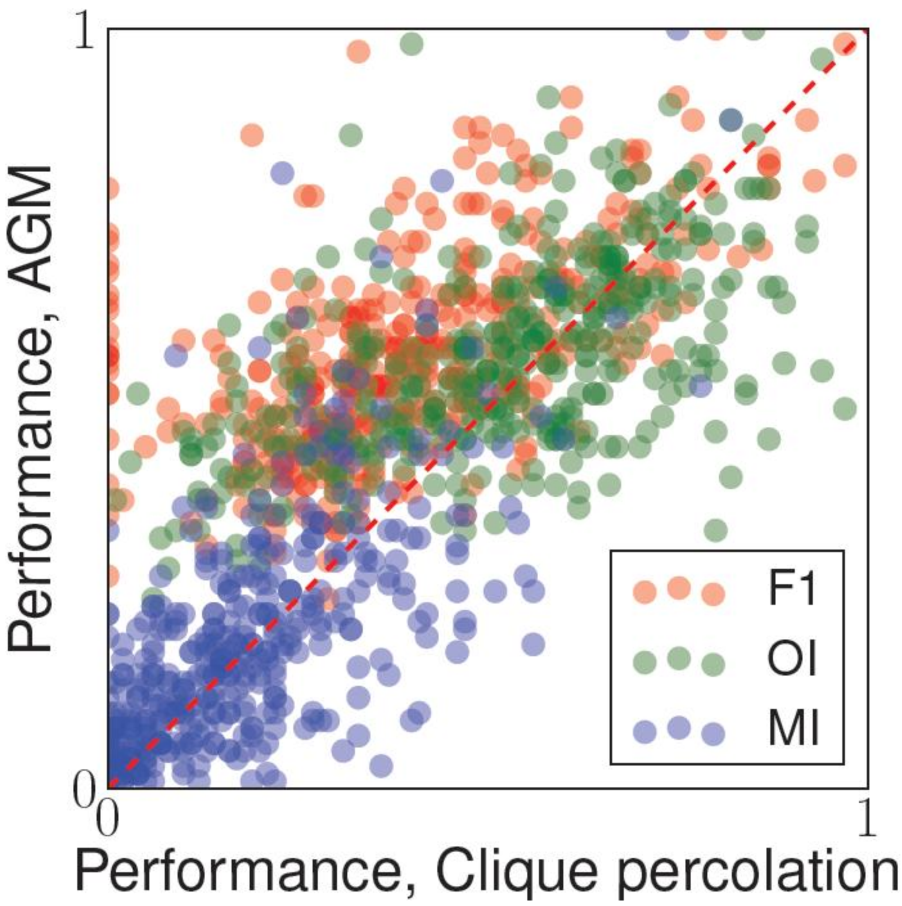
AGM vs. Clique Percolation

Score	LiveJ	Frster	Orkut	DBLP	Amzn	YouTube	Flickr	Improvement
F	0.44	0.44	0.44	0.42	1.76	1.28	0.32	**0.73**
Ω	0.21	0.08	0.1	0.46	6.7	0.17	0.16	**1.12**
MI	0.07	0.11	0.05	-0.4	0.78	0.29	-0.1	**0.12**
$\|C^*-C\|/C$	0.73	0.71	1.1	3.03	0.76	0.63	0.64	**1.09**

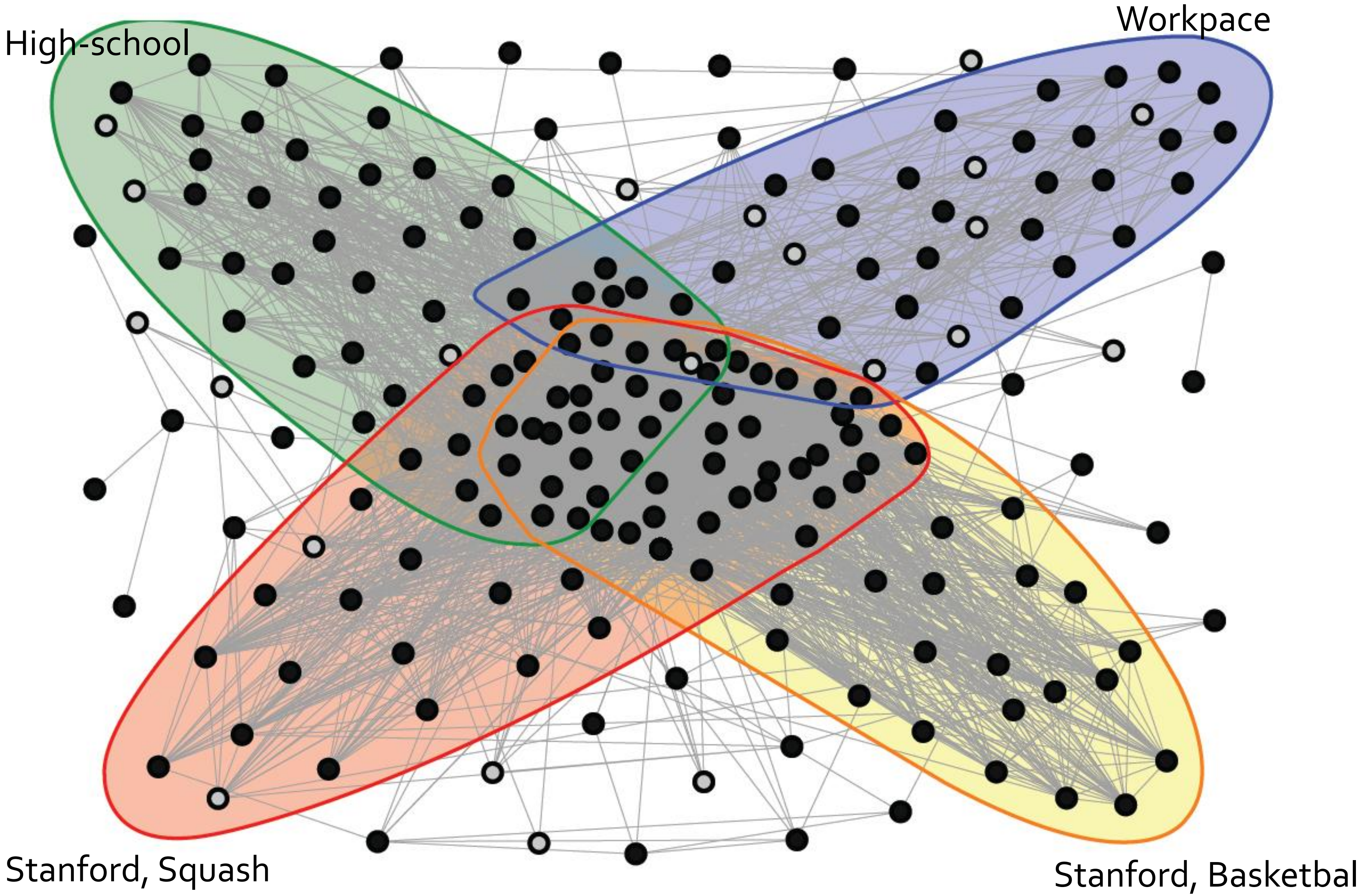
Relative improvement of our method over Clique Percolation

Experiments



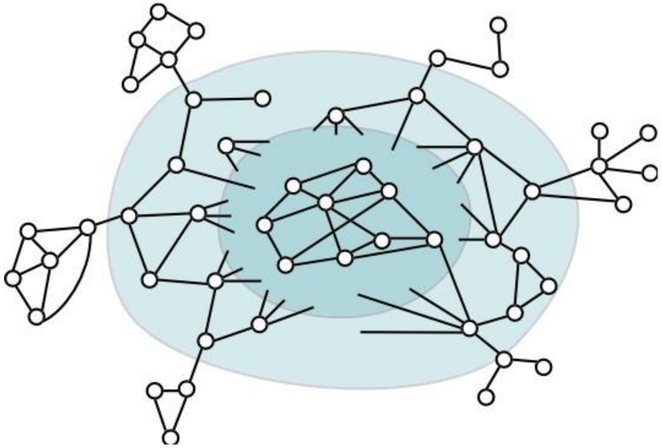


Example: Facebook



Conclusion

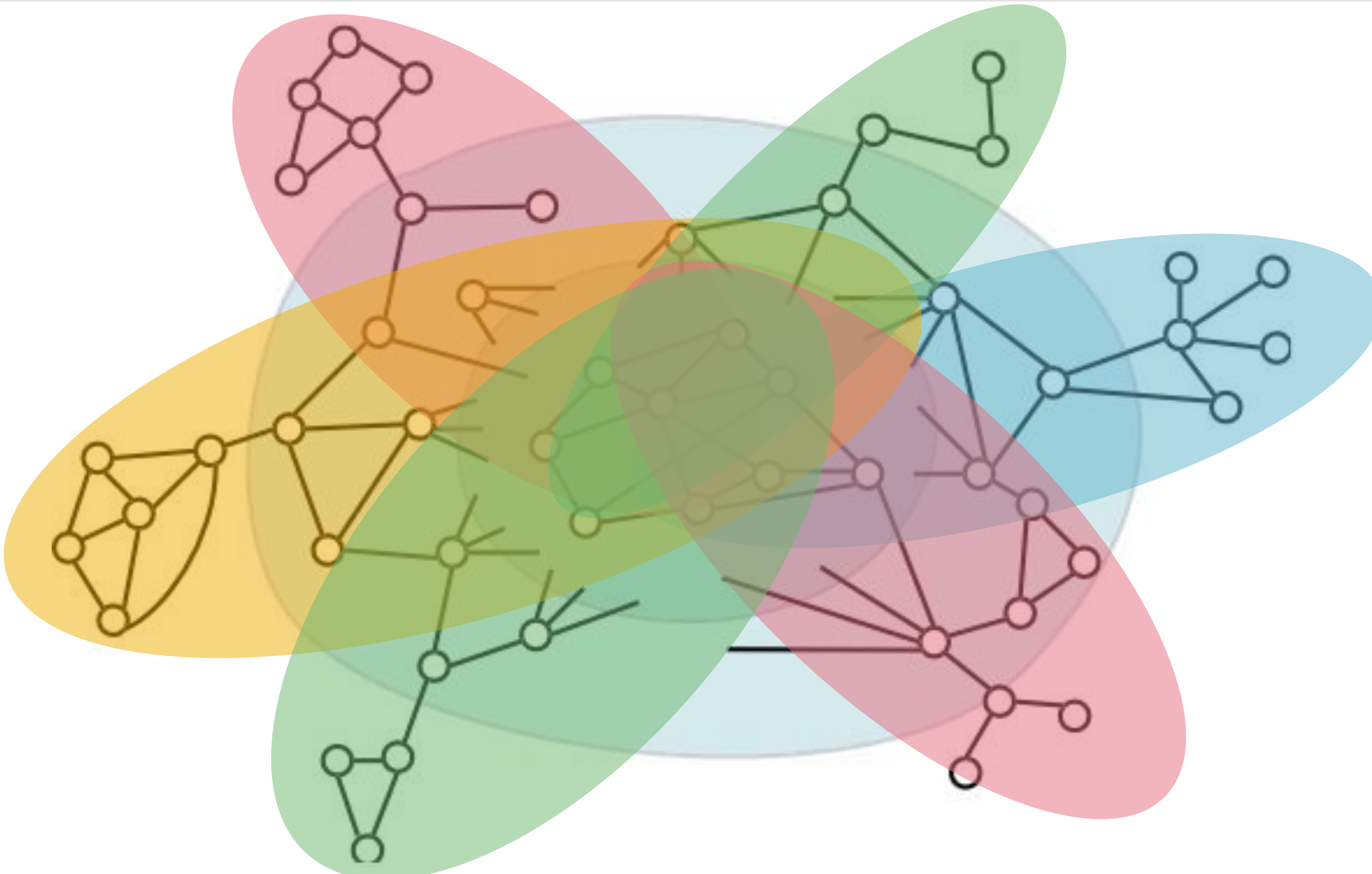
- **NCP plot** is a way to analyze network community structure
- Our results agree with previous work on small networks
- **But** we need to examine **massive networks** to observe the **nested core-periphery**



Conclusion

- Ground-Truth Communities
  $\Rightarrow$ Empirical insight --- Overlaps are **denser**
- Community-Affiliation Graph Model
  $\Rightarrow$ Model-based Community Detection
  - **Outperforms state-of-the-art:**
    - **30% over Link-Clustering**
    - **60% over Clique Percolation**
    - **2 to 70x better estimation of the number of communities**

Connections: Core-Periphery



THANKS!
http://snap.stanford.edu