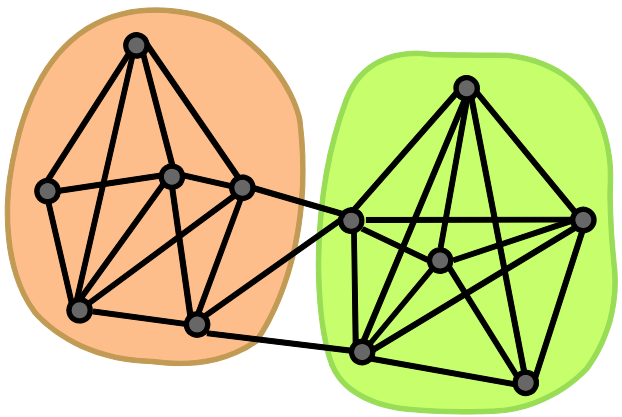
Community Detection: Overlapping Communities

CS224W: Social and Information Network Analysis

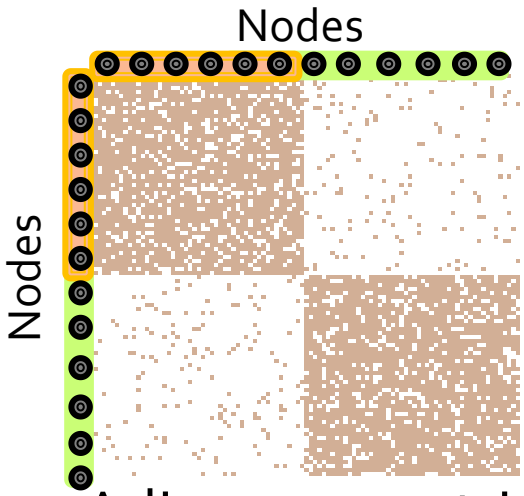
Jure Leskovec, Stanford University

http://cs224w.stanford.edu

Non-overlapping Communities



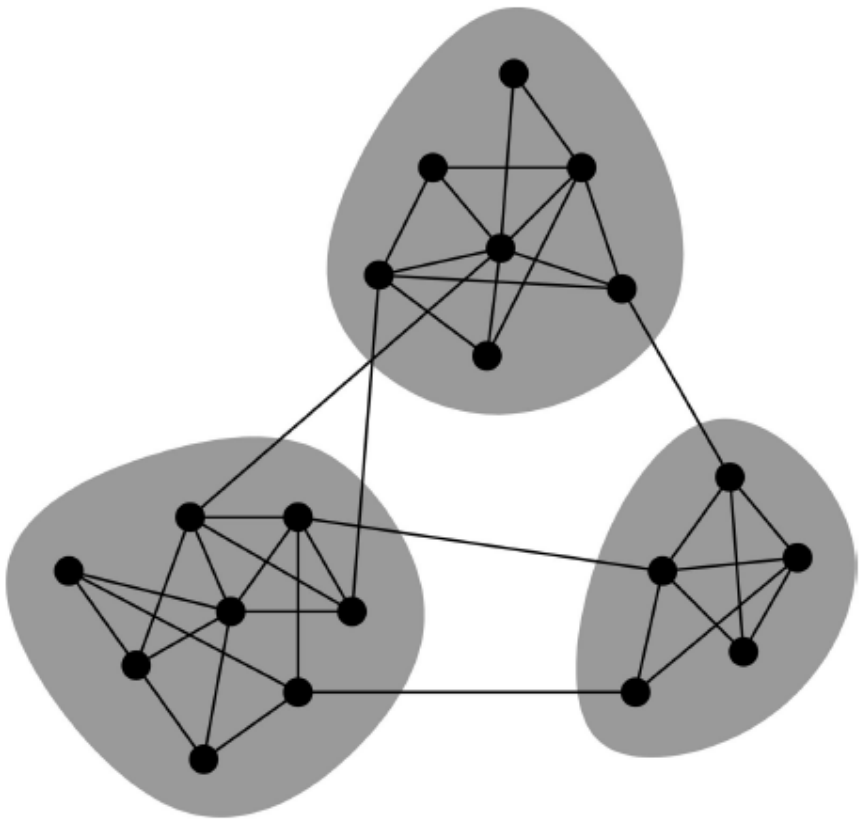
Network

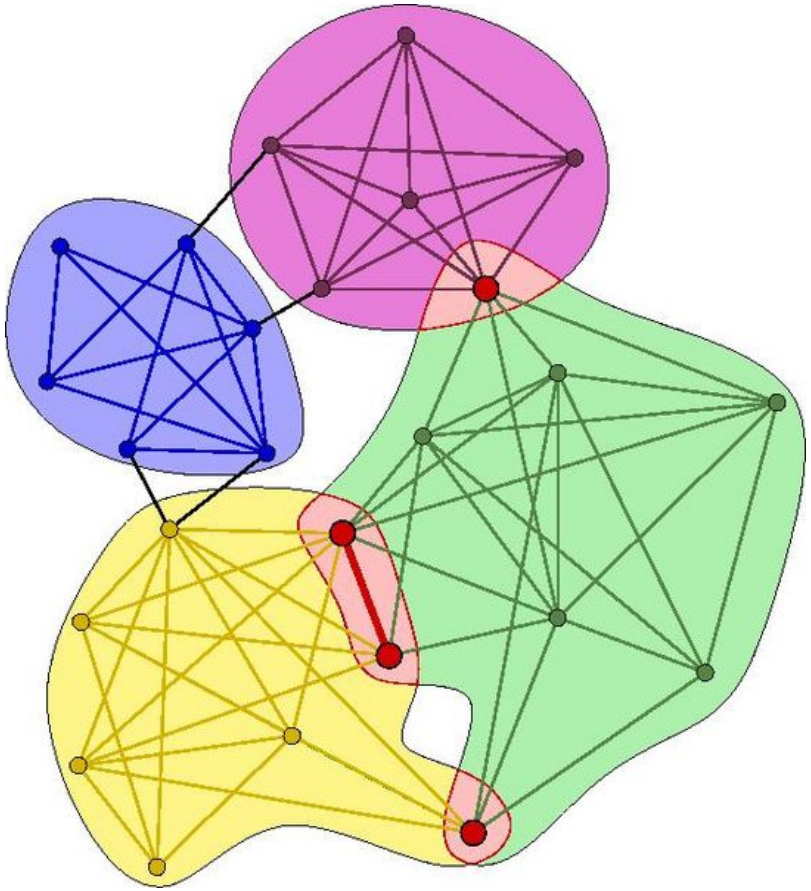


Adjacency matrix

Overlapping Communities

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

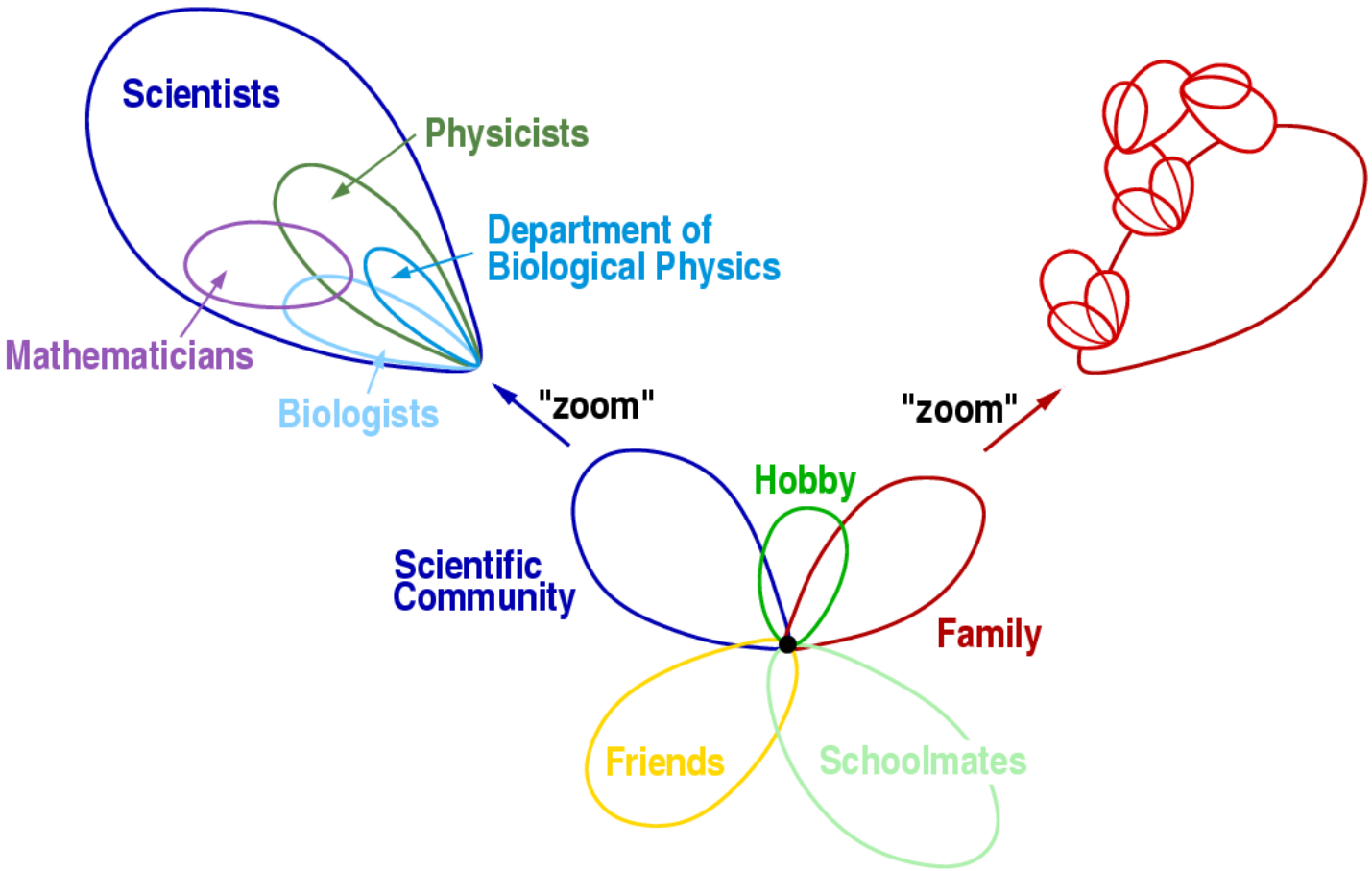




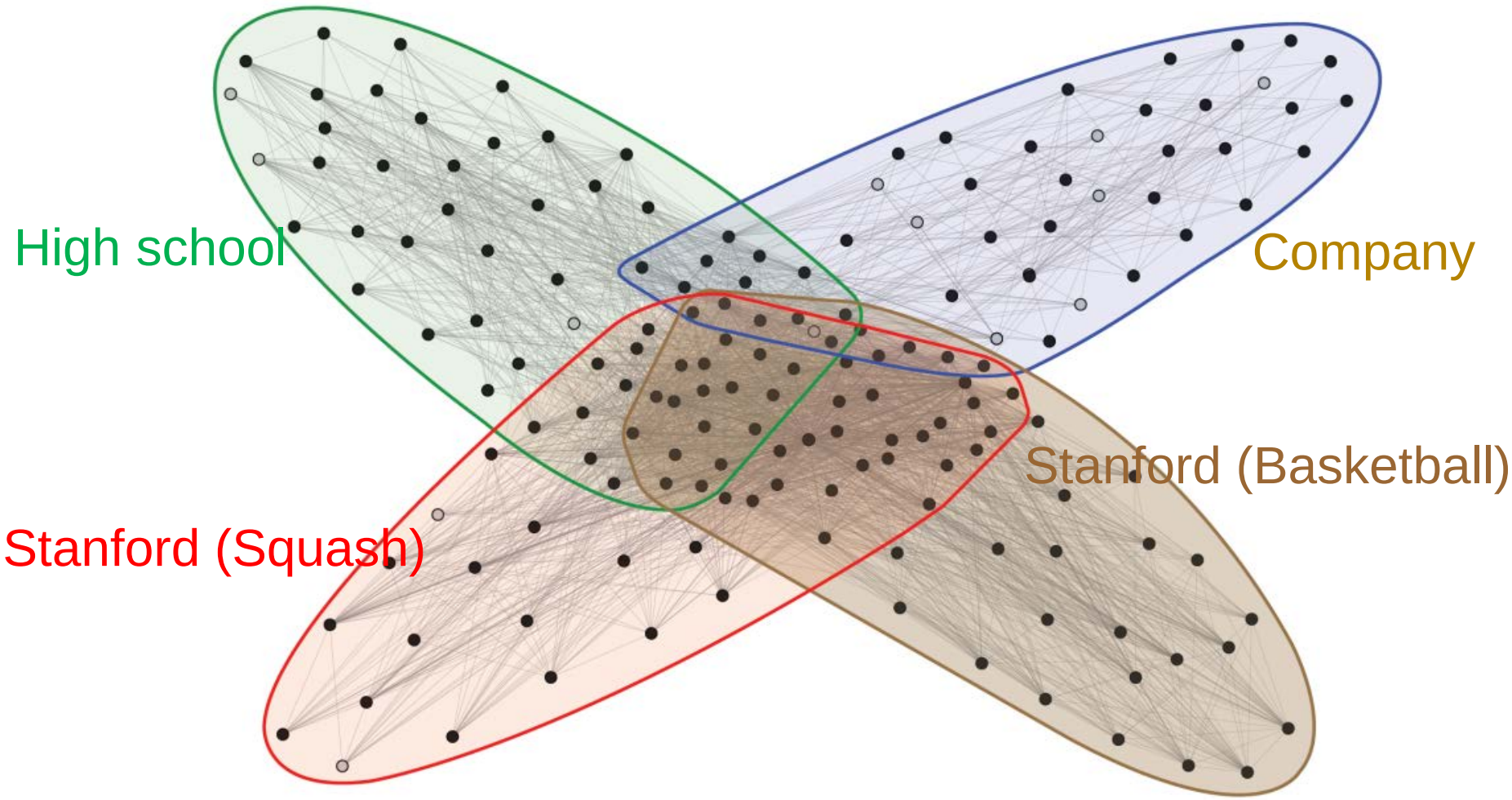
[Palla et al., '05]

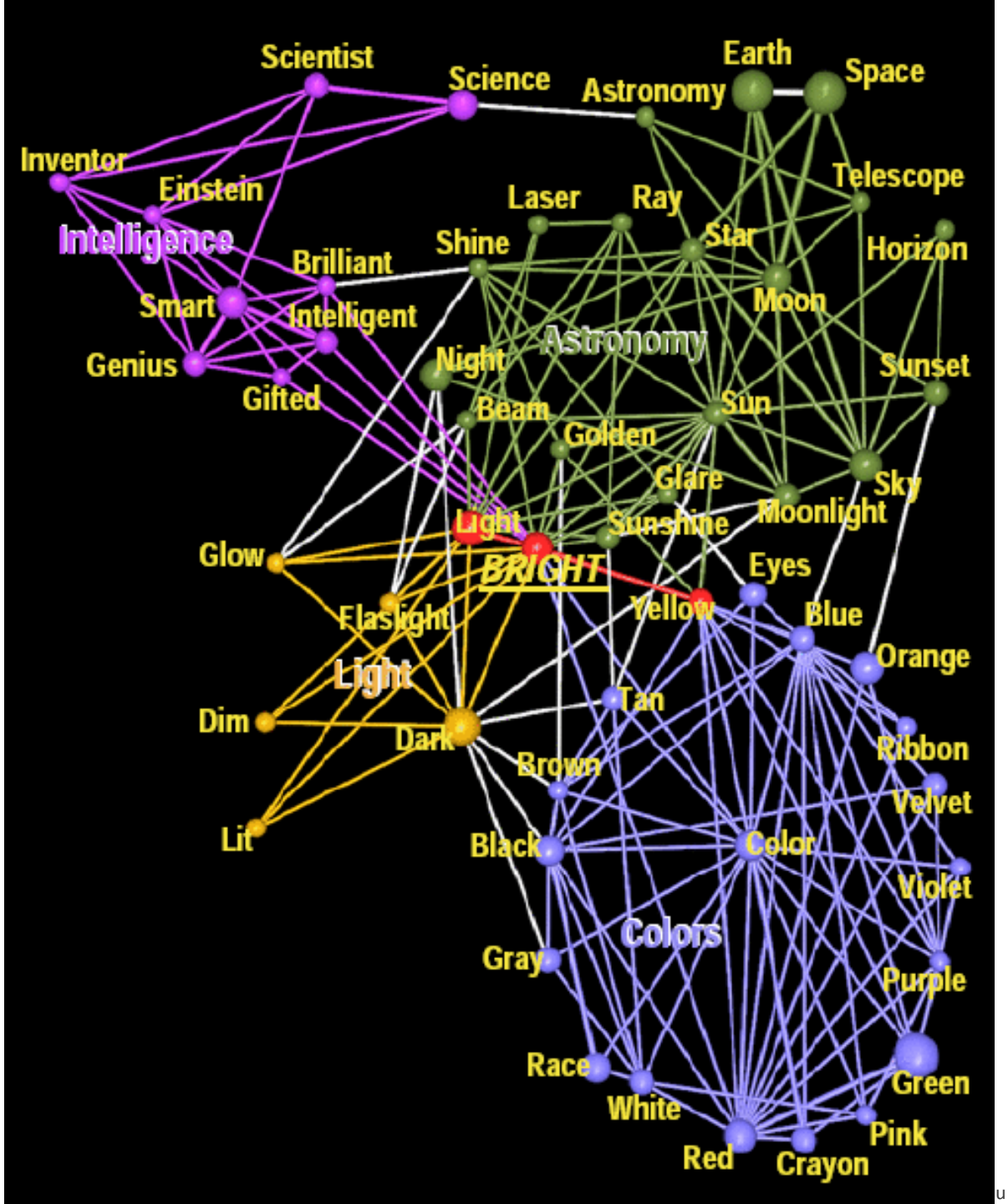
Overlaps of Social Circles

- **A node belongs to many social "circles'**



What if communities overlap?

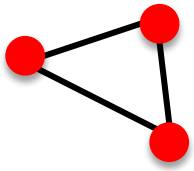



Scientist
Science
Astronomy
Earth
Space
Inventor
Einstein
Intelligence
Brilliant
Smart
Intelligent
Genius
Gifted
Laser
Ray
Shine
Star
Telescope
Horizon
Moon
Astronomy
Night
Sunset
Beam
Sun
Golden
Glare
Sky
Moonlight
Light
Sunshine
Glow
BRIGHT
Eyes
Flashlight
Yellow
Blue
Light
Orange
Tan
Dim
Dark
Ribbon
Brown
Velvet
Lit
Black
Color
Violet
Colors
Gray
Purple
Race
Green
White
Pink
Red
Crayon

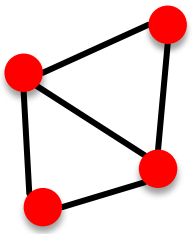
[Palla et al., '05]

Clique Percolation Method (CPM)

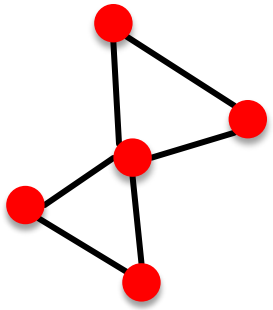
- **Two nodes belong to the same community if they can be connected through adjacent k-cliques:**
  - ***k*-clique:**
    - Fully connected graph on k nodes
  - **Adjacent *k*-cliques:**
    - overlap in k-1 nodes
- ***k*-clique community**
  - Set of nodes that can be reached through a sequence of adjacent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



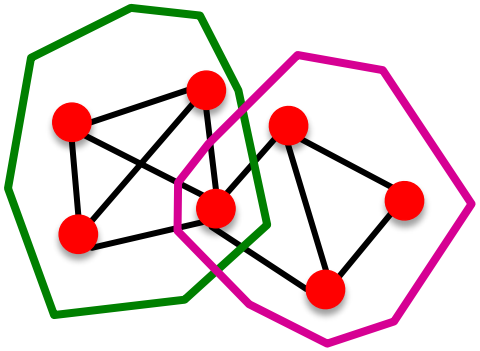
3-clique



Adjacent 3-cliques



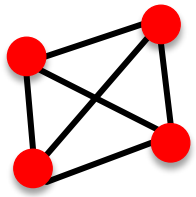
Non-adjacent 3-cliques



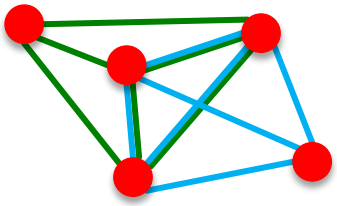
[Palla et al., '05]

Clique Percolation Method (CPM)

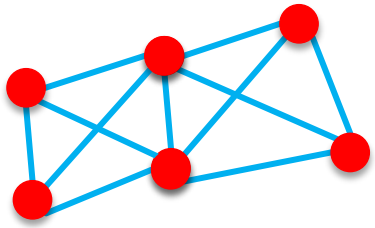
- **Two nodes belong to the same community if they can be connected through adjacent *k*-cliques:**



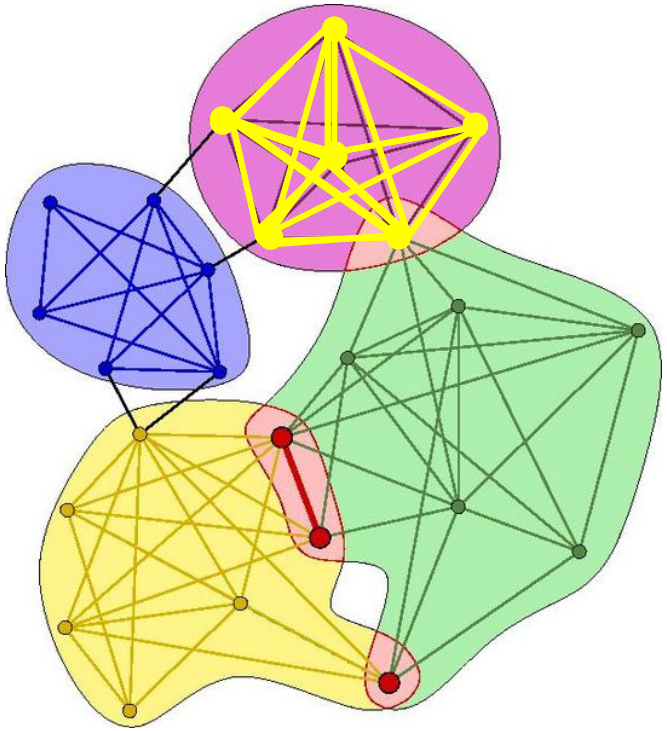
4-clique



Adjacent 4-cliques



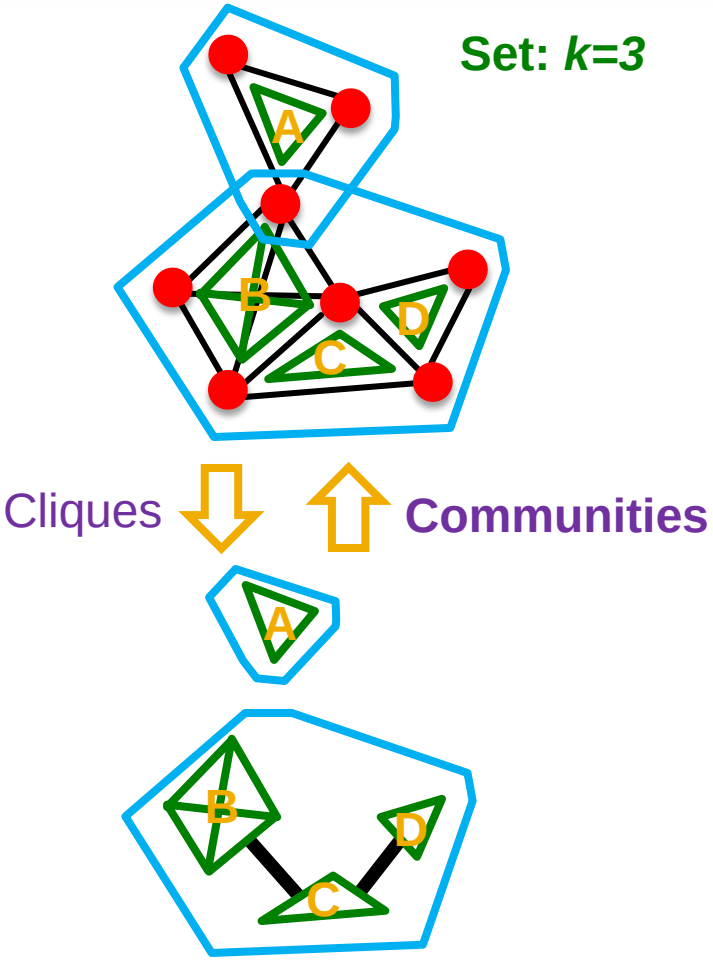
Non-adjacent 4-cliques



Communities for k=4

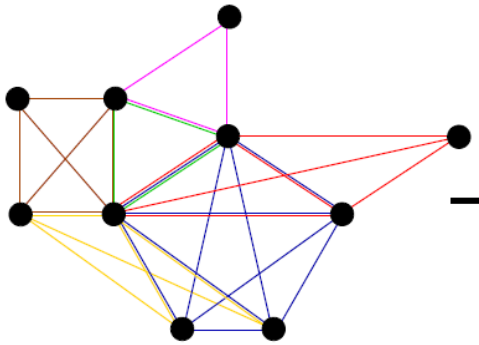
CPM: Steps

- **Clique Percolation Method:**
  - **Find maximal-cliques** (not k-cliques!)
  - **Clique overlap graph:**
    - 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
- **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



(1) Graph

Overlap size

Cliques

Cliques

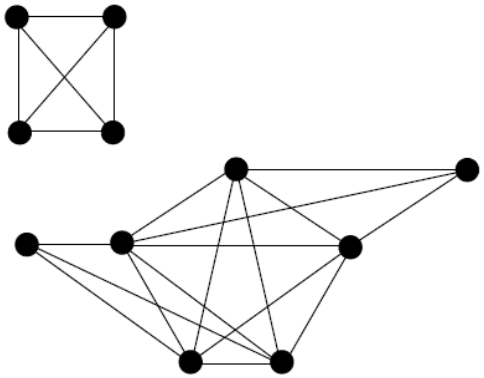
5	3	2	1	3	1
3	4	2	1	1	1
2	2	3	2	1	2
1	1	2	3	0	1
3	1	1	0	4	2
1	1	2	1	2	4

(2) Clique overlap matrix

k=4

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

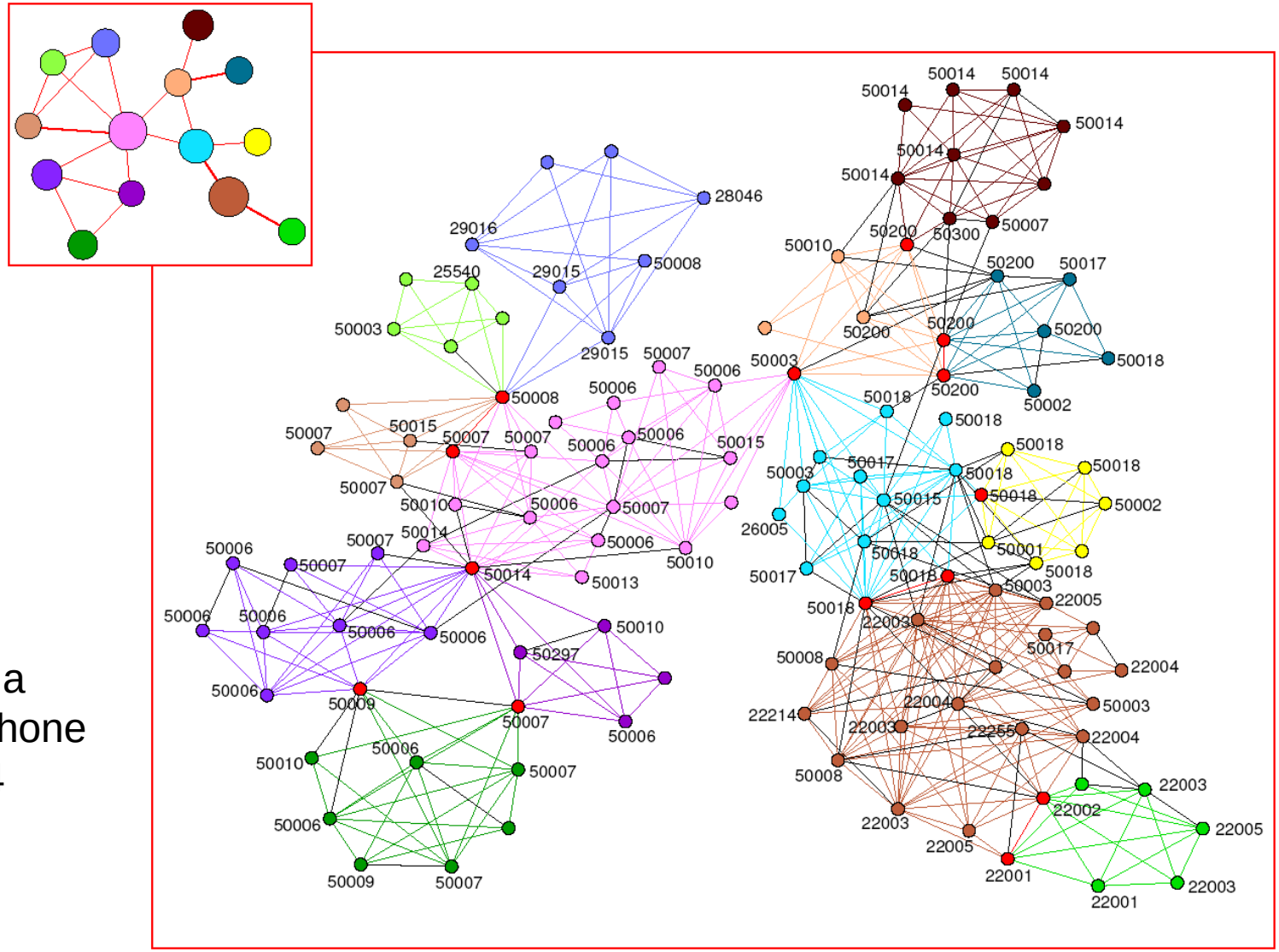
(3) Thresholded matrix at 3



(4) Communities (connected components)

[Palla et al., '07]

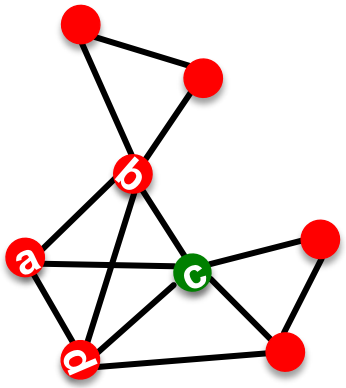
Example: Phone-Call Network



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

How to Find Maximal Cliques?

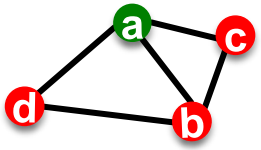
- **No nice way, hard combinatorial problem**
- **Maximal clique:** Clique that can't be extended
  - $\{a, b, c\}$ is a clique but not maximal clique
  - $\{a, b, c, d\}$ is maximal clique



- **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



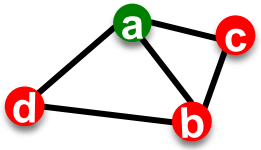
Q=
R=

Steps of the recursive algorithm

Γ(u)…neighbor set of *u*

How to Find Maximal Cliques?

- 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

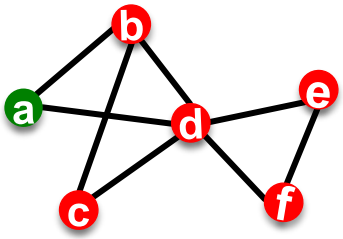
Q=	{a}	{a,b}	$\boxed{\{a,b,c\}}$	bktrack	$\boxed{\{a,b,d\}}$
R=	{<u>b</u>,c,d}	{b,c,d} $\cap\Gamma$(b)={<u>c</u>,d}	{d}$\cap\Gamma$(c)={}		{c}$\cap\Gamma$(d)={}

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** $R_p \neq \{\}$: Expand(R_p, Q_p)
      else: output Q_p
    - R = R – {p}



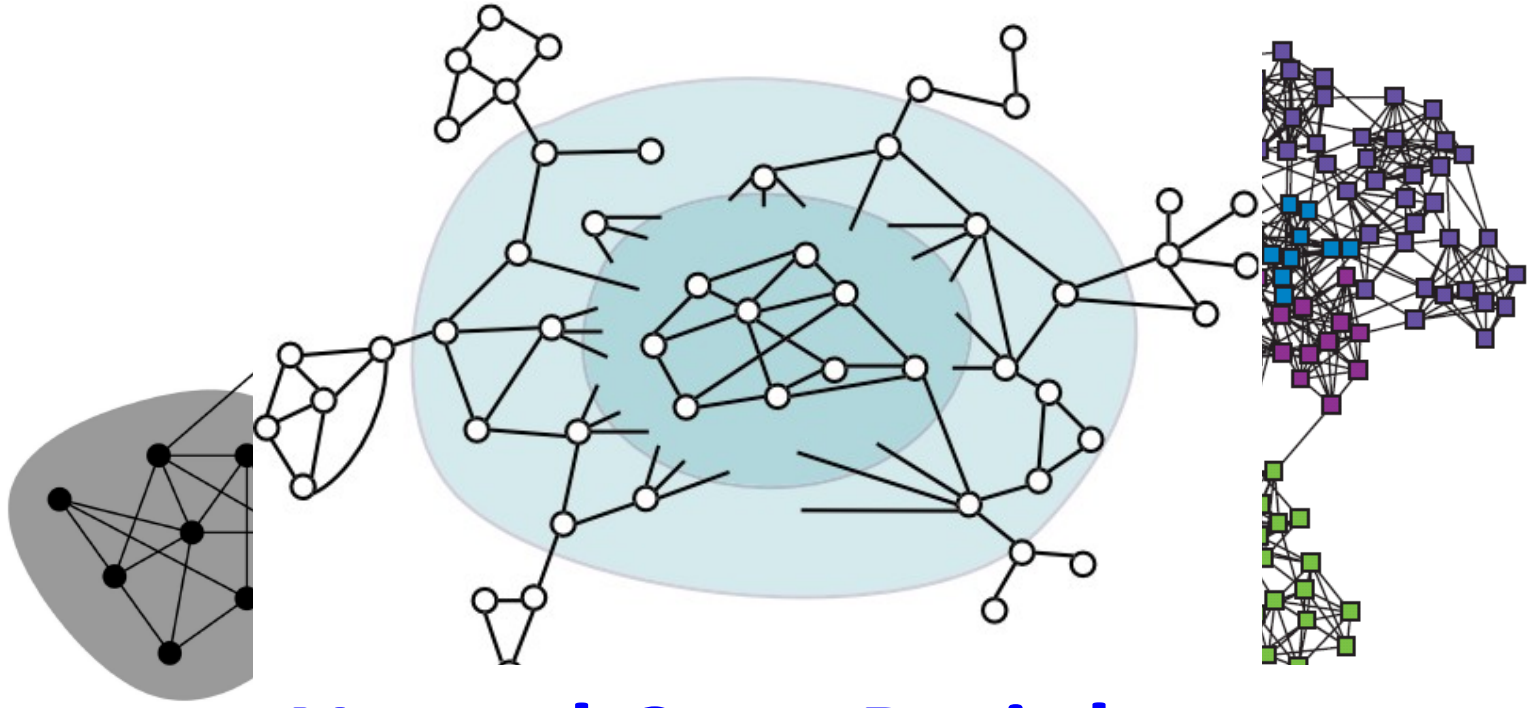
Start: Expand(V, {})
R={a…f}, Q={}
p = {a}
Q_p = {a}
R_p = {b,d}
Expand(R_p, Q):
R = {b,d}, Q={a}
p = {b}
Q_p = {a,b}
R_p = {d}
Expand(R_p, Q):
R = {d}, Q={a,b}
p = {d}
Q_p = {a,b,d}
R_p = {} : **output {a,b,d}**
p = {d}
Q_p = {a,d}
R_p = {b}
Expand(R_p, Q):
R = {b}, Q={a,d}
p = {b}
Q_p = {a,d}
R_p = {} : **output {a,d,b}**

How to Model Networks with Communities?

Announcement: On Thu we will have **Lars Backstrom** from Facebook Data Science team give a guest lecture!

Network and Communities

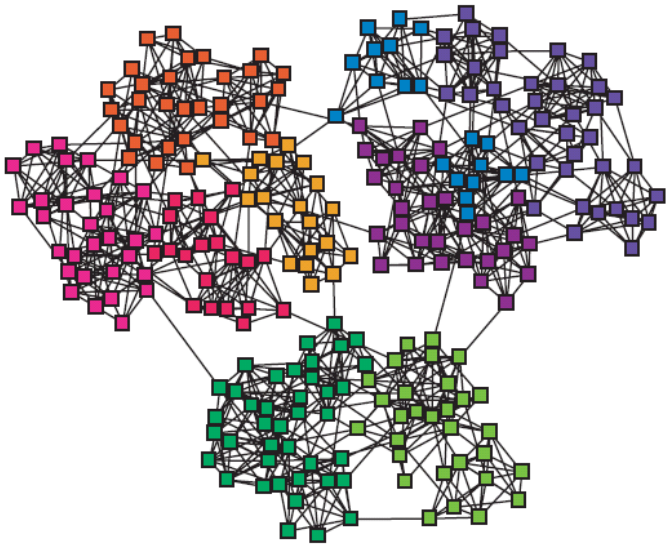
- **How should we think about large scale organization of clusters in networks?**
  - **Finding:** There is a lack of (traditional) clustering



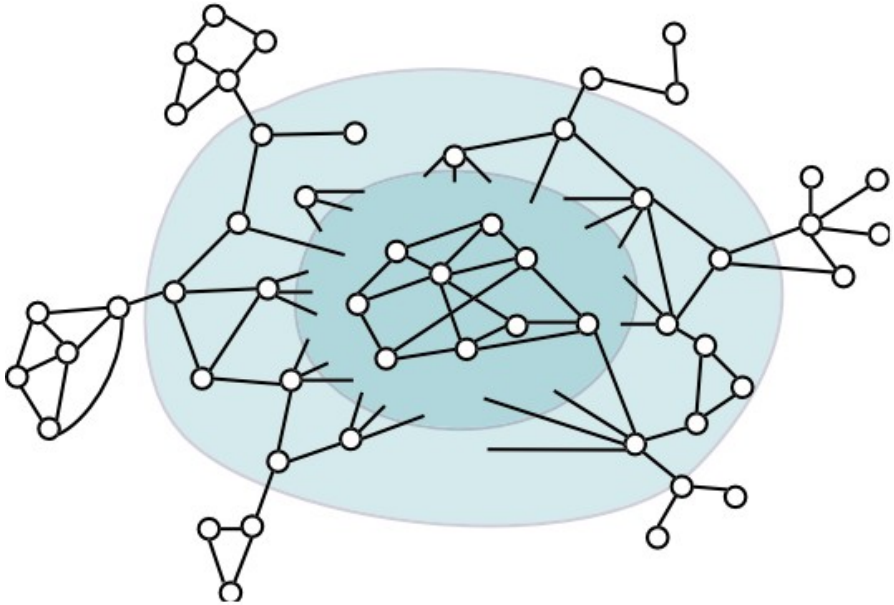
Nested Core-Periphery

Network and Communities

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

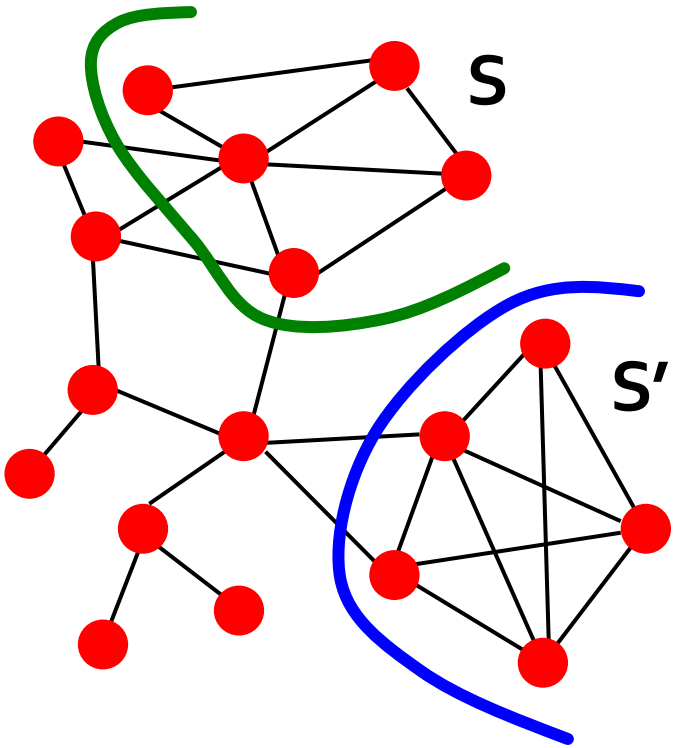


vs.



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

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

[WWW '08]

Network Community Profile Plot

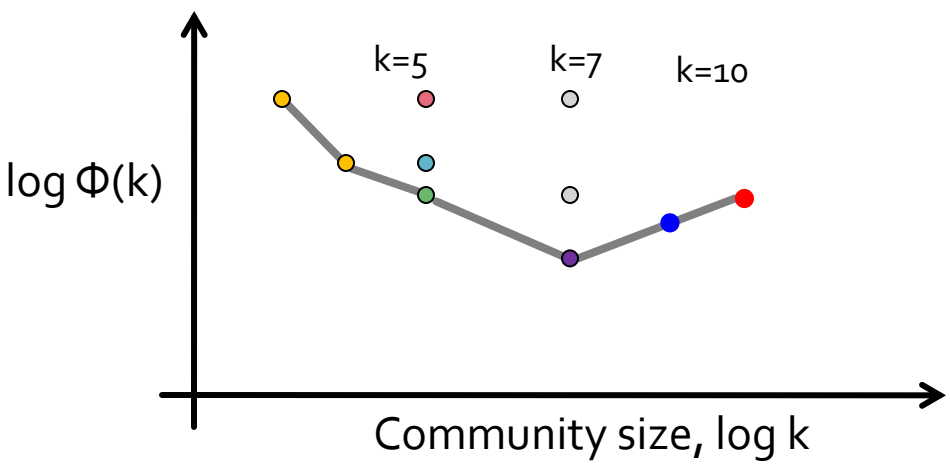
(Note |S| < |V|/2)

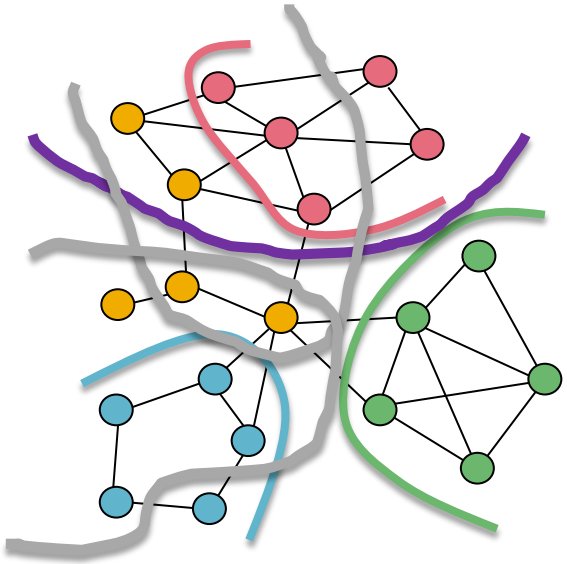
- <u>Define:</u>

  Network community profile (**NCP**) plot

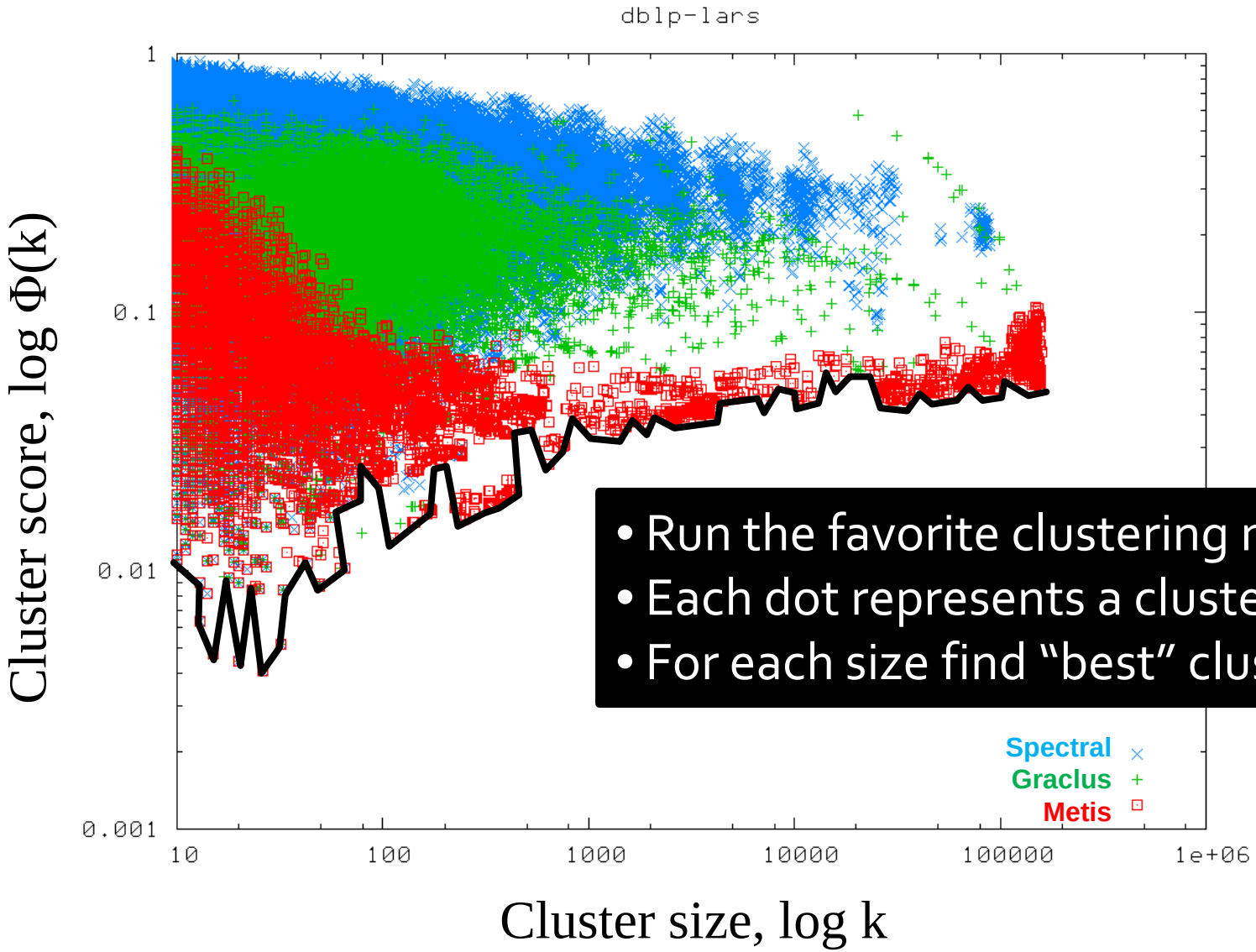
  Plot the score of **best** community of size *k*

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





How to (Really) Compute NCP?

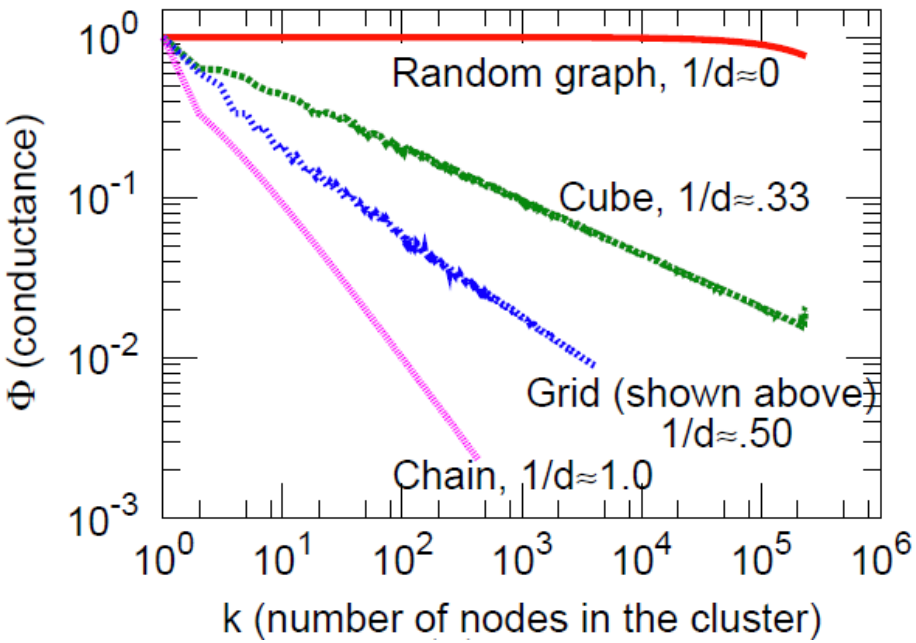


- Run the favorite clustering method
- Each dot represents a cluster
- For each size find "best" cluster

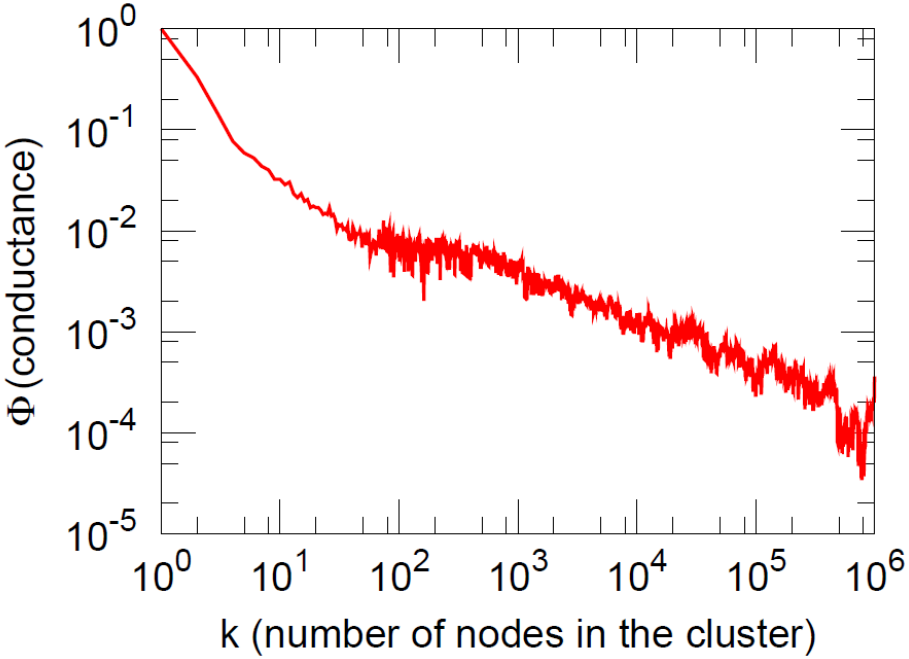
[WWW '08]

NCP Plot: Meshes

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



d-dimensional meshes

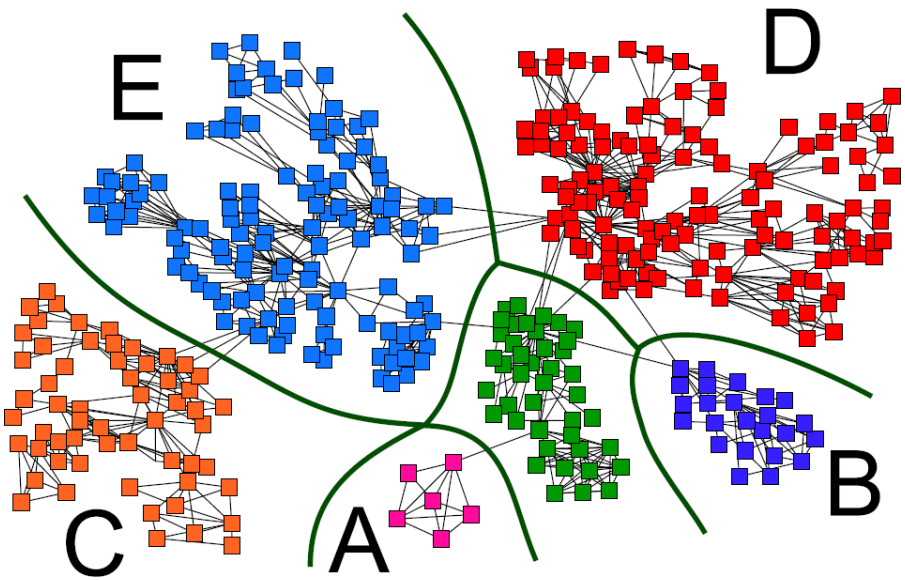


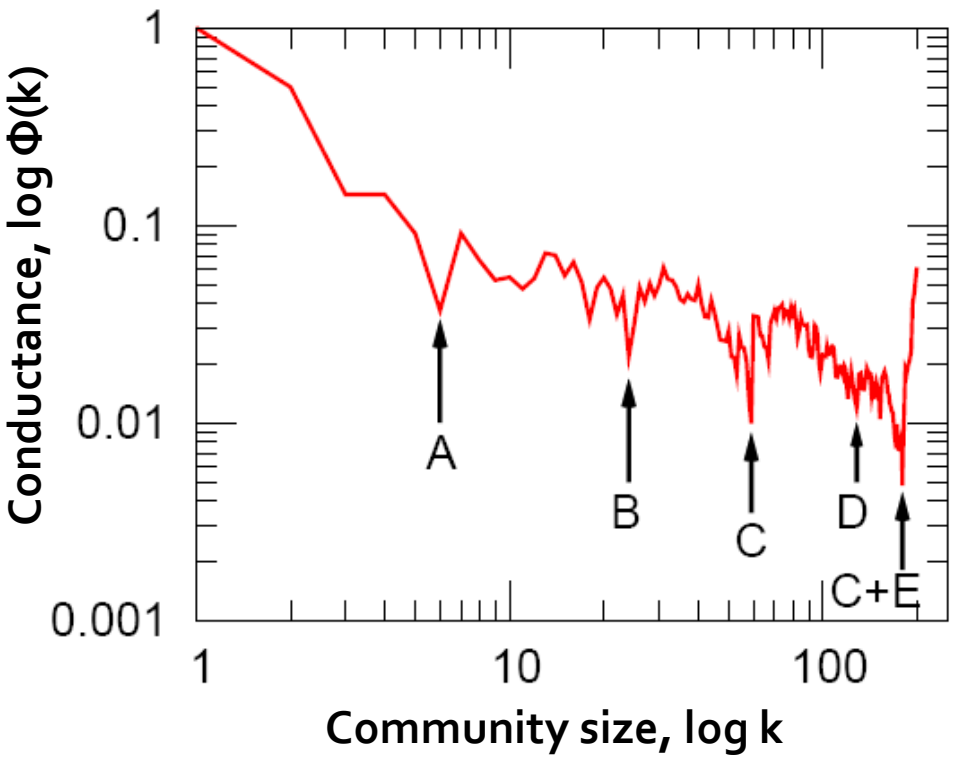
California road network

[WWW '08]

NCP plot: Network Science

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

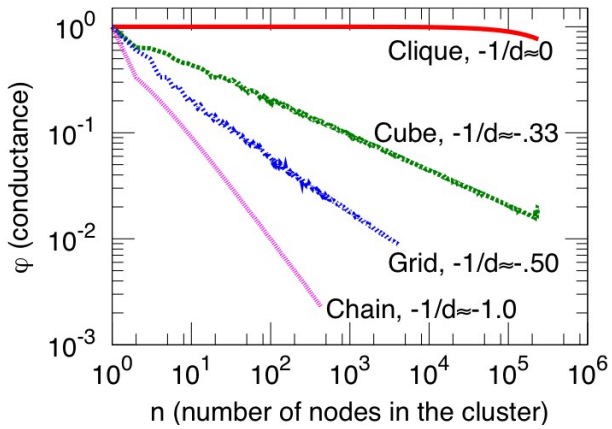




Natural Hypothesis

Natural hypothesis about NCP:

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



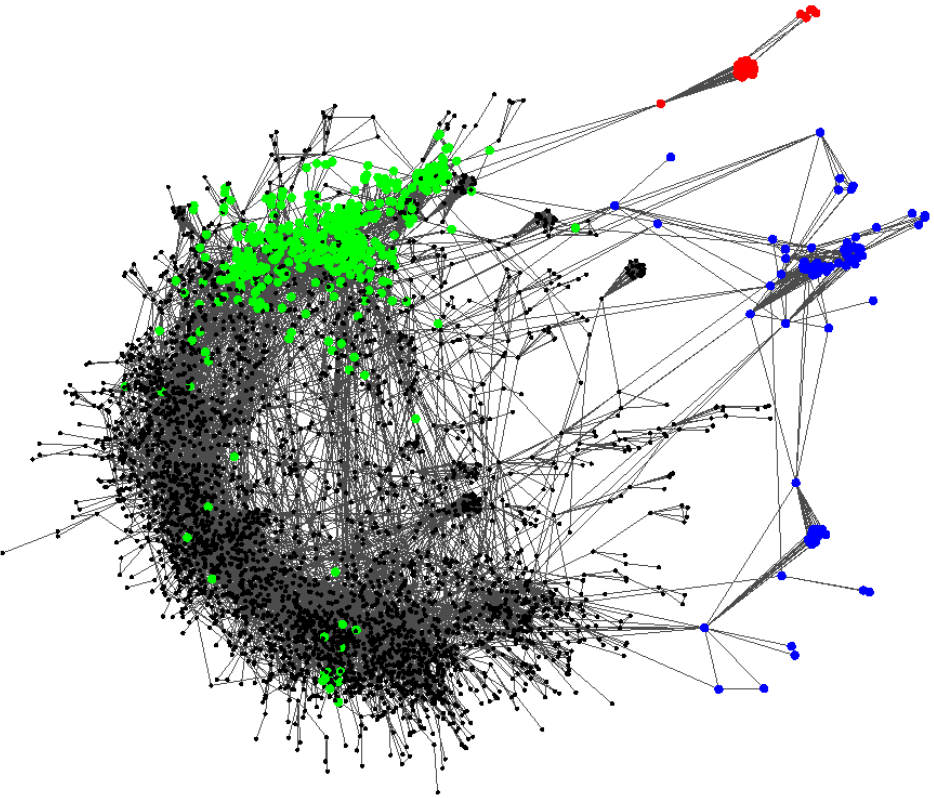
What about large networks?

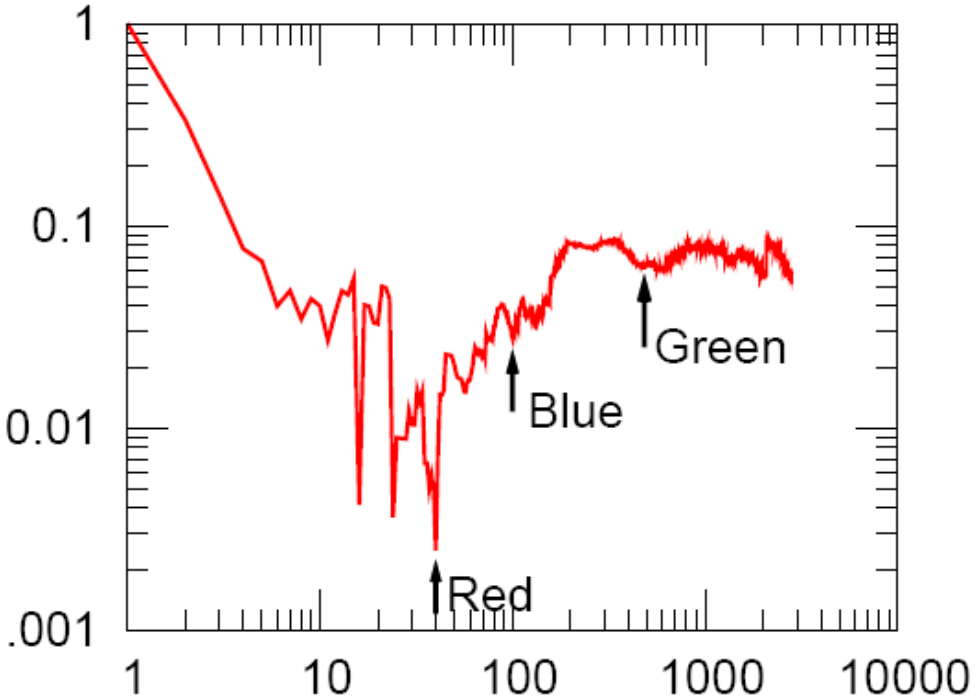
• Social nets	Nodes	Edges	Description
LiveJournal	4,843,953	42,845,684	Blog friendships [5]
Epinions	75,877	405,739	Trust network [28]
CA-DBLP	317,080	1,049,866	Co-authorship [5]
• Information (citation) networks			
Cit-hep-th	27,400	352,021	Arxiv hep-th [14]
AmazonProd	524,371	1,491,793	Amazon products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

[Internet Mathematics '09]

Large Networks: Very Different

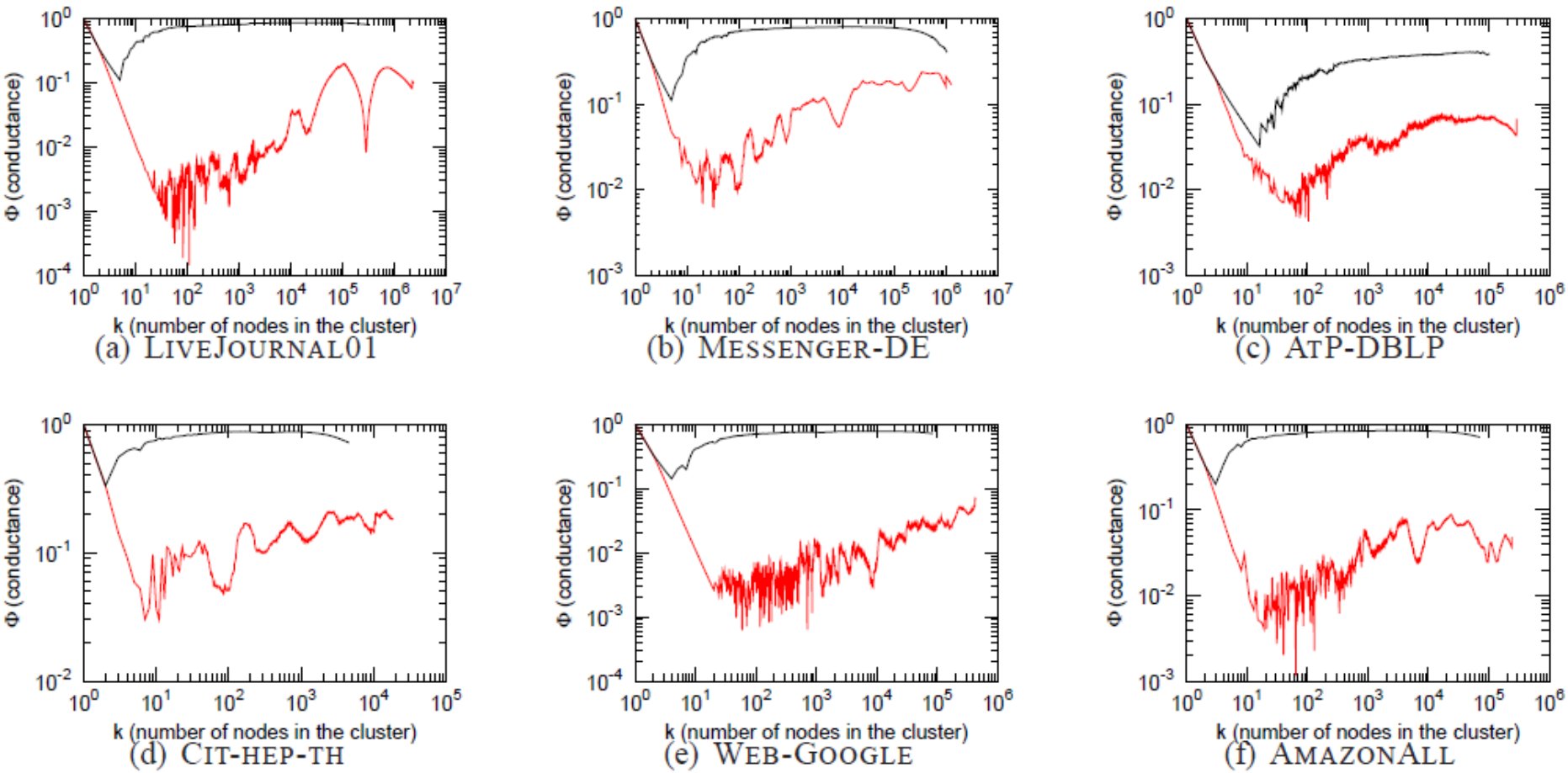
Typical example: General Relativity collaborations (n=4,158, m=13,422)



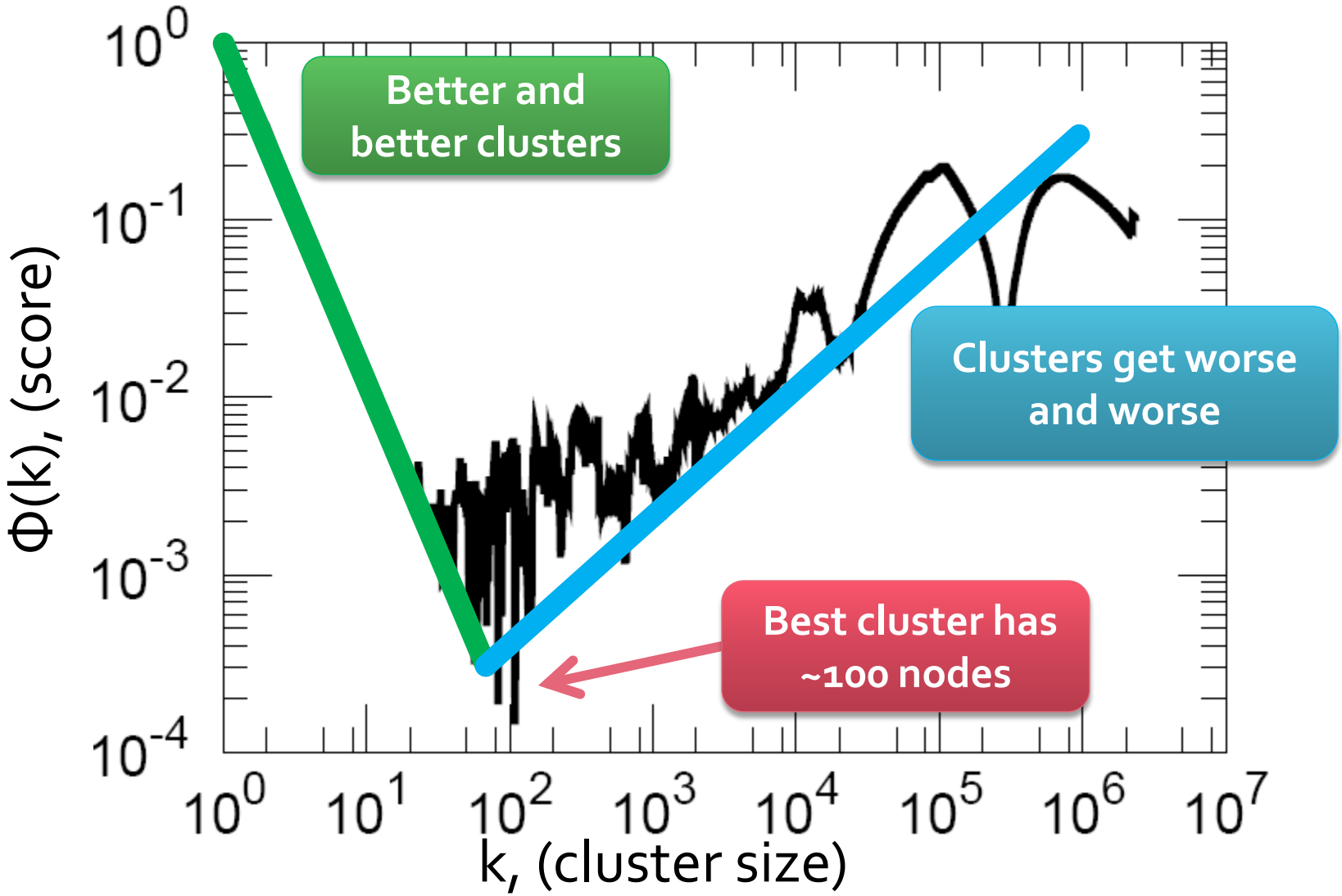


[Internet Mathematics '09]

More NCP Plots of Networks

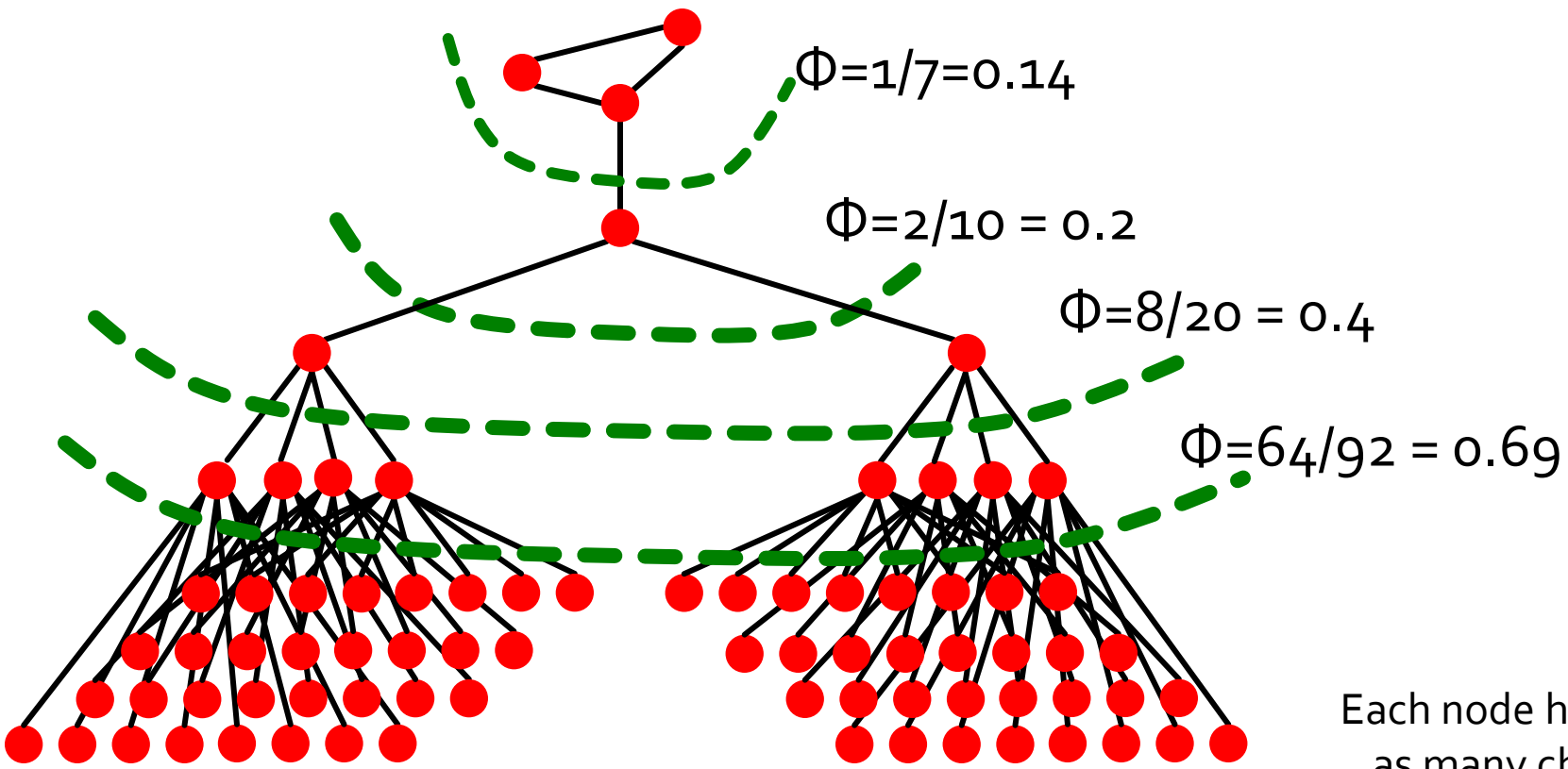


NCP: LiveJournal (n=5m, m=42m)



Explanation: The Upward Part

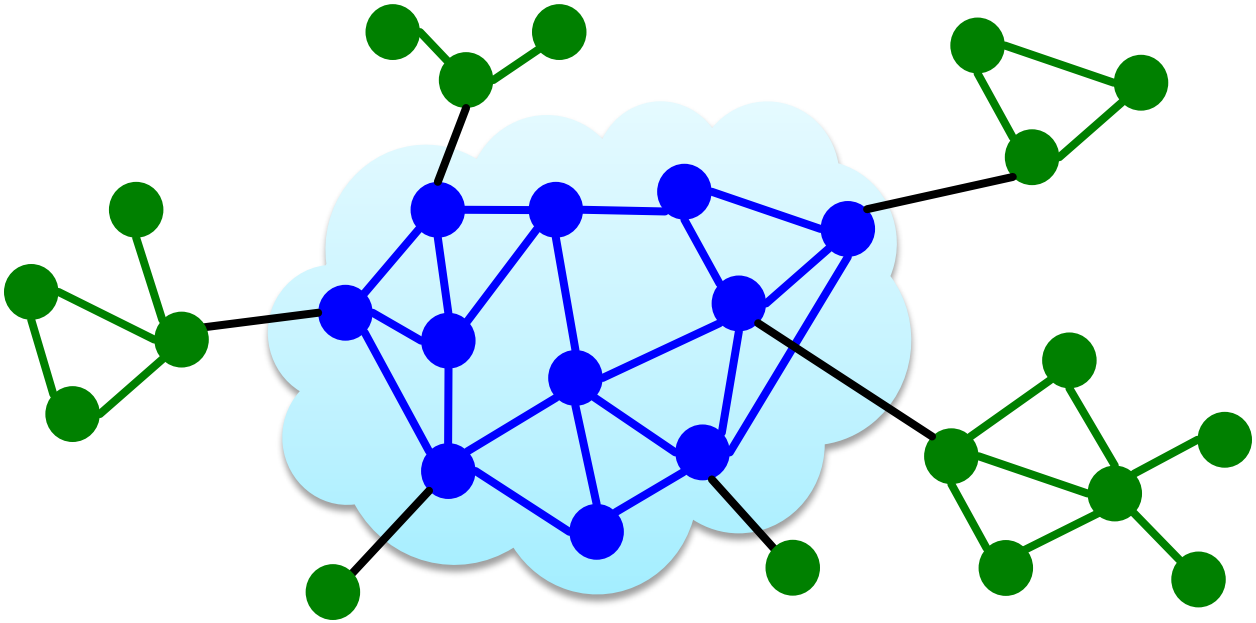
- As clusters grow the number of edges inside grows **slower** that the number crossing

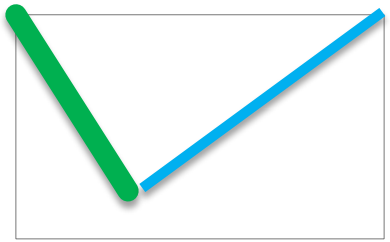


Each node has twice as many children

Explanation: Downward Part

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

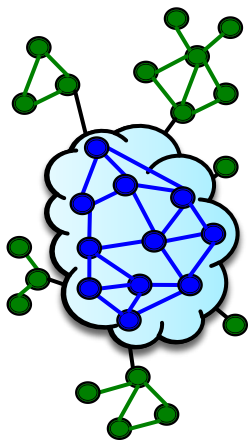


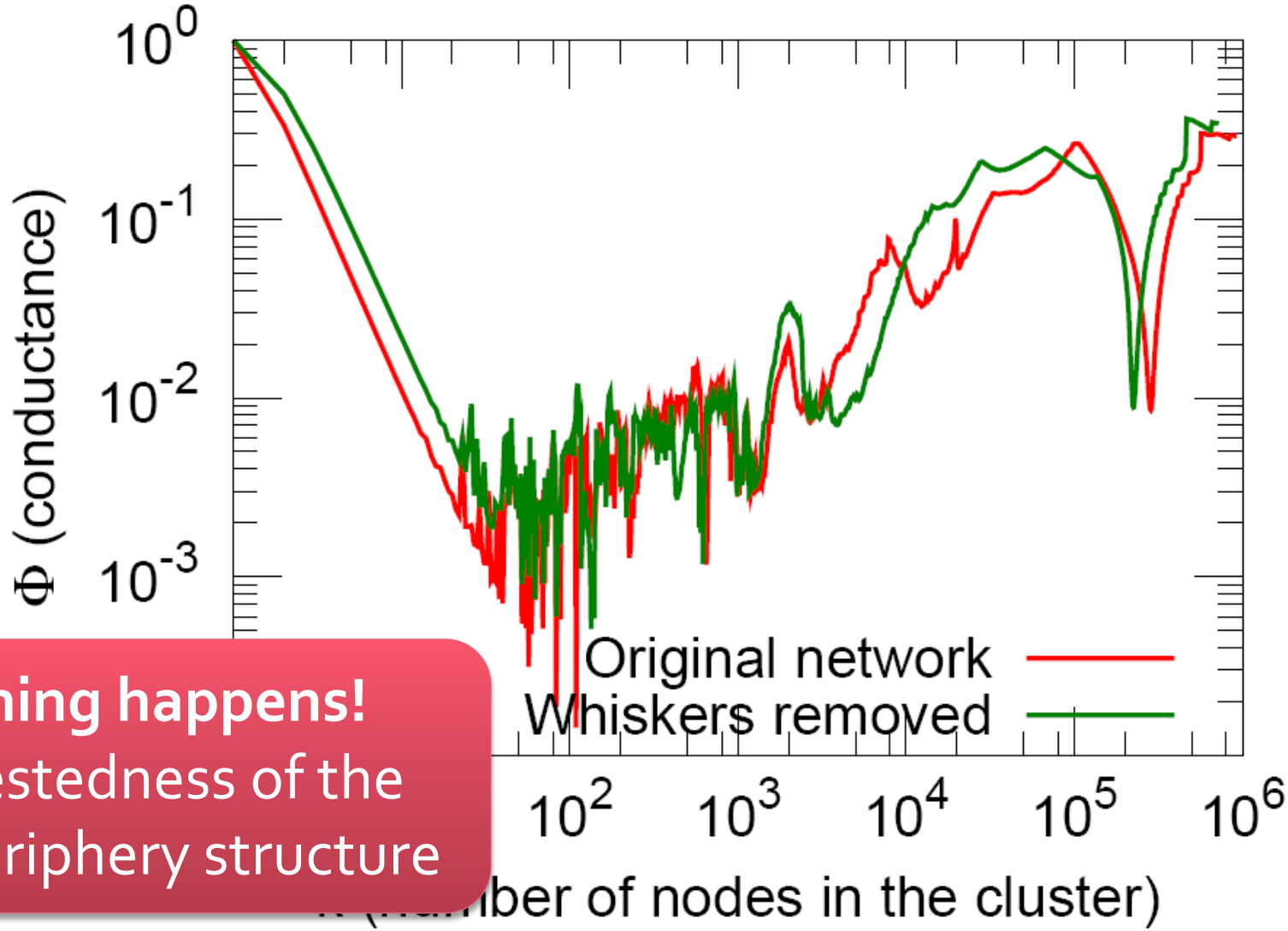


NCP plot

⇒ Core-periphery structure

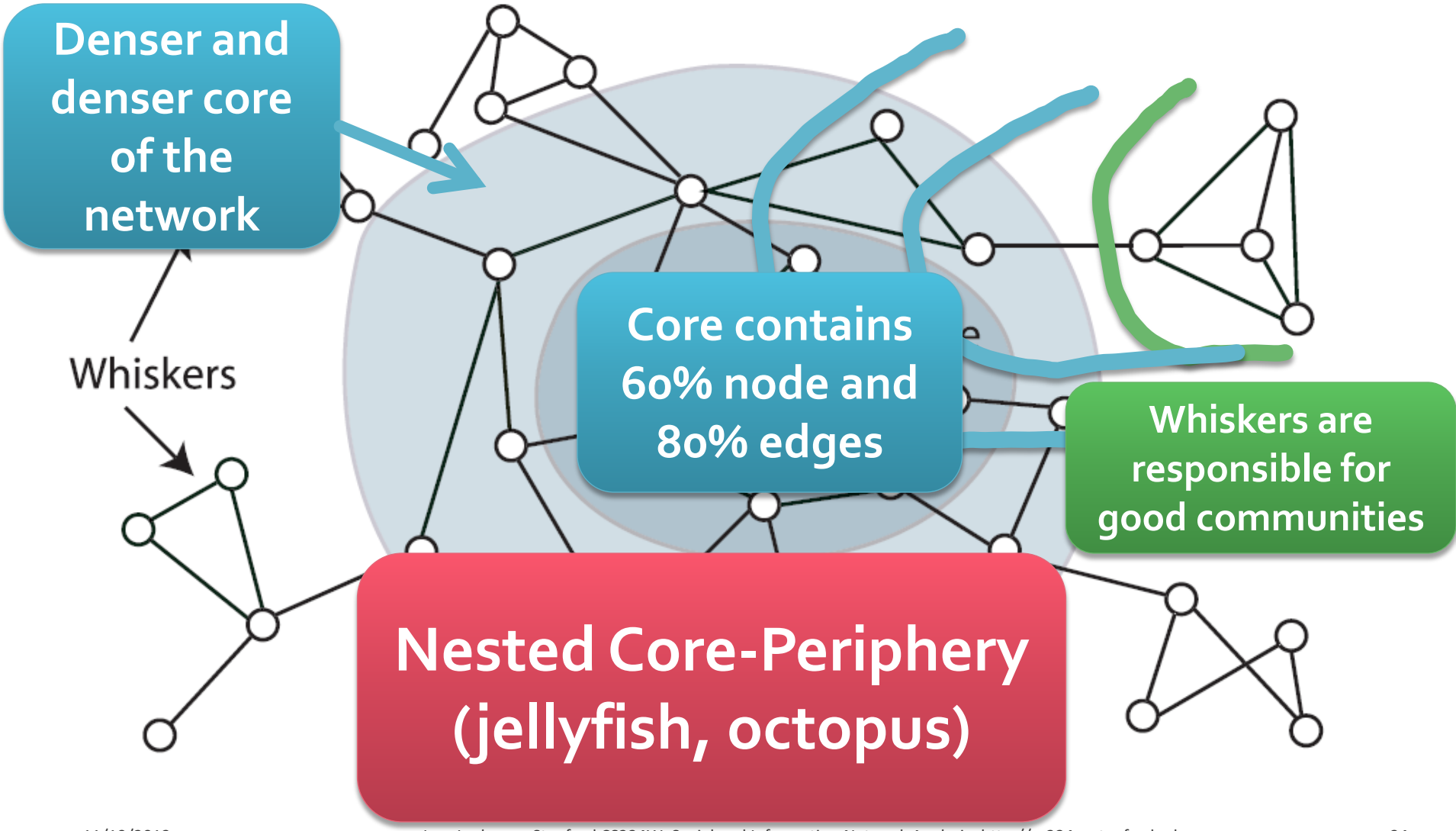
What If We Remove Good Clusters?



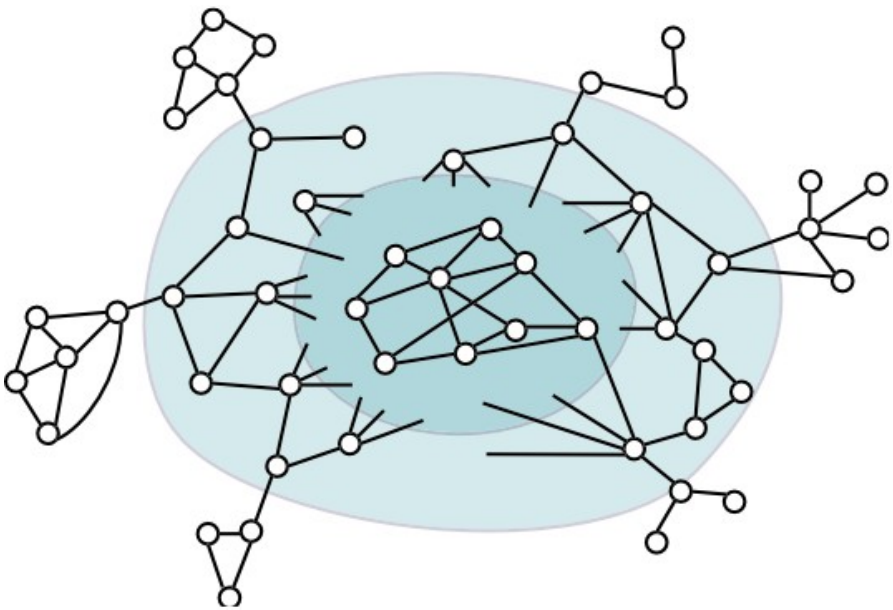


Nothing happens!
⇒ Nestedness of the core-periphery structure

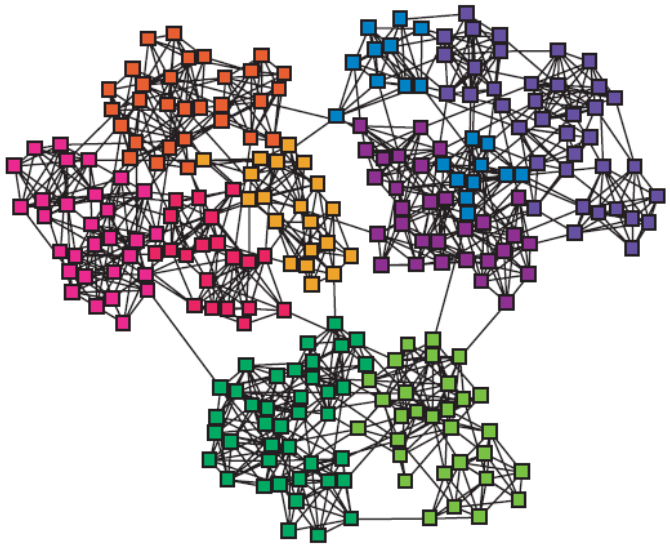
Suggested Network Structure



Part 2: Explanation



vs.



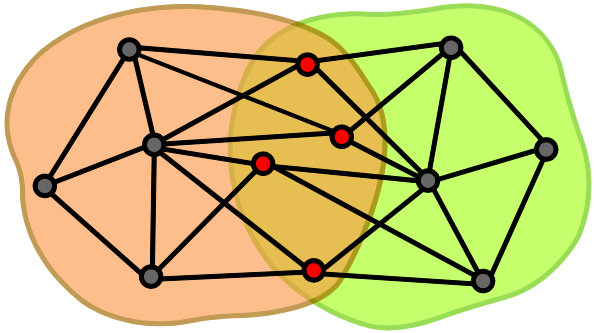
How do we reconcile these two views?

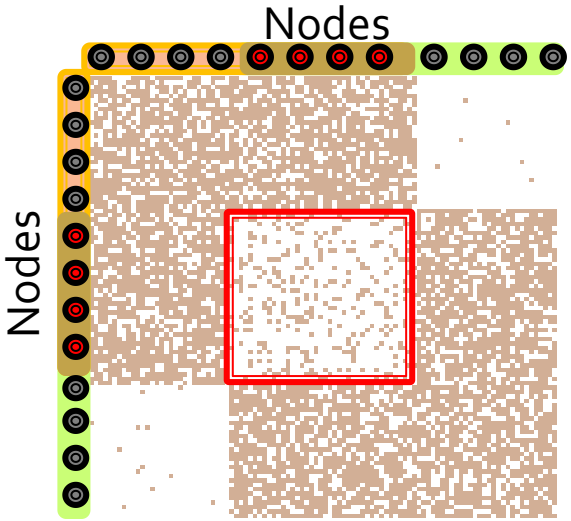
Overlapping Community Detection

- **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**



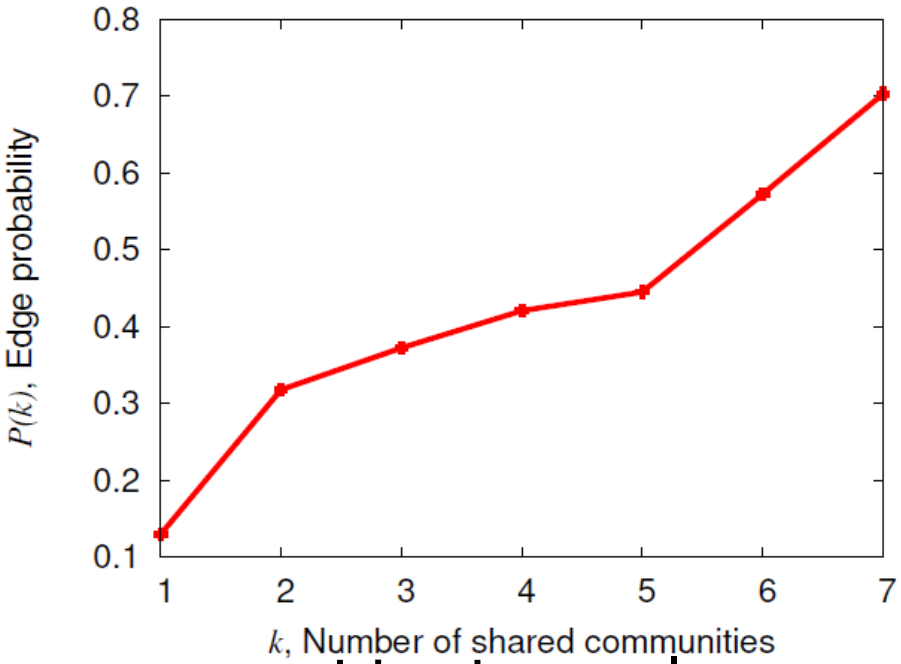


y matrix

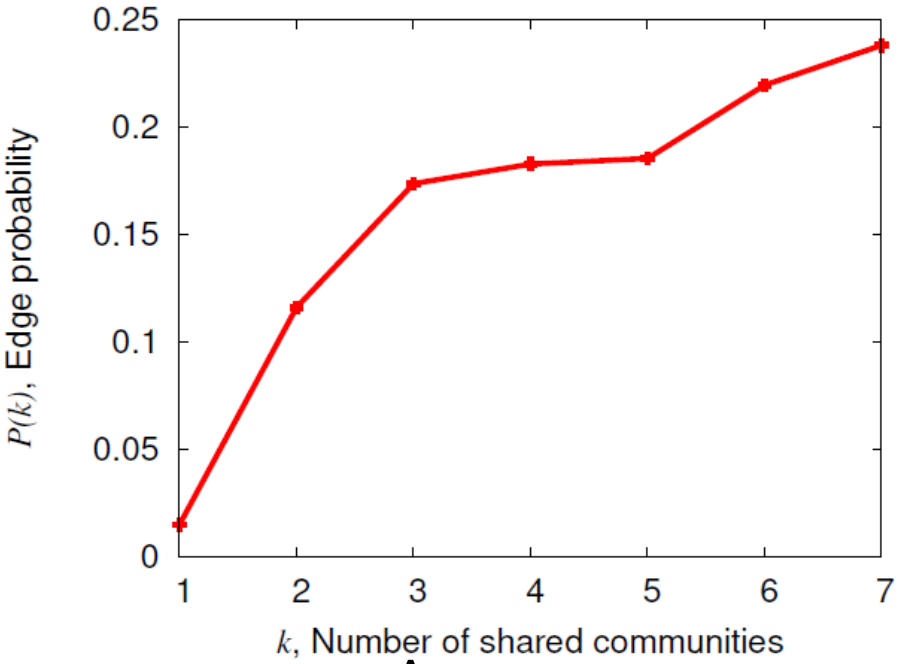
Is this true?

Ground-truth Communities

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



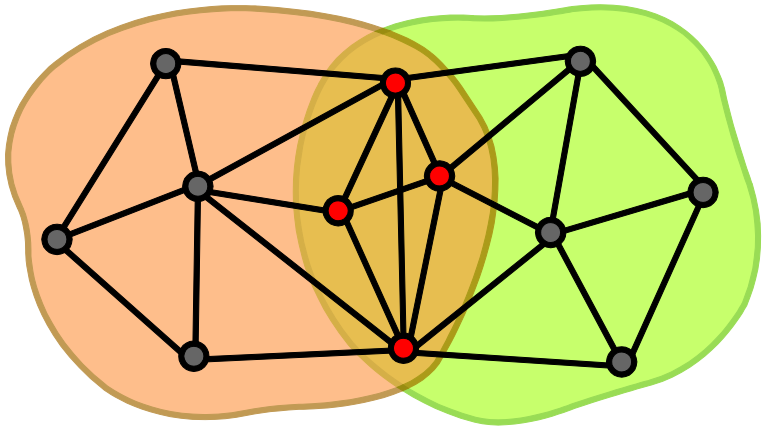
LiveJournal
social network

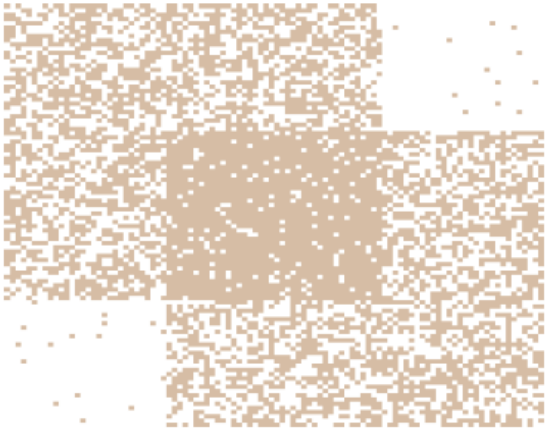


Amazon
product network

Communities as Tiles!

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

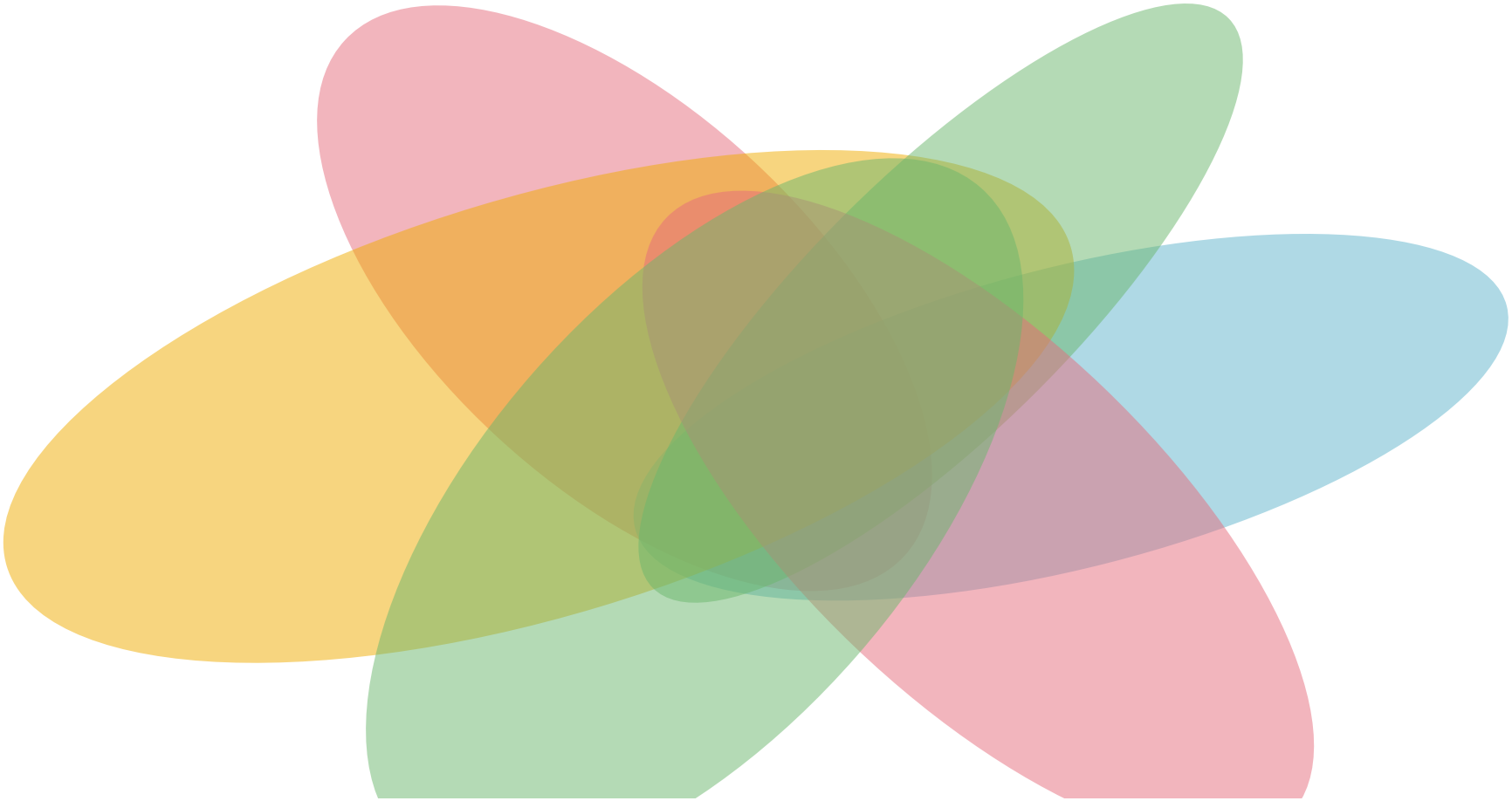




"The more different foci (communities) that two individuals share, the more likely it is that they will be tied" - S. Feld, 1981

Communities as "tiles"

Communities as Tiles/Circles

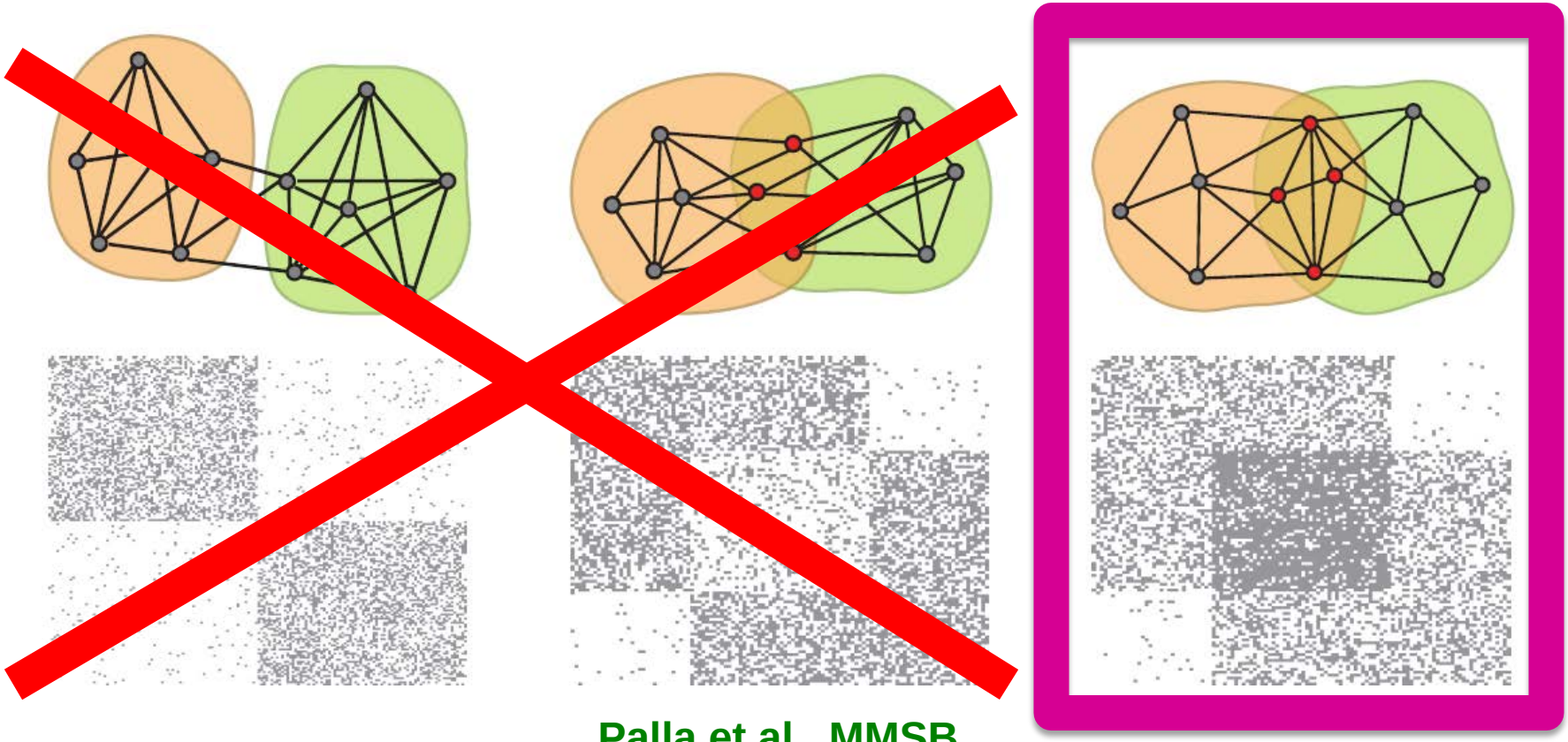


Communities as overlapping tiles

Web of affiliations [Simmel '64]

Communities in Networks

What does this mean?

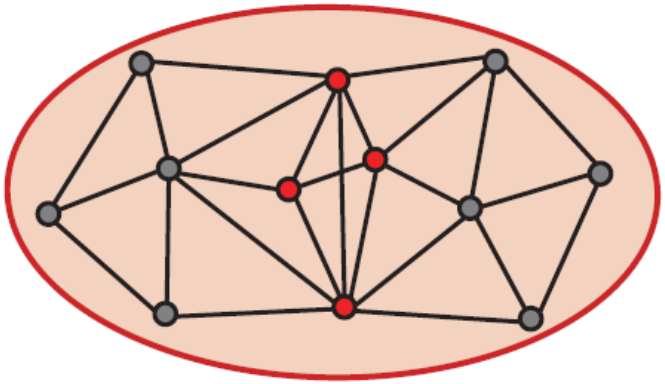


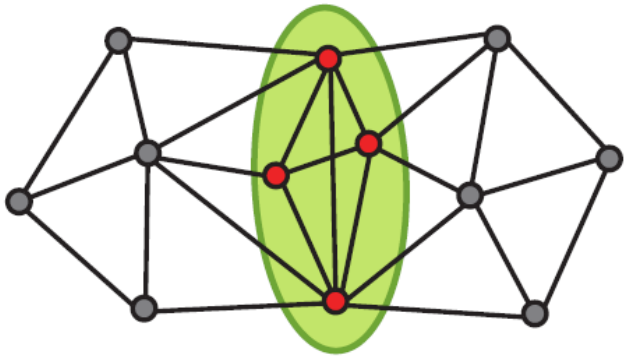
Ganovetter and all non-overlapping methods

Palla et al., MMSB and other overlapping methods as well

Many Methods Fail

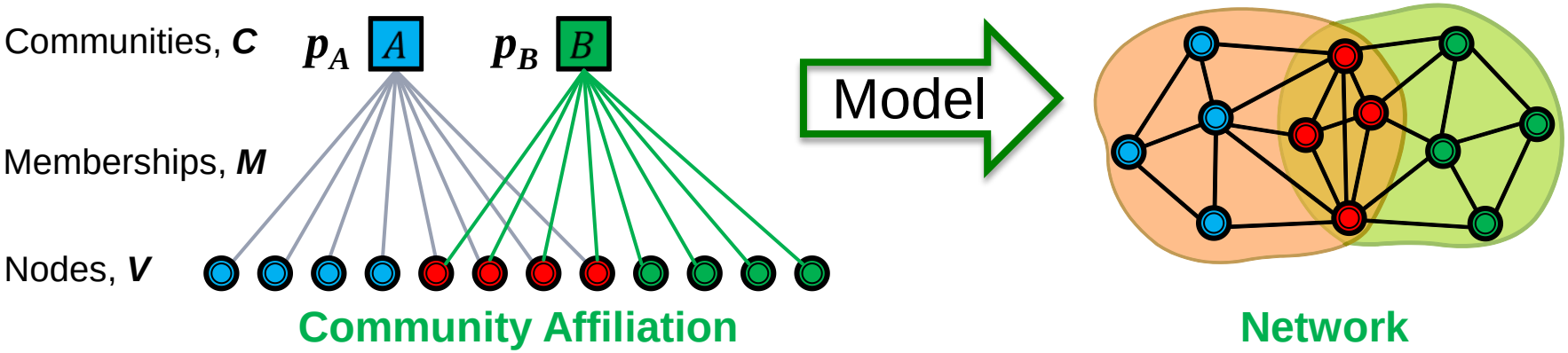
- **Many methods fail to detect dense overlaps:**
  - Clique percolation, …





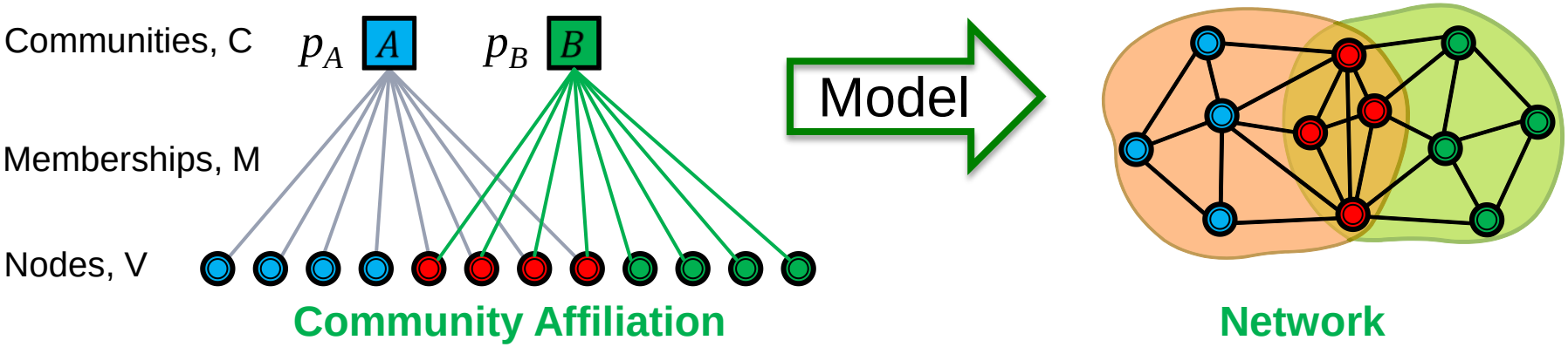
Clique percolation

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

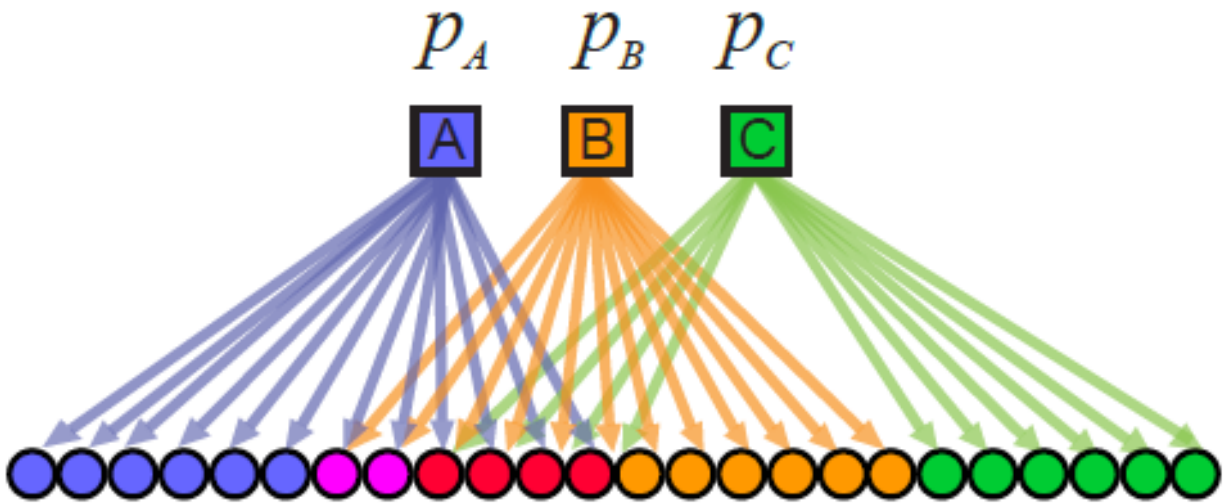
AGM: Generative Process



- **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

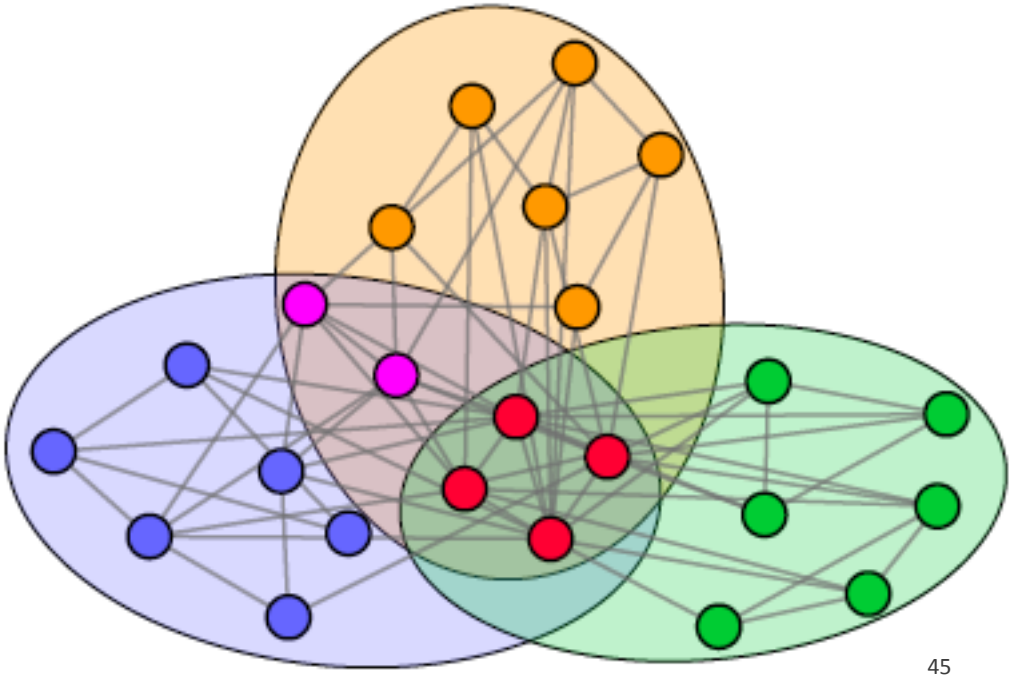
$$p(u,v) = 1 - \prod_{c \in M_u \cap M_v} (1 - p_c)$$

AGM: Dense Overlaps



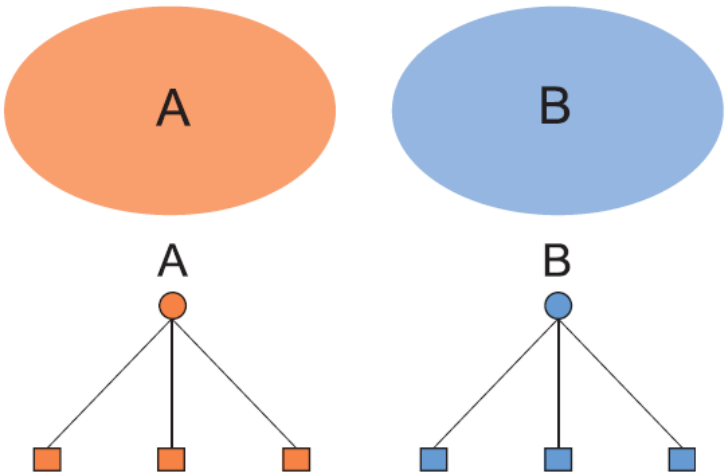
Model

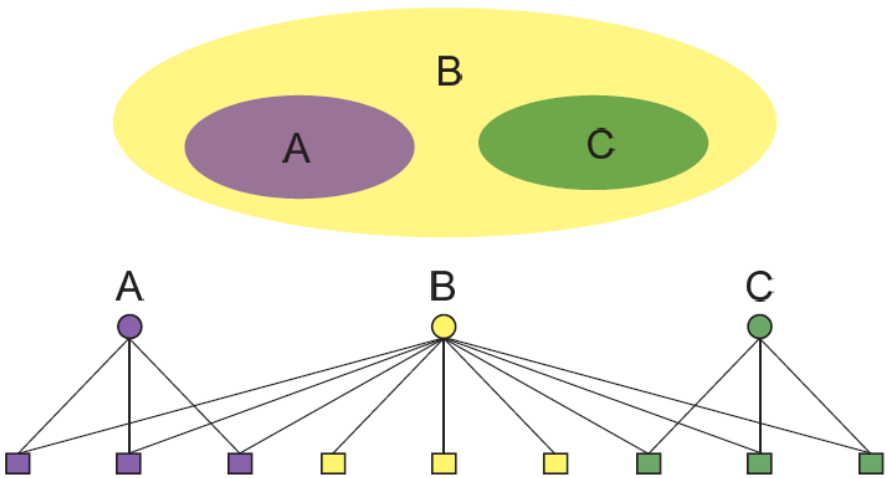
Network

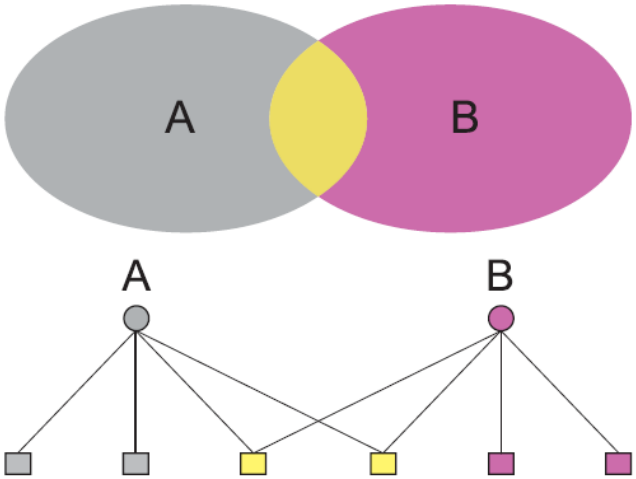


Community-Affiliation Graph Model

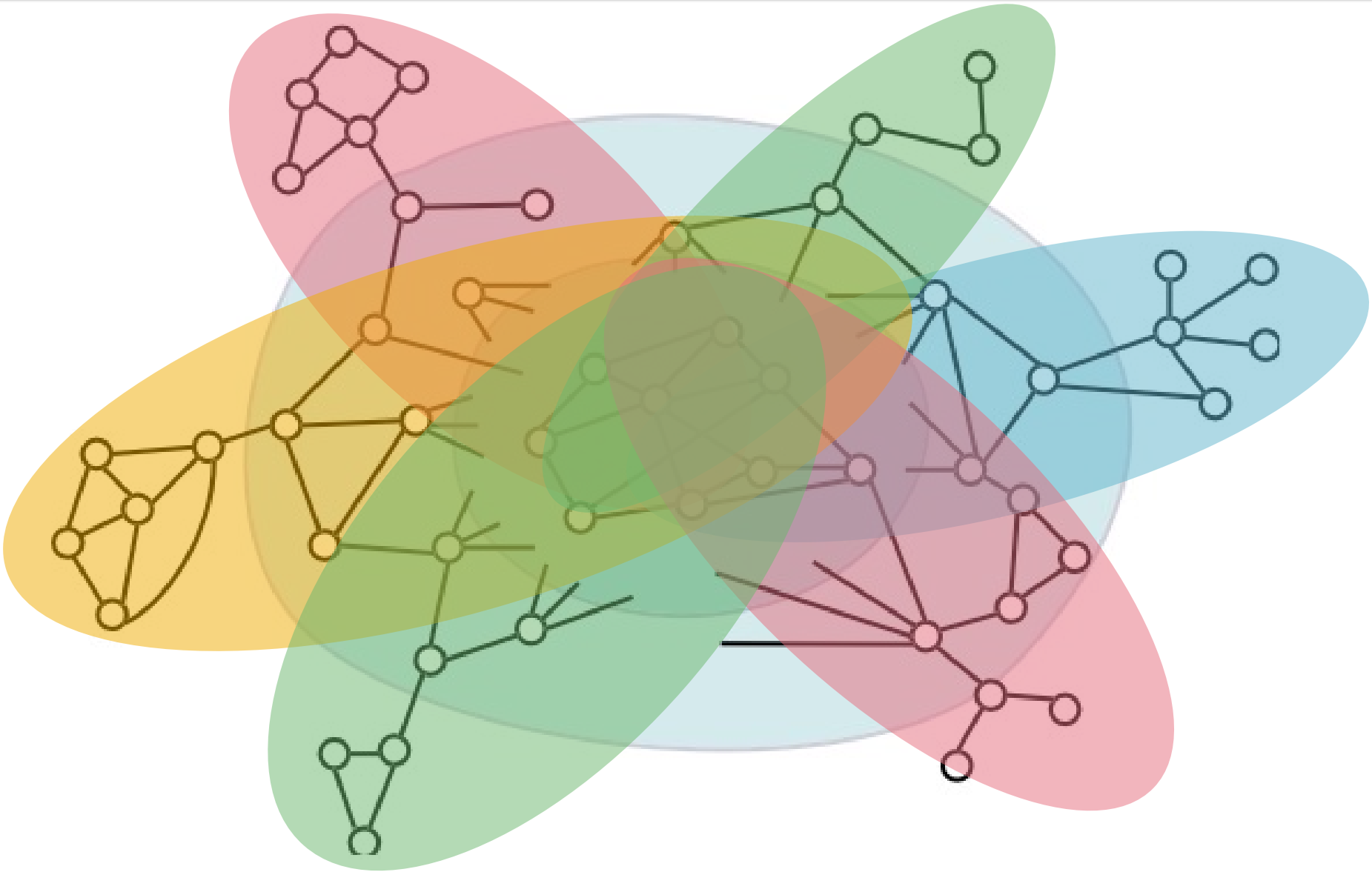
- **AGM is flexible and can express variety of network structures:** Non-overlapping, Nested, Overlapping



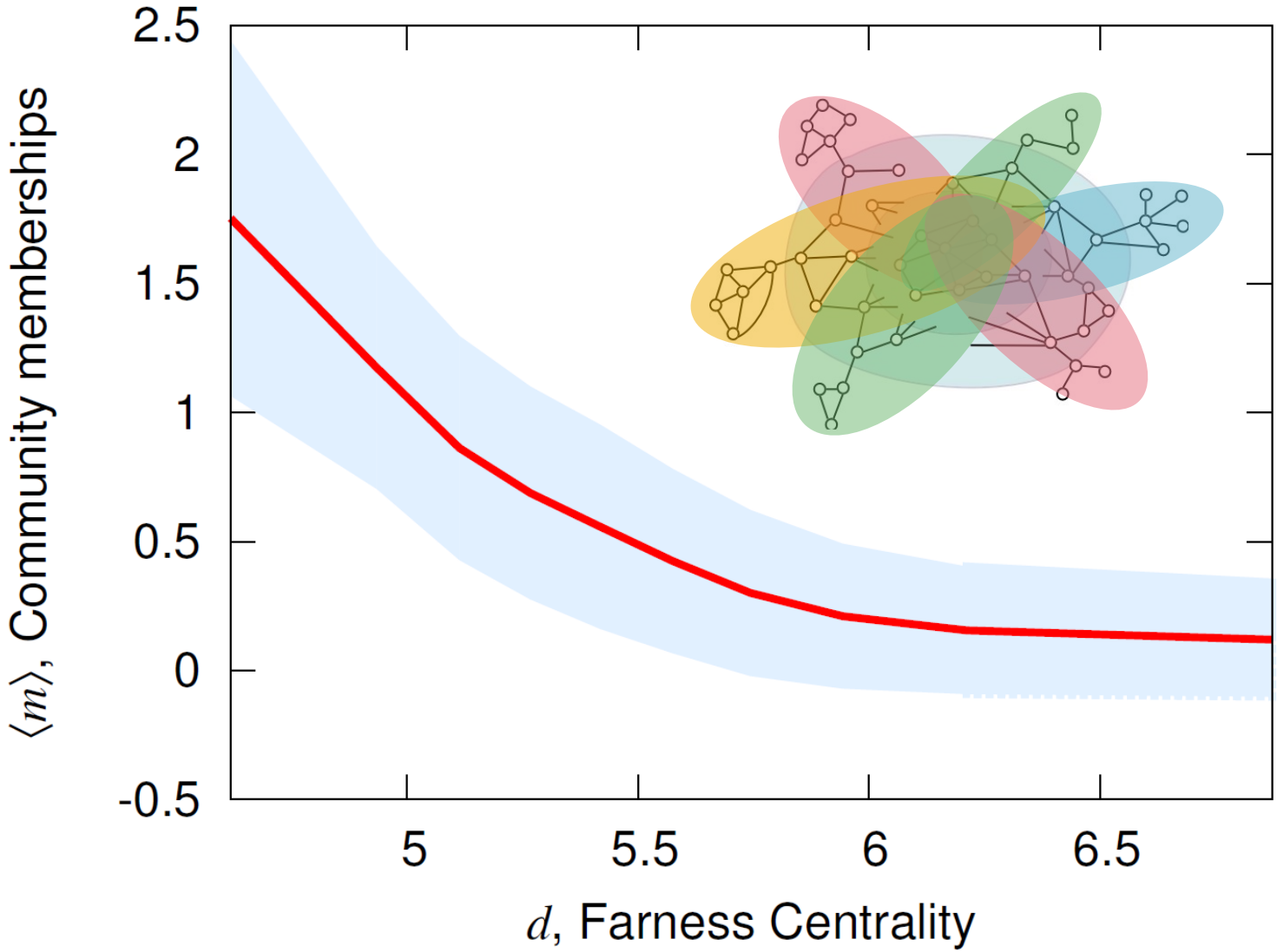




Connections: Core-Periphery

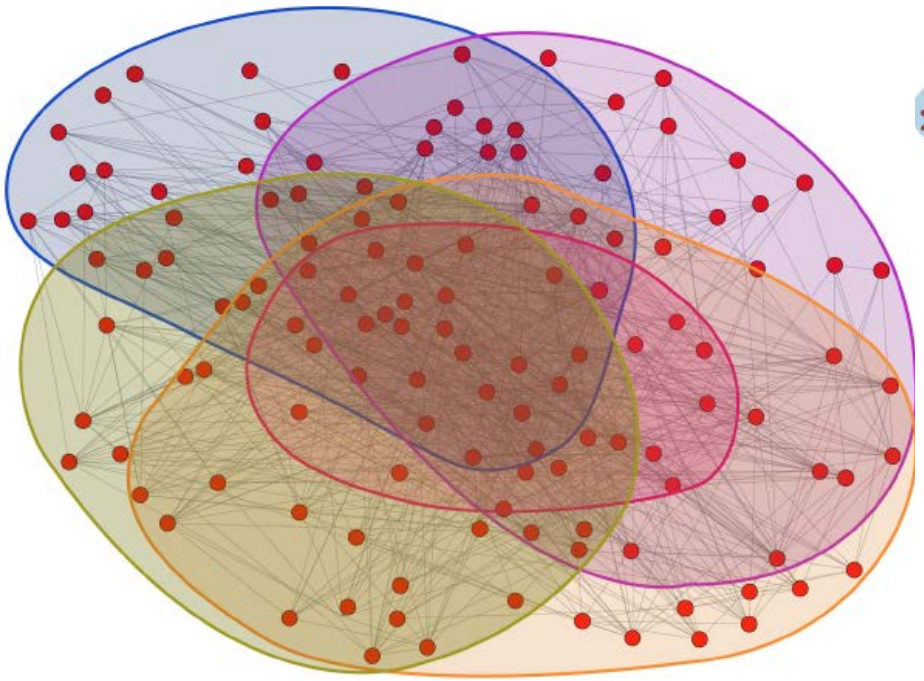


Test of the Conjecture

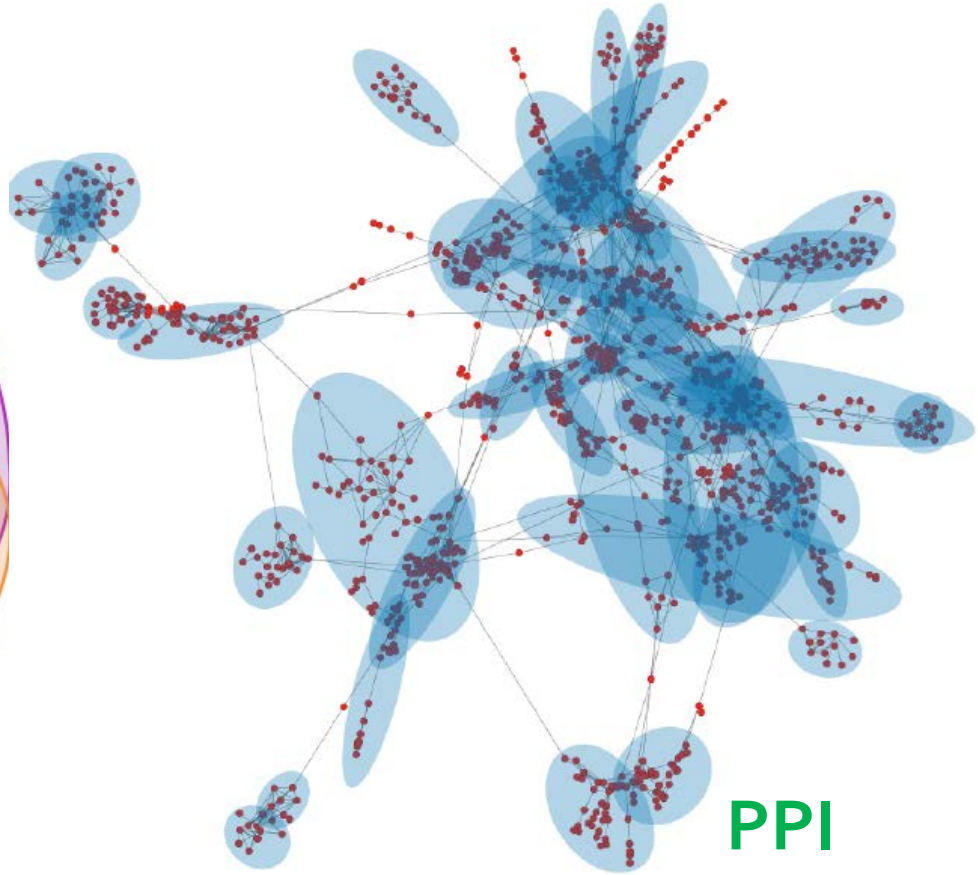


LiveJournal social network

Primary & Secondary Cores



Foodweb



PPI

- **Primary core:** Foodwebs, Web-graph, Social
- **Secondary cores:** PPI, Products