Community Structure in Large Social and Information Networks

Michael W. Mahoney

(Joint work at Yahoo with Kevin Lang and Anirban Dasgupta,
and also Jure Leskovec of CMU.)

(For more info, see: http://www.cs.yale.edu/homes/mmahoney.)

Workshop on Algorithms for Modern Massive Data Sets - MMDS 2008

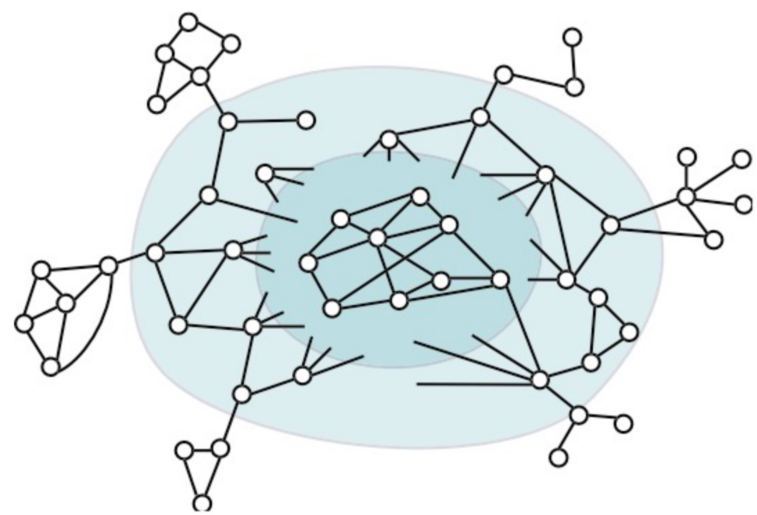
Networks and networked data

Lots of "networked" data!!

- technological networks
  - AS, power-grid, road networks
- biological networks
  - food-web, protein networks
- social networks
  - collaboration networks, friendships
- information networks
  - co-citation, blog cross-postings, advertiser-bidded phrase graphs...
- language networks
  - semantic networks...
- ...

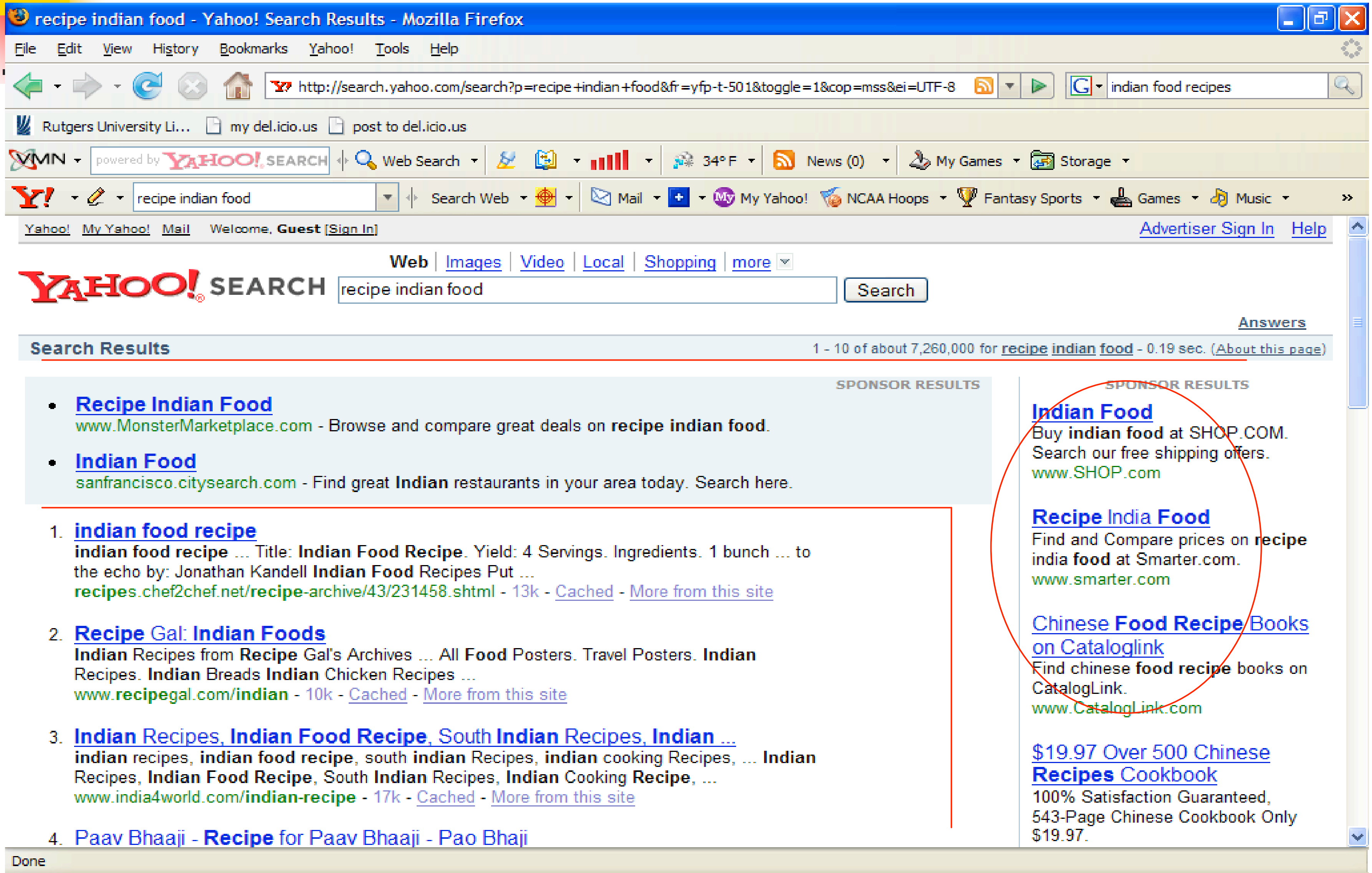
Interaction graph model of networks:

- Nodes represent "entities"
- Edges represent "interaction" between pairs of entities

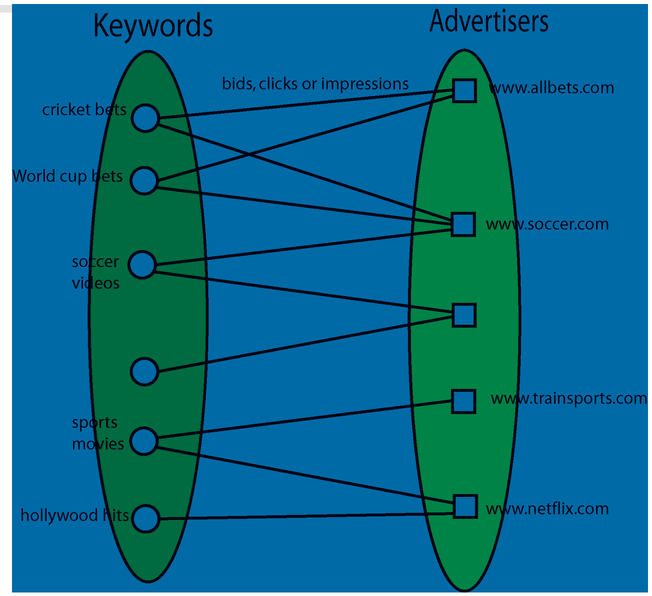


Sponsored ("paid") Search

Text-based ads driven by user query



Sponsored Search Problems



Keyword-advertiser graph:

- provide new ads
- maximize CTR, RPS, advertiser ROI

"Community-related" problems:

- Marketplace depth broadening:

  find new advertisers for a particular query/submarket

- Query recommender system:

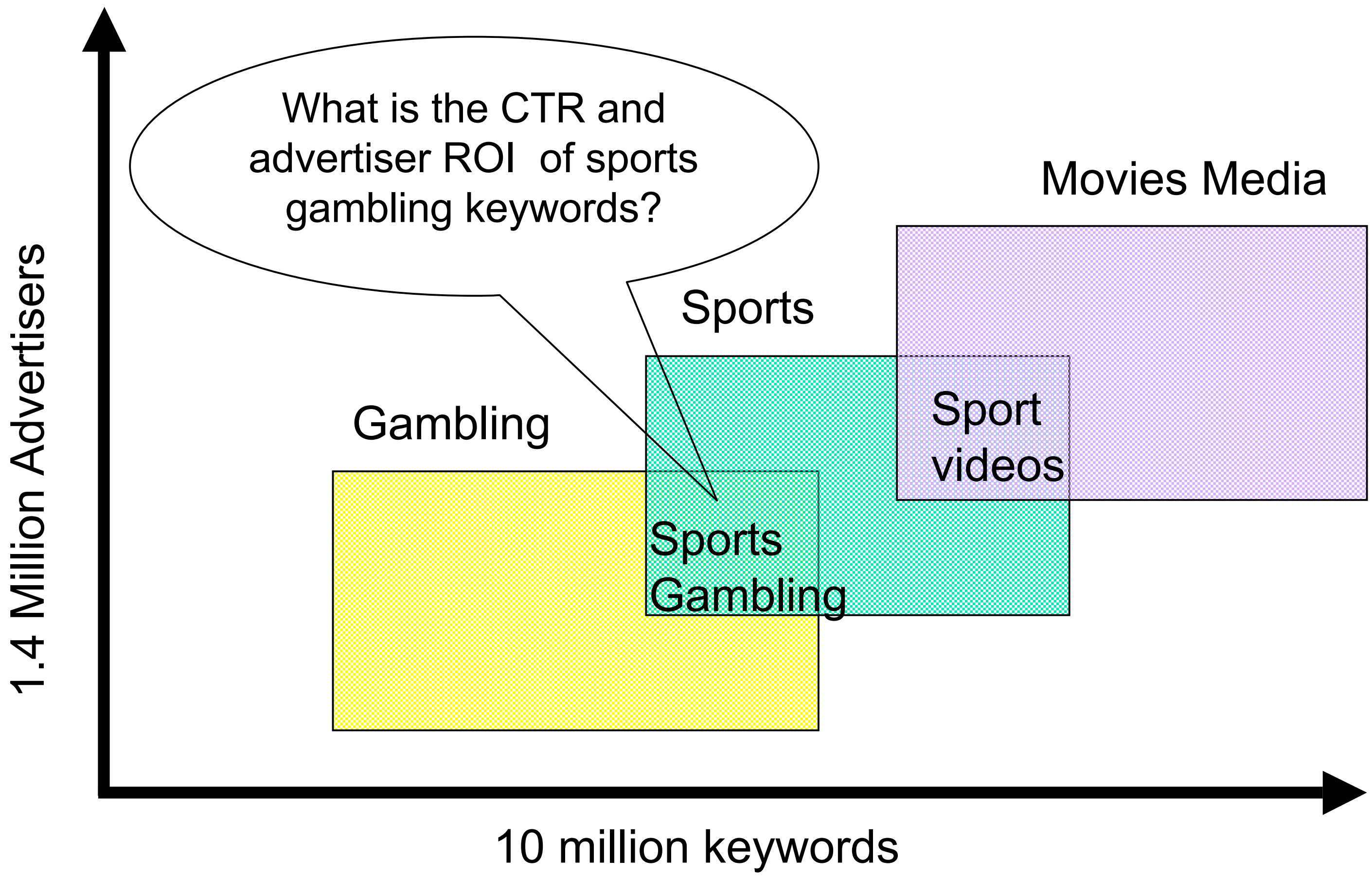
  suggest to advertisers new queries that have high probability of clicks

- Contextual query broadening:

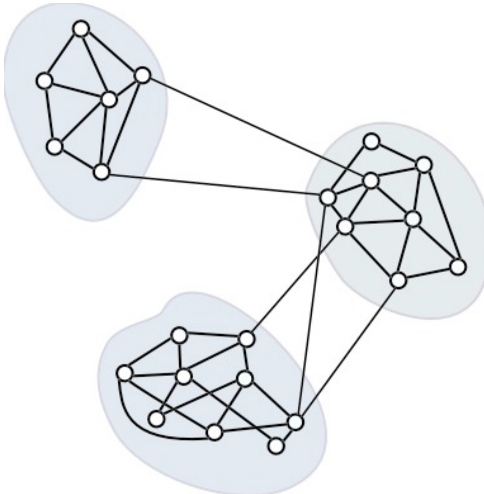
  broaden the user's query using other context information

Micro-markets in sponsored search

Goal: Find *isolated* markets/clusters with *sufficient money/clicks* with *sufficient coherence.*
Ques: Is this even possible?



Clustering and Community Finding



- Linear (Low-rank) methods

  If Gaussian, then low-rank space is good.

- Kernel (non-linear) methods

  If low-dimensional manifold, then kernels are good

- Hierarchical methods

  Top-down and botton-up -- common in the social sciences

- Graph partitioning methods

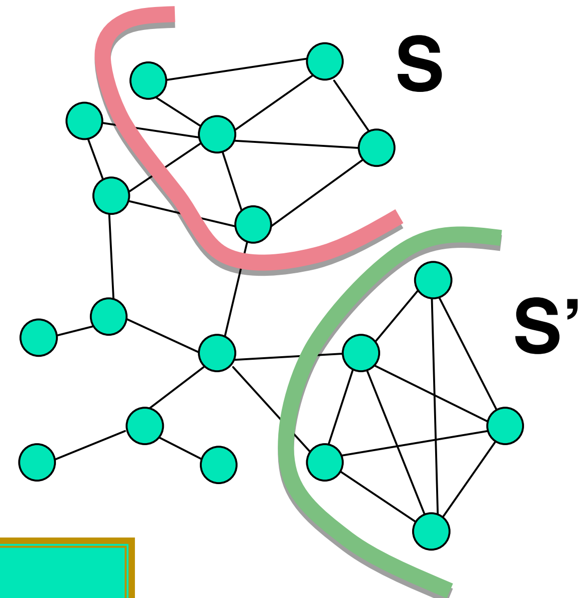
  Define "edge counting" metric -- conductance, expansion, modularity, etc. -- in interaction graph, then optimize!

"It is a matter of common experience that communities exist in networks ... Although not precisely defined, communities are usually thought of as sets of nodes with better connections amongst its members than with the rest of the world."

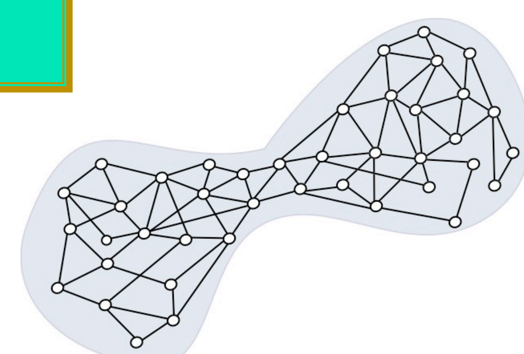
Community Score: Conductance

- How community like is a set of nodes?
- Need a natural intuitive measure:
- **Conductance** (normalized cut)

$\phi(S)$ = # edges cut / # edges inside

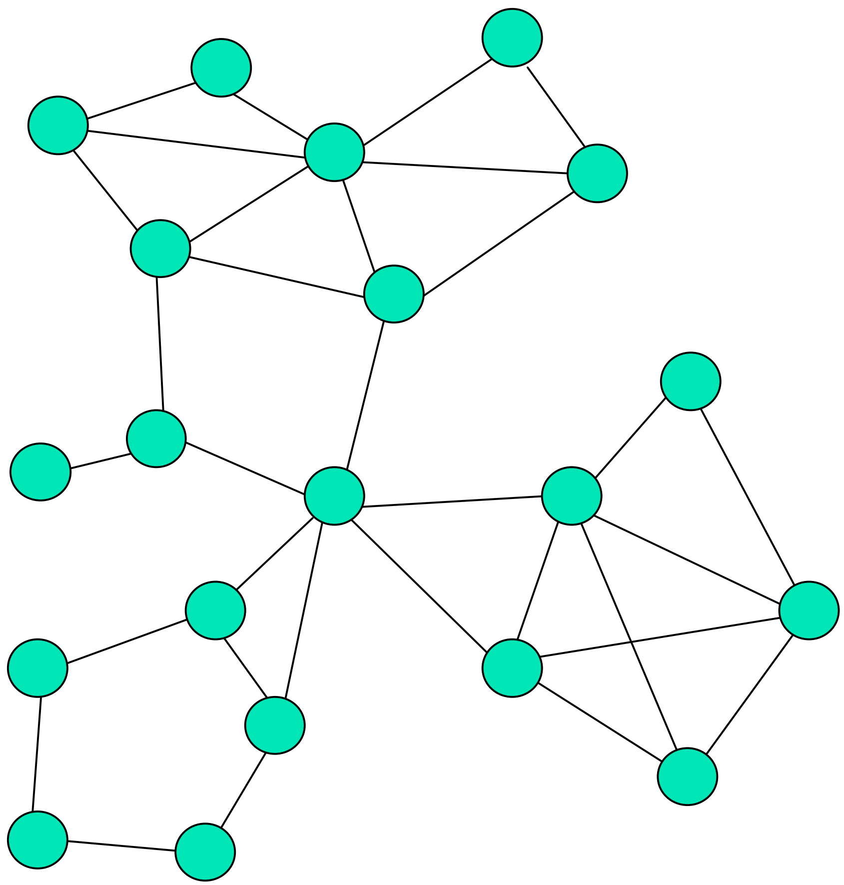


- Small $\phi(S)$ corresponds to more community-like sets of nodes



Community Score: Conductance

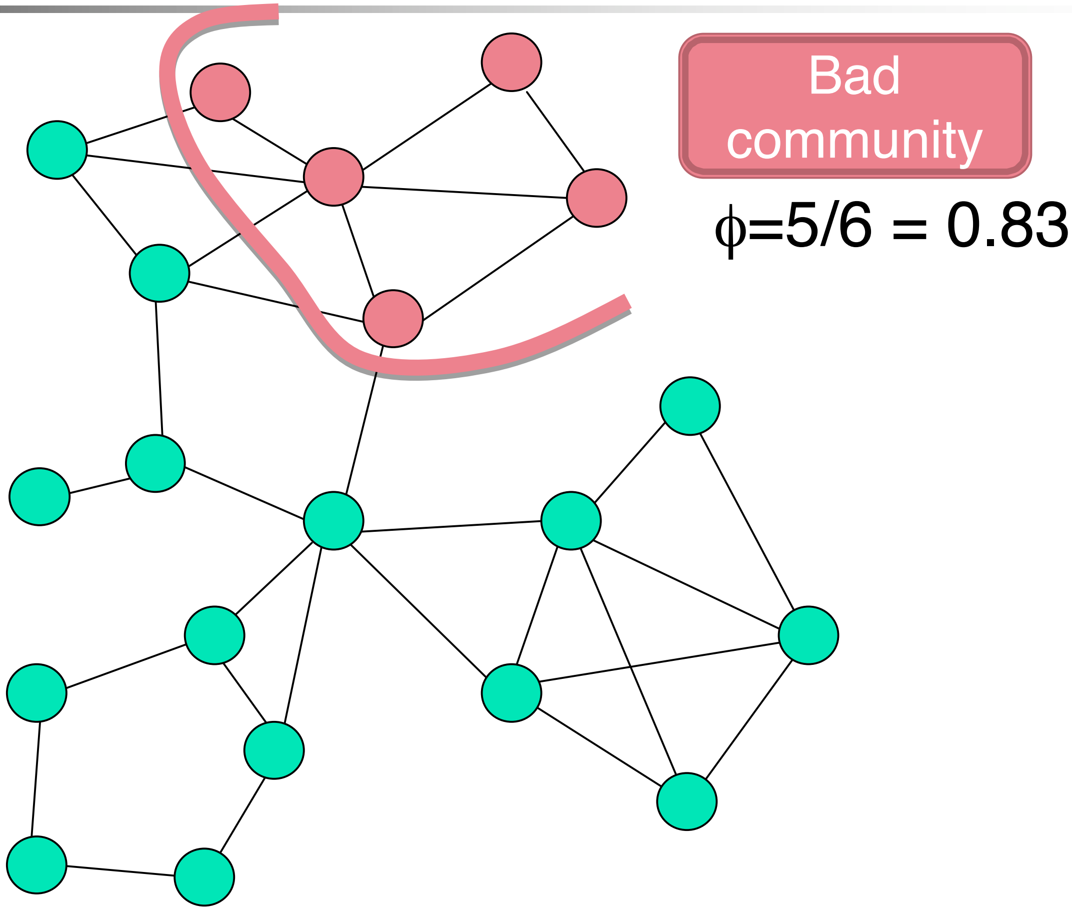
What is "best" community of 5 nodes?



Score: $\phi(S)$ **= # edges cut / # edges inside**

Community Score: Conductance

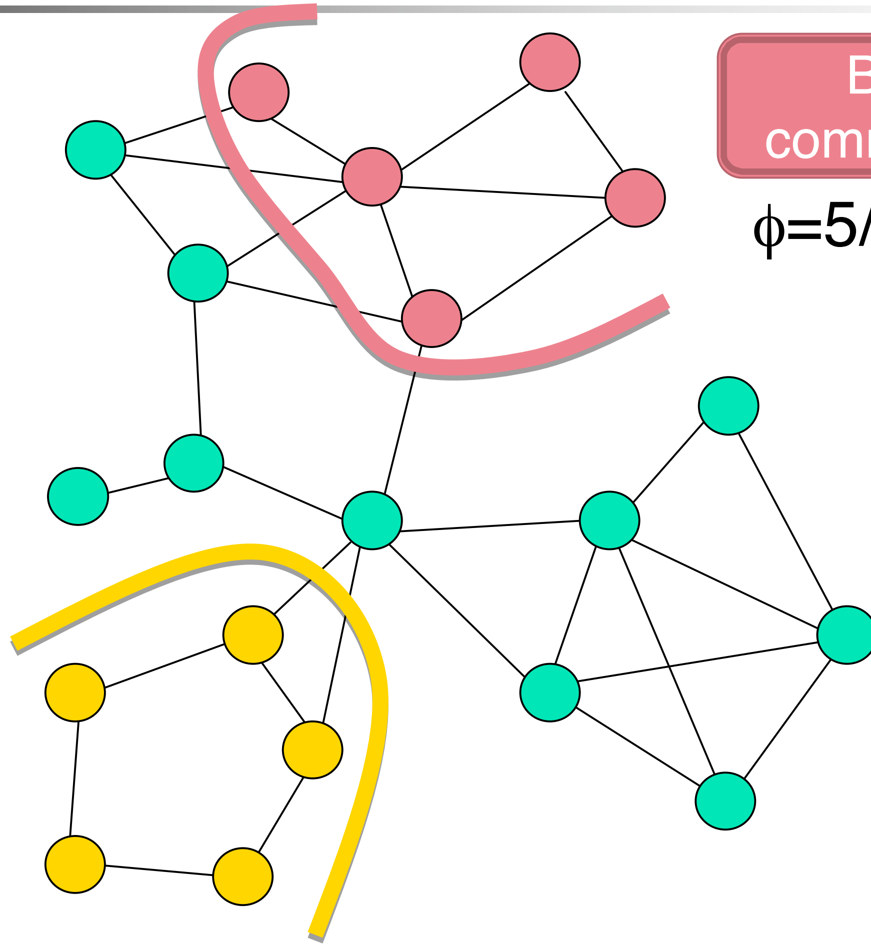
What is "best" community of 5 nodes?



Score: $\phi(S)$ **= # edges cut / # edges inside**

Community Score: Conductance

What is "best" community of 5 nodes?



Bad community

ϕ=5/6 = 0.83

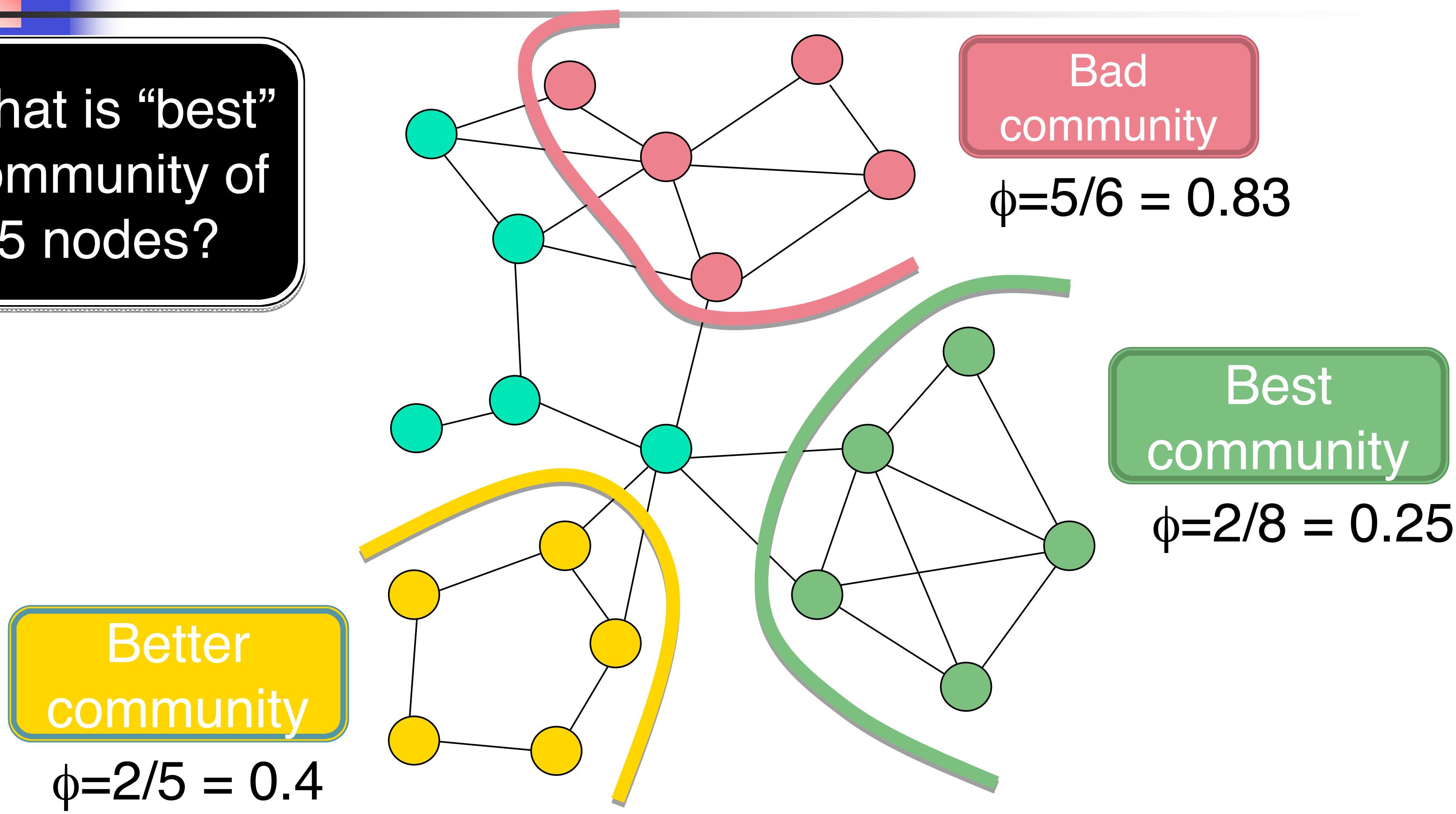
Better community

ϕ=2/5 = 0.4

Score: ϕ(S) = # edges cut / # edges inside

Community Score: Conductance

What is "best" community of 5 nodes?



Score: $\phi(S)$ = # edges cut / # edges inside

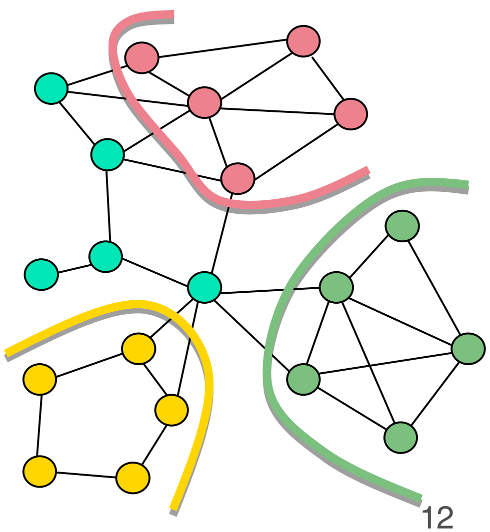
Network Community Profile Plot

- We define:

  Network community profile (NCP) plot

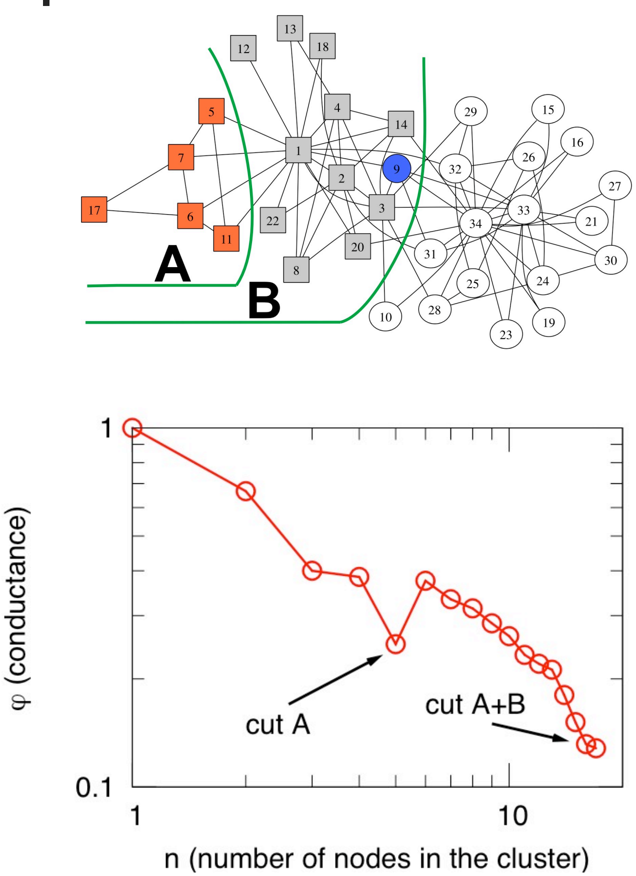
  Plot the score of best community of size k

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

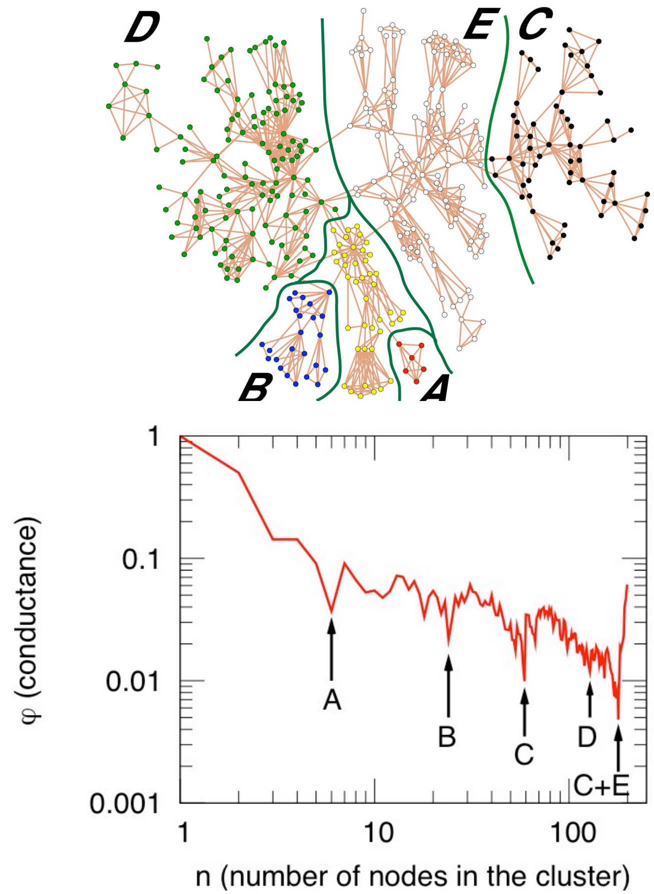


- Search over all subsets of size k and find best: $\phi(k=5) = 0.25$
- NCP plot is intractable to compute
- Use approximation algorithms

Widely-studied small social networks

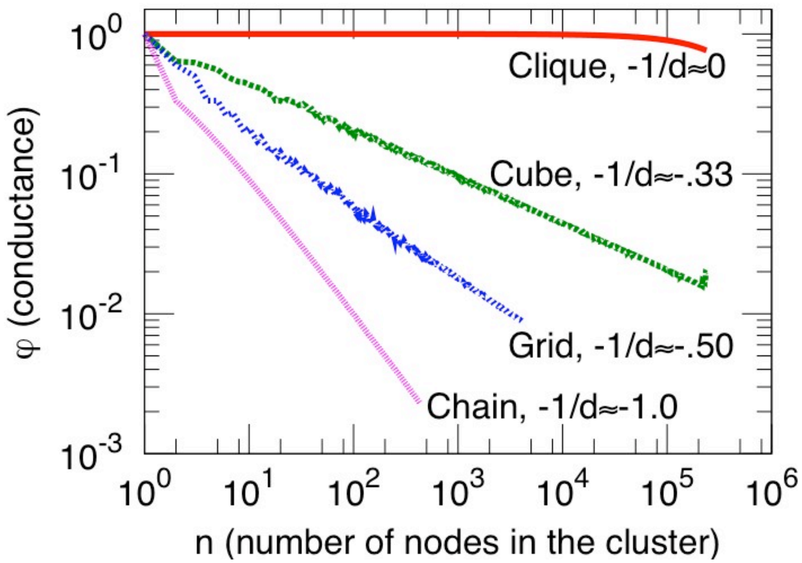


Zachary’s karate club

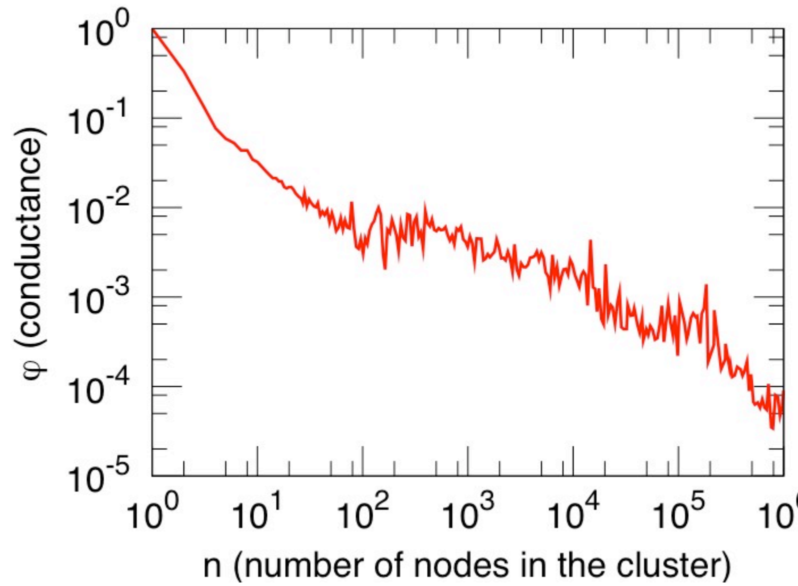


Newman’s Network Science

"Low-dimensional" graphs (and expanders)

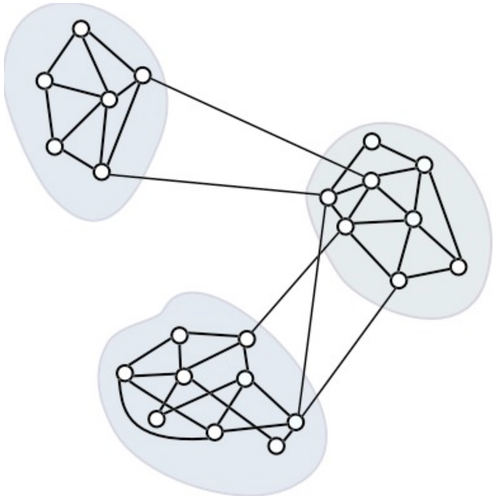


d-dimensional meshes



RoadNet-CA

What do large networks look like?



Downward sloping NCPP

- small social networks (validation)
- "low-dimensional" networks (intuition)
- hierarchical networks (model building)

Natural interpretation in terms of isoperimetry

- implicit in modeling with low-dimensional spaces, manifolds, k-means, etc.

Large social/information networks are very *very* different

- We examined more than 70 large social and information networks
- We developed principled methods to interrogate large networks
- Previous community work: on small social networks (hundreds, thousands)

Large Social and Information Networks

• Social nets	Nodes	Edges	Description
LiveJournal	4,843,953	42,845,684	Blog friendships [4]
Epinions	75,877	405,739	Who-trusts-whom [35]
Flickr	404,733	2,110,078	Photo sharing [21]
Delicious	147,567	301,921	Collaborative tagging
CA-DBLP	317,080	1,049,866	Co-authorship (CA) [4]
CA-cond-mat	21,363	91,286	CA cond-mat [25]
• Information networks			
Cit-hep-th	27,400	352,021	hep-th citations [13]
Blog-Posts	437,305	565,072	Blog post links [28]
• Web graphs			
Web-google	855,802	4,291,352	Web graph Google
Web-wt10g	1,458,316	6,225,033	TREC WT10G web
• Bipartite affiliation (authors-to-papers) networks			
AtP-DBLP	615,678	944,456	DBLP [25]
AtP-astro-ph	54,498	131,123	Arxiv `astro-ph` [25]
• Internet networks			
AS	6,474	12,572	Autonomous systems
Gnutella	62,561	147,878	P2P network [36]

Table 1: Some of the network datasets we studied.

Probing Large Networks with Approximation Algorithms

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

Spectral - (quadratic approx) - confuses "long paths" with "deep cuts"

Multi-commodity flow - (log(n) approx) - difficulty with expanders

SDP - (sqrt(log(n)) approx) - best in theory

Metis - (multi-resolution for mesh-like graphs) - common in practice

X+MQI - post-processing step on, e.g., Spectral of Metis

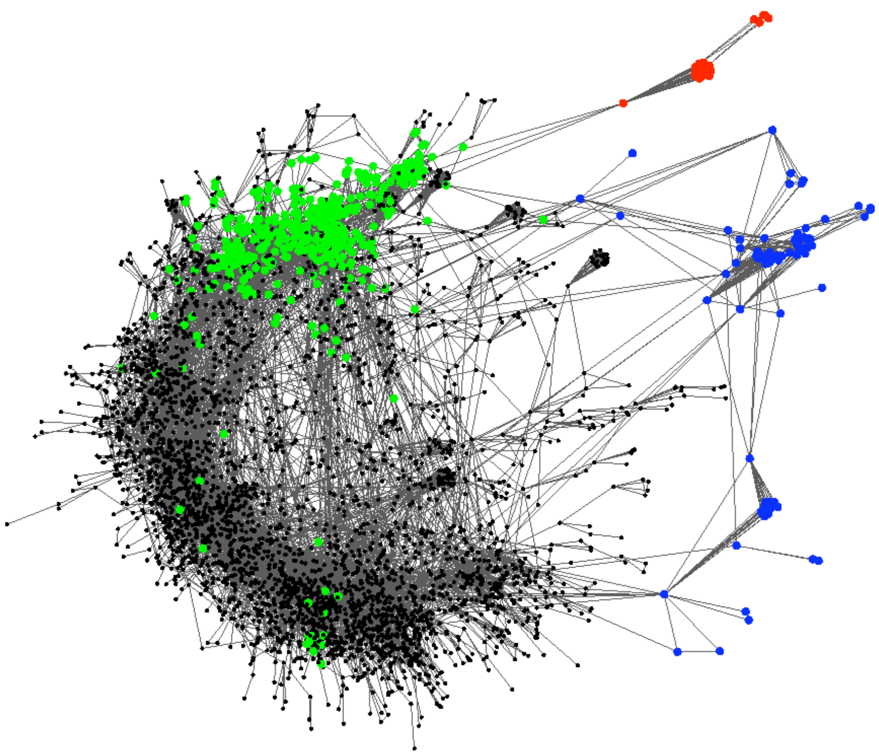
Metis+MQI - best conductance (empirically)

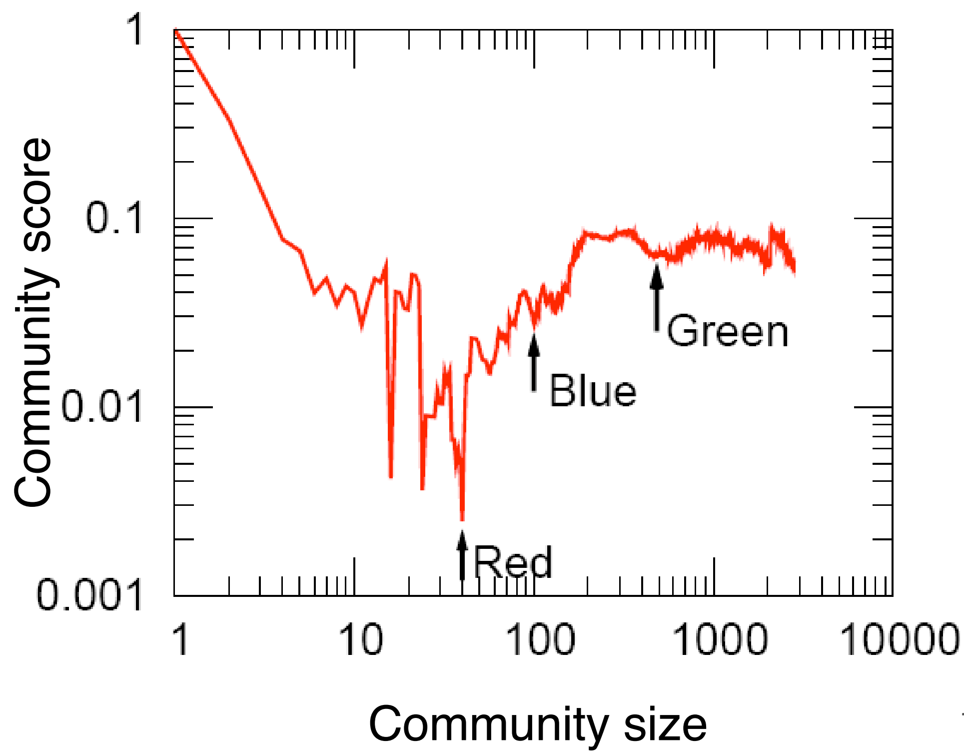
Local Spectral - connected and tighter sets (empirically, regularized communities!)

We are not interested in partitions per se, but in probing network structure.

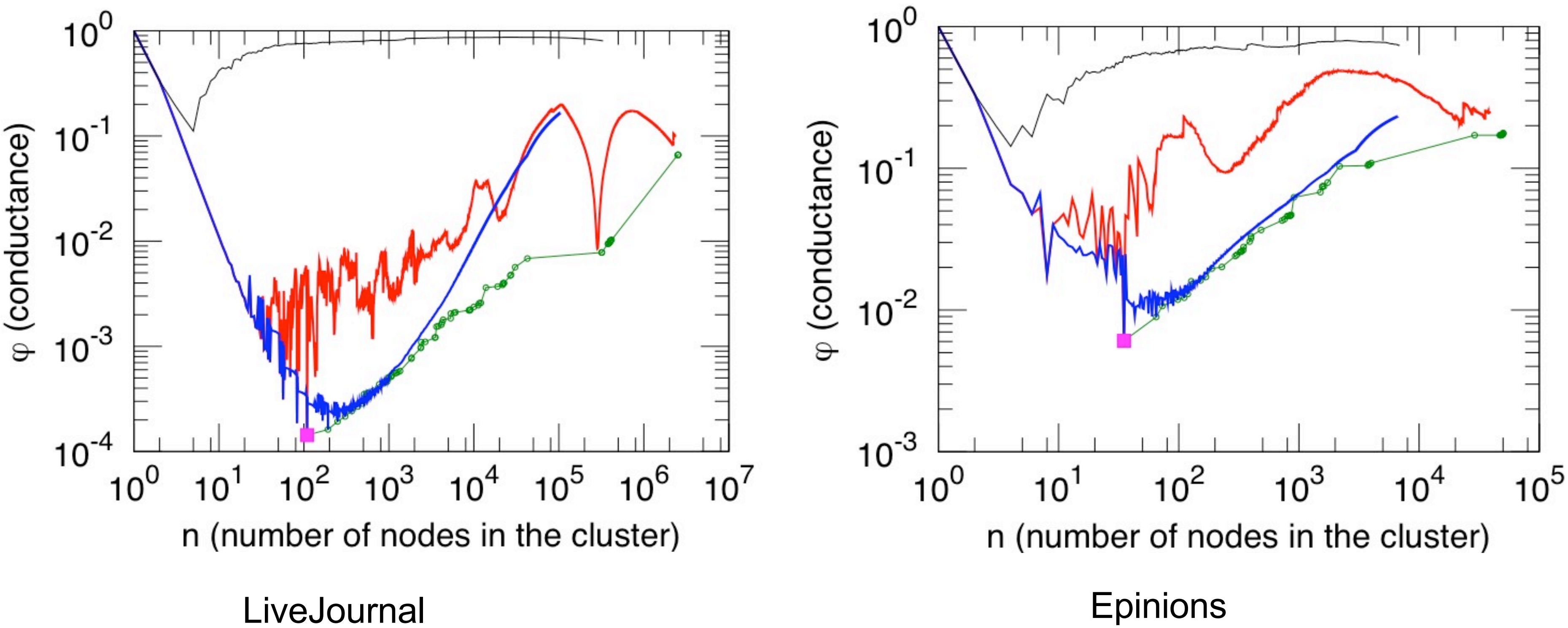
Typical example of our findings

General relativity collaboration network
(4,158 nodes, 13,422 edges)



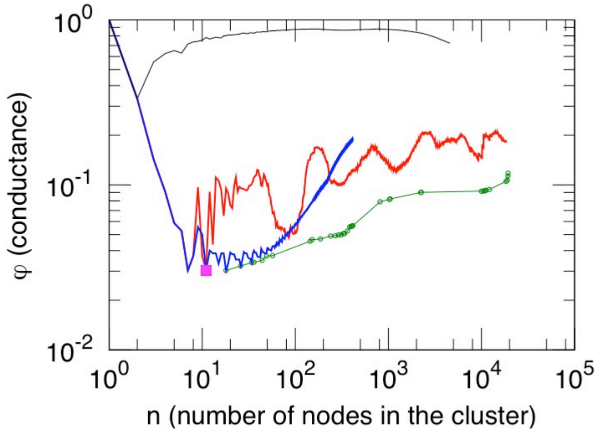


Large Social and Information Networks

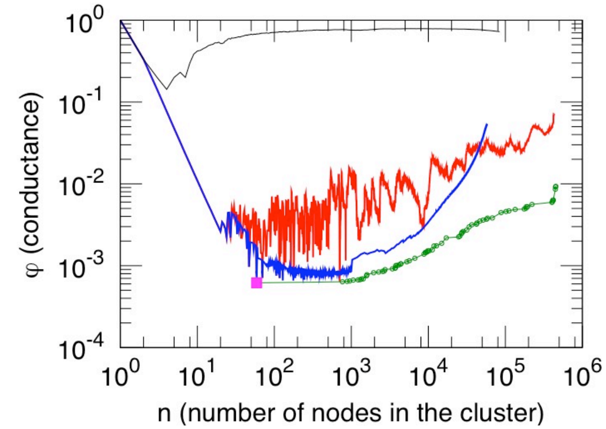


Focus on the red curves (local spectral algorithm) - blue (Metis+Flow), green (Bag of whiskers), and black (randomly rewired network) for consistency and cross-validation.

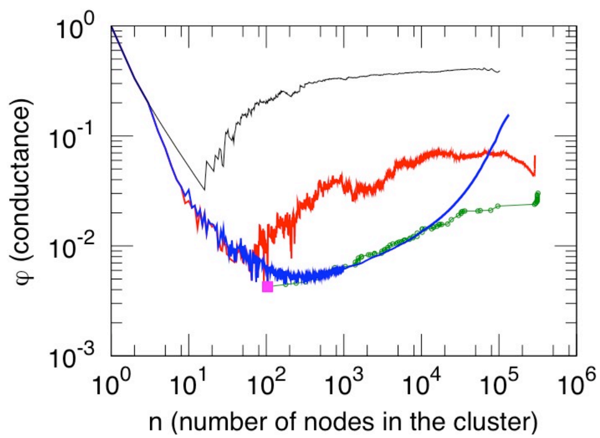
More large networks



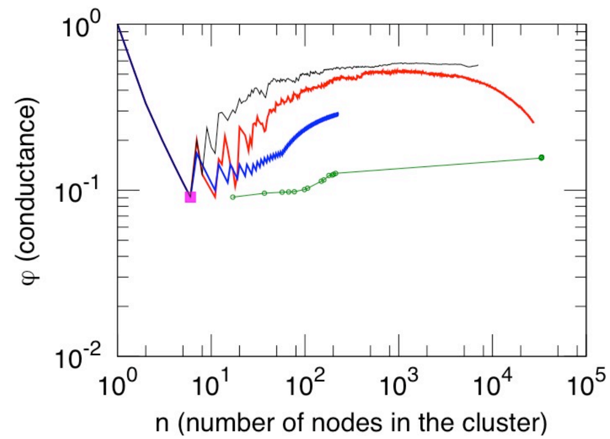
Cit-Hep-Th



Web-Google

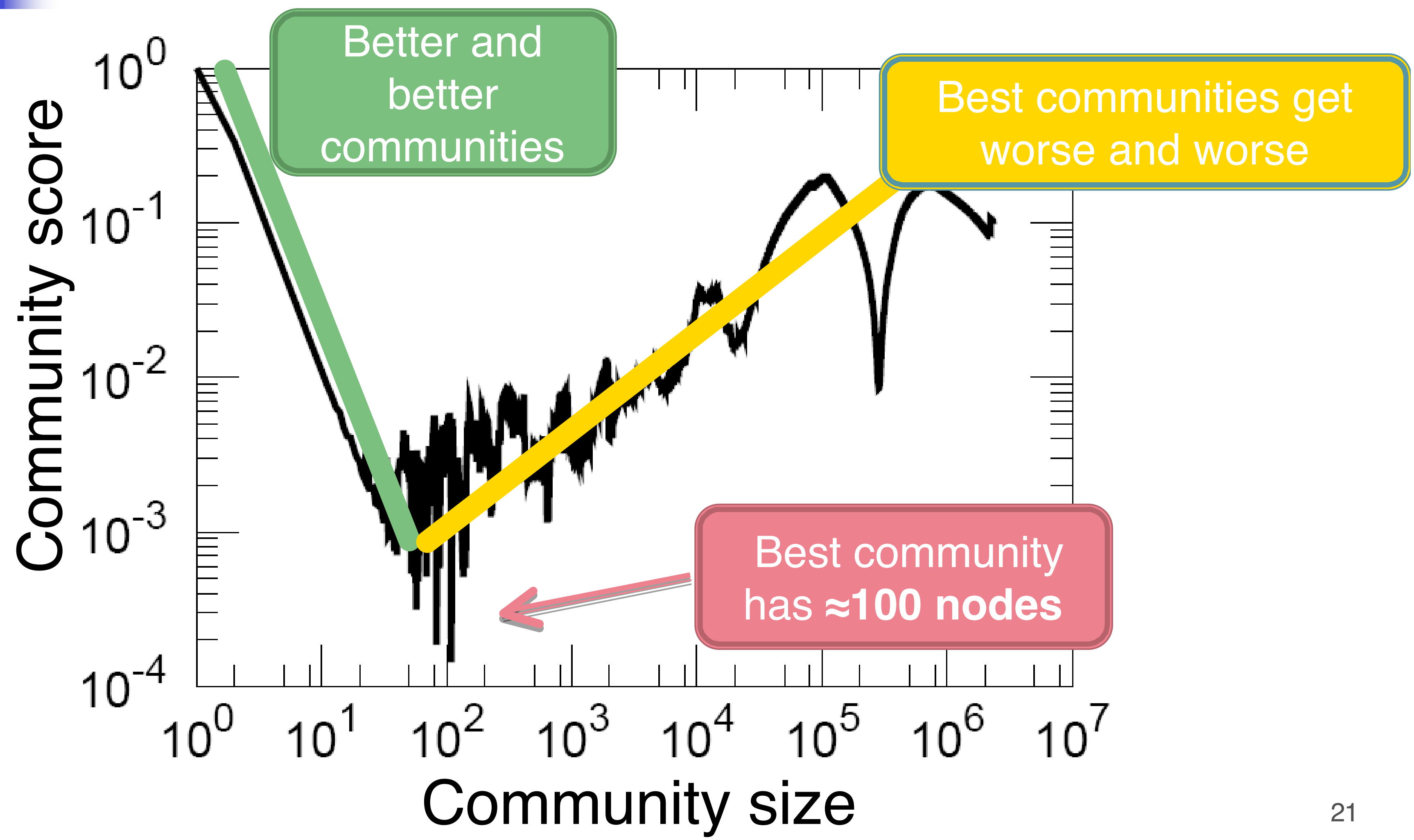


AtP-DBLP

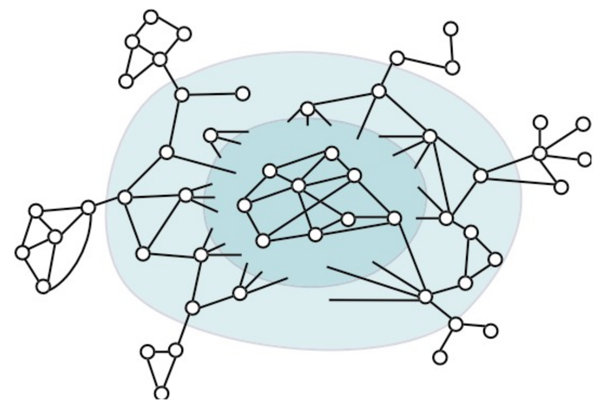


Gnutella

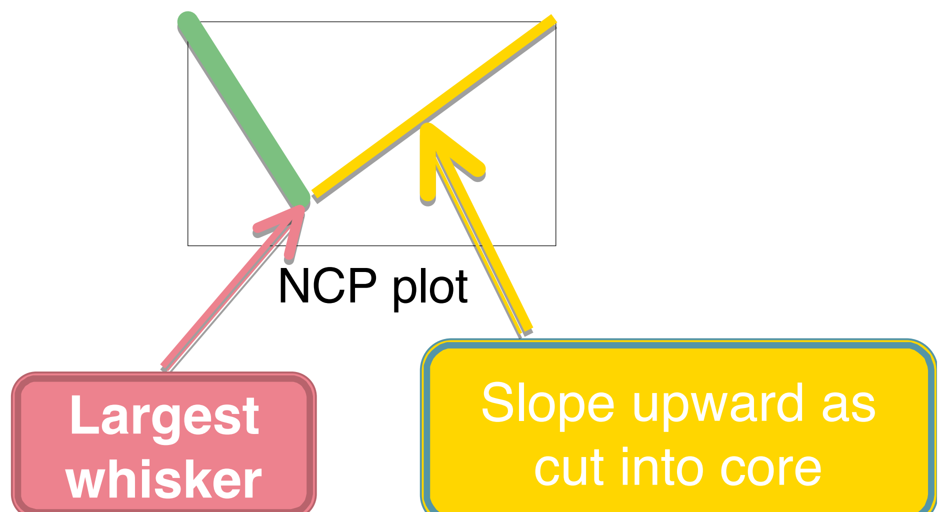
NCPP: LiveJournal (N=5M, E=43M)



“Whiskers” and the “core”



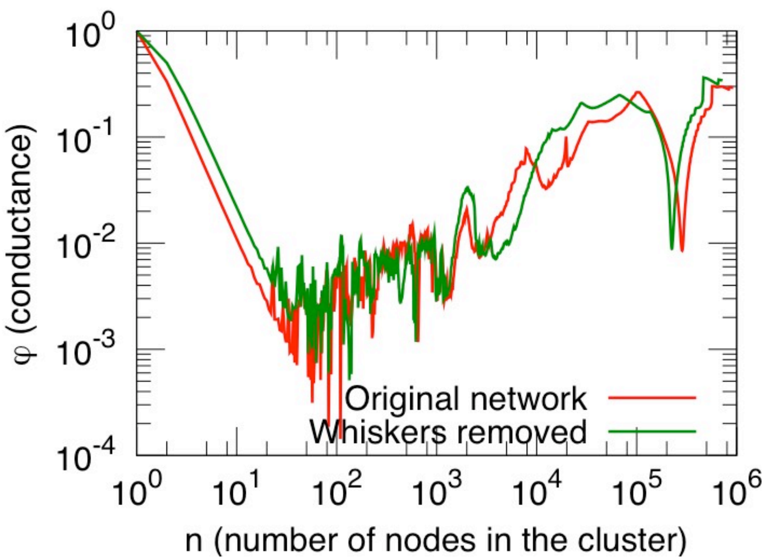
- “Whiskers”
  - maximal sub-graph detached from network by removing a single edge
  - contains 40% of nodes and 20% of edges
- “Core”
  - the rest of the graph, i.e., the 2-edge-connected core
- Global minimum of NCPP is a whisker



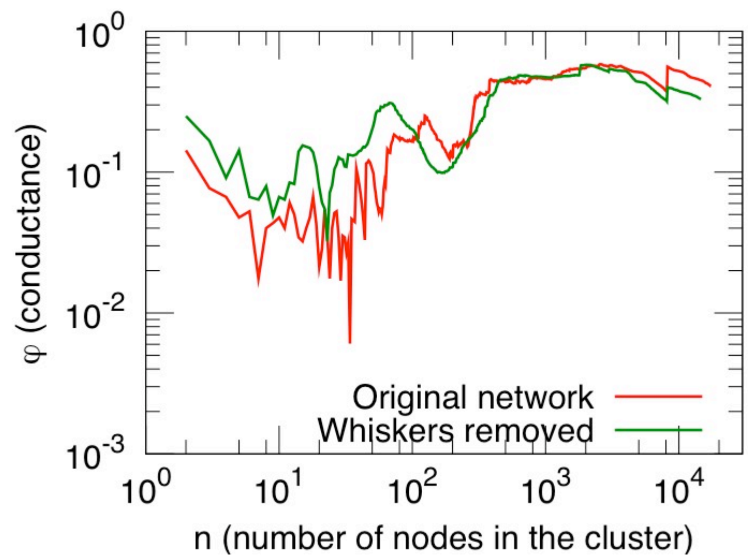
What if the "whiskers" are removed?

Then the lowest conductance sets - the "best" communities - are "2-whiskers."

(So, the "core" peels apart like an onion.)

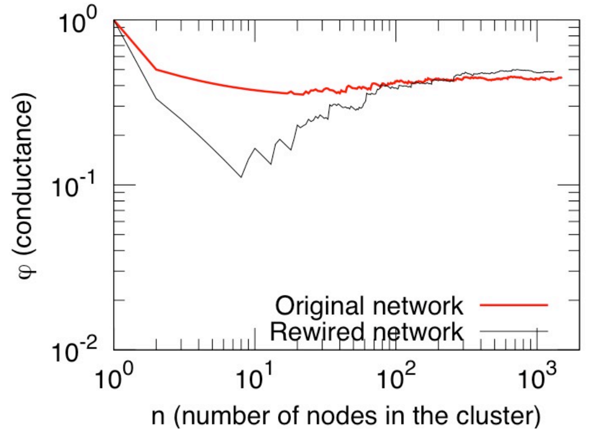


LiveJournal

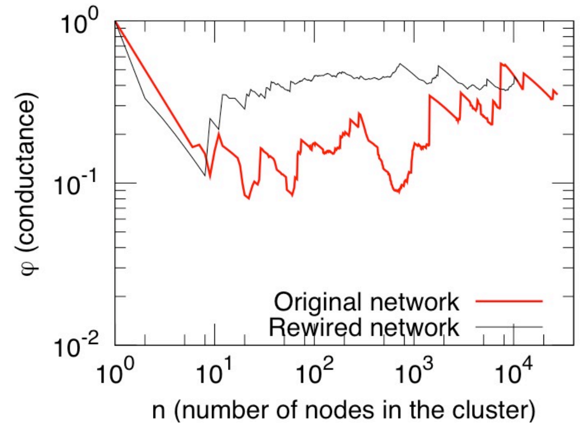


Epinions

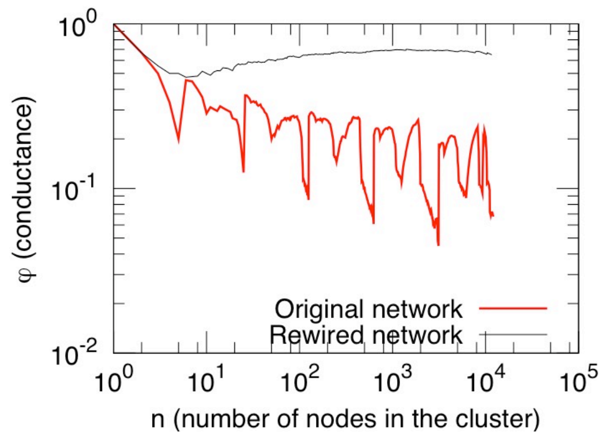
NCPP for common generative models



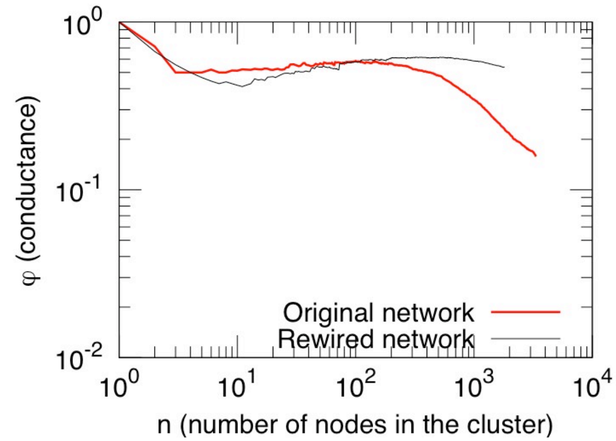
Preferential Attachment



Copying Model



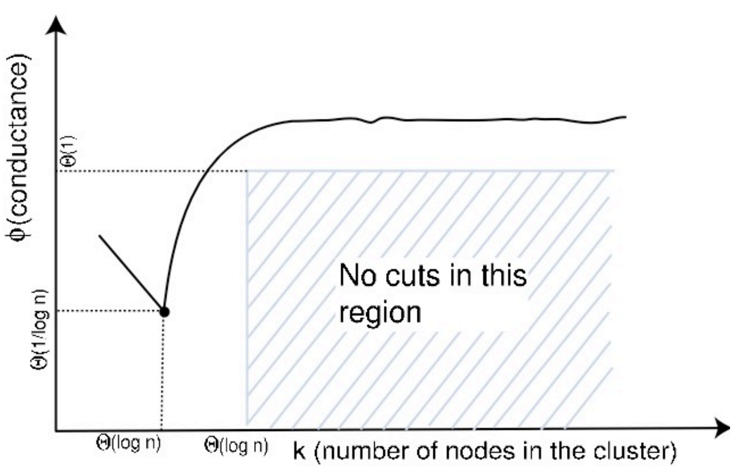
RB Hierarchical



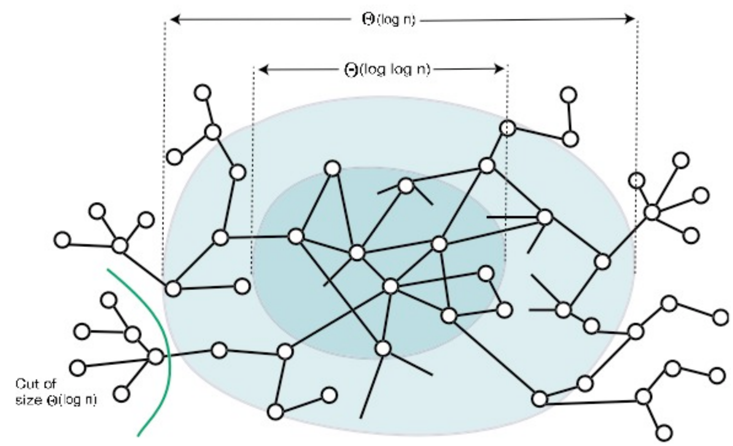
Geometric PA

A simple theorem on random graphs

Let $\mathbf{w} = (w_1, \ldots, w_n)$, where
$w_i = ci^{-1/(\beta-1)}, \quad \beta \in (2,3)$.
Connect nodes i and j w.p.
$p_{ij} = w_i w_j / \sum_k w_k$.



Power-law random graph with β ε (2,3).



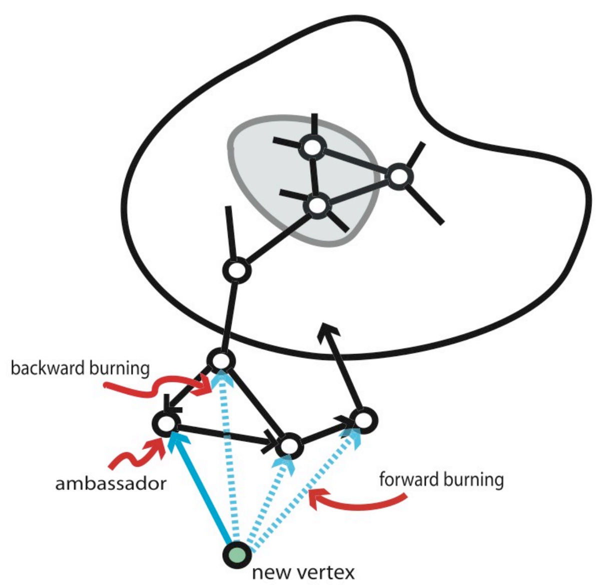
Structure of the G(w) model, with β ε (2,3).

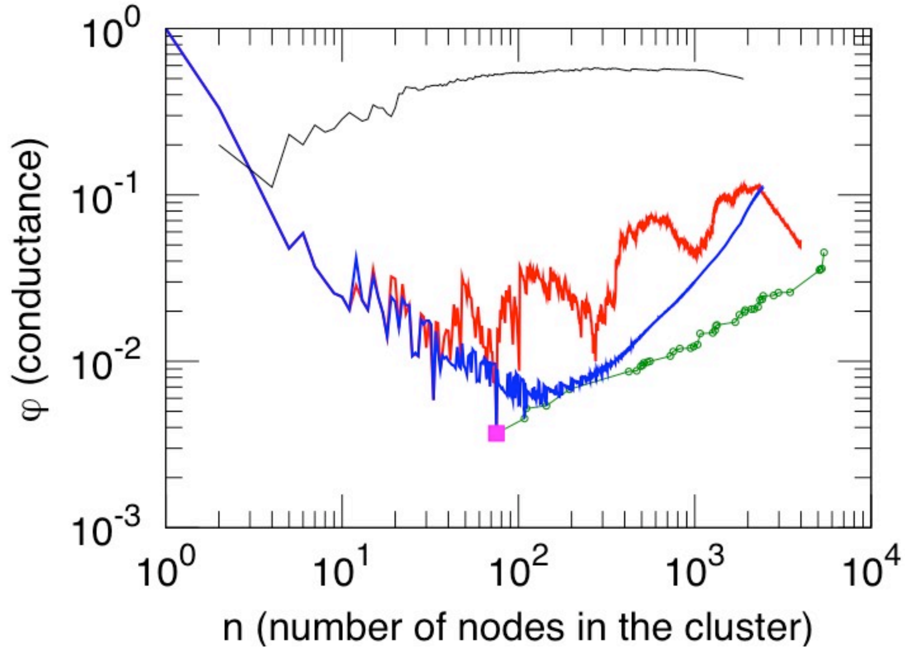
- Sparsity (coupled with randomness) is the issue, *not* heavy-tails.
- (Power laws with β ε (2,3) give us the appropriate sparsity.)

A "forest fire" model

Model of: Leskovec, Kleinberg, and Faloutsos 2005

At each time step, iteratively add edges with a "forest fire" burning mechanism.





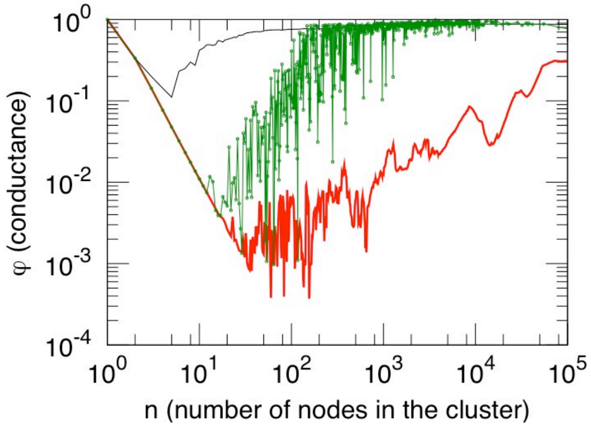
Also get "densification" and "shrinking diameters" of real graphs with these parameters (Leskovec et al. 05).

Comparison with "Ground truth" (1 of 2)

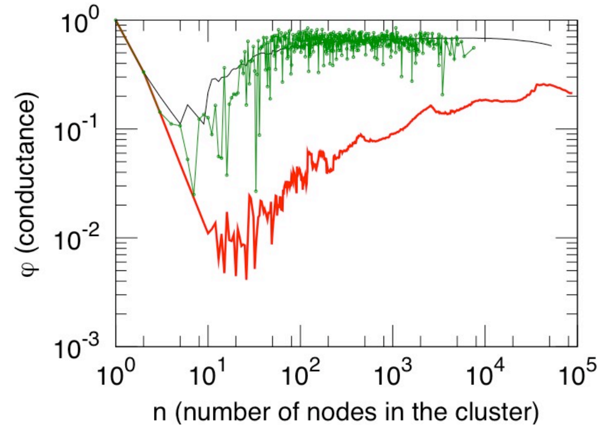
Networks with "ground truth" communities:

- LiveJournal12:
  - users create and explicitly join on-line groups
- CA-DBLP:
  - publication venues can be viewed as communities
- AmazonAllProd:
  - each item belongs to one or more hierarchically organized categories, as defined by Amazon
- AtM-IMDB:
  - countries of production and languages may be viewed as communities (thus every movie belongs to exactly one community and actors belongs to all communities to which movies in which they appeared belong)

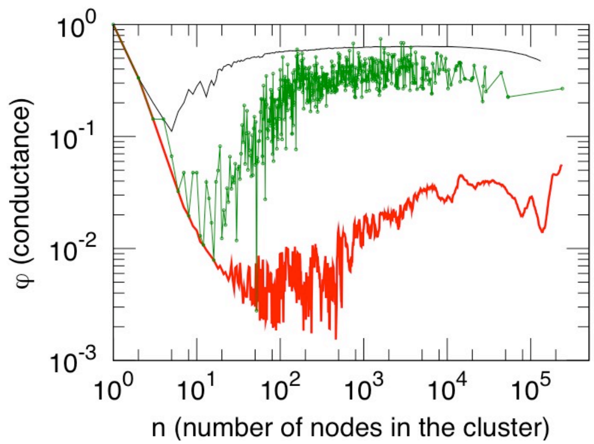
Comparison with “Ground truth” (2 of 2)



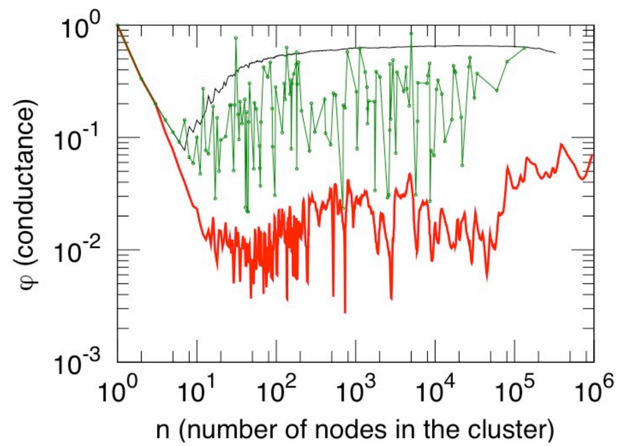
LiveJournal



CA-DBLP



AmazonAllProd



AtM-IMDB

Miscellaneous thoughts …

Sociological work on community size (Dunbar and Allen)

- 150 individuals is maximum community size
- Military companies, on-line communities, divisions of corporations all ≤ 150

Common bond vs. common identity theory

- Common bond - people are attached to individual community members
- Common identity - people are attached to the group as a whole

What edges "mean" and community identification

- social networks - reasons an individual adds a link to a friend very diverse
- citation networks - links are more "expensive" and semantically uniform.

Conclusions

Approximation algorithms as experimental probes!

- Hard-to-cut onion-like core with more structure than random
- Small well-isolated communities gradually blend into the core

Community structure in large networks is qualitatively different!

- Agree with previous results on small networks
- Agree with sociological interpretation (Dunbar's 150 and bond vs. identity)!

Common generative models don't capture community phenomenon!

- Graph locality - important for realistic network generation
- Local regularization - important due to sparsity