Community Structure in Large Social and Information Networks

Michael W. Mahoney

Yahoo Research

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

Joint work with: Jure Leskovec, Kevin Lang and Anirban Dasgupta

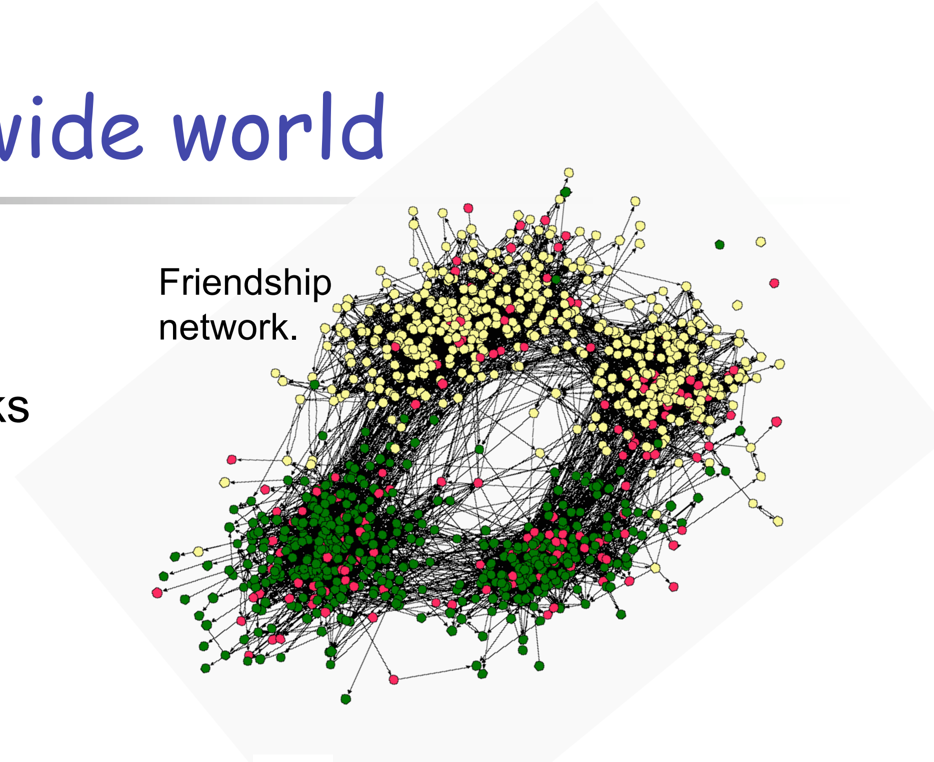
Lots and lots of large data!

- DNA micro-array data and DNA SNP data
- High energy physics experimental data
- Hyper-spectral medical and astronomical image data
- Term-document data
- Medical literature analysis data
- Collaboration and citation networks
- Internet networks and web graph data
- Advertiser-bidded phrase data
- Static and dynamic social network data

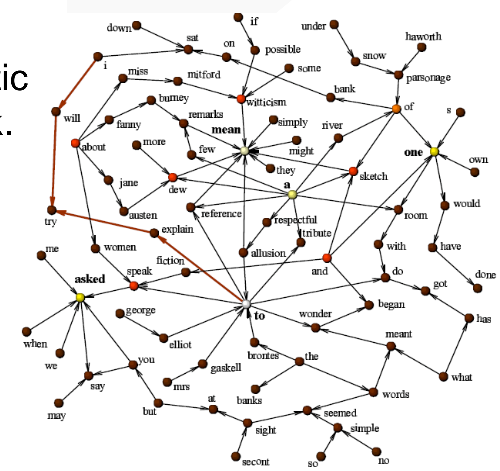
Networks in the wide world

- technological networks
  - AS, power-grid, road networks
- biological networks
  - food-web, protein networks
- social networks
  - collaboration networks, friendships
- language networks
  - semantic networks...
- ...

Friendship network.



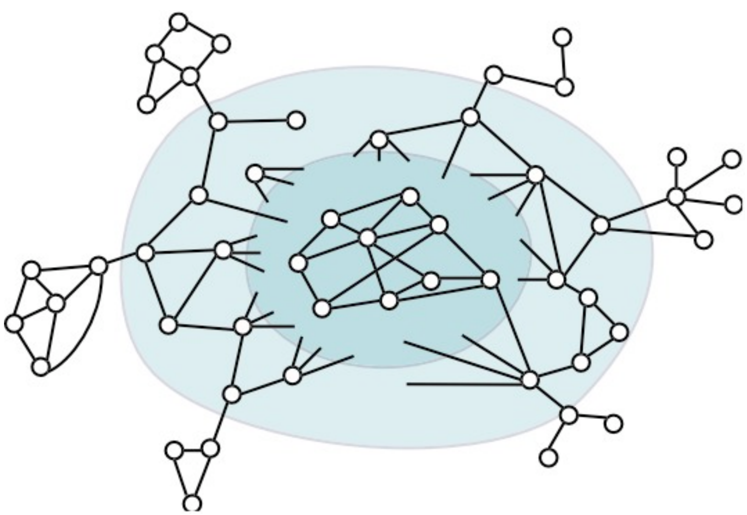
Semantic network.



Large Social and Information Networks

Interaction graph model of networks:

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

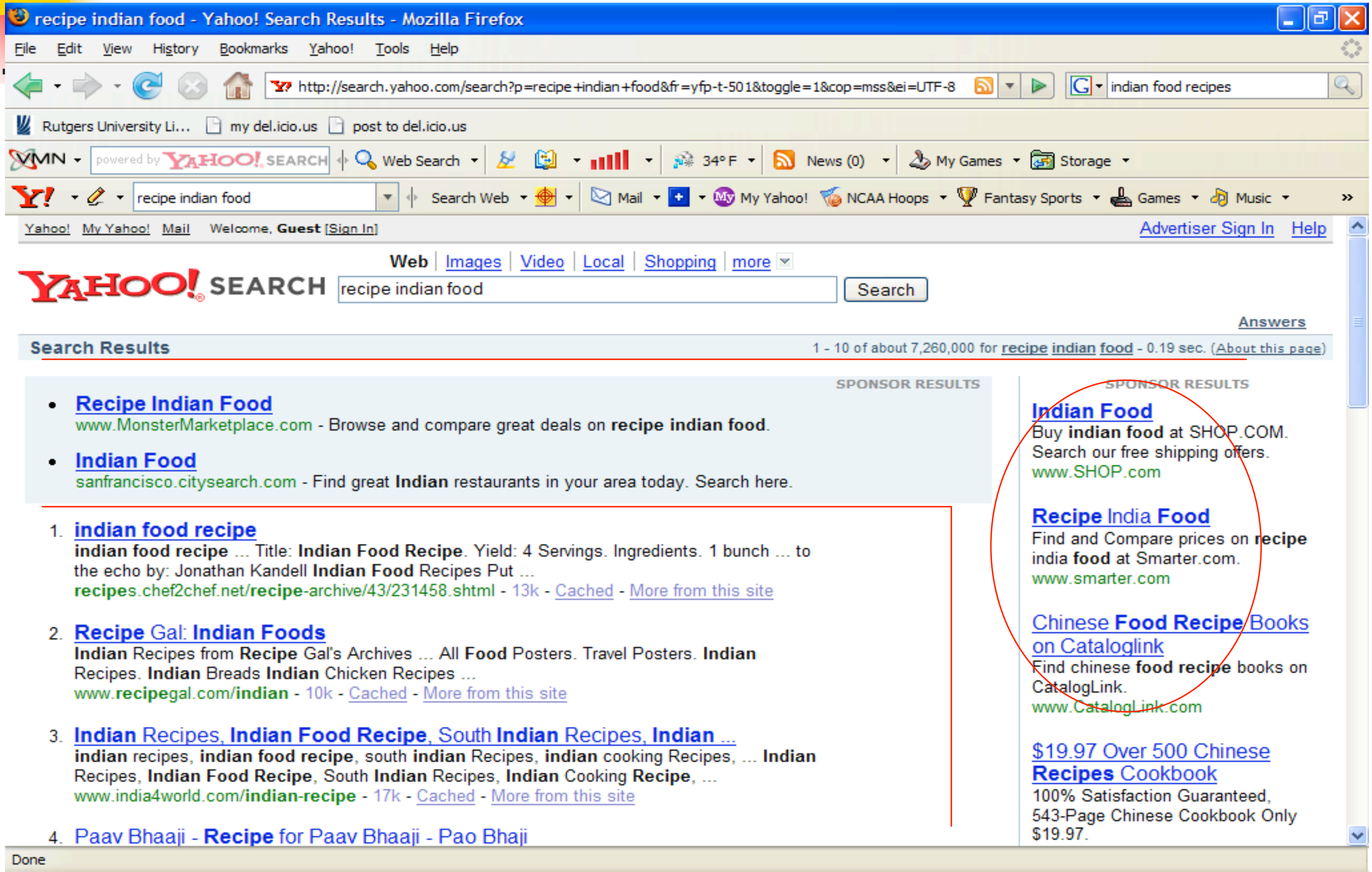


Large networks at Yahoo:

- social networks
  - –Y! messenger buddylist
  - –interaction on Y! Answers
  - –address book in Y! mail
- information networks
  - –advertiser-query
  - –user-query
  - –query-webpage
  - –webpage-webpage

Sponsored ("paid") Search

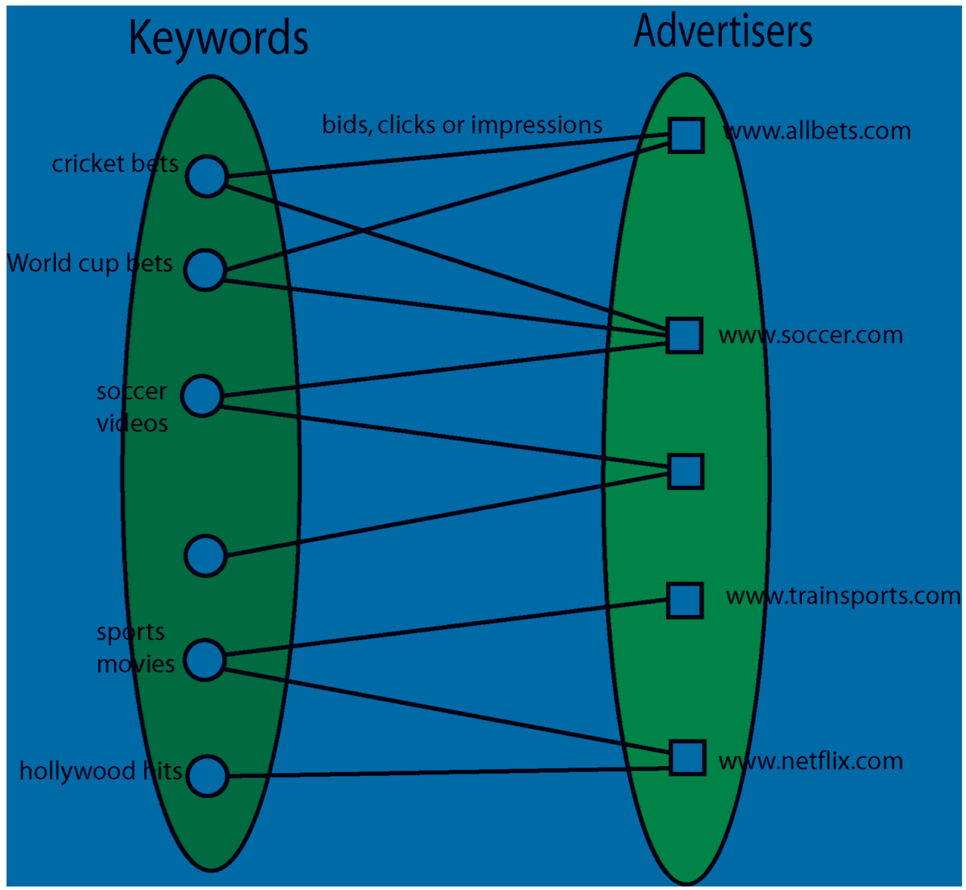
Text based ads driven by user specified query



Graphs and sponsored search data

- bid, click and impression information for "keyword x advertiser" pair

- mine information at query-time to provide new ads
  - maximize CTR, RPS, advertiser ROI



Sponsored Search Problems

- Marketplace depth broadening:

  find new advertisers for a particular query/submarket

- Query recommender system:

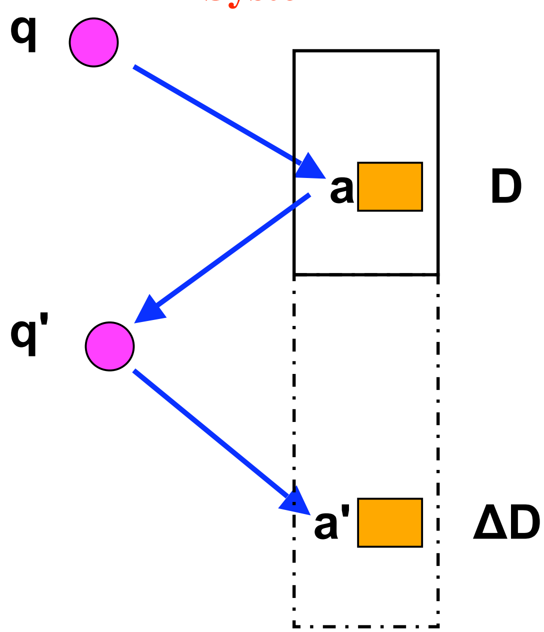
  suggest to advertisers new queries that have high probability of clicks

- Contextual query broadening:

  broaden the user's query using context information, e.g. past sessions, phrasing, etc.

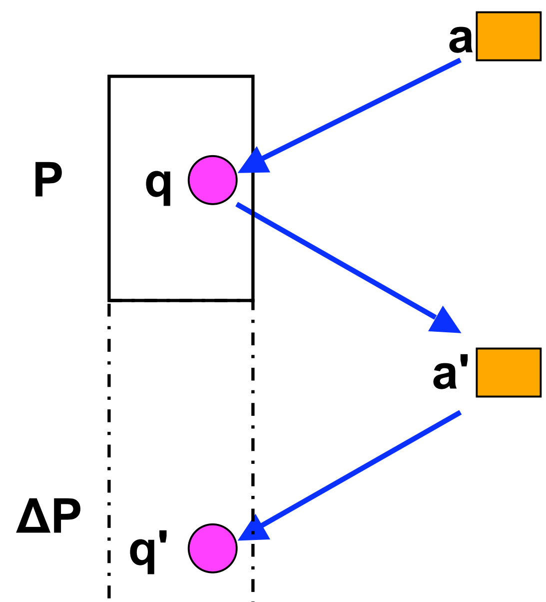
Graph Mining Recommendations

Marketplace Broadening System



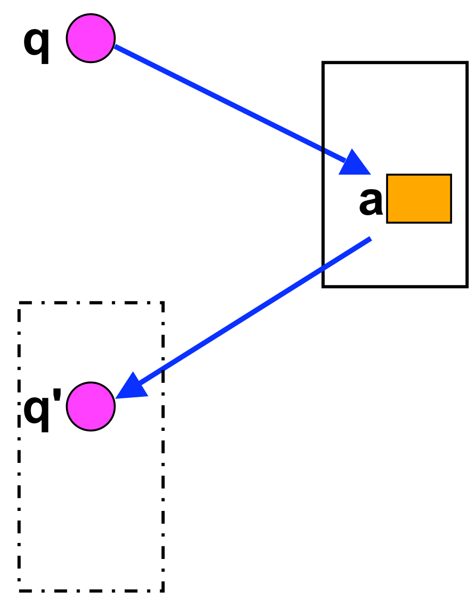
Depth Gain = $\frac{\Delta D*100}{D}$

Query Recommender System



Portfolio Gain = $\frac{\Delta P*100}{P}$

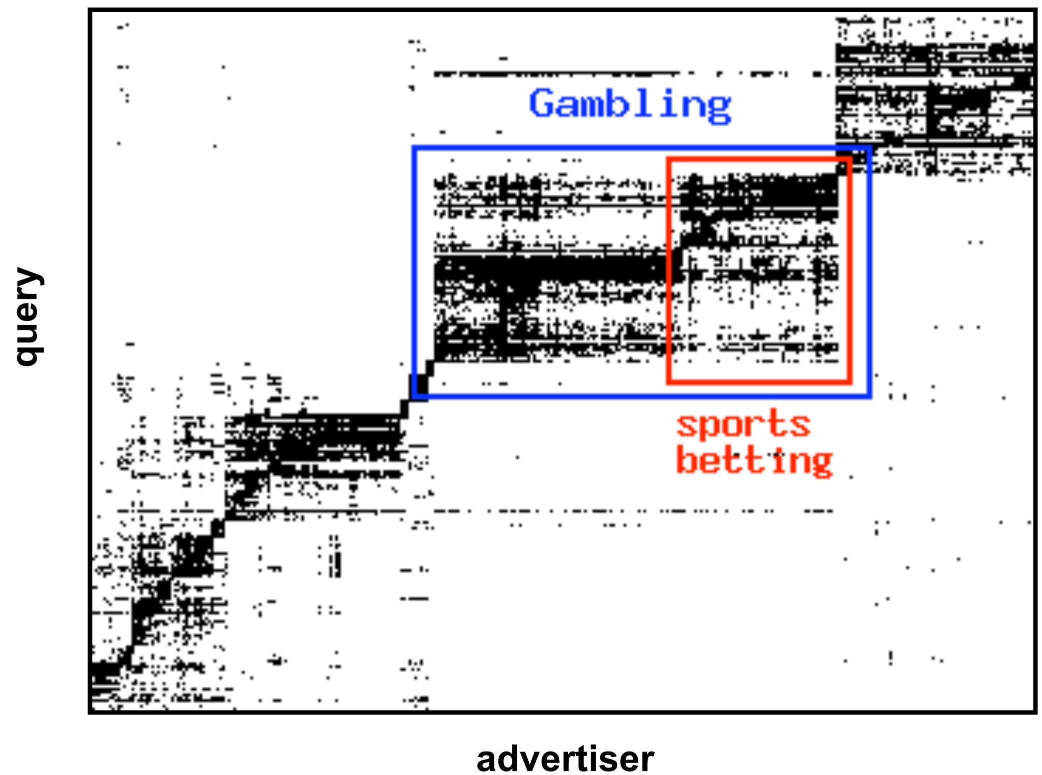
Dynamic Query Broadening



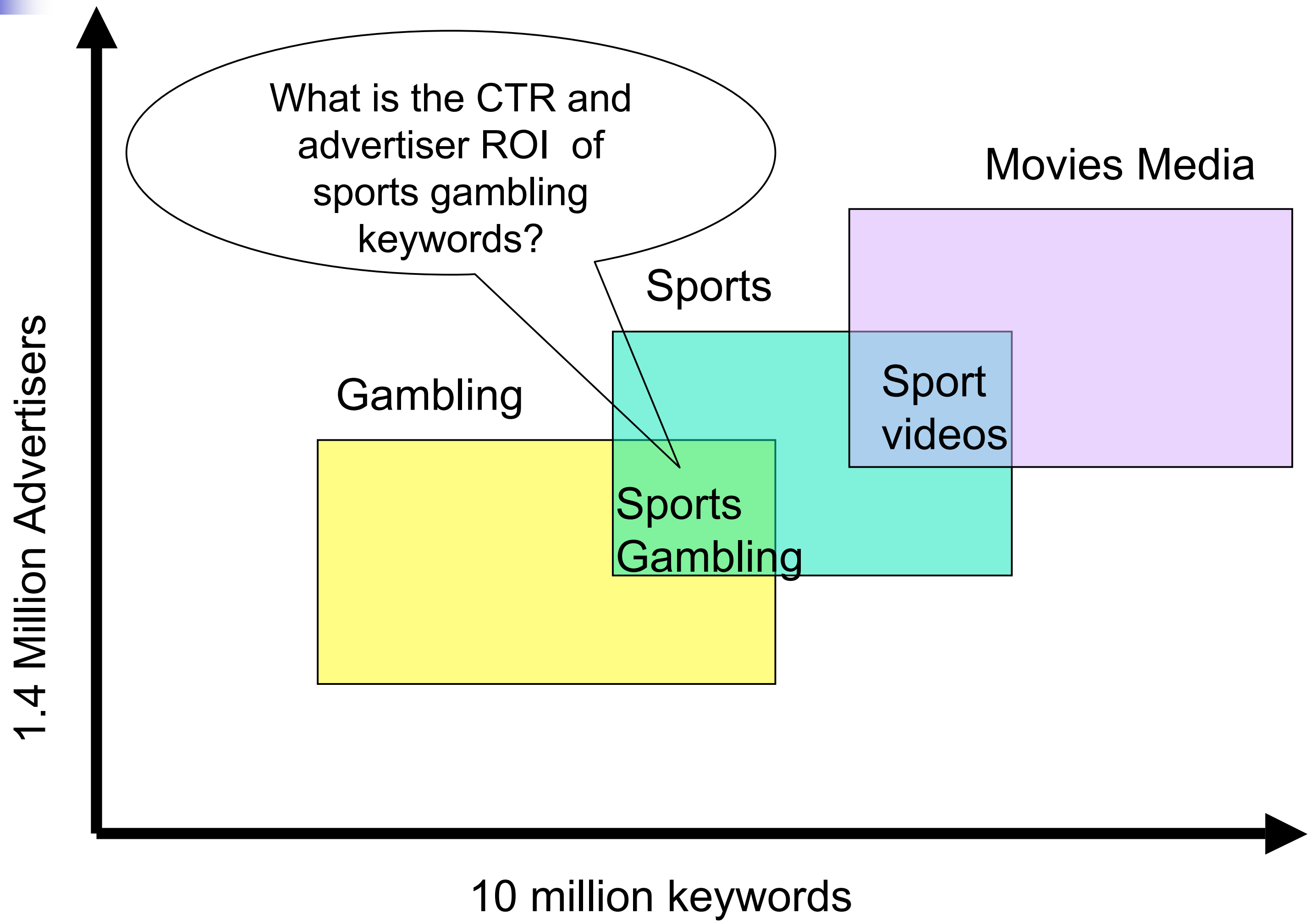
ΔQ

Micro-markets

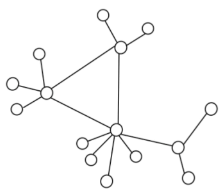
find micro-markets by partitioning the "query x advertiser" graph:

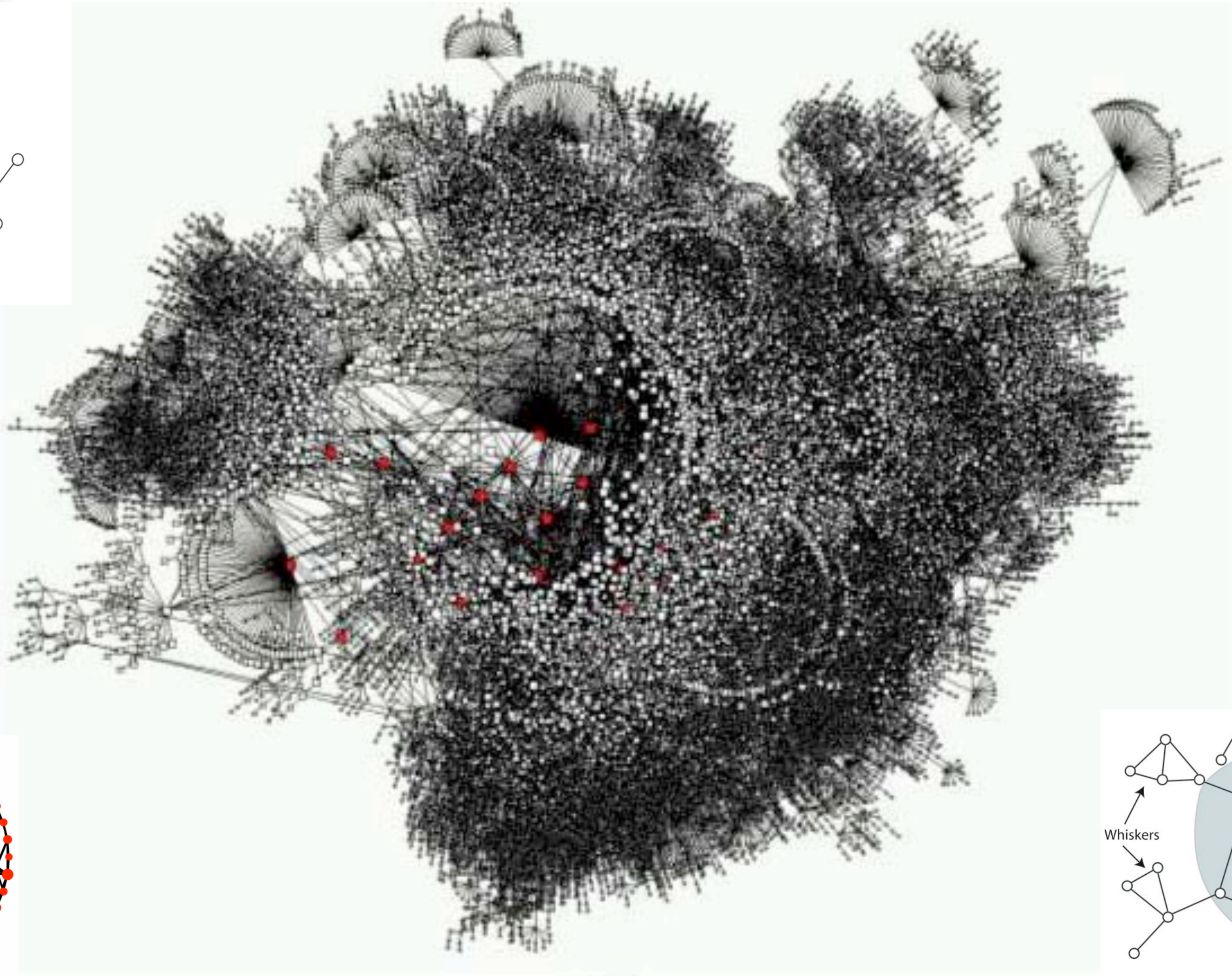


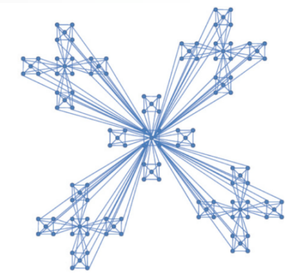
Micro-markets in sponsored search

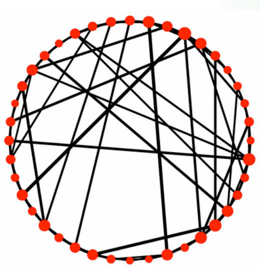


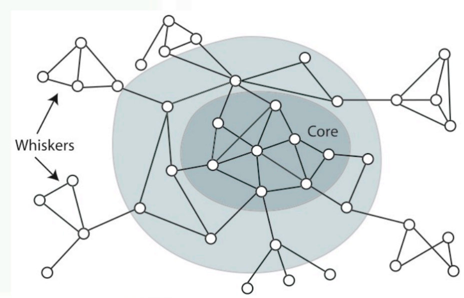
What do these networks "look" like?











Questions of interest ...

What are degree distributions, clustering coefficients, diameters, etc.?

Heavy-tailed, small-world, expander, geometry+rewiring, local-global decompositions, ...

Are there natural clusters, communities, partitions, etc.?

Concept-based clusters, link-based clusters, density-based clusters, ...

(e.g., *isolated* micro-markets with *sufficient money/clicks* with *sufficient coherence*)

How do networks grow, evolve, respond to perturbations, etc.?

Preferential attachment, copying, HOT, shrinking diameters, ...

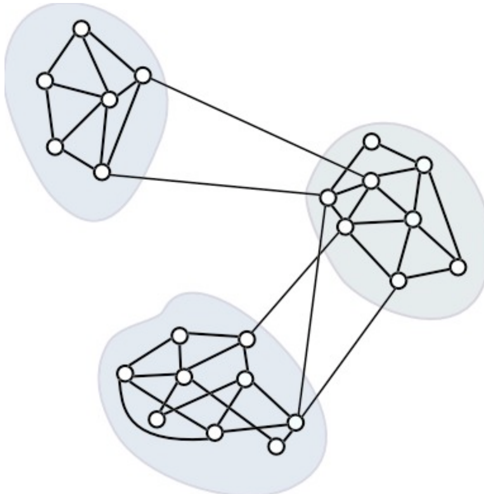
How do dynamic processes - search, diffusion, etc. - behave on networks?

Decentralized search, undirected diffusion, cascading epidemics, ...

How best to do learning, e.g., classification, regression, ranking, etc.?

Information retrieval, machine learning, ...

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

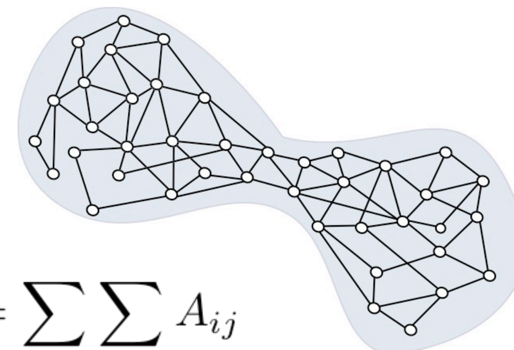
  Define "edge counting" metric -- conductance, expansion, modularity, etc. -- in interaction graph, then optimize!

"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."

Communities, Conductance, and NCPPs

Let A be the adjacency matrix of G=(V,E).

The conductance ϕ of a set S of nodes is:



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

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

The Network Community Profile (NCP) Plot of the graph is:

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

Just as conductance captures the "gestalt" notion of cluster/community quality, the NCP plot measures cluster/community quality as a function of size.

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

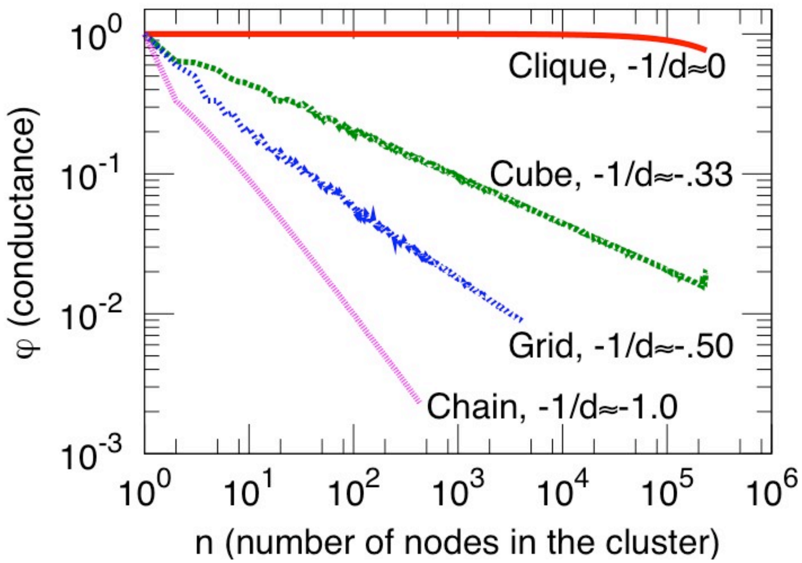
X+MQI - post-processing step on, e.g., Spectral of Metis

Metis+MQI - best conductance (empirically)

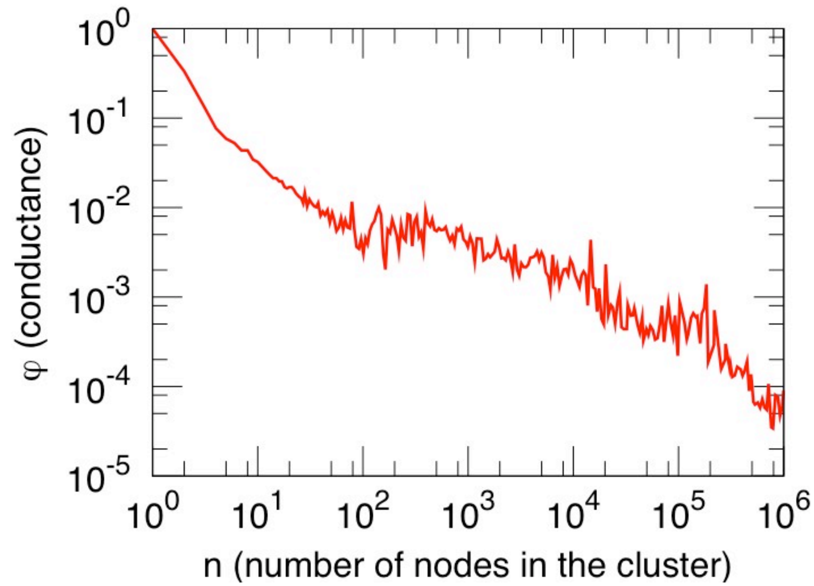
Local Spectral - connected and tighter sets (empirically)

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

Low-dimensional graphs and expanders

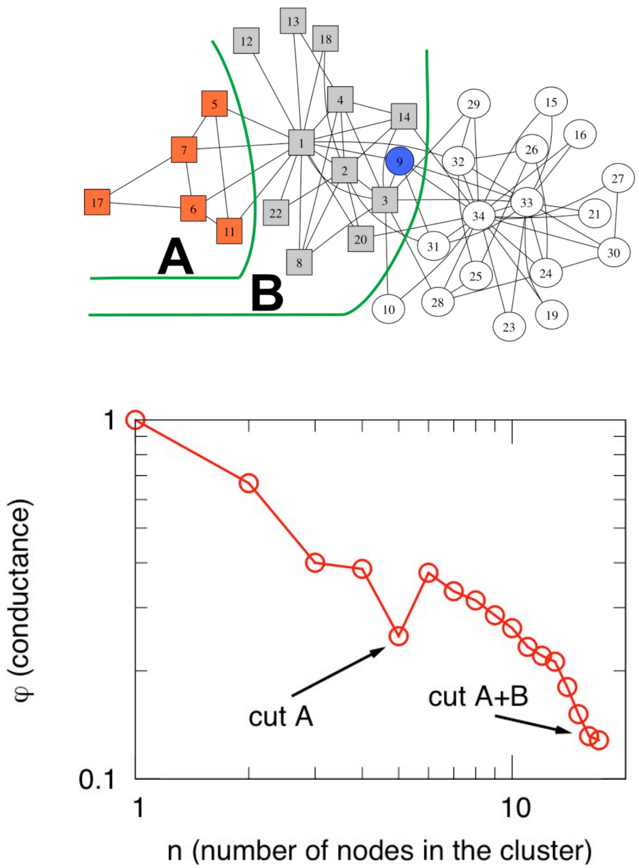


d-dimensional meshes

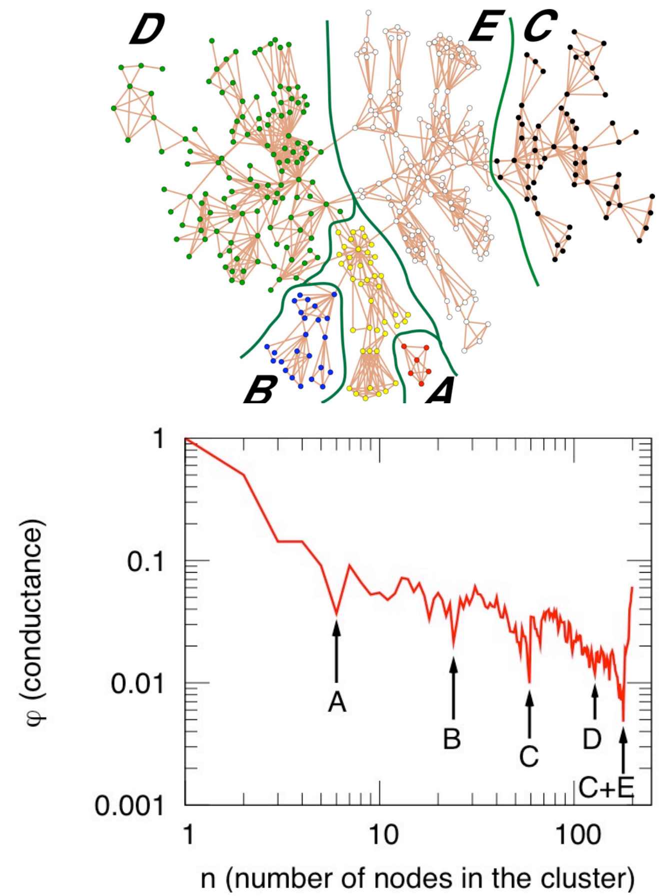


RoadNet-CA

Widely-studied small social networks



Zachary's karate club



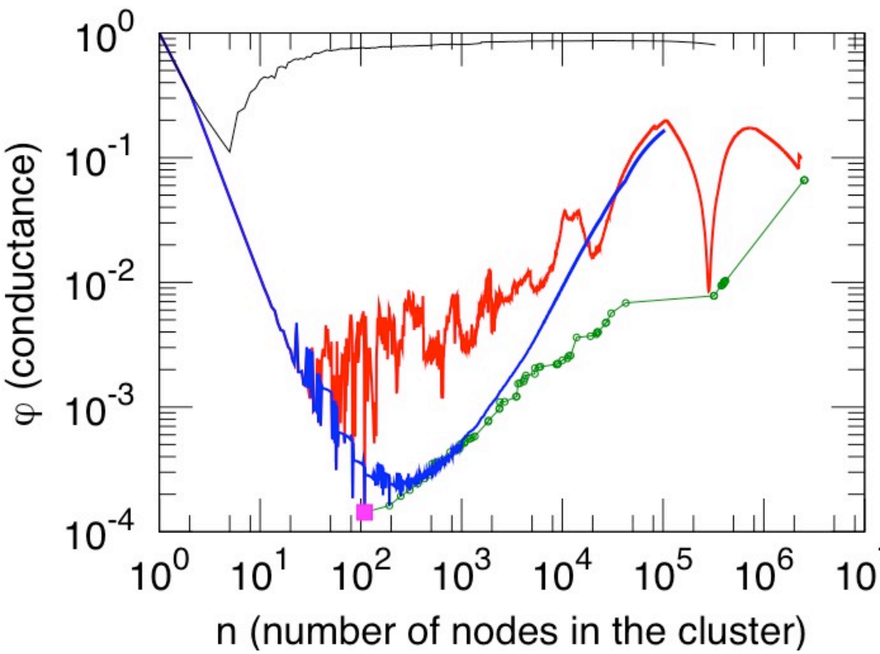
Newman's Network Science

Large Social and Information Networks

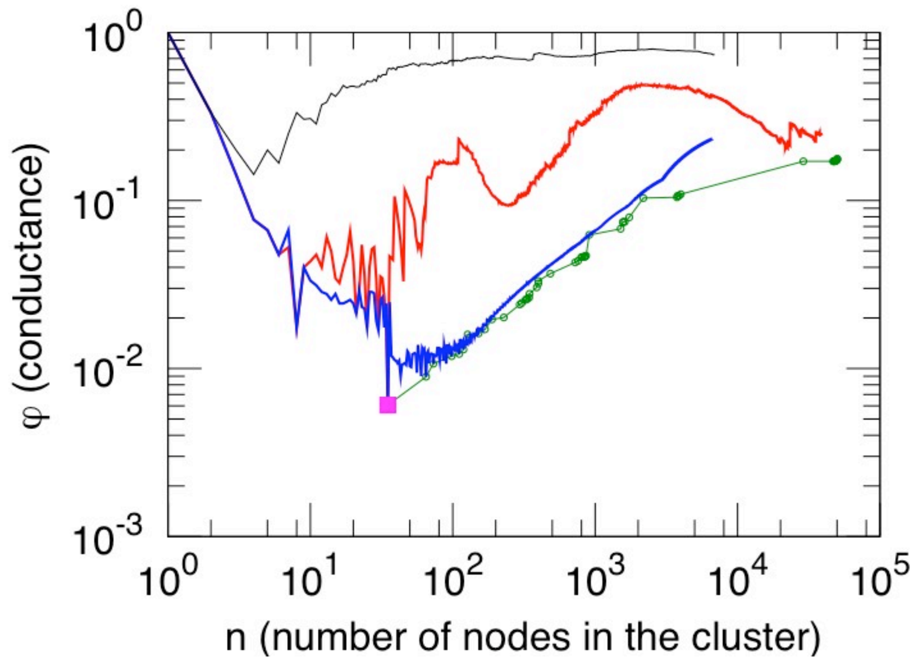
• Social nets	Nodes	Edges	Description
LiveJournal	4,843,953	42,845,684	Blog friendships [4]
Epinions	75,877	405,739	Who-trusts-whom [35]
Flickr	404,733	2,110,078	Photo sharing [21]
Delicious	147,567	301,921	Collaborative tagging
CA-DBLP	317,080	1,049,866	Co-authorship (CA) [4]
CA-cond-mat	21,363	91,286	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.

Large Social and Information Networks

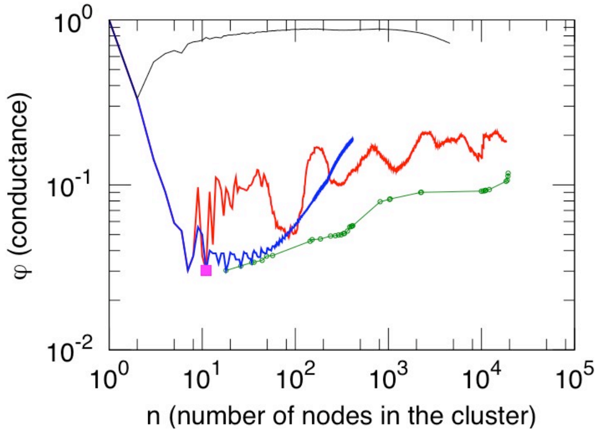


LiveJournal

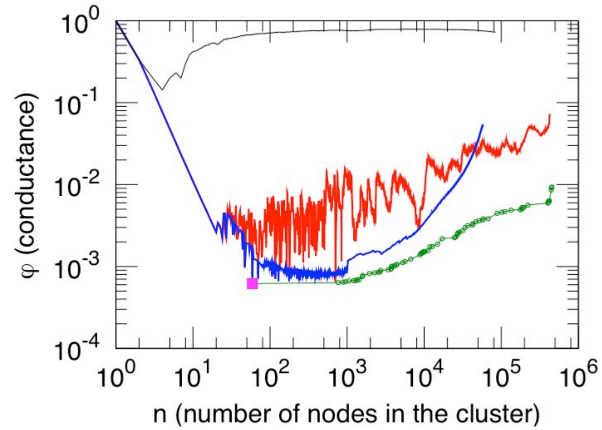


Epinions

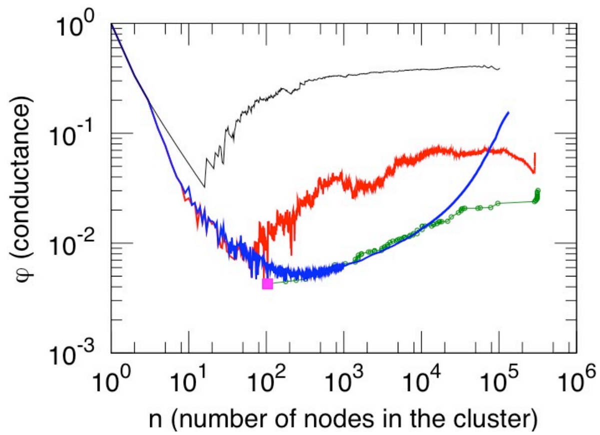
More large networks



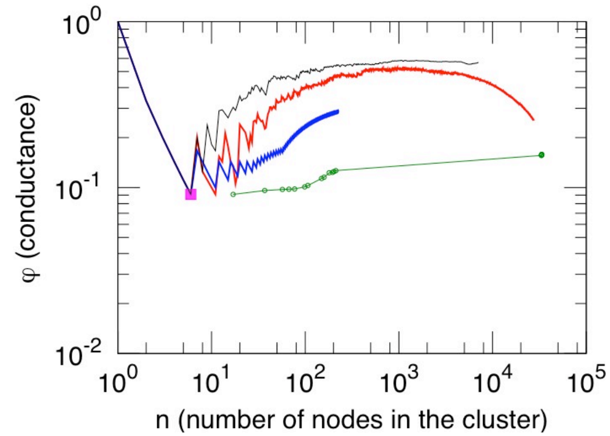
Cit-Hep-Th



Web-Google



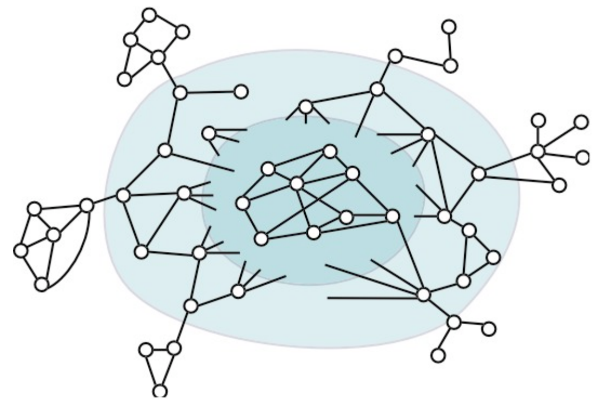
AtP-DBLP

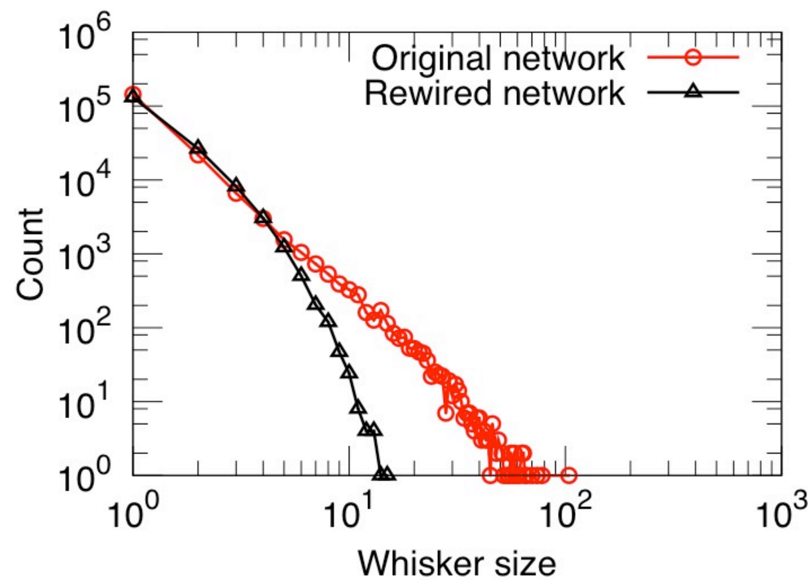


Gnutella

"Whiskers" and the "core"

- Whiskers
  - maximal sub-graph detached from network by removing a single edge
  - on average, contains 40% of nodes and 20% of edges
- Core
  - the rest of the graph, i.e., the 2-edge-connected core
  - on average, contains 60% of nodes and 80% of edges
- Global minimum of NCPP is a whisker

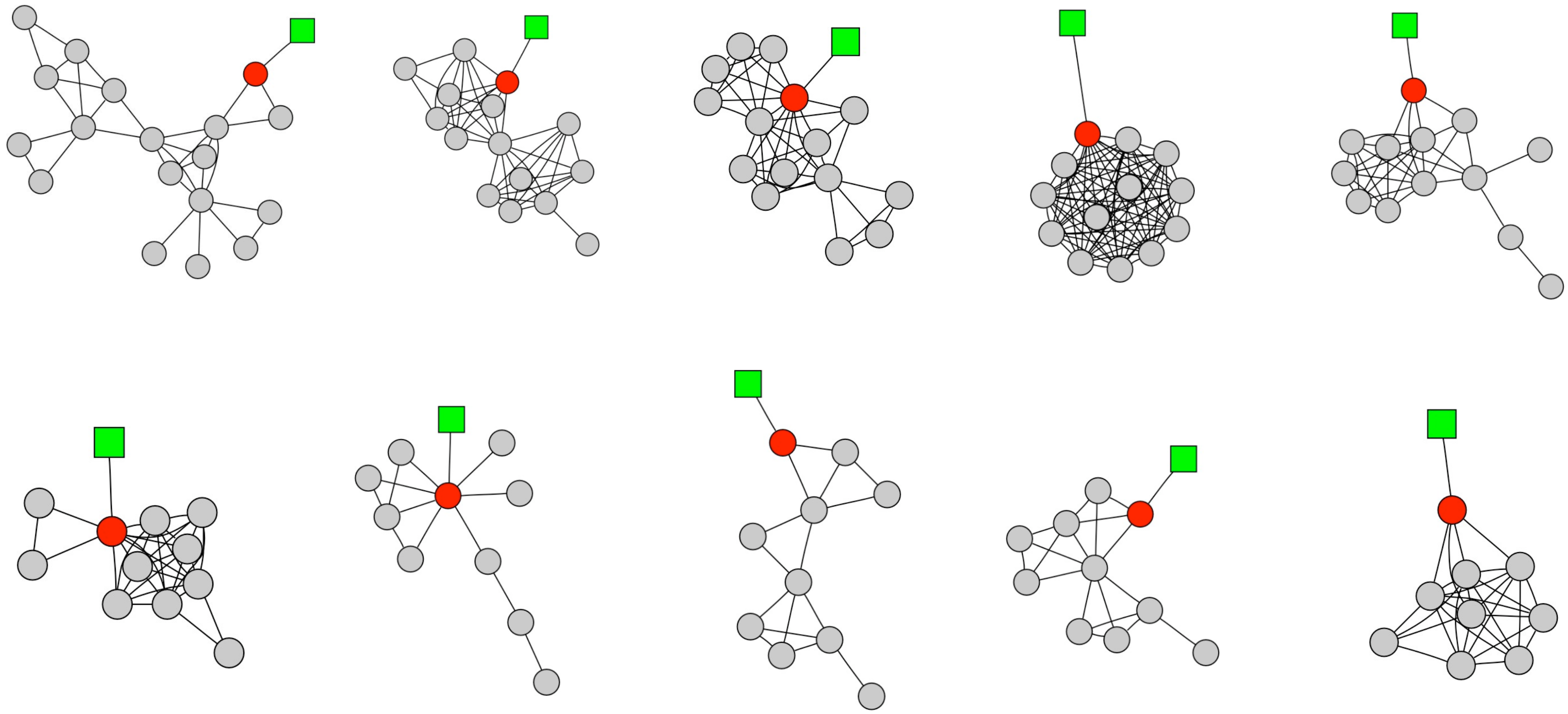




Distribution of "whiskers" for AtP-DBLP.

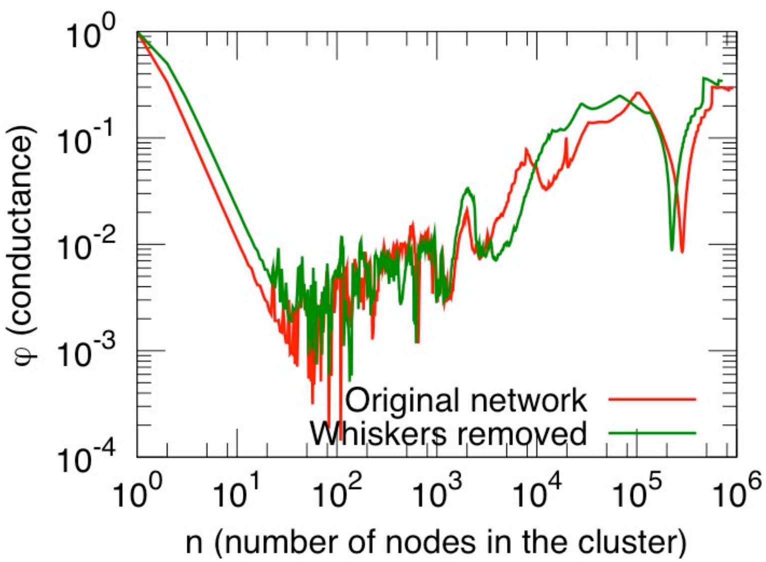
Examples of whiskers

Ten largest "whiskers" from CA-cond-mat.

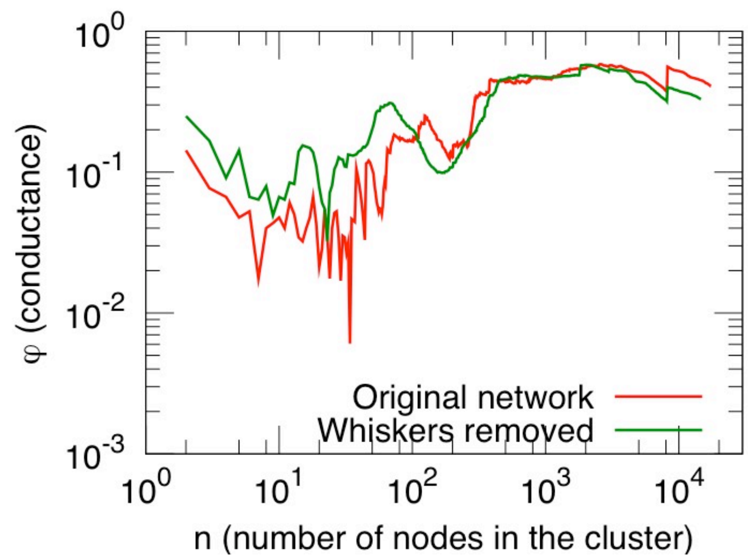


What if the “whiskers” are removed?

Then the lowest conductance sets - the “best” communities - are “2-whiskers”:



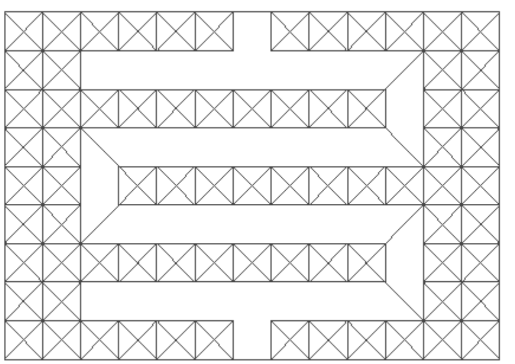
LiveJournal

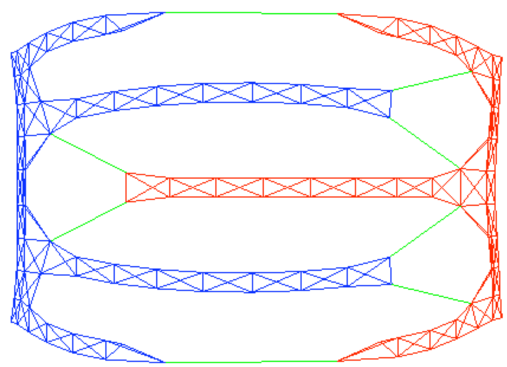


Epinions

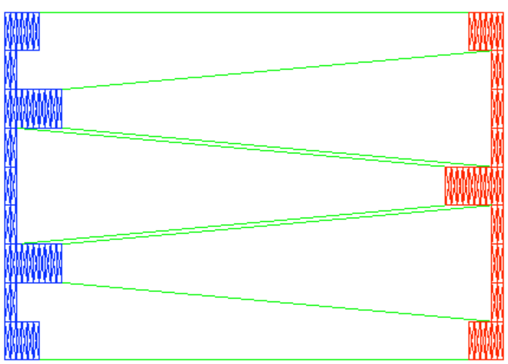
"Regularization" and spectral methods

- regularization properties: 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



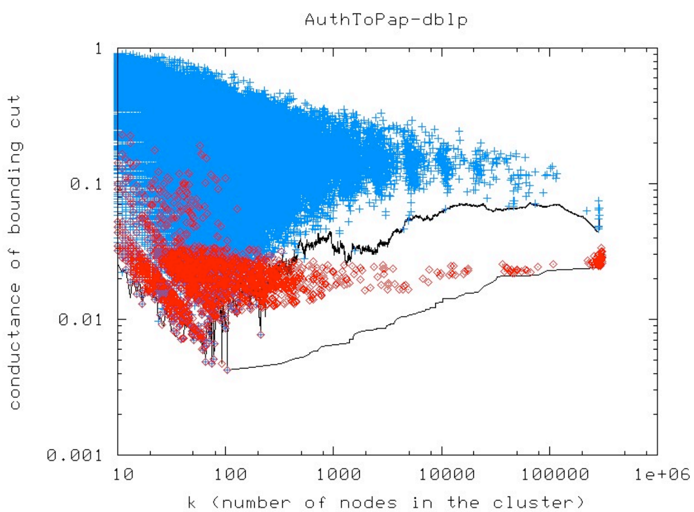


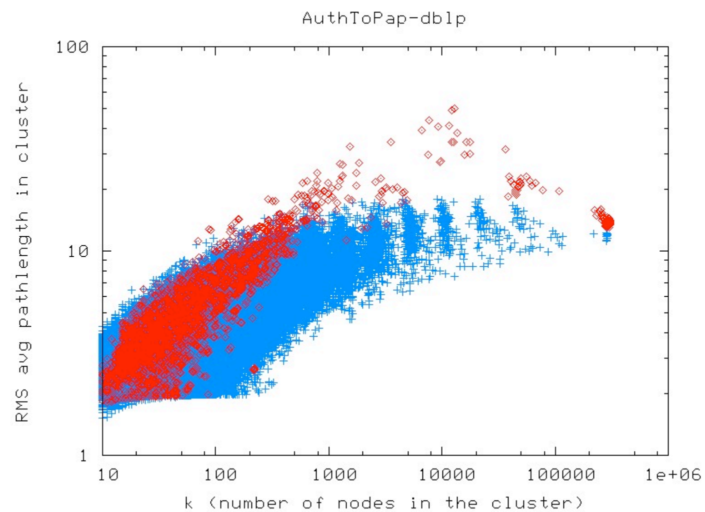
spectral embedding

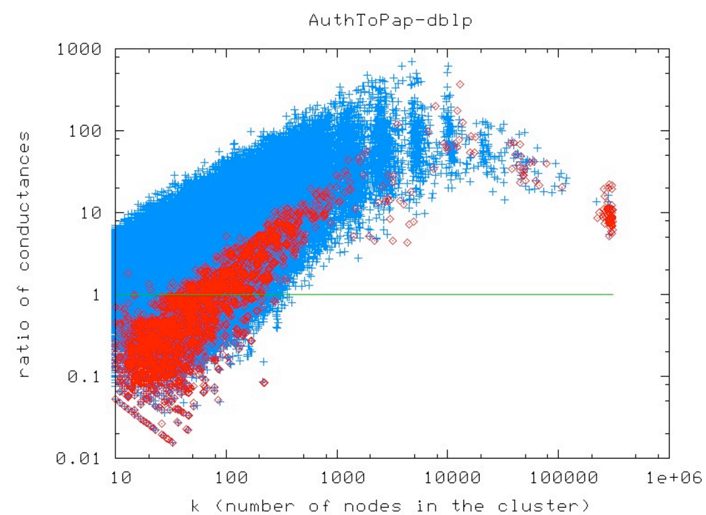


notional flow based embedding

Regularized and non-regularized communities (1 of 2)



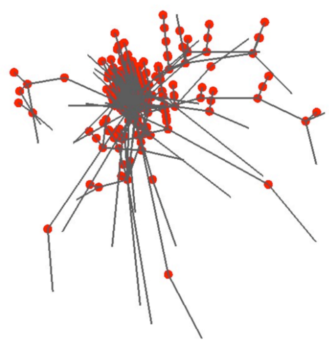


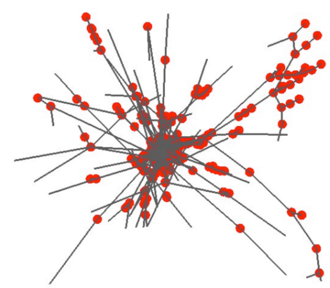


- Metis+MQI (red) gives sets with better conductance.
- Local Spectral (blue) gives tighter and more well-rounded sets.

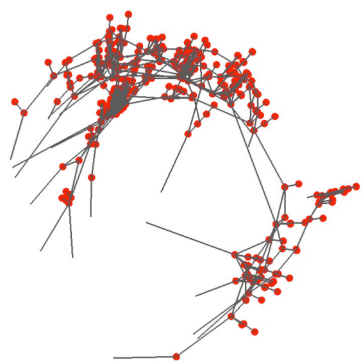
Regularized and non-regularized communities (2 of 2)

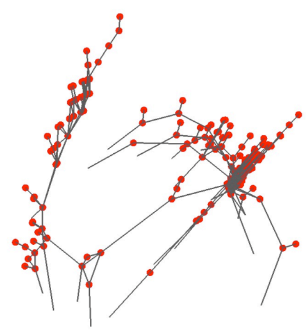
Two ca. 500 node communities from Local Spectral Algorithm:





Two ca. 500 node communities from Metis+MQI:





Lower Bounds …

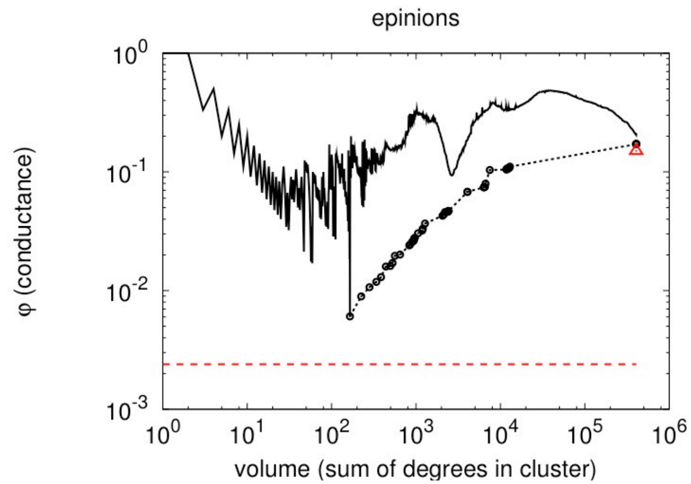
… can be computed from:

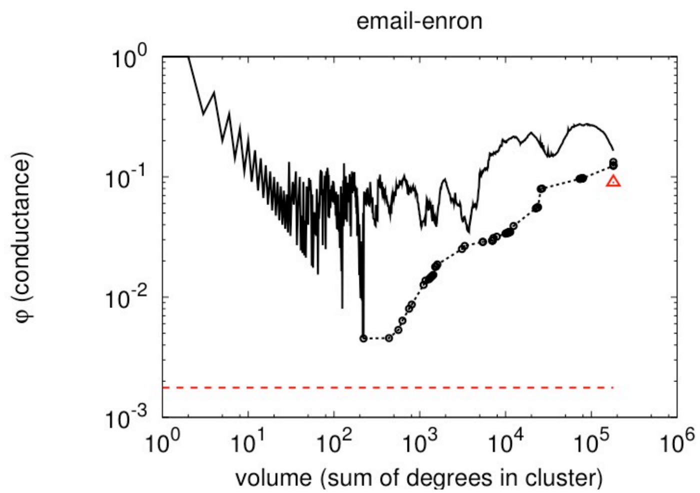
- Spectral embedding

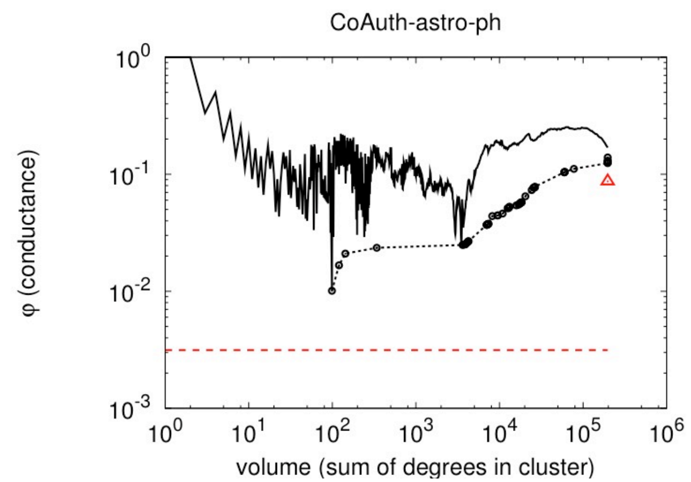
  (independent of balance)

- SDP-based methods

  (for volume-balanced partitions)







Lots of Generative Models

- Preferential attachment - add edges to high-degree nodes

  (Albert and Barabasi 99, etc.)

- Copying model - add edges to neighbors of a seed node

  (Kumar et al. 00, etc.)

- Hierarchical methods - add edges based on distance in hierarchy

  (Ravasz and Barabasi 02, etc.)

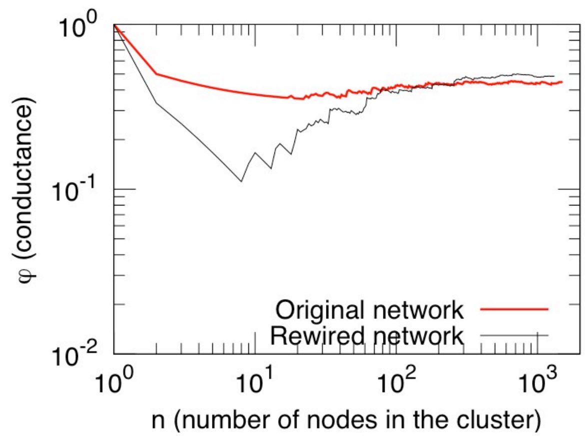
- Geometric PA and Small worlds - add edges to geometric scaffolding

  (Flaxman et al. 04; Watts and Strogatz 98; etc.)

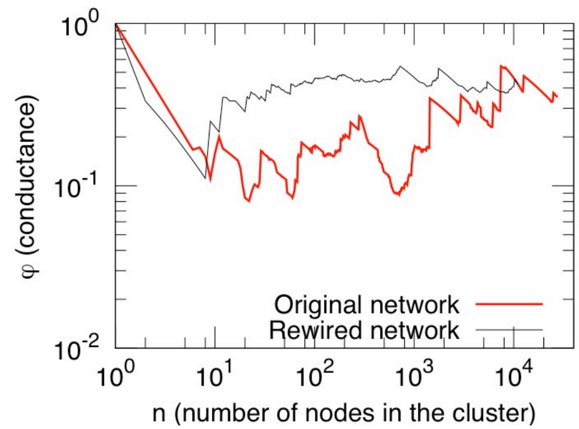
- Random/configuration models - add edges randomly

  (Molloy and Reed 98; Chung and Lu 06; etc.)

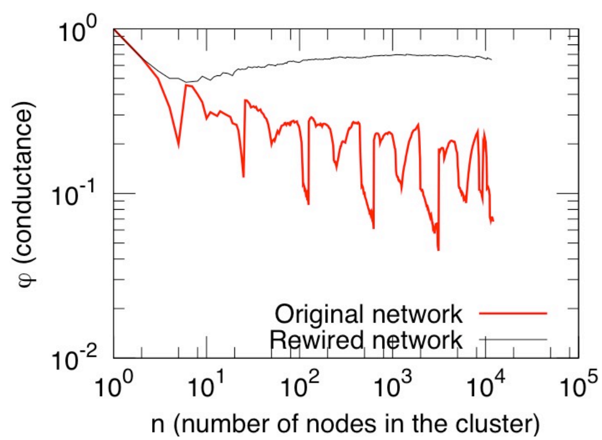
NCPP for common generative models



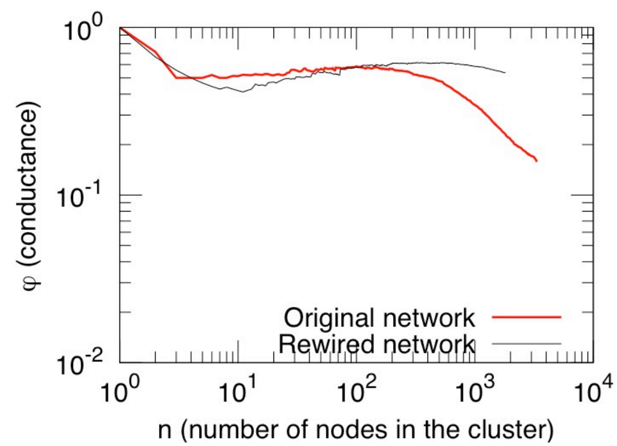
Preferential Attachment



Copying Model



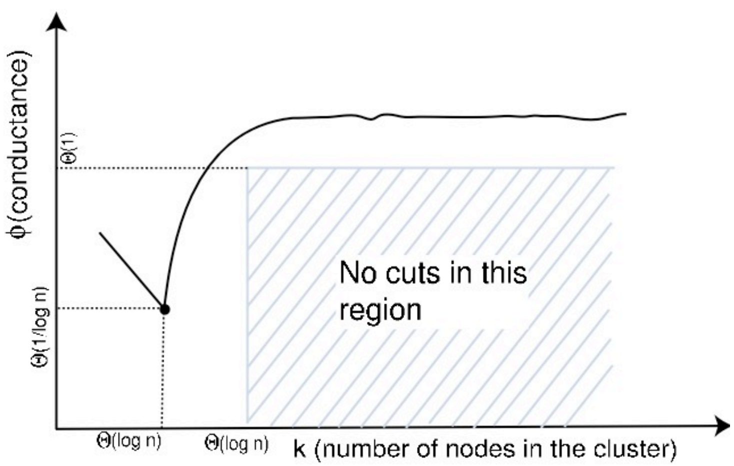
RB Hierarchical



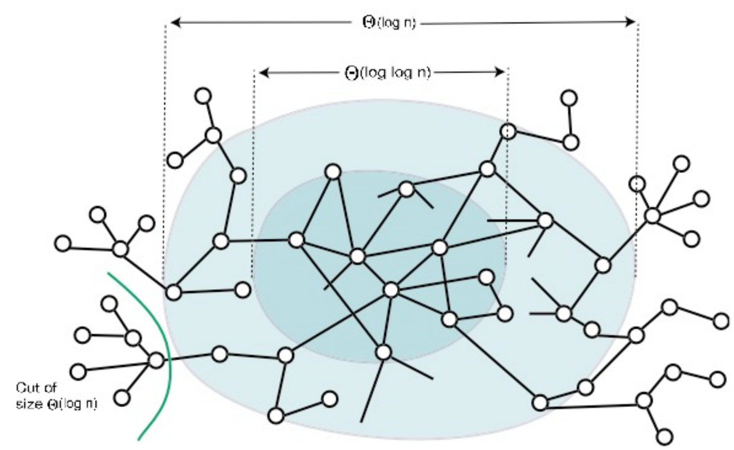
Geometric PA

A simple theorem: 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.
$p_{ij} = w_i w_j / \sum_k w_k.$



NCP Plot for G(w) model, with power-law degrees and β ε (2,3).



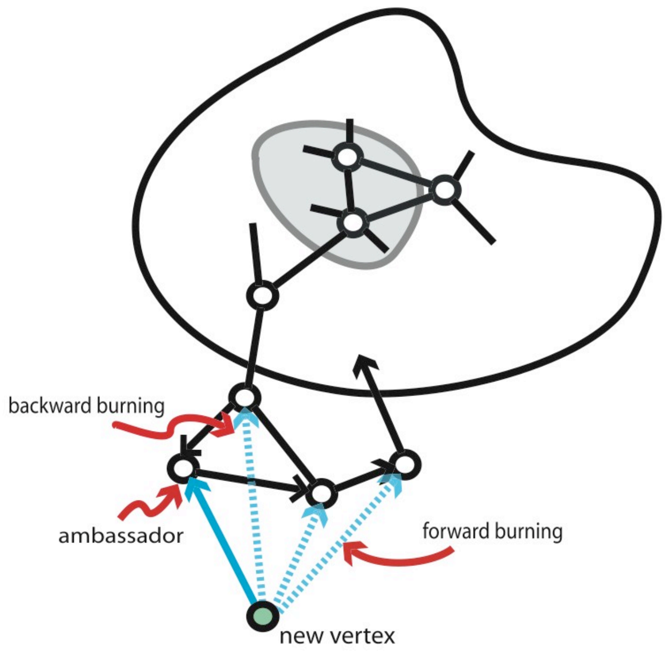
Structure of the G(w) model, with β ε (2,3).

Note: Sparsity is the issue, not heavy-tails per se. (Power laws with β ε (2,3) give us the appropriate sparsity.)

A "forest fire" model

Leskovec, Kleinberg, and Faloutsos 2005

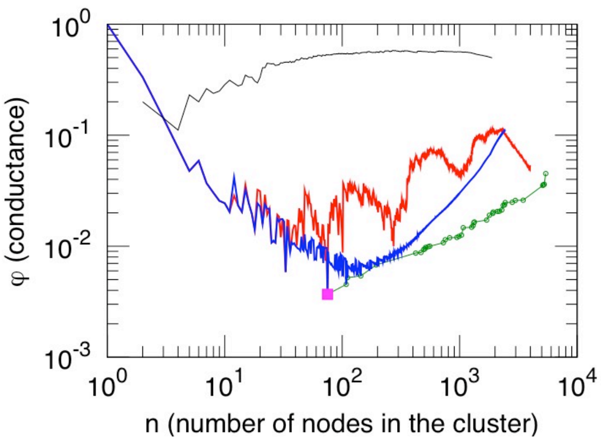
At each time step, a new node v_i:
- links to a "seed" node w, chosen uniformly at random.
- selects x "outlinks" and y "inlinks" of w at random.
- forms "outlinks," i.e., burns, to those selected nodes and then proceeds to burn recursively.

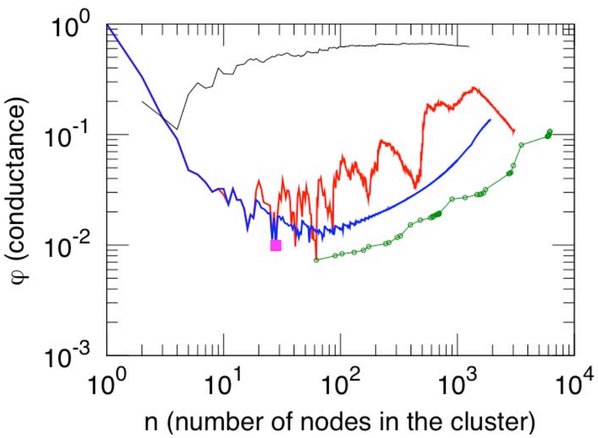


Notes:
- Preferential attachment flavor - second neighbor is not uniform at random.
- Copying flavor - since burn seed's neighbors.
- Hierarchical flavor - seed is parent.
- "Local" flavor - burn "near" -- in a diffusion sense -- the seed vertex.

NCPP Of the FF Model

Two different parameter values:





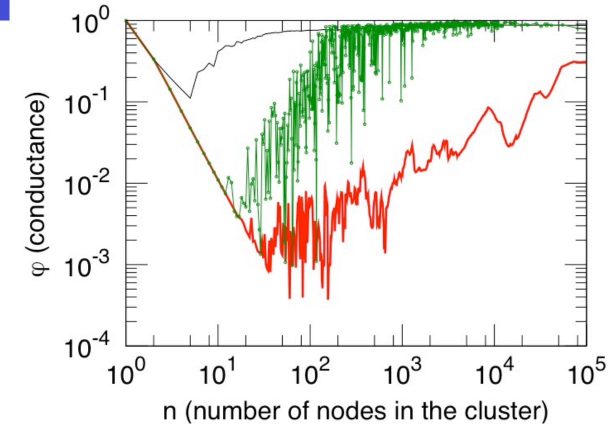
Note: for these parameters, this model also reproduces "densification" and "shrinking diameters" of real graphs (Leskovec et al. 05).

Comparison with "Ground truth" (1 of 2)

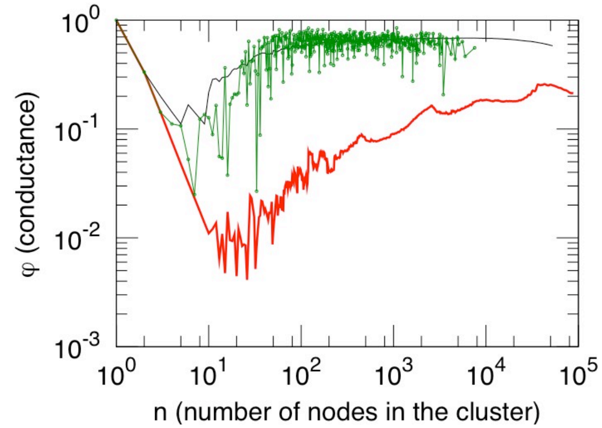
Networks with "ground truth" communities:

- 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)

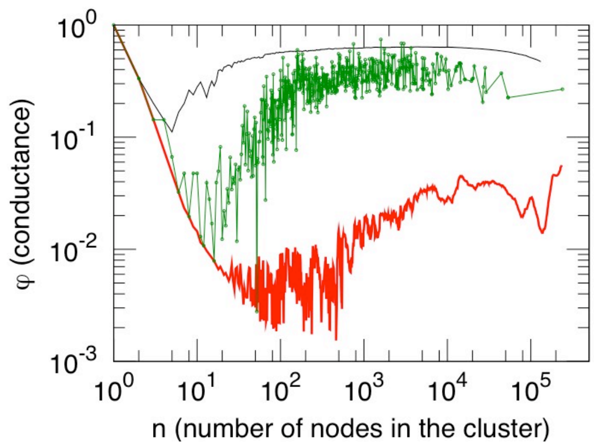
Comparison with "Ground truth" (2 of 2)



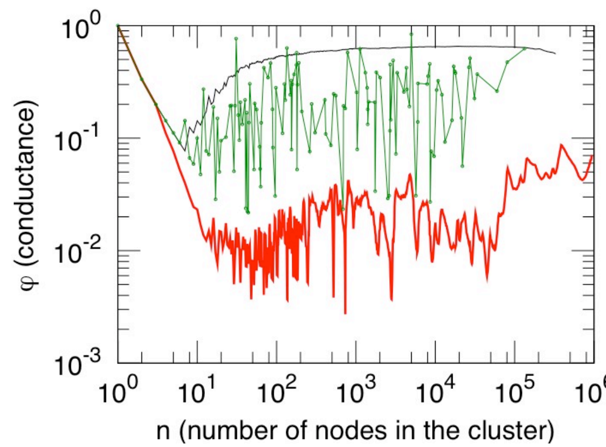
LiveJournal



CA-DBLP



AmazonAllProd



AtM-IMDB

Miscellaneous thoughts …

Sociological work on community size (Dunbar and Allen)

- 150 individuals is maximum community size
- On-line communities have 60 members and break down at 80
- Military companies, divisions of corporations, etc. - close to the Dunbar's 150

Common bond vs. common identity theory

- Common bond (people are attached to individual community members) are smaller and more cohesive
- Common identity (people are attached to the group as a whole) focused around common interest and tend to be larger and more interpersonally diverse

What edges "mean" and community identification

- social networks - reasons an individual adds a link to a friend can vary enormously
- citation networks or web graphs - links are more "expensive" and are more semantically uniform.

Conclusions

- about networks and data:

  Can use approximation algorithms as experimental probes

  "Best" communities get less and less "community-like"

  "Octopus" or "Jellyfish" model - with "whiskers" and "core"

- about modeling these networks:

  Common generative models don't capture community phenomenon

  Graph locality - important for realistic network generation

  Local regularization - important due to sparsity

Workshop on "Algorithms for Modern Massive Data Sets"

(http://mmds.stanford.edu)

Stanford University and Yahoo! Research, June 25-28, 2008

Objectives:

- Address algorithmic, mathematical, and statistical challenges in modern statistical data analysis.

- Explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinear-structured data.

- Bring together computer scientists, mathematicians, statisticians, and data analysis practitioners to promote cross-fertilization of ideas.

Organizers: M. W. Mahoney, L-H. Lim, P. Drineas, and G. Carlsson.

Sponsors: NSF, Yahoo! Research, PIMS, DARPA.