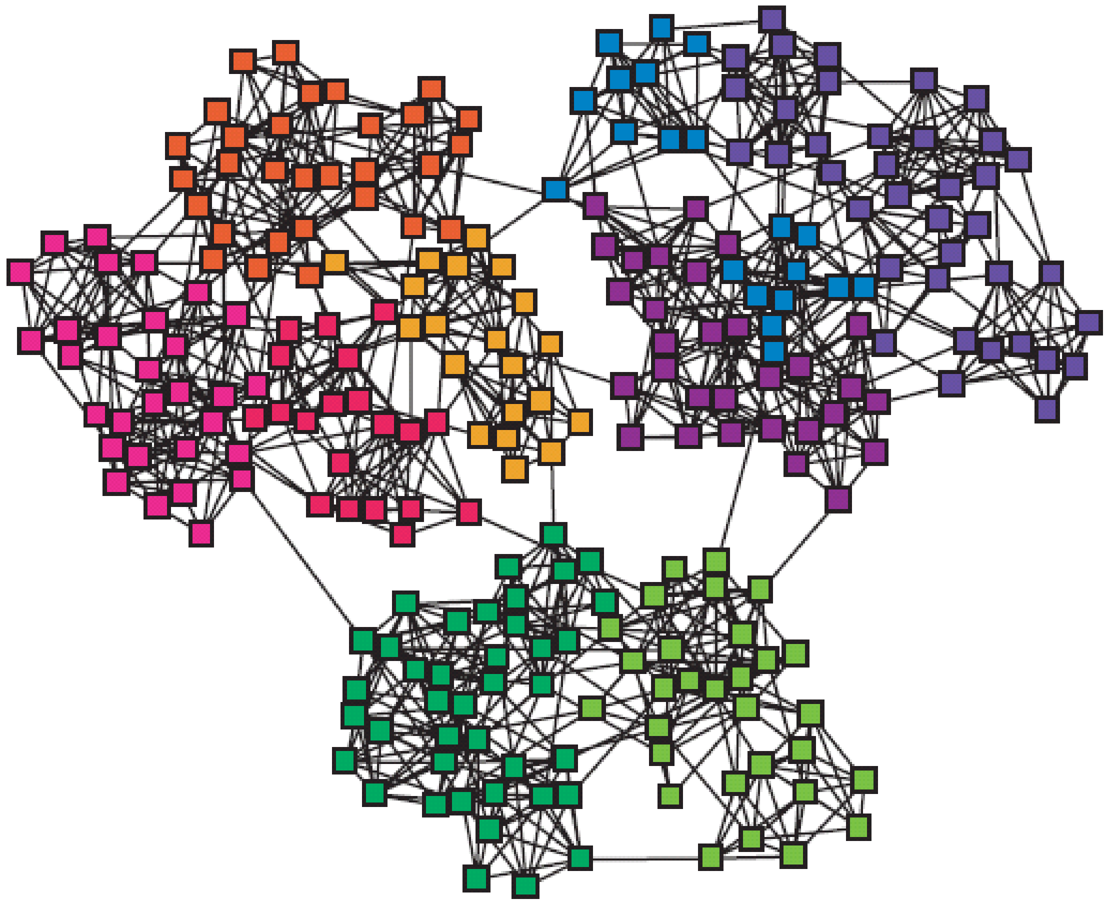
Empirical Comparison of Algorithms for Network Community Detection

Jure Leskovec (Stanford)
Kevin Lang (Yahoo! Research)
Michael Mahoney (Stanford)

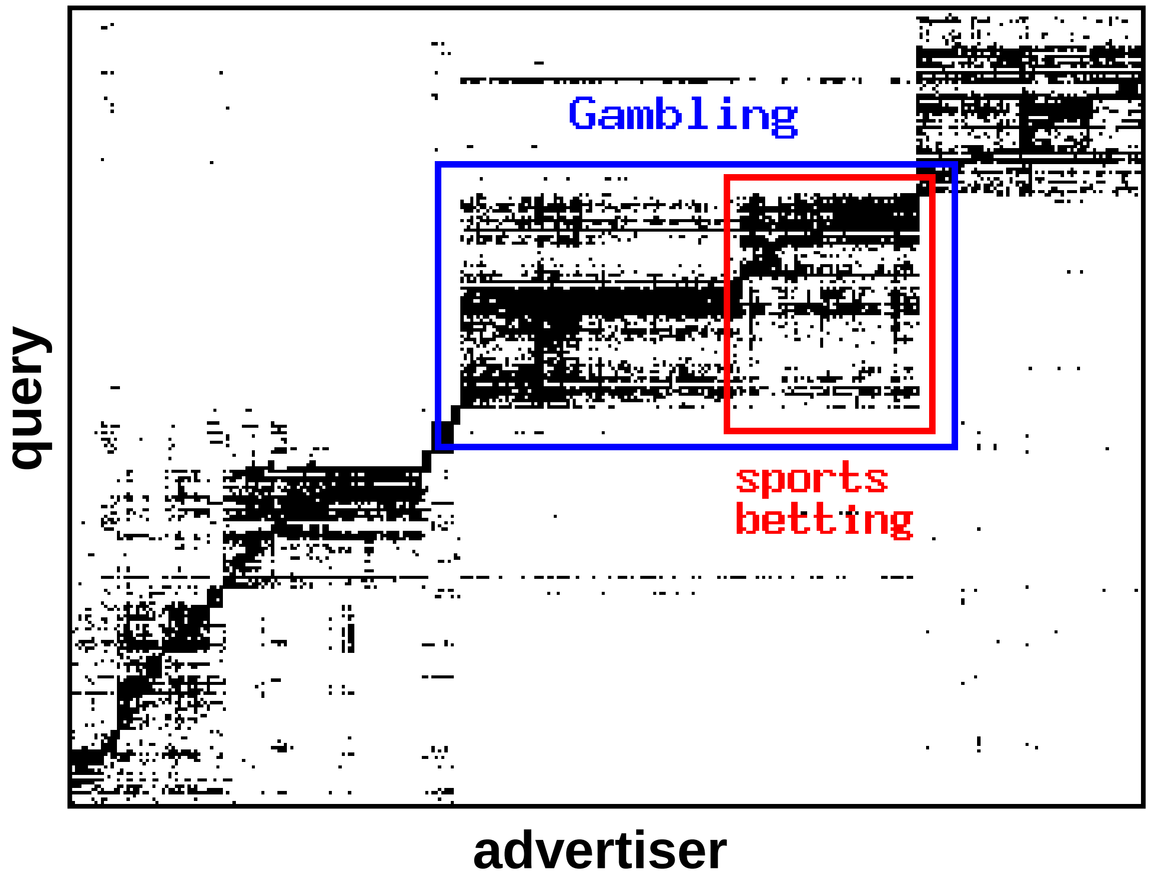
How we think about networks?



Communities, clusters, groups, modules

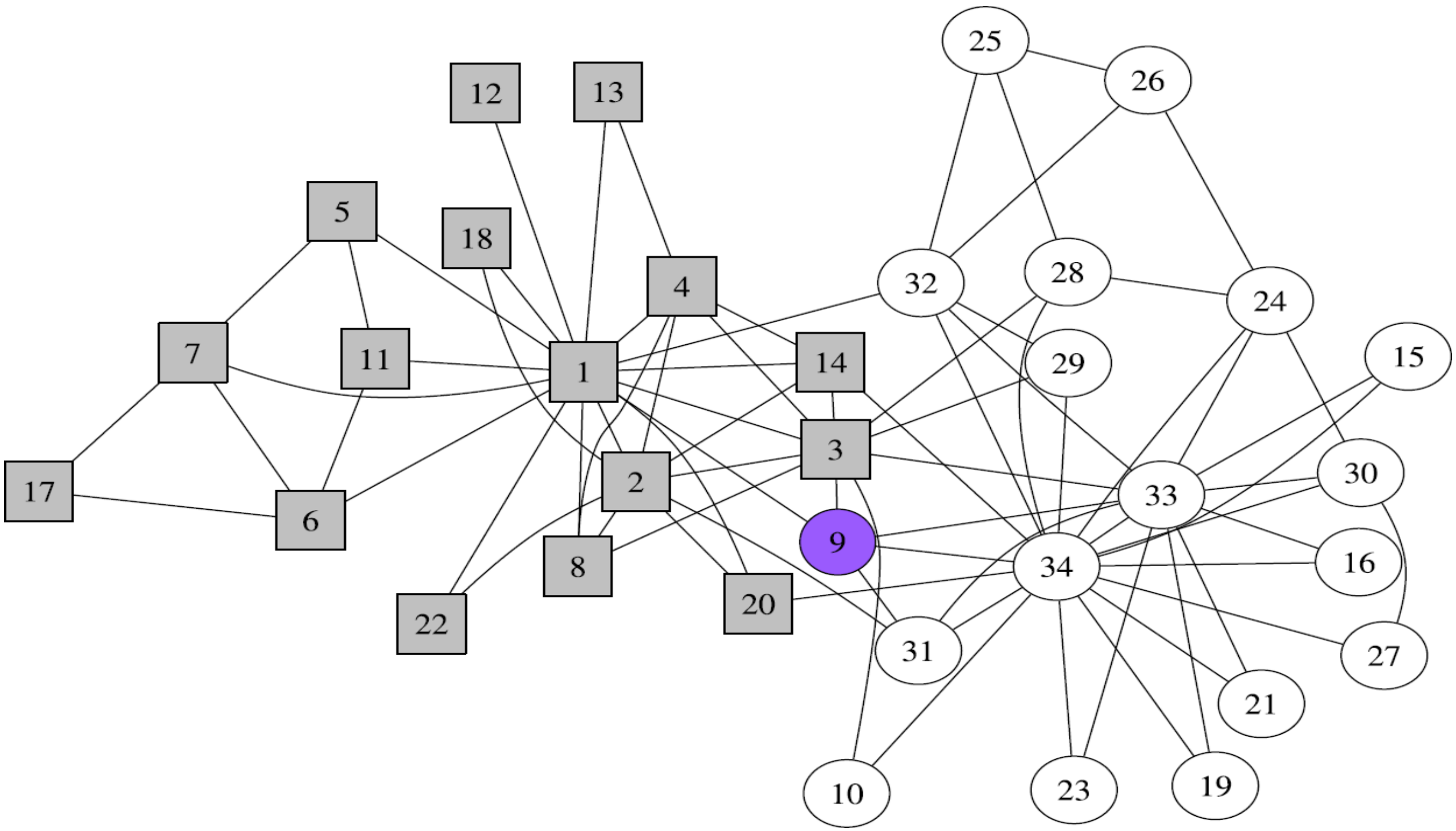
Micro-markets in sponsored search

- Micro-markets in "query × advertiser" graph



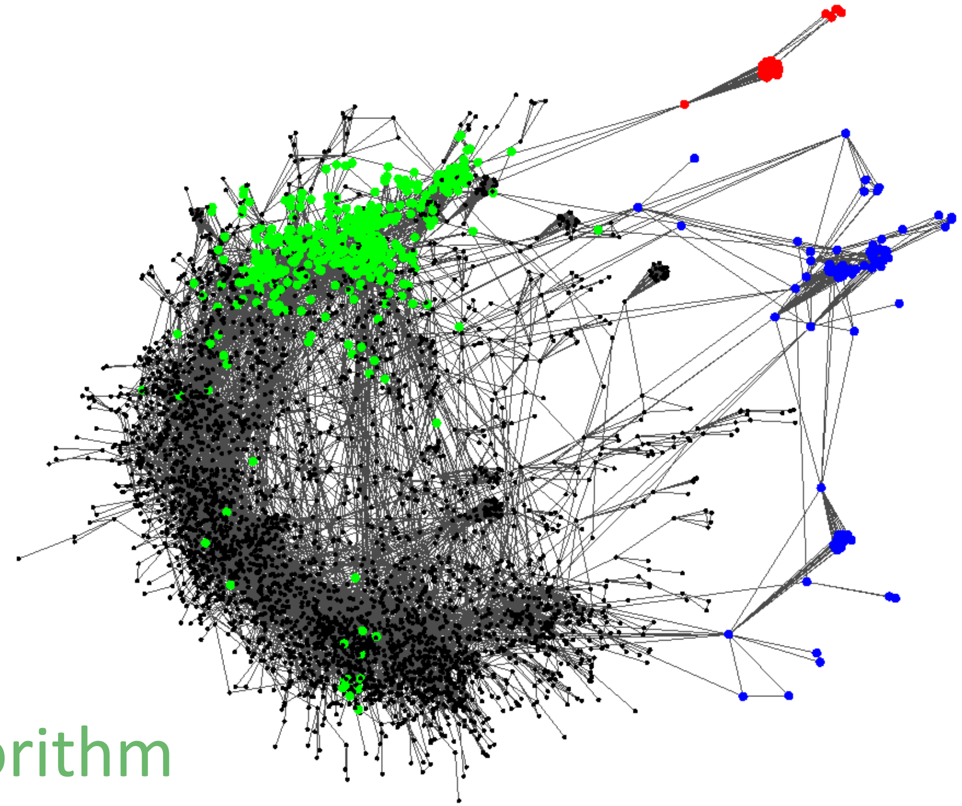
Social Network Data

- Zachary's Karate club network:



Finding communities

- Given a network:



- Want to find clusters!
- Need to:
  - Formalize the notion of a cluster
  - Need to design an algorithm that will find sets of nodes that are “good” clusters

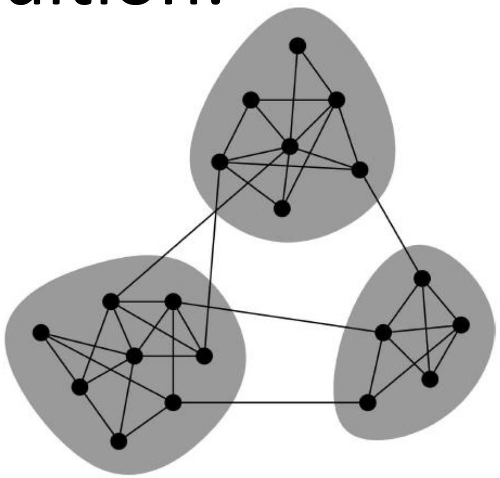
This talk: Focus and issues

- Our focus:
  - Objective functions that formalize notion of clusters
  - Algorithms/heuristic that optimize the objectives
- We explore the following issues:
  - Many different formalizations of clustering objective functions
  - Objectives are NP-hard to optimize exactly
  - Methods can find clusters that are systematically "biased"
  - Methods can perform well/poorly on some kinds of graphs

This talk: Comparison

- Our plan:
  - 40 networks, 12 objective functions, 8 algorithms
  - Not interested in “best” method but instead focus on finer differences between methods
- Questions:
  - How well do algorithms optimize objectives?
  - What clusters do different objectives and methods find?
  - What are structural properties of those clusters?
  - What methods work well on what kinds of graphs?

Clustering objective functions



- Essentially all objectives use the intuition:
  A good cluster S has
  - Many edges internally
  - Few edges pointing outside
- Simplest objective function:
  Conductance

  $$\Phi(S) = \#\text{edges outside } S \;/\; \#\text{edges inside } S$$

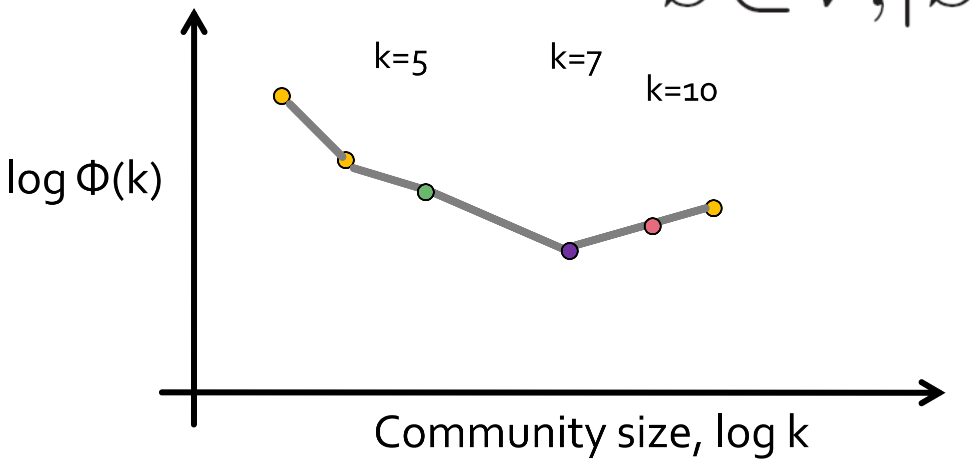
  - **Small conductance** corresponds to good clusters
- Many other formalizations of basically the same intuition (in a couple of slides)

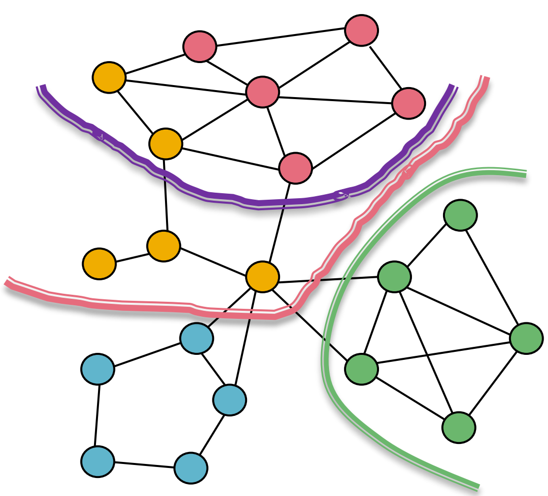
Experimental methodology

- How to quantify performance:
  - What is the score of clusters across a range of sizes?
- Network Community Profile (**NCP**) [Leskovec et al. '08]

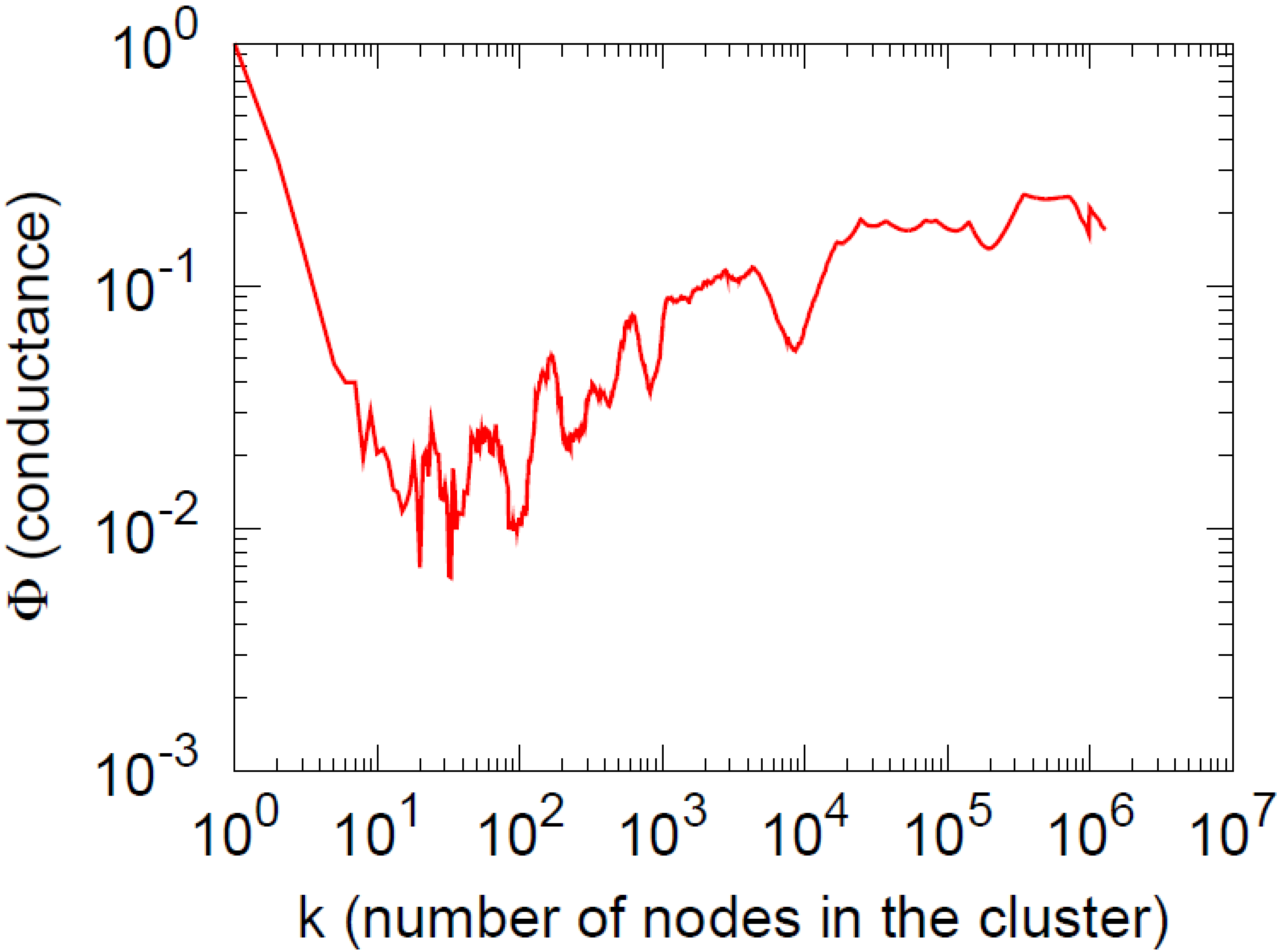
The score of best cluster of size *k*

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





Typical NCP



10^0
10^{-1}
10^{-2}
10^{-3}
Φ (conductance)
10^0 10^1 10^2 10^3 10^4 10^5 10^6 10^7
k (number of nodes in the cluster)

Plan for the talk

- Comparison of algorithms
  - Flow and spectral methods
  - Other algorithms
- Comparison of objective functions
  - 12 different objectives
- Algorithm optimization performance
  - How good job do algorithms do with optimization of the objective function

Many classes of algorithms

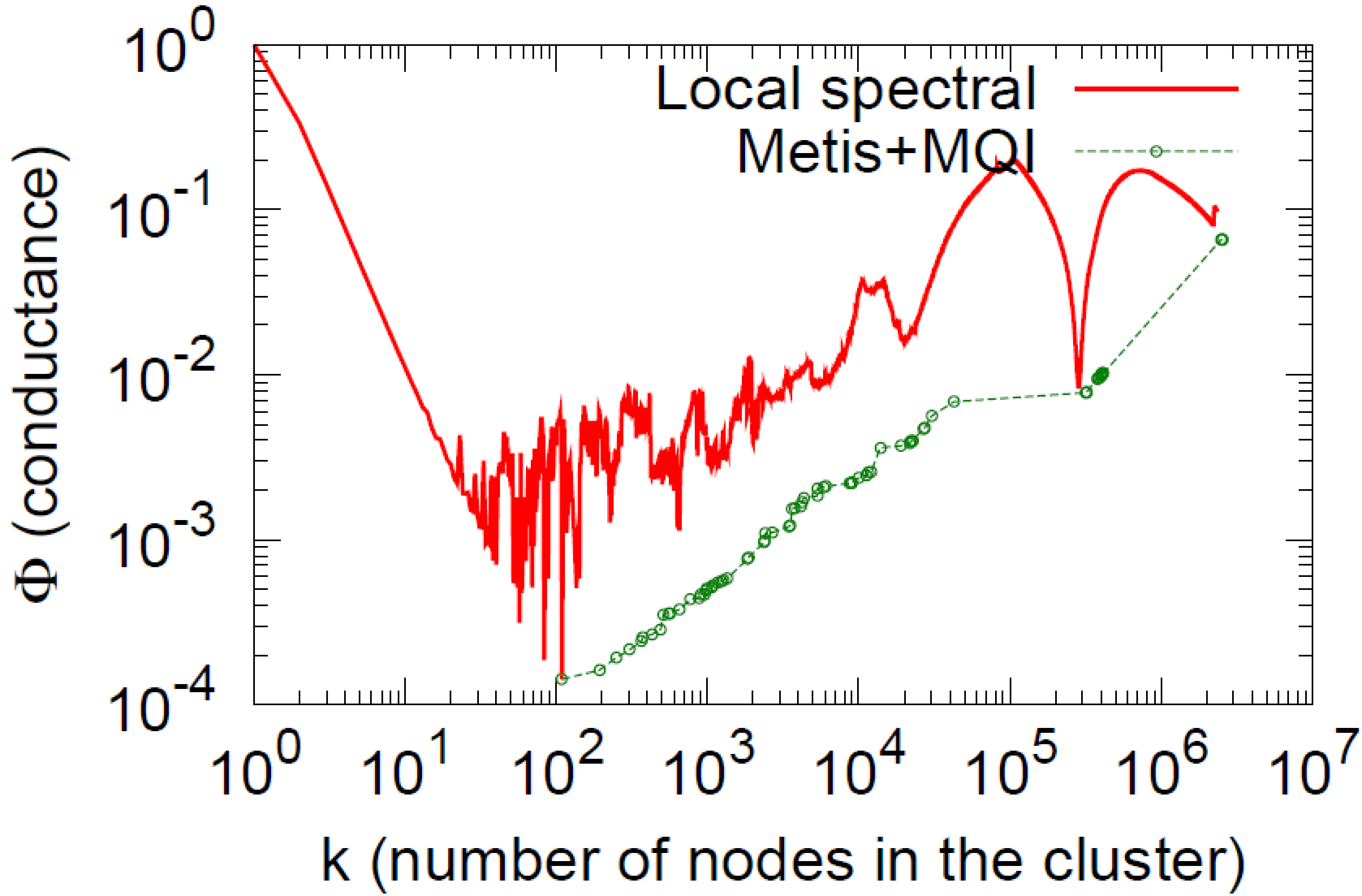
Many algorithms to extract clusters:

- Heuristics:
  - Metis, Graclus, Newman’s modularity optimization
    - Mostly based on local improvements
  - MQI: flow based post-processing of clusters
- Theoretical approximation algorithms:
  - Leighton-Rao: based on multi-commodity flow
  - Arora-Rao-Vazirani: semidefinite programming
  - Spectral: most practical but confuses “long paths” with “deep cuts”

Clusters based on conductance

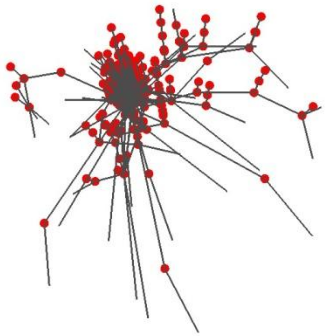
- Practical methods for finding clusters of good conductance in large graphs:
  - Heuristic:
    - Metis+MQI [Karypis-Kumar ‘98, Lang-Rao ‘04]
  - Spectral method:
    - Local Spectral [Andersen-Chung ’06]
- Questions:
  - How well do they optimize conductance?
  - What kind of clusters do they find?

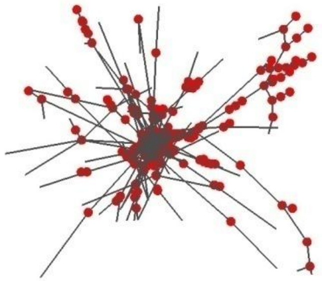
Results (LiveJournal network)



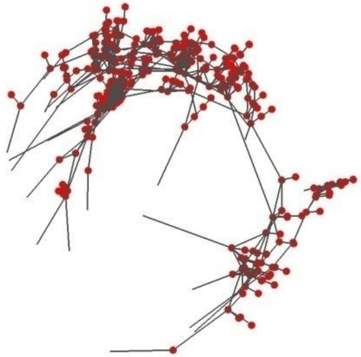
Properties of clusters (1)

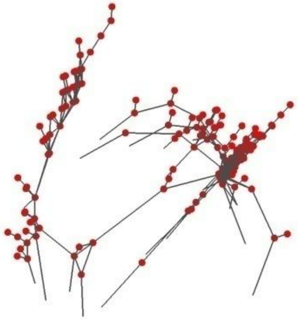
500 node communities from Local Spectral:



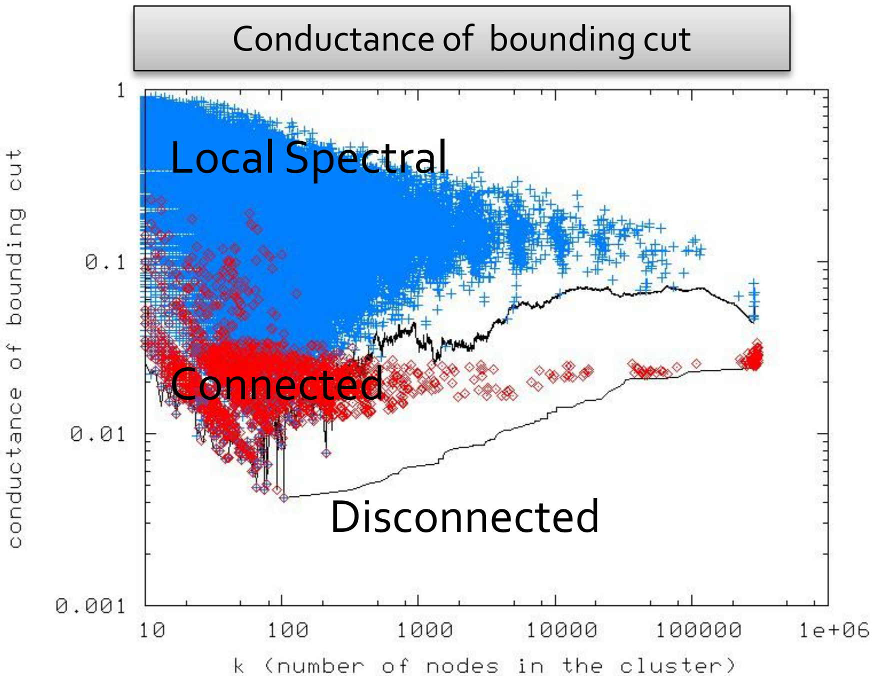


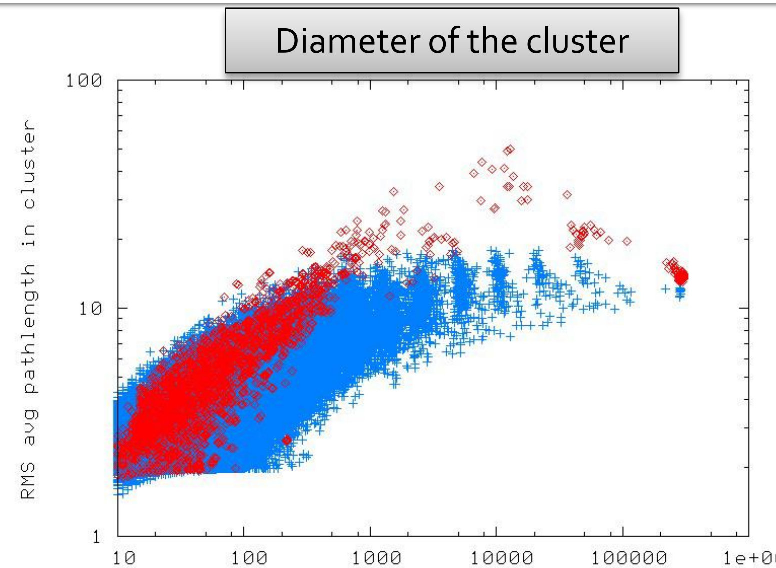
500 node communities from Metis+MQI:

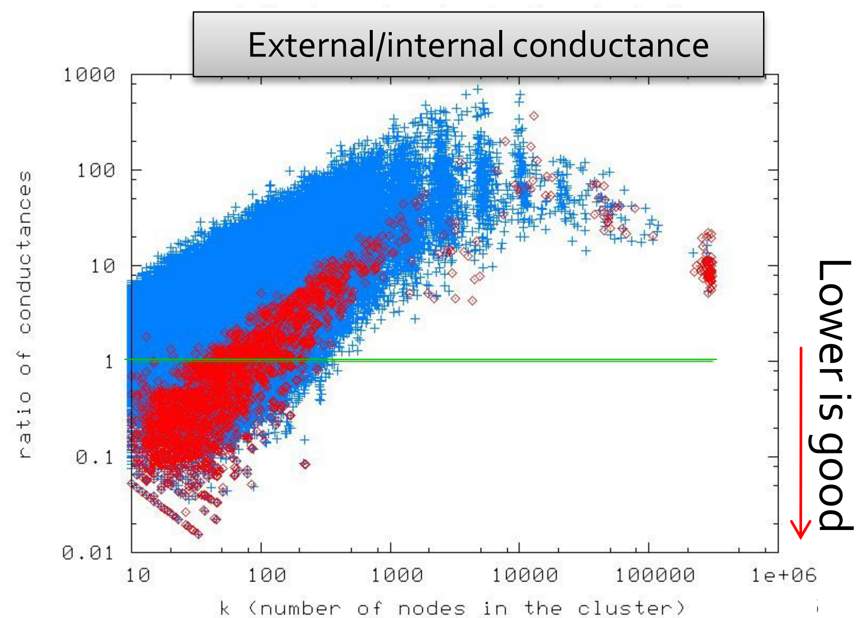




Properties of clusters (2)



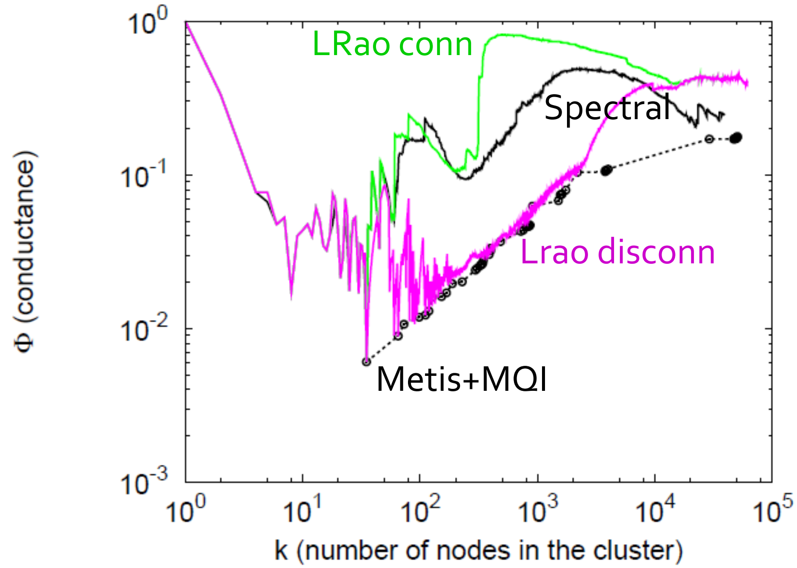


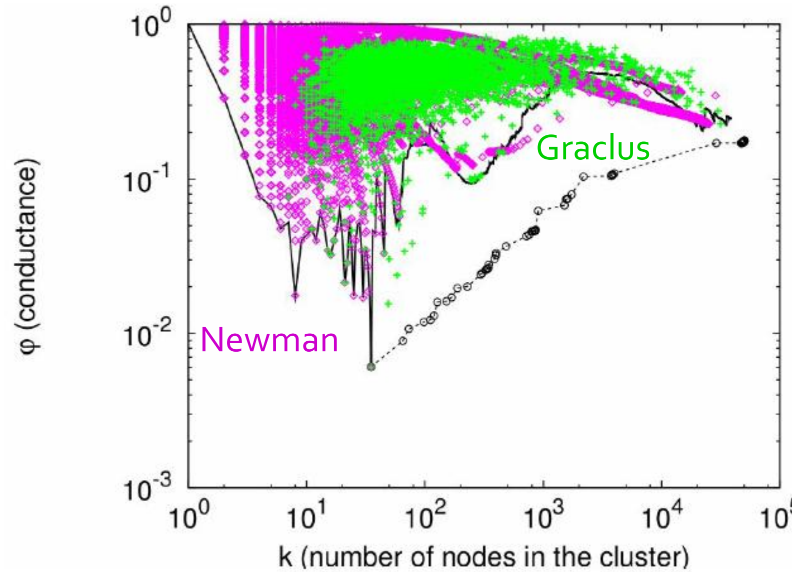


- Metis+MQI (red) gives sets with better conductance
- Local Spectral (blue) gives tighter and more well-rounded sets.

Other clustering methods

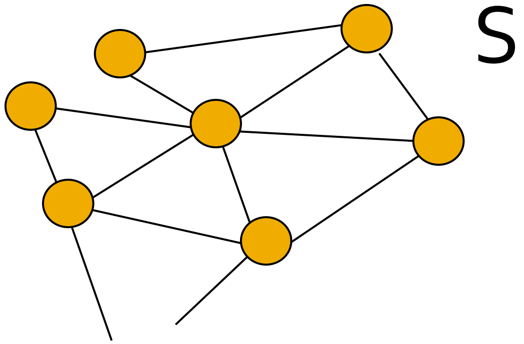
- LeightonRao: based on multi-commodity flow
  - Disconnected clusters vs. Connected clusters
- Graclus prefers larger clusters
- Newman's modularity optimization similar to Local Spectral





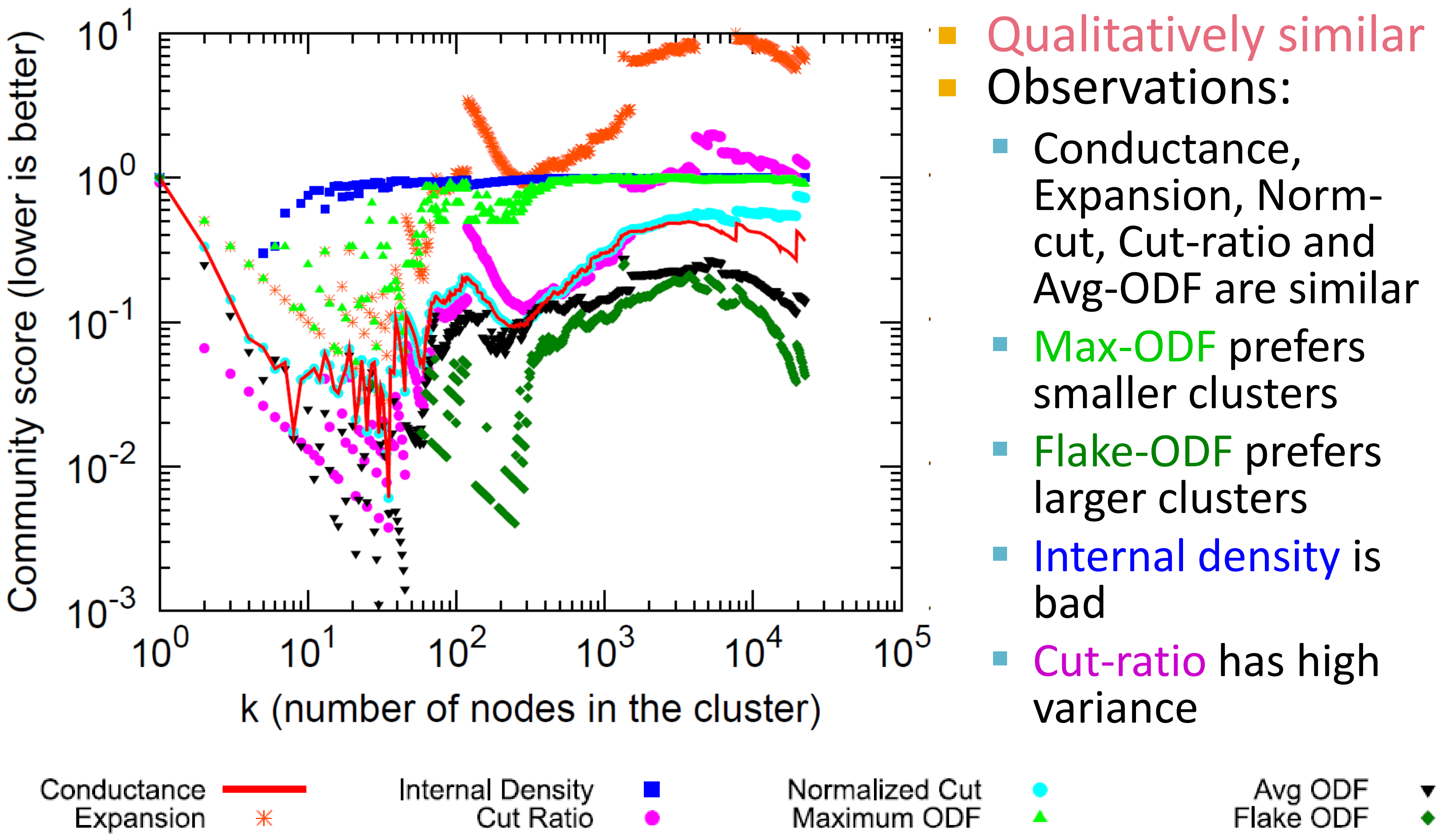
8 objective functions

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



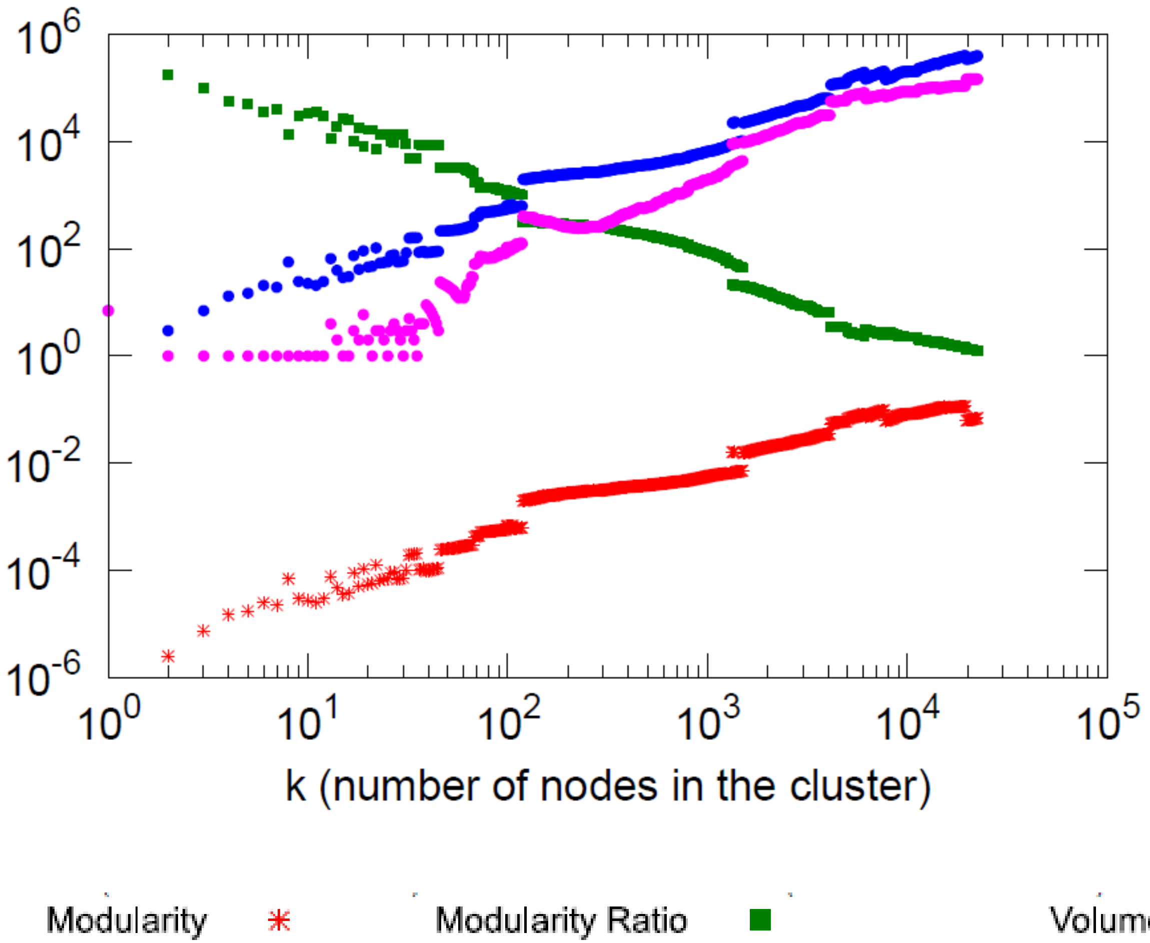
n: nodes in S
m: edges in S
c: edges pointing outside S

Multi-criterion objectives



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

Single-criterion objectives

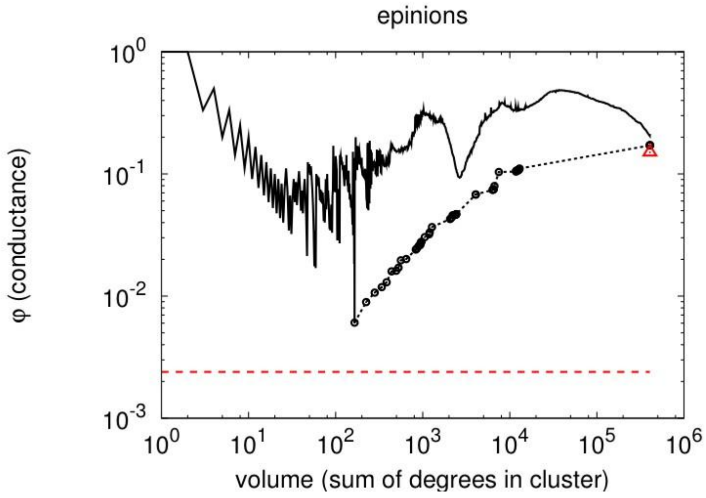


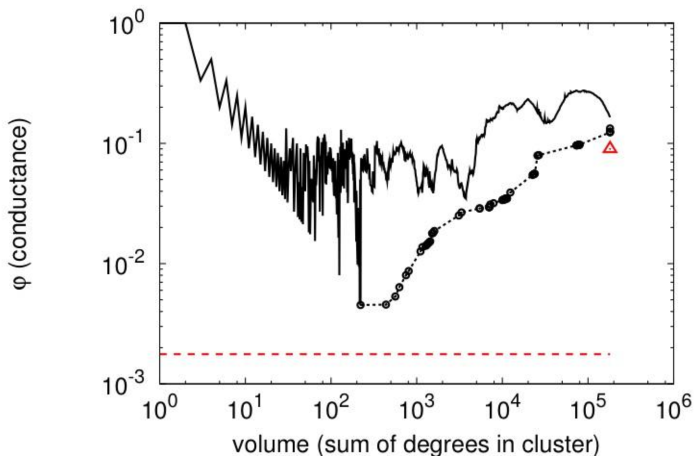
Observations:

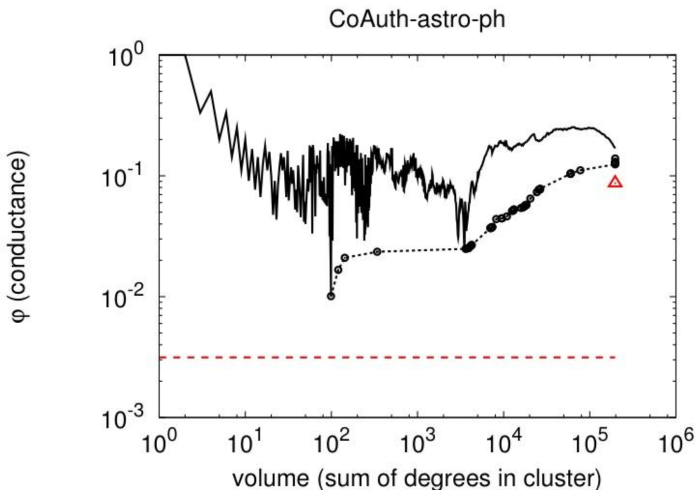
- All measures are monotonic
- Modularity
  - prefers large clusters
  - Ignores small clusters

Lower and upper bounds

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







Conclusion

- NCP reveals global network community structure:
  - Good small clusters but no big good clusters
- Community quality objectives exhibit similar qualitative behavior
- Algorithms do a good job with optimization
- Too aggressive optimization of the objective leads to “bad” clusters

THANKS!

Data + Code:

http://snap.stanford.edu