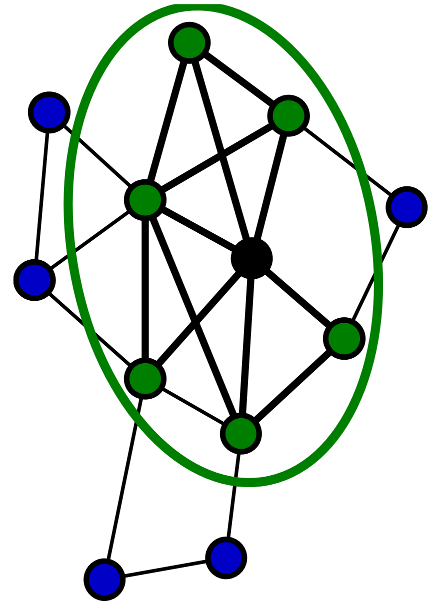
SAND2013-6812P

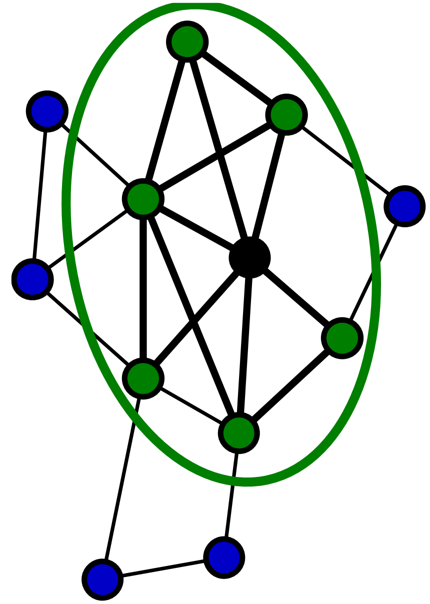
Vertex Neighborhoods, Low Conductance Cuts, and Good Seeds for Local Community Methods (KDD 2012)



DAVID F. GLEICH
PURDUE

SESH COMANDUR
SANDIA - LIVERMORE

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000."



A vertex neighborhood is a “good” conductance community in a graph with a heavy-tailed degree distribution and large clustering coefficient.

Our contributions

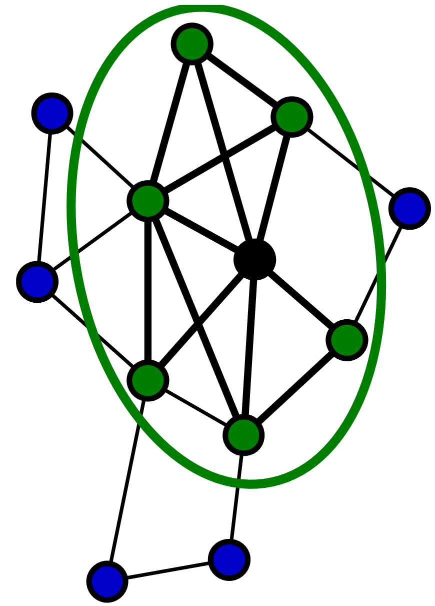
1. The previous theorem and its proof. This shows that good communities are *expected and easy to find* in modern networks with heavy-tailed degrees and large clustering.

2. An empirical evaluation of neighborhood communities that shows vertex neighborhoods are the “backbone” of the network community profile.

Formal background for the theorem

1. Vertex neighborhoods
2. Low conductance cuts
3. Clustering coefficients

Vertex neighborhoods



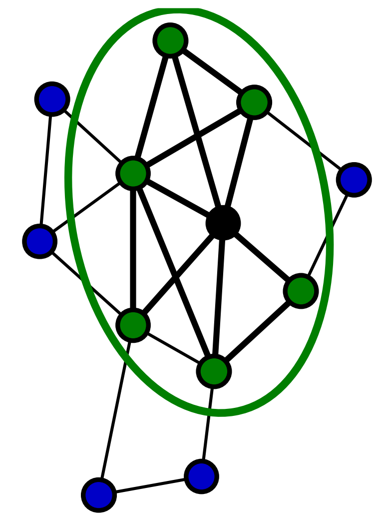
The set of a vertex and all its neighborhood

Also called an "egonet"

Prior research on egonets of social networks from the "structural holes" perspective *[Burt95,Kleinberg08]*.

Used for anomaly detection *[Akoglu10]*, community seeds *[Huang11,Schaeffer11]*, overlapping communities *[Schaeffer07,Rees10]*.

Conductance communities



Conductance is one of the most important community scores [Schaeffer07]

The conductance of a set of vertices is the ratio of edges leaving to total edges:

$$\phi(S) = \frac{\text{cut}(S)}{\min\left(\text{vol}(S), \text{vol}(\bar{S})\right)}$$

(edges leaving the set)

(total edges in the set)

$\text{cut}(S) = 7$

$\text{vol}(S) = 33$

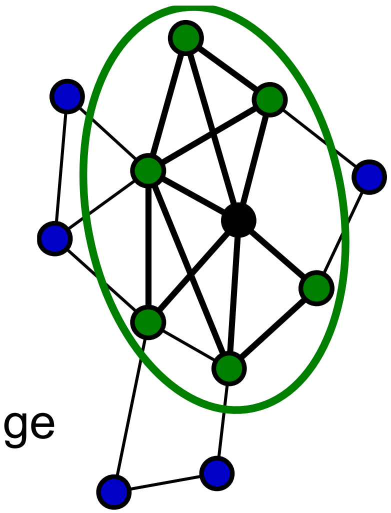
$\text{vol}(\bar{S}) = 11$

$\phi(S) = 7/11$

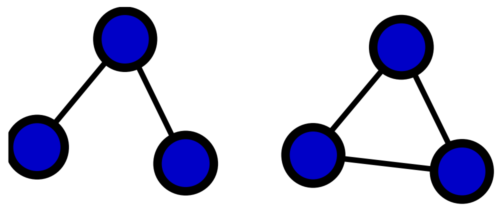
Equivalently, it's the probability that a random edge leaves the set.

Small conductance ⇔ Good community

Clustering coefficients



Wedge



Global clustering coefficient

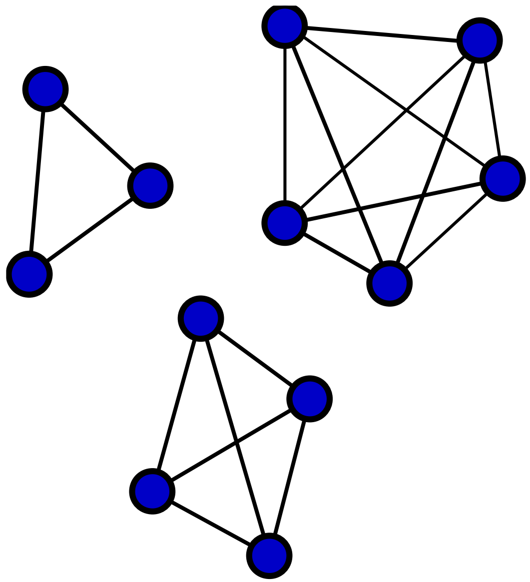
$$\kappa = \frac{\text{number of closed wedges}}{\text{number of wedges}}$$

Probability that a random wedge is closed

Simple version of theorem

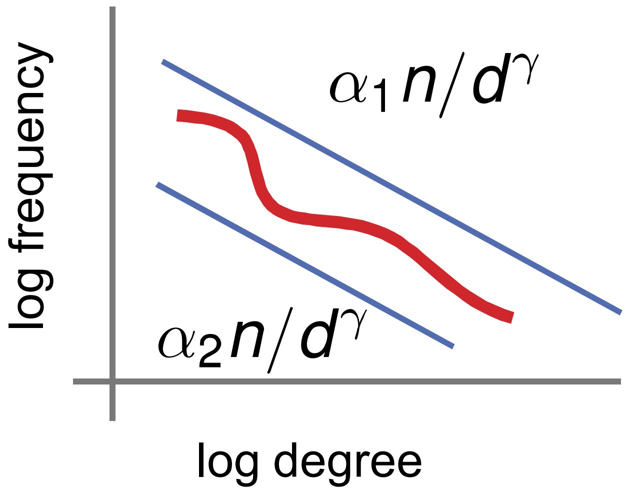
If global clustering coefficient = 1, then the graph is a disjoint union of cliques.

Vertex neighborhoods are optimal communities!



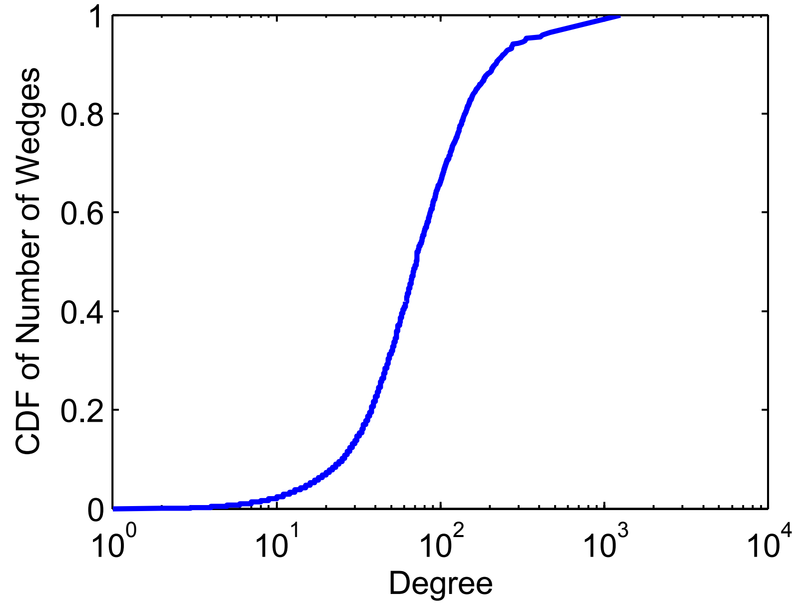
Theorem

Condition: Let graph G have clustering coefficient κ and have vertex degrees bounded by a power-law function with exponent γ less than 3.



Theorem: Then there exists a vertex neighborhood with conductance $\leq 4(1-\kappa)/(3-2\kappa)$

Proof Sketch



1) Large clustering coefficient
⇒ many wedges are closed

2) Heavy tailed degree dist
⇒ a few vertices have a very large degree

3) Large degree ⇒ O(d^2) wedges ⇒ “most” of wedges

Thus, there must exist a vertex with a high edge density ⇒ “good” conductance

Use the probabilistic method to formalize

Confession
The theory is really weak

$$\phi(S) \leq 4(1-\kappa)/(3-2\kappa)$$

This bound is useless unless $\kappa \geq 1/2$

Graph	Verts	Edges	κ	$\bar{C}$
ca-AstroPh	17903	196972	0.318	0.633
email-Enron	33696	180811	0.085	0.509
cond-mat-2005	36458	171735	0.243	0.657
arxiv	86376	517563	0.560	0.678
dblp	226413	716460	0.383	0.635
hollywood-2009	1069126	56306653	0.310	0.766
fb-Penn94	41536	1362220	0.098	0.212
fb-A-oneyear	1138557	4404989	0.038	0.060
fb-A	3097165	23667394	0.048	0.097
soc-LiveJournal1	4843953	42845684	0.118	0.274
oregon2-010526	11461	32730	0.037	0.352
p2p-Gnutella25	22663	54693	0.005	0.005
as-22july06	22963	48436	0.011	0.230
itdk0304	190914	607610	0.061	0.158

Collaboration networks
$\kappa \sim [0.1 – 0.5]$

Social networks
$\kappa \sim [0.05 – 0.1]$

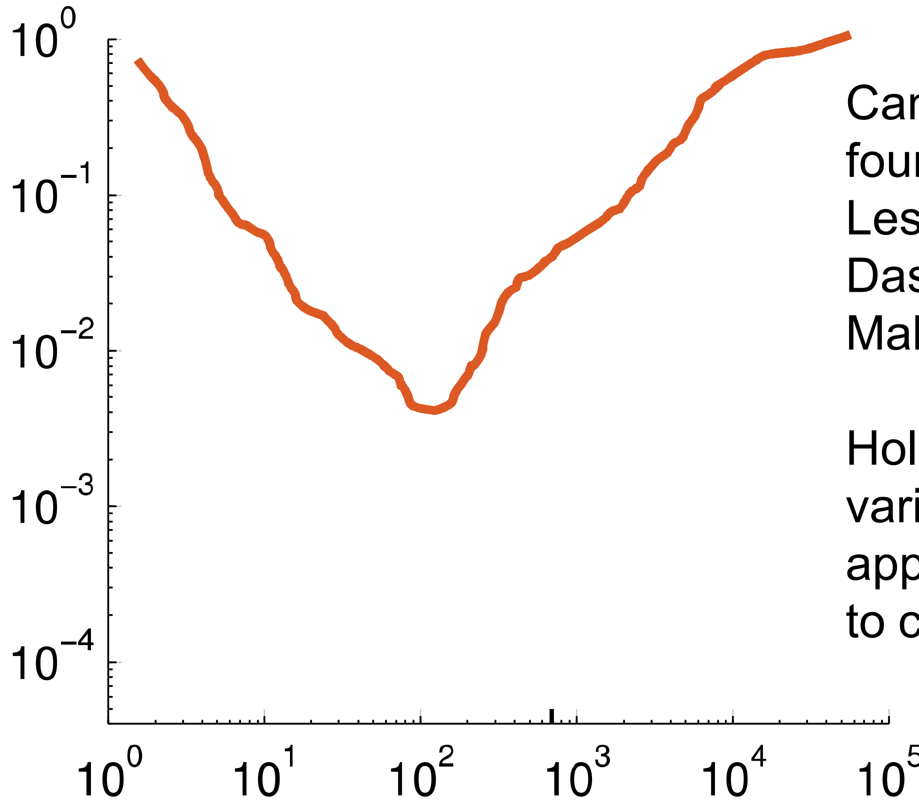
Tech. networks
$\kappa \sim [0.005 – 0.05]$

We view this theory as

“intuition for the truth”

Network Community Profiles

Minimum conductance for any community of the given size



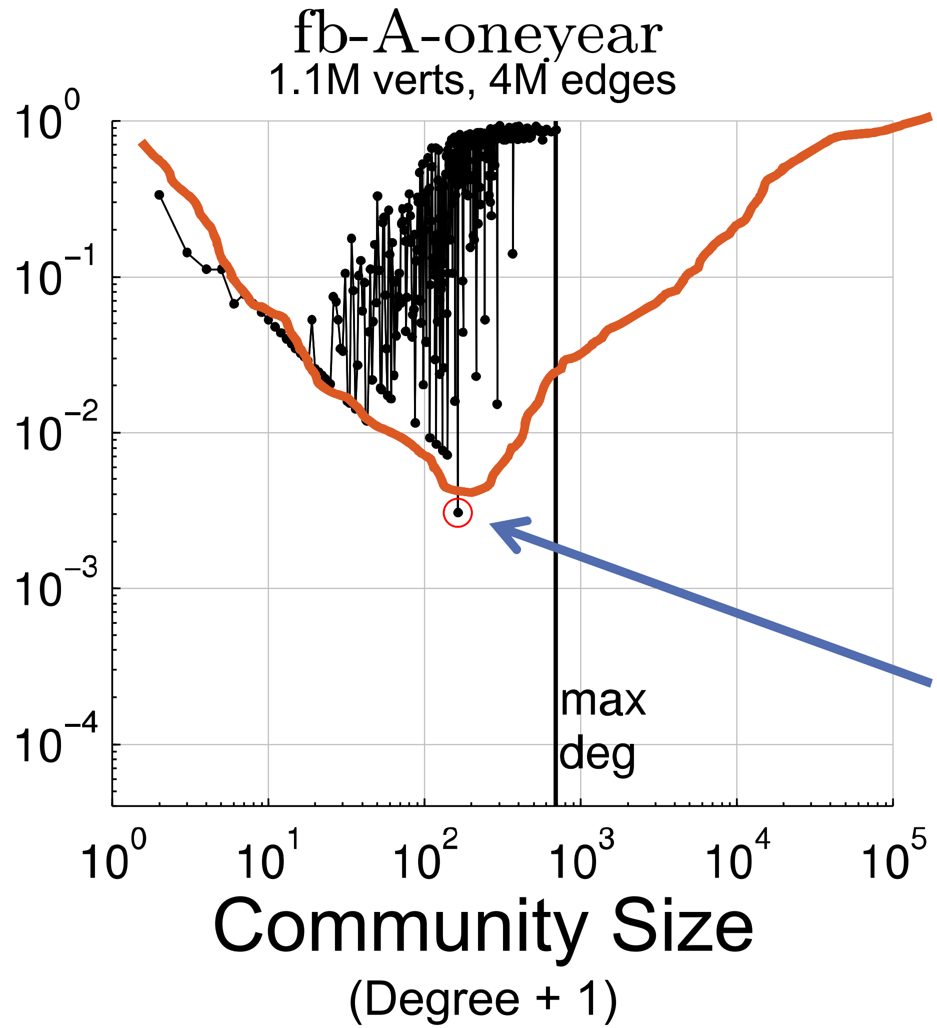
Community Size

Canonical shape found by Leskovec, Lang, Dasgupta, and Mahoney

Holds for a variety of approximations to conductance.

Network Community Profiles

Minimum conductance for any ~~community~~ neighborhood of the given size



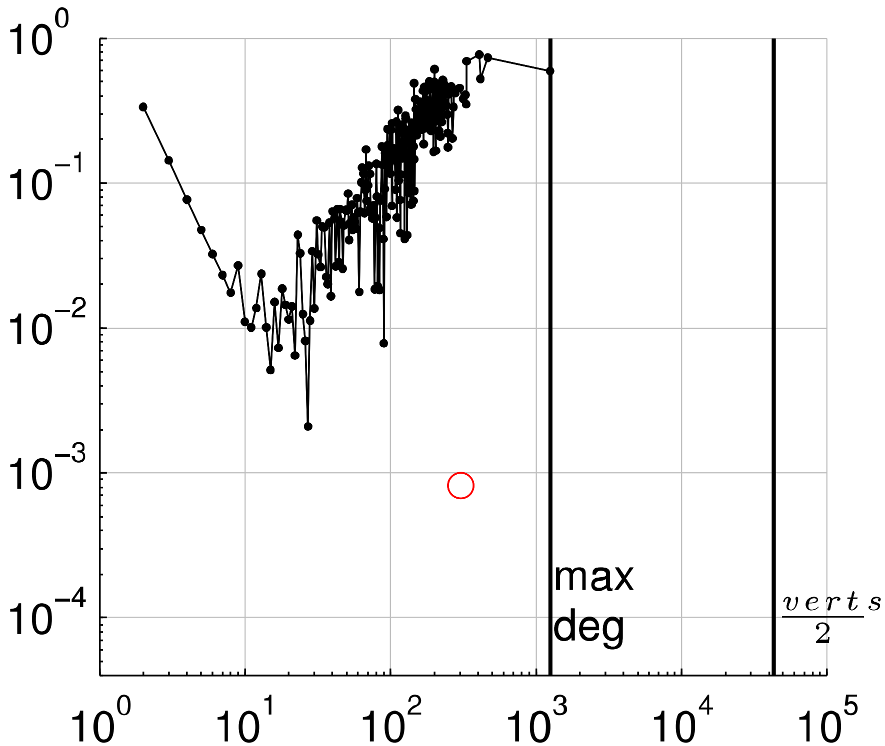
Facebook data from Wilson et al. 2009

"Egonet community profile" shows the same shape, 3 secs to compute.

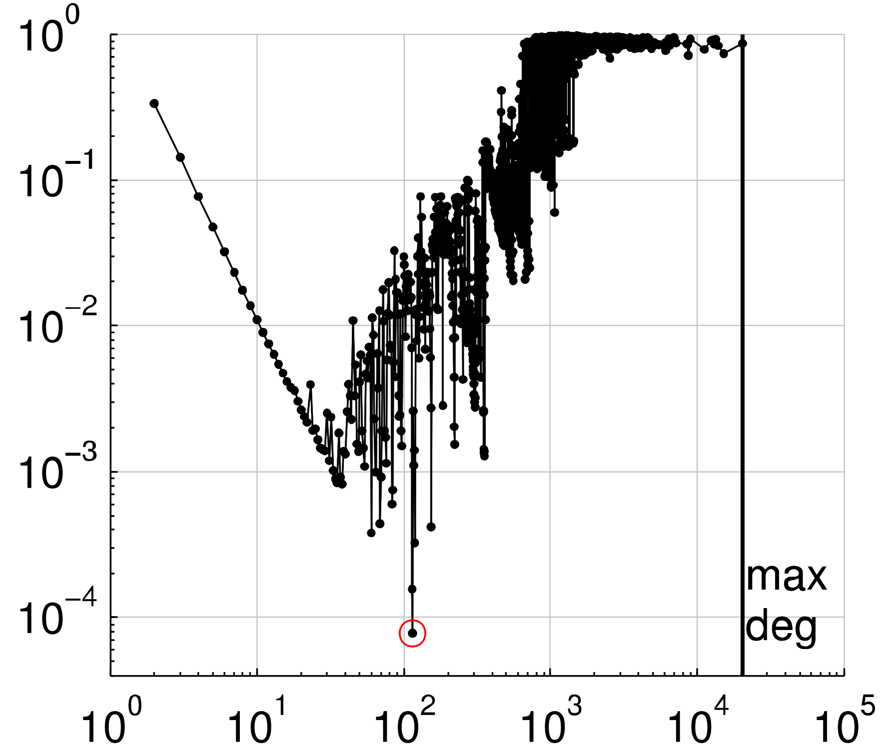
The Fiedler community computed from the normalized Laplacian is a neighborhood!

Not just one graph

arXiv – 86k verts, 500k edges



soc-LiveJournal – 5M verts, 42M edges



15 more graphs available

www.cs.purdue.edu/~dgleich/codes/neighborhoods

Communities [Andersen06]

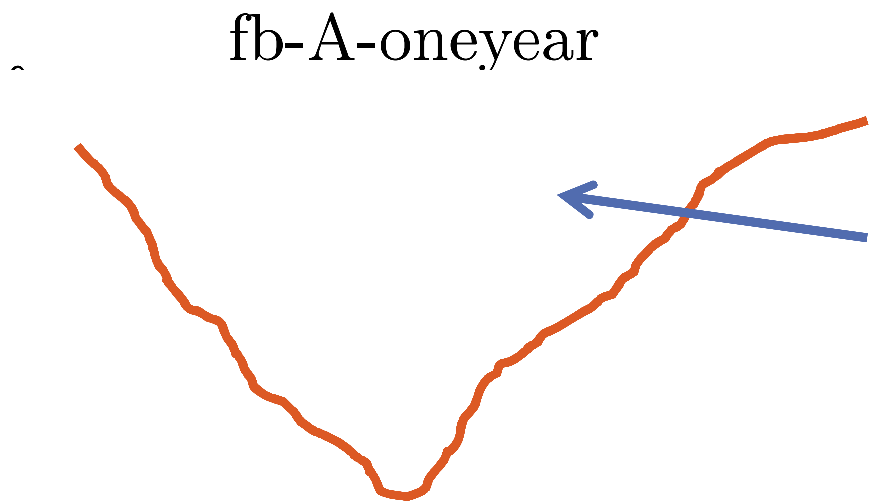
To find the canonical NCP structure, Leskovec et al. used a personalized PageRank based community finder.

These start with a single vertex seed, and then expand the community based on the solution of a personalized PageRank problem.

The resulting community satisfies a local Cheeger inequality.

This needs to run thousands of times for an NCP

Network Community Profile



Minimum conductance for any community of the given size

7807 seconds

This region fills when using the PPR method (like now!)

Community Size

5

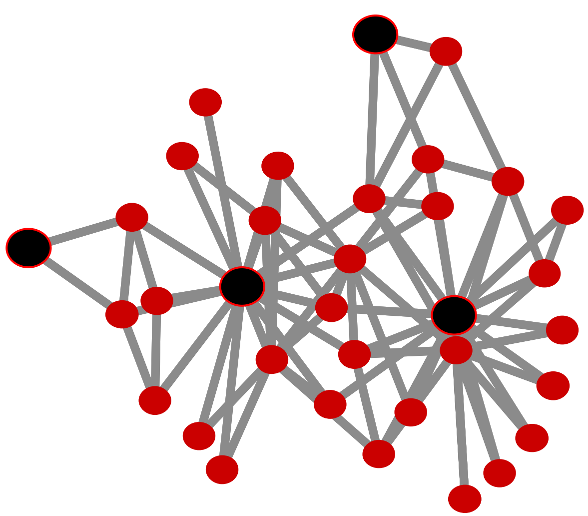
Vertex Neighborhoods, Low Conductance Cuts, and Good Seeds for Local Community Methods

Locally Minimal Communities

"My conductance is the best locally."

$$\phi(N(v)) \leq \phi(N(w))$$

w adjacent to v

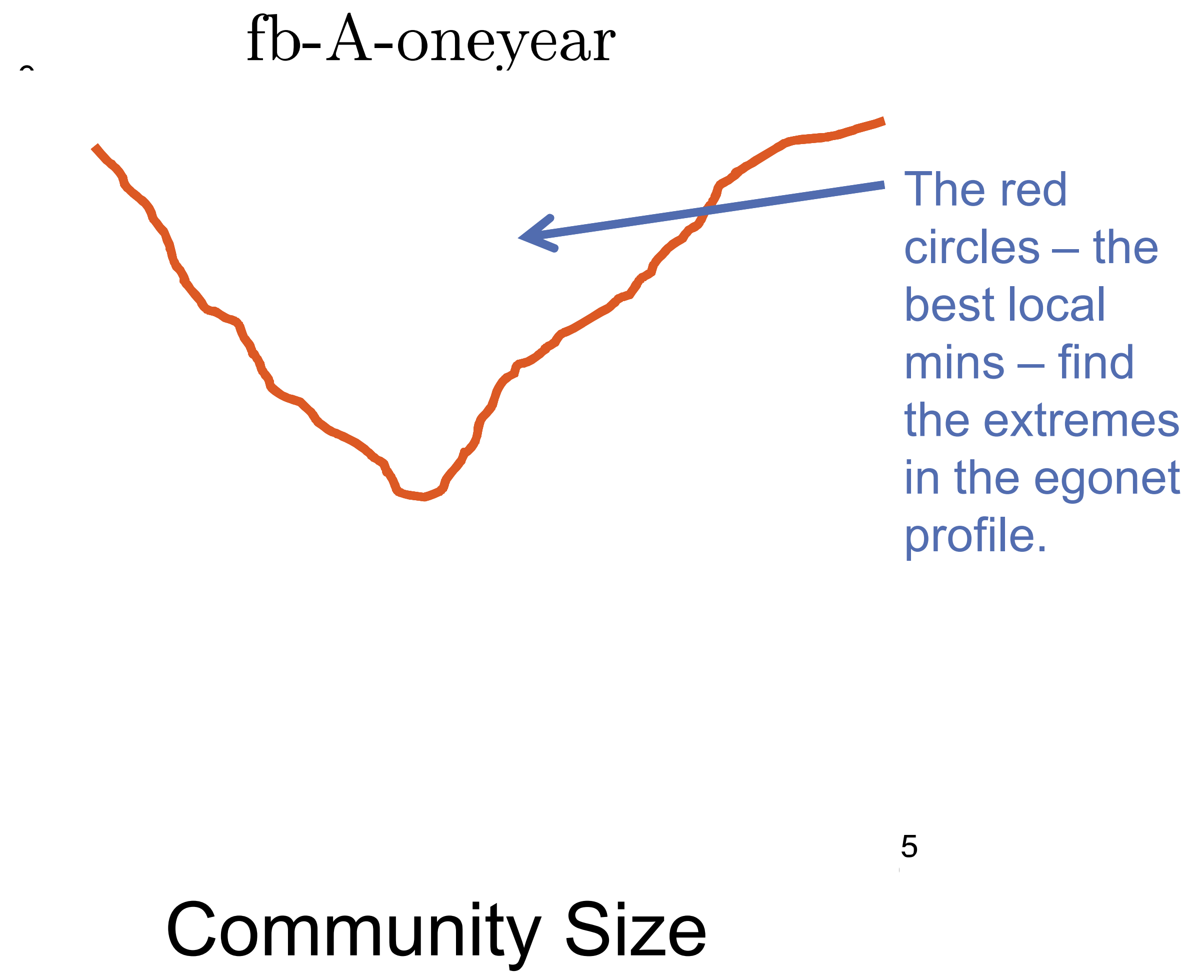


In Zachary's Karate Club network, there are four locally minimal communities

capture extremal neighborhoods

Red dots are conductance and size of a locally minimal community

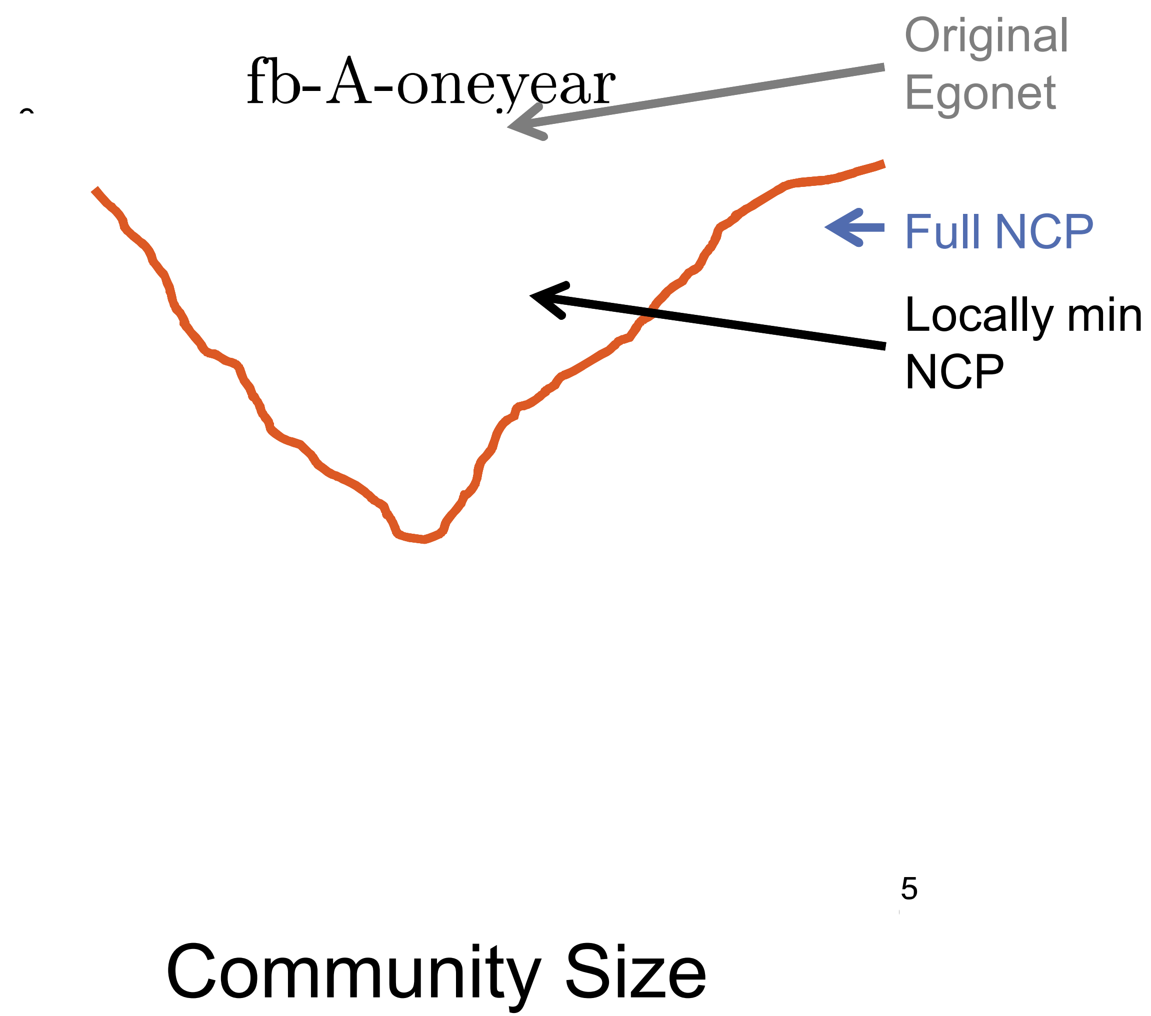
Usually about 1% of # of vertices.



Growing locally minimal comm.

Growing only locally minimal communities

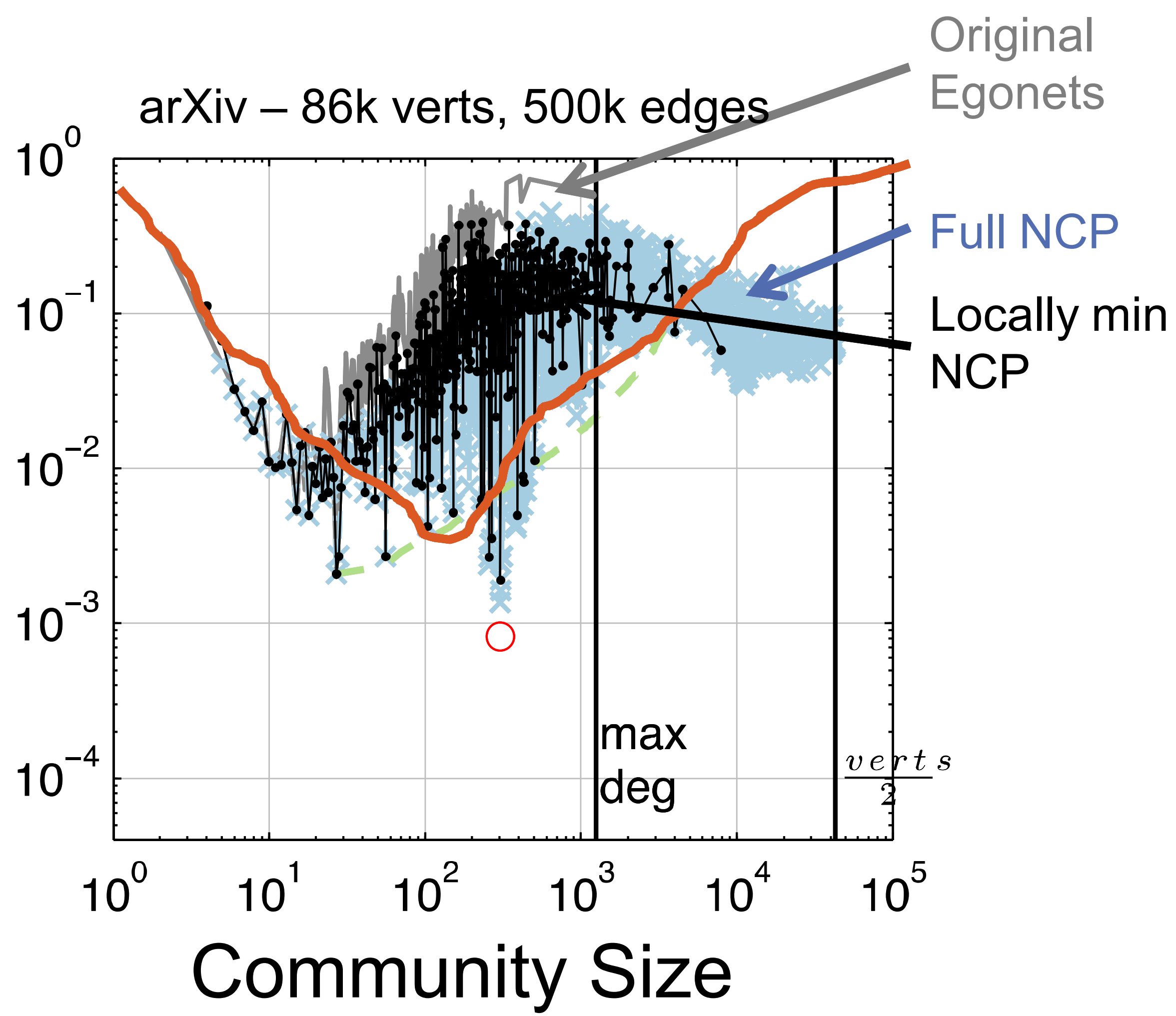
283 seconds
vs.
7807 seconds



Growing locally minimal comm.

Growing only locally minimal communities

143 seconds vs. 2211 seconds



Recap

Acknowledgements

David supported by NSF CAREER award 1149756-CCF.

Sesh supported by the Sandia LDRD program (project 158477) and the applied mathematics program at the Dept. of Energy.

A theorem relating clustering, heavy-tailed degrees, and low-conductance cuts of vertex neighborhoods.

Empirical evaluation of vertex neighborhoods.

More on k-cores in the paper.

$\Rightarrow$ Many communities are *easy to find*!

$\Rightarrow$ Explains success of community detection?