SAND2016-7336PE

Social Network Analysis Research at Sandia (a non-comprehensive survey)

Jon Berry (Sandia National Laboratories)

July 21, 2016

LDRD
LABORATORY DIRECTED RESEARCH & DEVELOPMENT

NNSA
National Nuclear Security Administration

CCR
Center for Computing Research

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.

Sandia National Laboratories

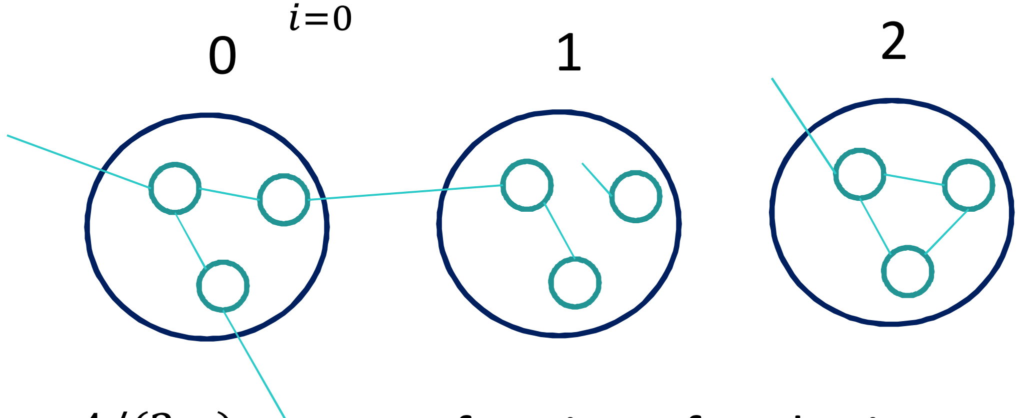
Outline

- Studying the structure of electronic social networks
  – Identifying, understanding, modeling
- Designing algorithms for electronic social networks
  – Basic, distributed, streaming, sampling, benchmarking
- "Cleaning" electronic social networks
  – Non-human activity violating social scientific assumptions
- Computing with electronic social networks
  – Multi-core, GPU, HPC, cloud

Studying the Structure

- Consider a network with *n* vertices and *m* edges
- "Communities": the most familiar "structure"
- "Community detection": the most familiar problems
  - "Modularity": the most familiar way of measuring comm. Str.

$$\sum_{i=0}^{c} (e_{ii} - a_i^2)$$



$e_{00} = 4/(2m)$

$a_0 = 7/(2m)$

e_{ii}: fraction of endpoints w/in comm. i

a_i: fraction of endpoints touching comm. i

Modularity Maximization

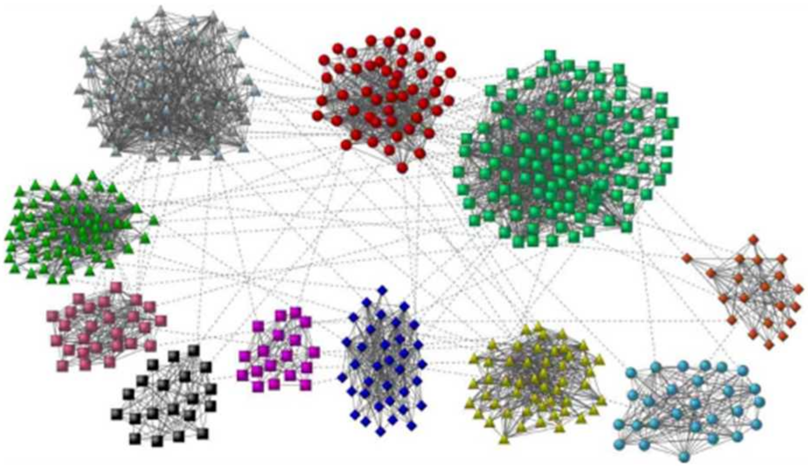


Image: Lancichinetta, Fortunato, Radicchi, Physical Review E (78) 046110, 2008

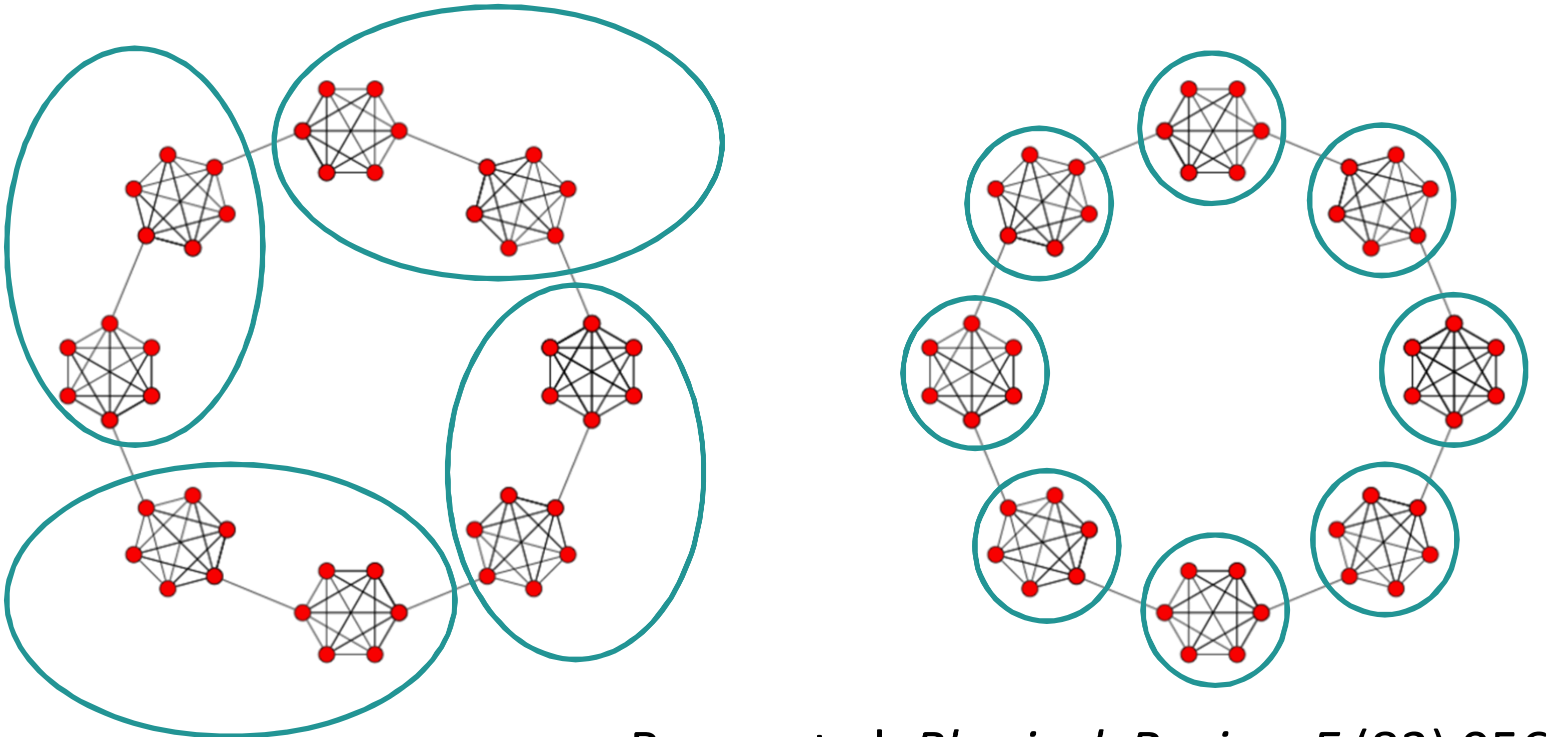
Thousands of algorithms, any of which suffers a "Resolution limit"

Cannot "resolve" communities with fewer than $\sqrt{\frac{m}{2}}$ edges

(Fortunato and Barthelemy, PNAS 2007)

Sandia Work: "Tolerate" the Resolution Limit

The resolution limit

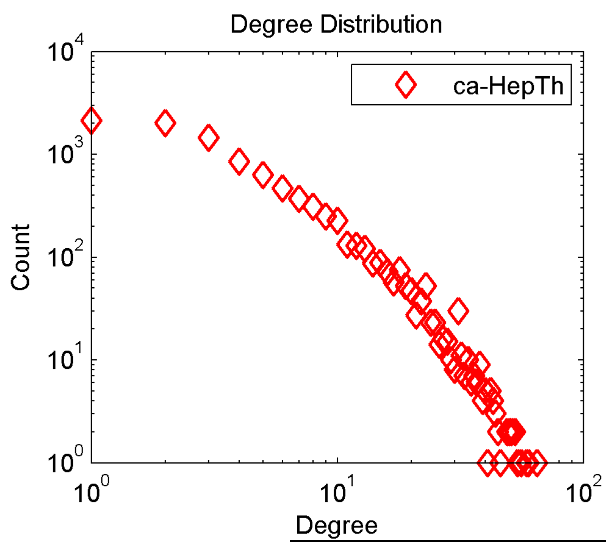


Berry, et al. *Physical Review E* (83) 056119, 2011

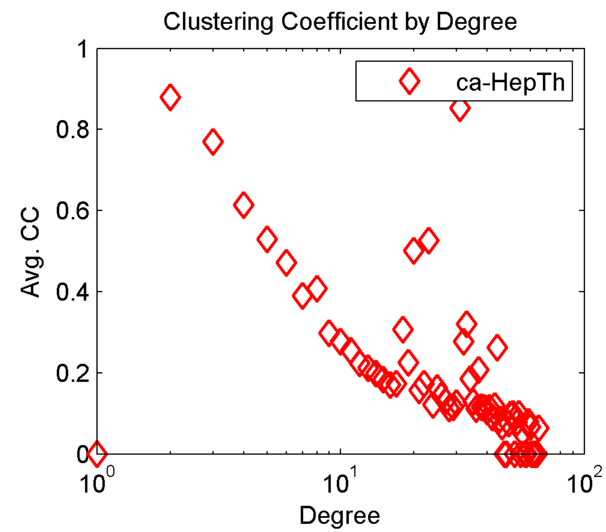
Weight edges, then resolve to $\sqrt{\frac{W}{\varepsilon}}$ where ε bounds inter-comm. edges

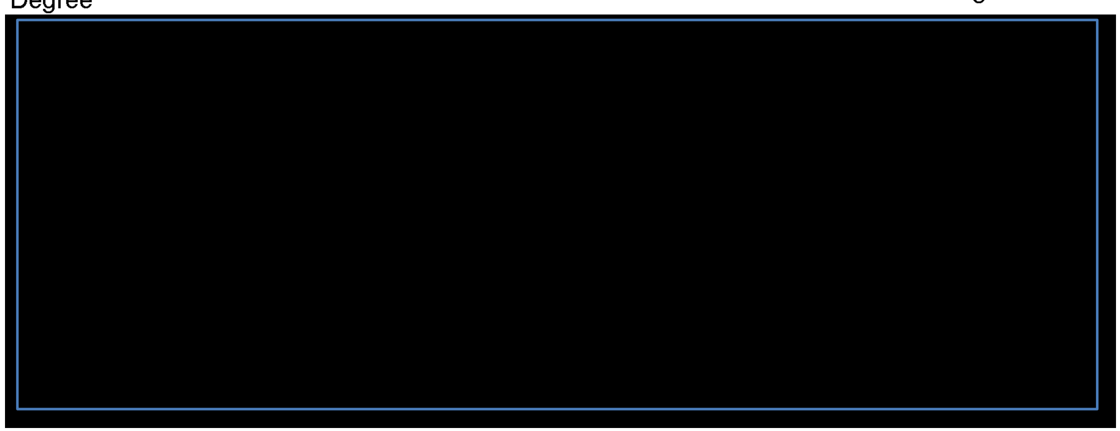
Now We'll Consider More Fundamental Structural Properties

Vertex degree distribution



Clustering coefficient distribution





Current Network Models Cannot Match Both Degree & Clust. Comp. Dists.

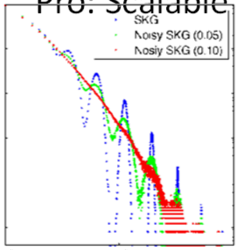
- **Erdös-Rényi** (1960)
  - All edges have equal probability
  - Con: Poisson degree distribution

- **Preferential Attachment** (Barabási-Albert 1999)
  - Nodes join the graph sequentially
  - Prefer nodes of higher degree
  - Pro: Power-law degree distribution
  - Con: Too few triangles

- **Stochastic Blockmodel** (Holland et al. 1983)
  - Each node belongs to a block
  - Edge probability between blocks
  - Pro: Explicit community structure
  - Con: Wrong degree distribution

$$\begin{bmatrix} 0.6 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}$$

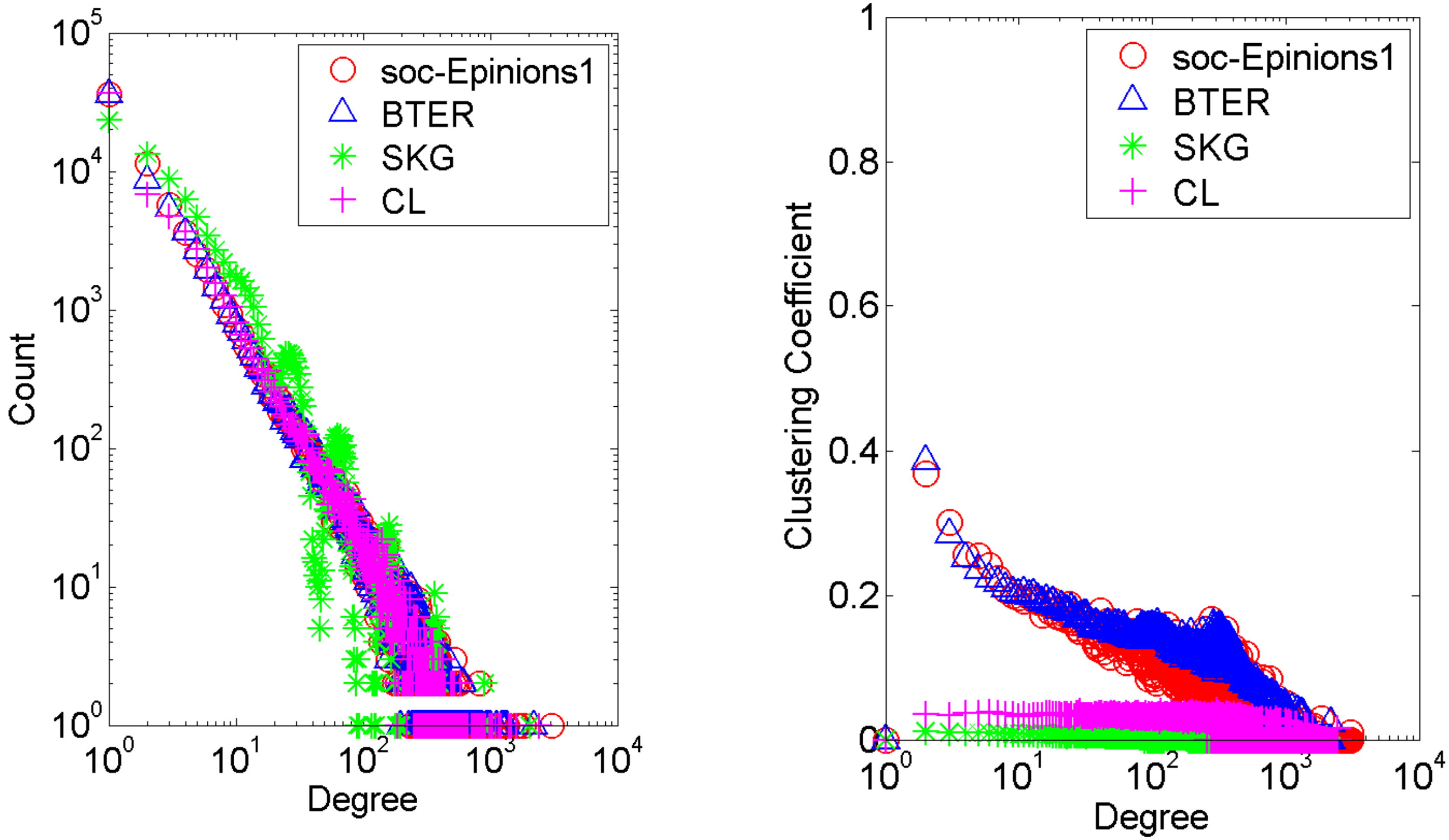
- **Stochastic Kronecker**, aka R-MAT (Chakrabarti et al. 2004)
  - Edge probabilities defined by Kronecker products of generator matrices
  - Pro: Scalable



  - Con: Wrong degree distribution
  - Con: Too few triangles

- **Chung-Lu** (2002), aka Configuration Model
  - Edge probabilities defined by desired degree of endpoints
  - Pro: Scalable
  - Pro: Matches many degree distributions
  - Con: Too few triangles

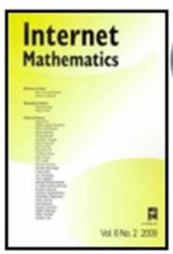
Sandia Work: "BTER Model" Captures Clustering Coefficients



Seshadhri, Kolda, Pinar *(*Phys. Rev. E 2012)
Kolda, Plantenga, Pinar, Seshadhri (SISC 2014)

Sandia Work: Quantify Triangle Counts

- The 4/3-moment of the degree distribution is the expected value of $d_v^{4/3}$ for any vertex v
- Sandia theoretical computer scientists, working with Iowa State statisticians, showed that if this moment is bounded by a constant, the number of triangles in a network is linear (efficiently listed!)



CrossMark
click for updates

Original Articles

Why Do Simple Algorithms for Triangle Enumeration Work in the Real World?

DOI: 10.1080/15427951.2015.1037030

Jonathan W. Berry[a], Luke A. Fostvedt[b], Daniel J. Nordman[a], Cynthia A. Phillips[c], C. Seshadhri[d*] & Alyson G. Wilson[e]

pages 555-571

Sandia National Laboratories

Designing Algorithms

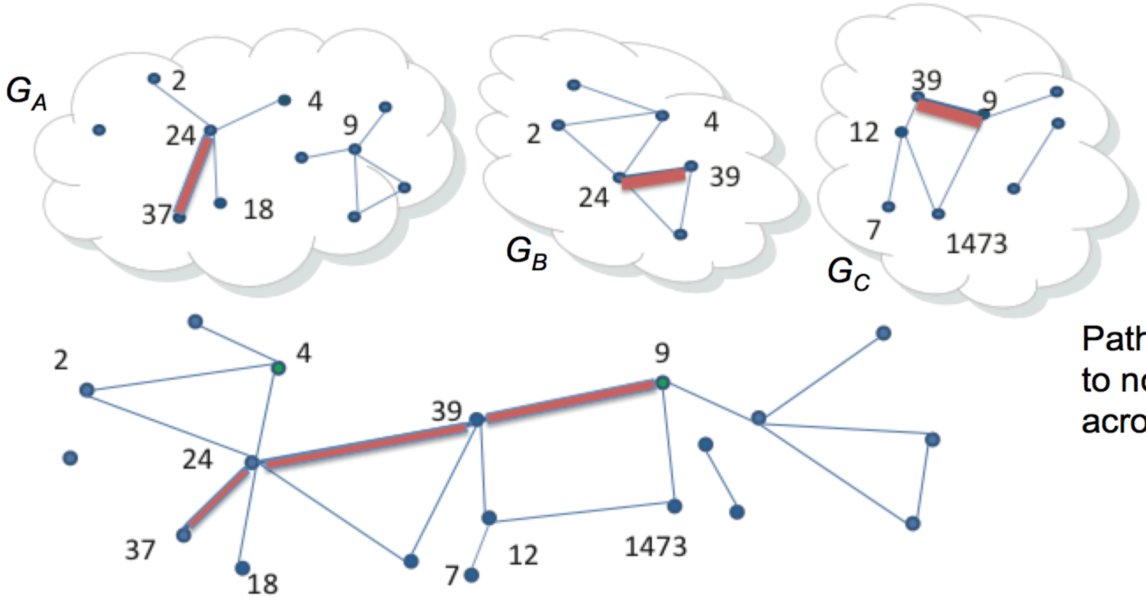
- Tamara Kolda, C. Seshadhri, A. Pinar, G. Ballard, K. Matulef, and other Sandia/CA staff have designed many efficient sampling algorithms for:
  - Wedges (paths with three vertices and two edges)
  - Triangles (3-cycles)
  - Diamonds (4-cycles)
  - See: http://www.sandia.gov/~tgkolda/pubs/index.html
- I'll focus on work in NM with Cindy Phillips
  – Distributed graph algorithms
  – "Cleaning" social networks

A New Distributed Computing Model

Alice and Bob (or more) independently create social graphs G_A and G_B.

- Alice and Bob each know nothing of the other's graph.
- Shared namespace. Overlap at nodes.

Goal: Cooperate to compute algorithms over G_A union G_B with limited sharing: $O(\log^k n)$ total communication for size n graphs, constant k



s-t connectivity
For 3 centers

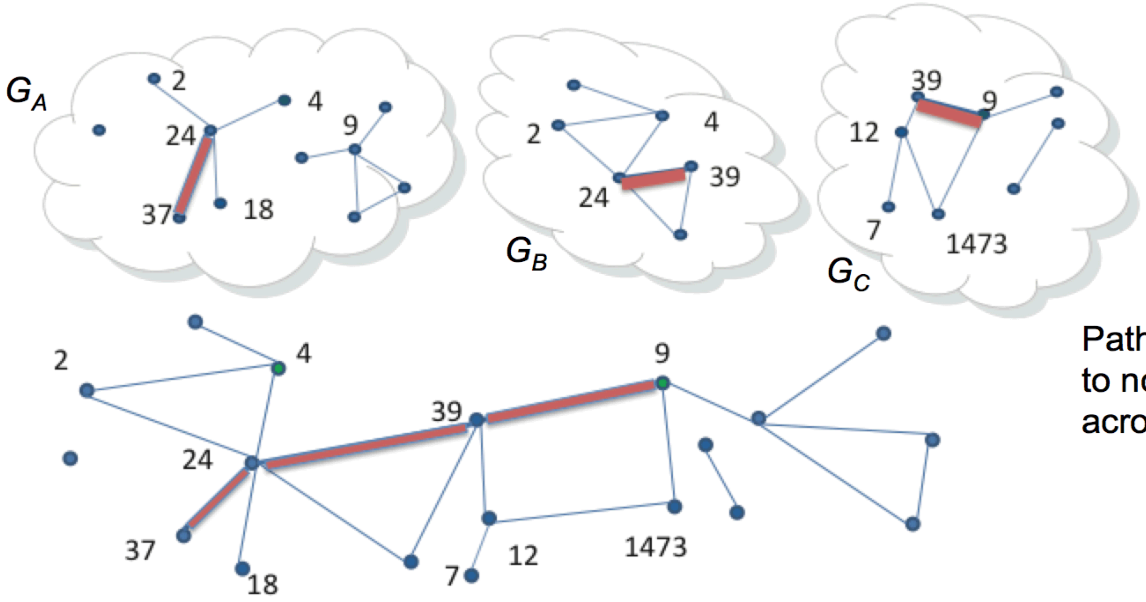
Path from node s=37
to node t=9 is split
across all three centers

Another Limited Sharing Model

Goal: Cooperate to compute algorithms over $G_A \cup G_B (\cup G_C \ldots)$

Alice gets no information beyond answer in honest-but-curious model.

- Secure multiparty computation
  - *Few players, large data (this context is new)*



s-t connectivity
For 3 centers

Path from node s=37 to node t=9 is split across all three centers

Motivation

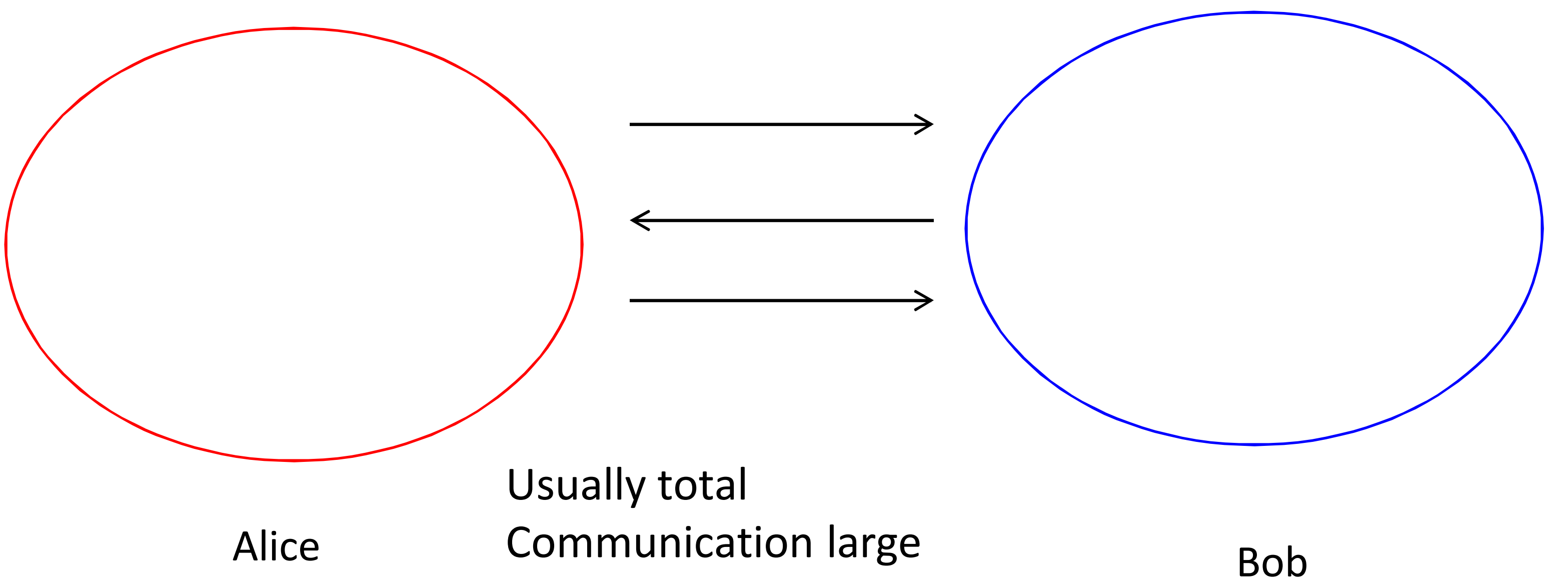
- Company mergers
- National security: connect-the-dots for counterterrorism
- Nodes are people
  - Exploit structure of social networks

Topics

- s-t connectivity
- Planted clique
- Engineering better test sets

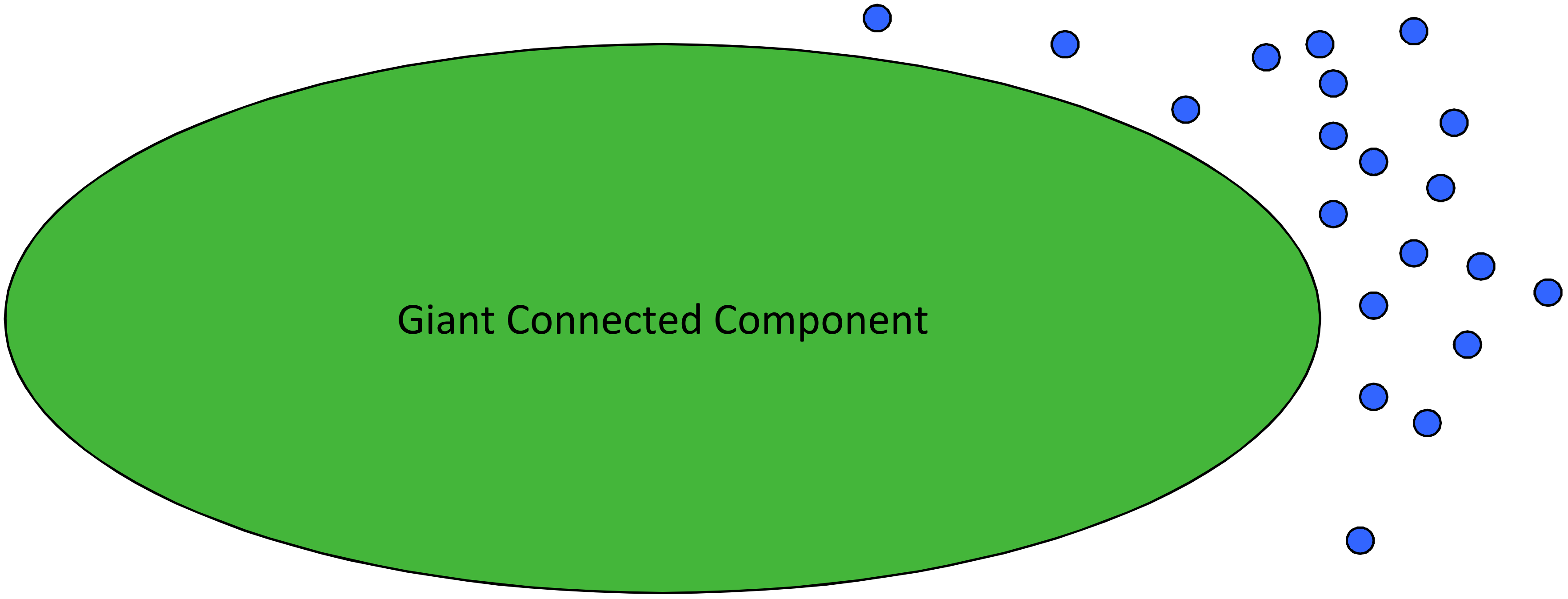
Result: Low-communication s-t Connectivity

- s-t connectivity for social graphs: $O(\log^2 n)$ bits for n-node graphs
- $\Omega(n \log n)$ lower bound for general graphs (Hajnal, Maass, Turàn)
  - Edges partitioned, 2 parties



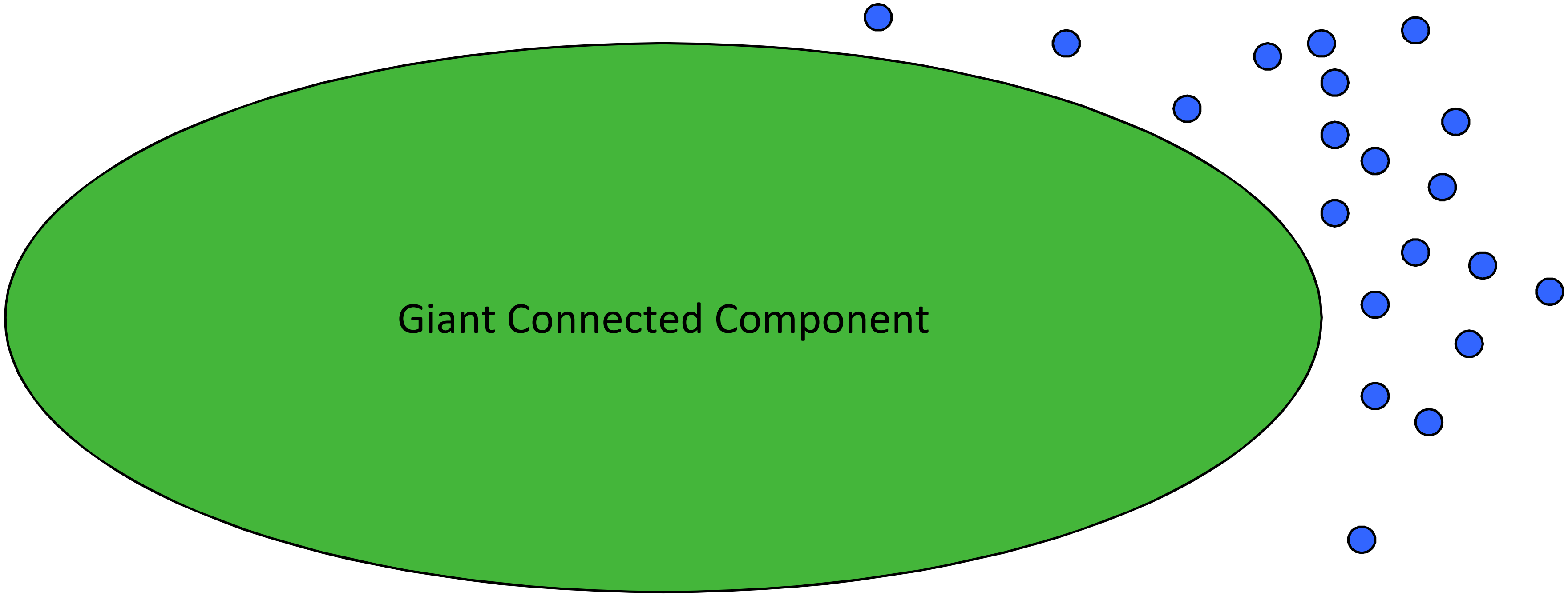
Social Network Structure

- Social networks have a giant component: second smallest component of size O(log n)



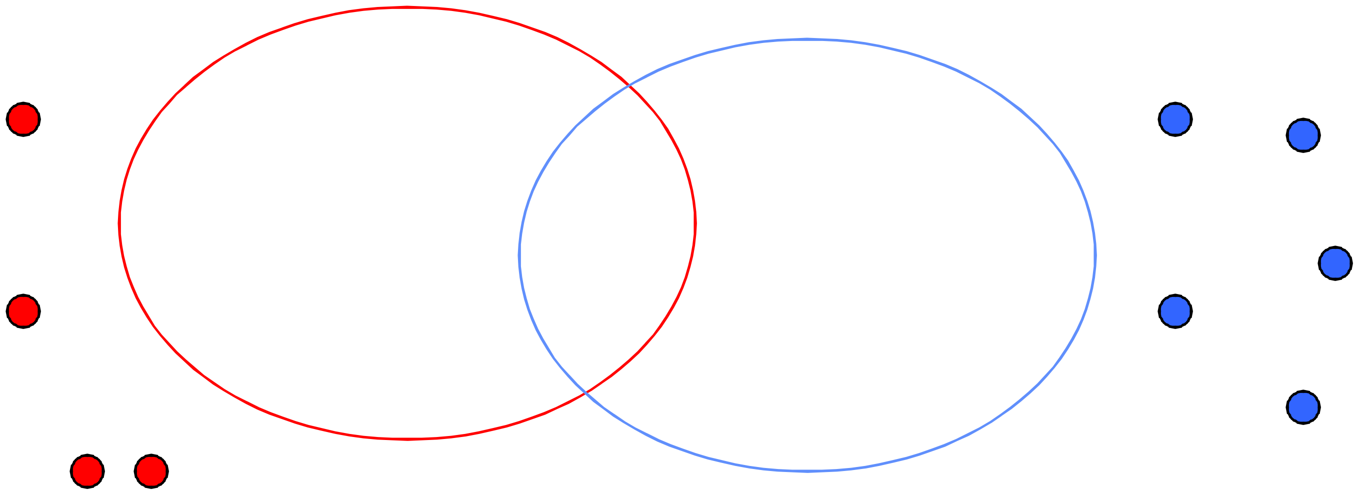
Social Network Structure

- Normal connection growth (Easley and Kleinberg)
- Observed in social networks (long distance phone call, linkedin, etc)
- Theoretically in Chung-Lu graphs with power law exponent between $1+\varepsilon$ and 3.47



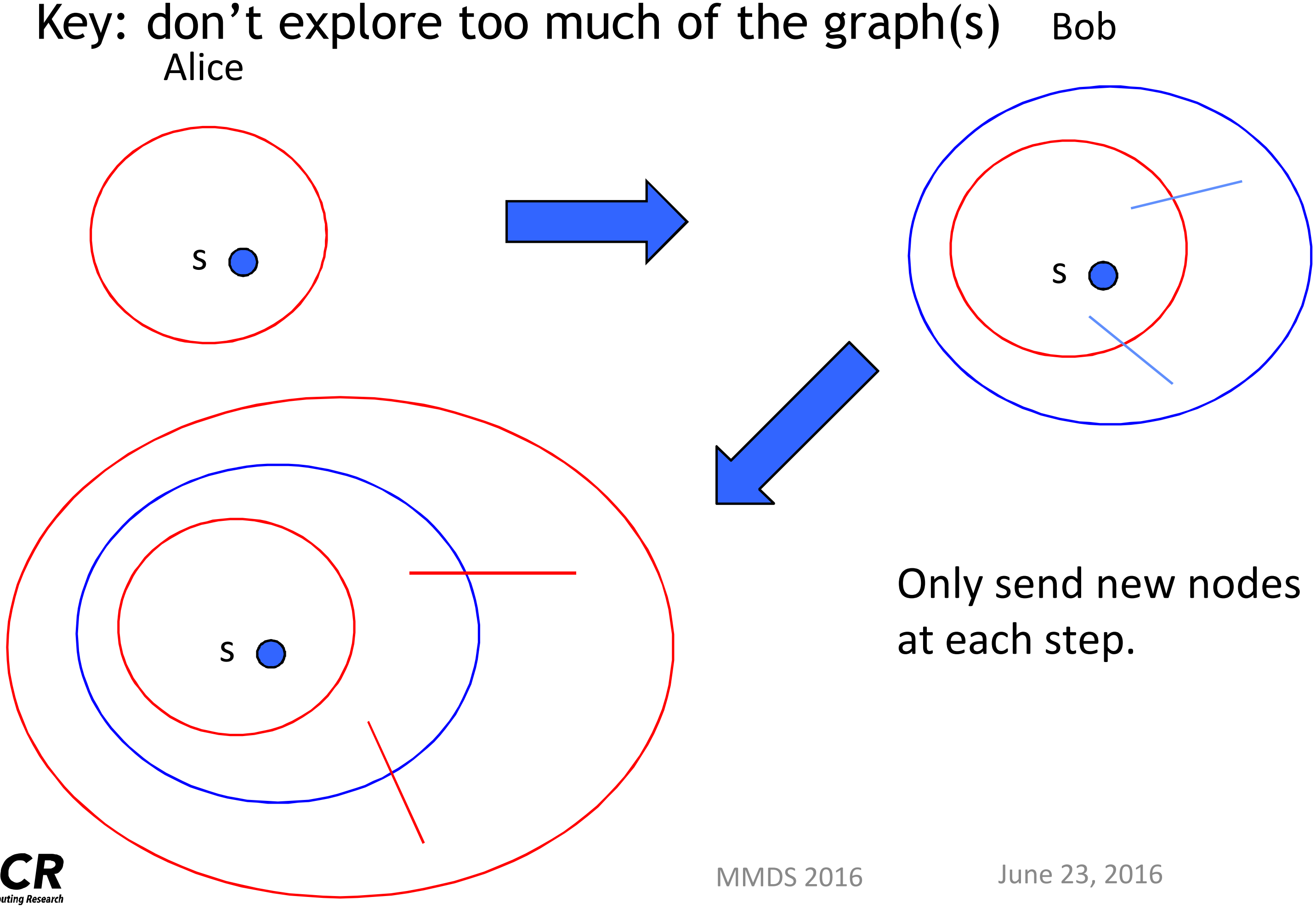
Assumptions

- Alice's graph G_A and Bob's graph G_B both have giant components
- These giant components intersect
  - Can verify with $O(\log^2 n)$ communication with high probability if intersect by a constant fraction (say 1%)



Shell expansion

- Like breadth-first-search, “layer” is connected piece in G_A or G_B
- Key: don’t explore too much of the graph(s)



Only send new nodes at each step.

Low-Sharing s-t Connectivity Algorithm

- Alice and Bob agree on a value γ (polylog in n)
  - Algorithm is correct iff γ at least size of 2nd largest component
- Do shell expansion (BFS) from both s and t
- Stopping criteria:
  1. s shell merges with t shell (yes)
  2. No new nodes added in some step (no)
  3. Shell merges with giant component of G_A or G_B (yes)
  4. Shell size exceeds γ . Stop before sending. (yes)

- With a good guess, $\gamma = O(\log n)$, so $O(\log^2 n)$ bits communicated

Also: Secure multi-party communication version of S-T connectivity (IEEE/IPDPS 2015)

S-T connectivity (yes/no) without revealing node names

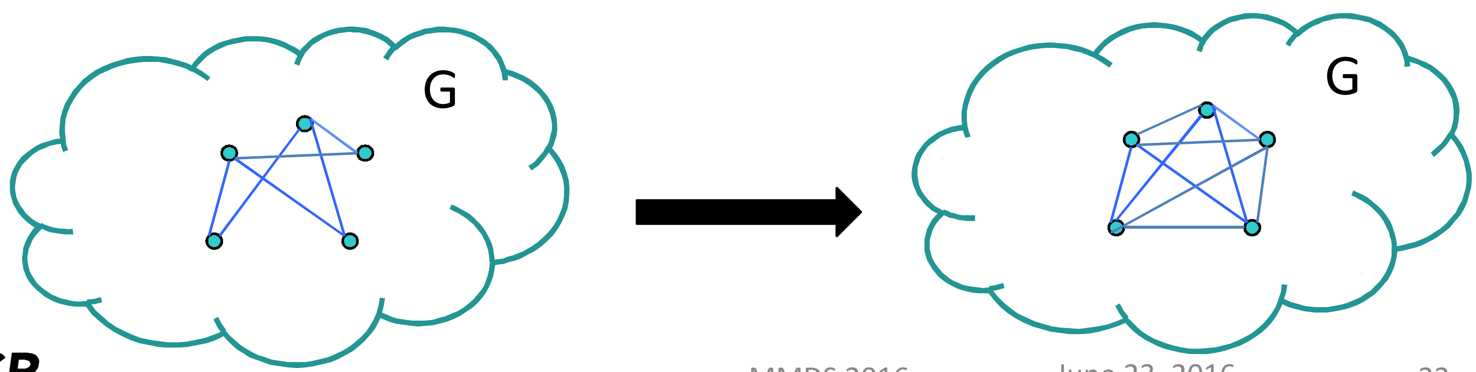
Topics

- s-t connectivity
- Planted clique
- Engineering better test sets

J. Berry, M. Collins, Aaron Kearns, C. Phillips, J. Saia, R. Smith, “Cooperative computing for autonomous data centers,” *Proceedings of the IEEE International Parallel and Distributed Processing Sympoisum*, May 2015.

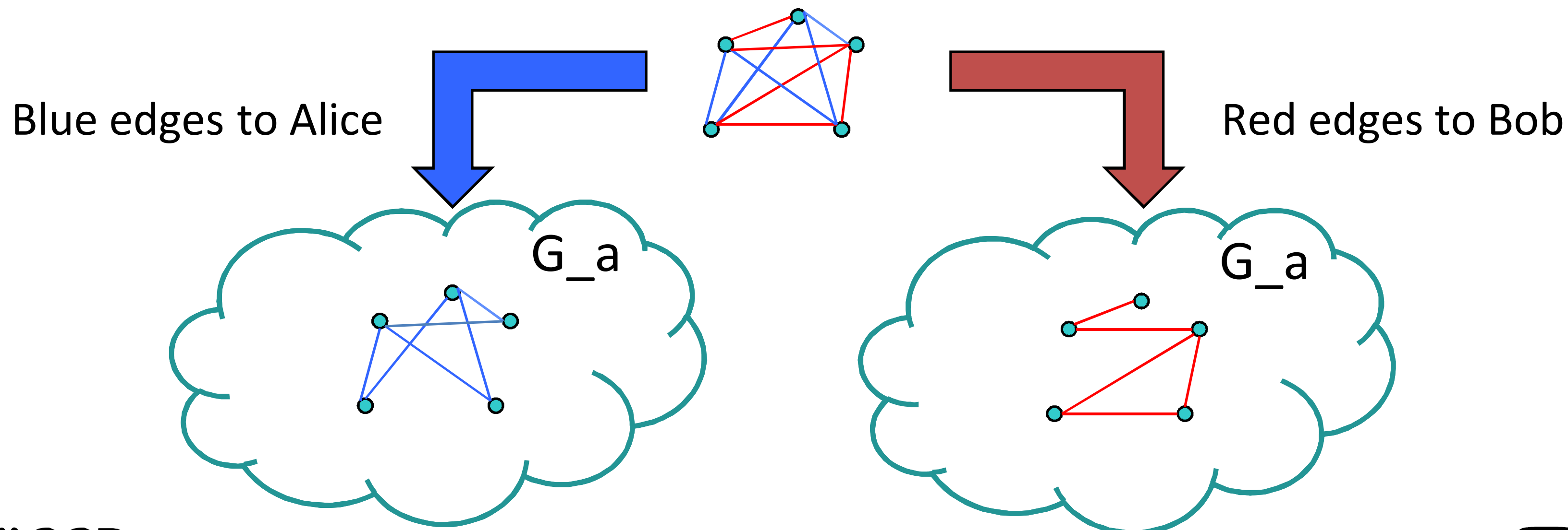
The Planted Clique Problem

- **Find a clique that has been artificially added to a graph**
  - Given graph, choose nodes randomly and build a clique
- **Can we find a clique that's a little larger than "native" clique size?**
- **For Erdos-Renyi, native is log n, can find** $\sqrt{n/e}$
  - (Deshpande and Montanari 2013, Alon, Krivelevich, Sudakov, 1998)
- **A form of anomaly detection, with other theoretical applications**



The Distributed Planted Clique Problem

- **When can social network structure help in solving a problem?**
- **Find a clique that has been artificially added to a graph**
  - *O(log n)* nodes chosen randomly and builds a clique
  - Adversary assigns clique edges to Alice or Bob
- **Can we find a clique that's a little larger than "native" clique size?**



Exploiting Social Network Structure

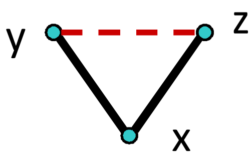
- Two key assumptions (*n*-node graph)
  1. Maximum degree is $O(n^{1-\epsilon})$
  2. Clustering coefficient for degree-*d* nodes is $O\left(\frac{1}{d^2}\right)$

These two assumptions lead to a polynomial-time, polylog-communication algorithm for finding an O(log n)-size planted clique.

Clustering Coefficient Assumption: Social Science Justification (slide 1)

Assumption: Clustering coefficient for degree-*d* nodes is $O\left(\frac{1}{d^2}\right)$

- **Strong triadic closure (Easley, Kleinberg):** two strong edges in a wedge implies (at least weak) closure.
  - Reasons: opportunity, trust, social stress



- **Converse of strong triadic closure:** not (both edges strong) implies not (more than coincidental closures)
  - experimental evidence: Kossinets, Watts 2006

Clustering Coefficient Assumption: Social Science Justification (slide 2)

Bounded number of strong human interactions even with social media (Dunbar 2012)

- so bounded number of strong wedges.
- As degree increases, more wedges involve weak pairs
- Social reasons for triadic closure all reduced as strength decreases

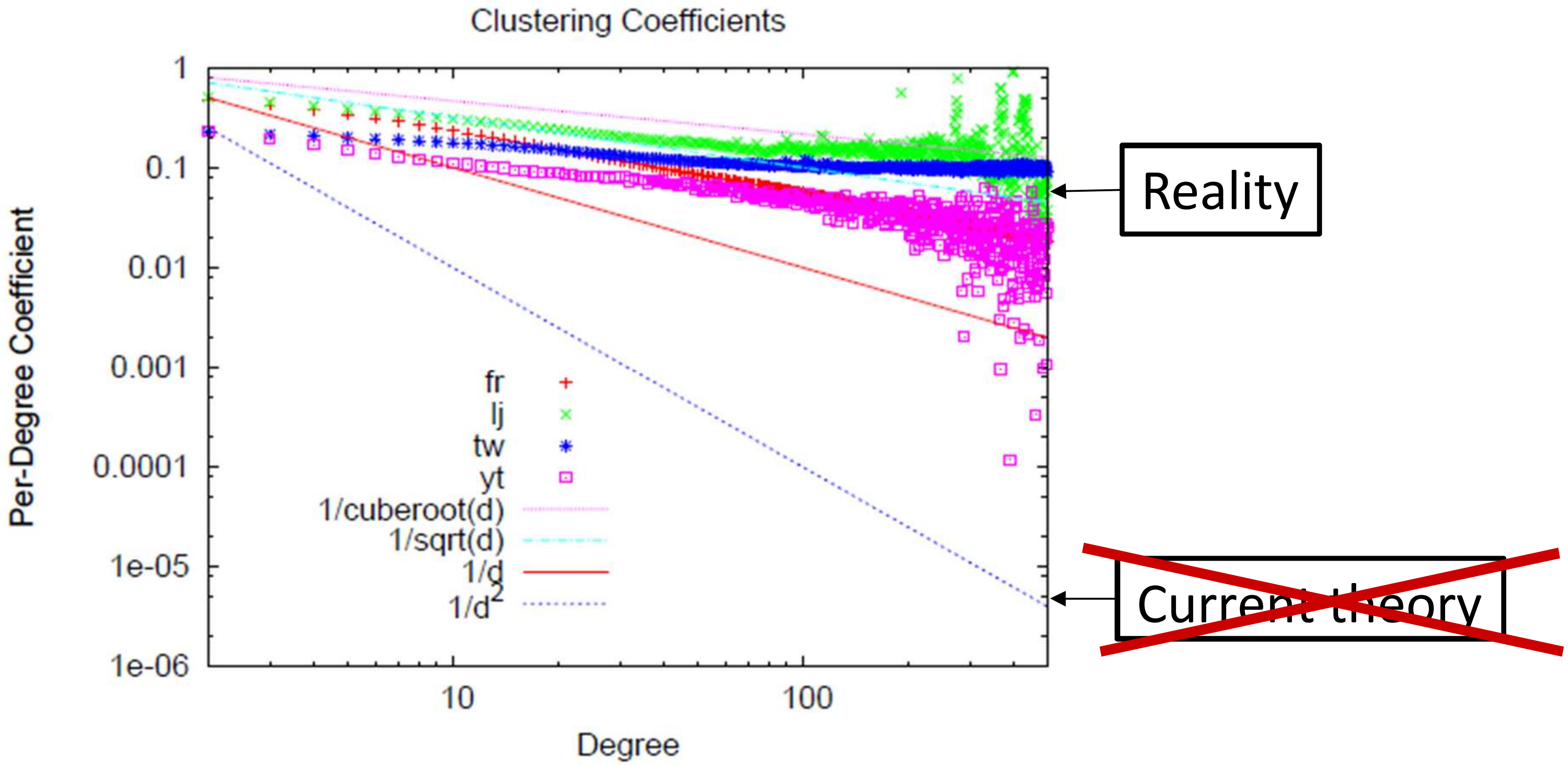
- Assumption is implied on average whp by Kolda et al. (SISC), where ξ fit from global CC: $c_{\mathrm{avg}}(d) = c_{\max} \exp(-(d-1) \cdot \xi)$

But the assumption actually isn't justified at all!

Problems

Experimental validation on some public social networks failed!

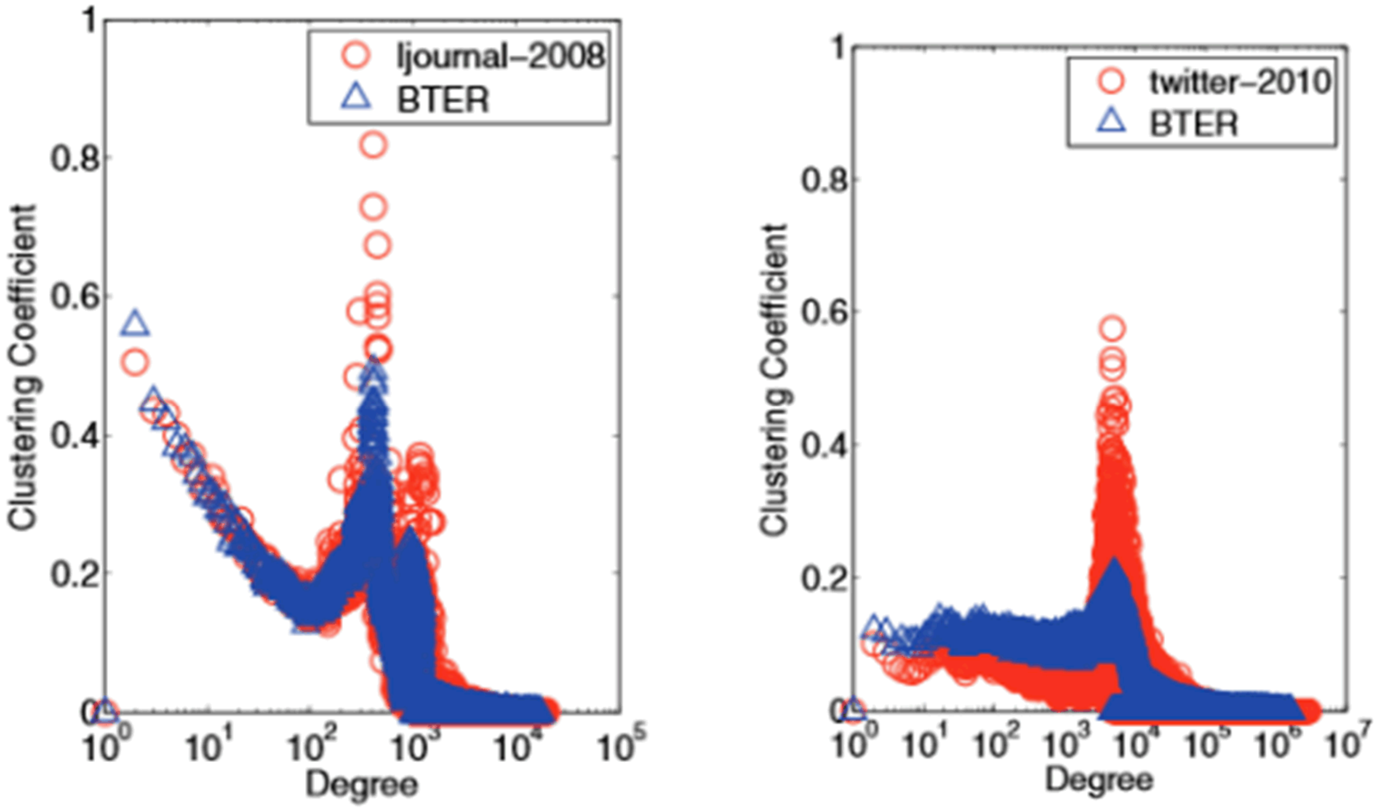
Why? Because the clustering coefficient assumption doesn't hold.



Topics

- s-t connectivity
- Planted clique
- Engineering better test sets

Clustering Coefficient "Rhino Horn"



Images from Kolda, et al. SIAM J. Computing 2014

Human vs Automated

- **Networks like Twitter contain a vast amount of non-human behavior**
  - You can buy 500 followers for $5 US
  - Economic incentives to manipulate connections
- **For applications, we assume that the network owners (e.g. law-enforcement agencies) will have human-only networks**
  - Their networks are not public where entities can sign up
  - No cleaning problem
  - Will our distributed algorithms work?
- **Our work uses data from SNAP, LAW**
  - What cleaning of these networks can we justify?

Human vs Automated

Goal: Clean (enough) non-human behavior to test our algorithms

- **Limitation: we have only topology**
- **Dunbar: Real human relationships require attention**
  - Attention can be divided
  - Total attention, time of day, etc, is limited
- **Communities with too many "strong" connections may not be human.**
  - E.g.: in Twitter-2010, there is a 317-clique of mutual follower relations (with no apparent common ground among nodes)

Some Test Network Desired Properties

- **Automated sub-networks are not present**
- **Edges plausibly represent a social bond**
  - Even better if the relationship requires time/effort
- **Large size (millions/billions of nodes/edges)**
- **Approximates a full network snapshot**
  - *Not ego-networks*

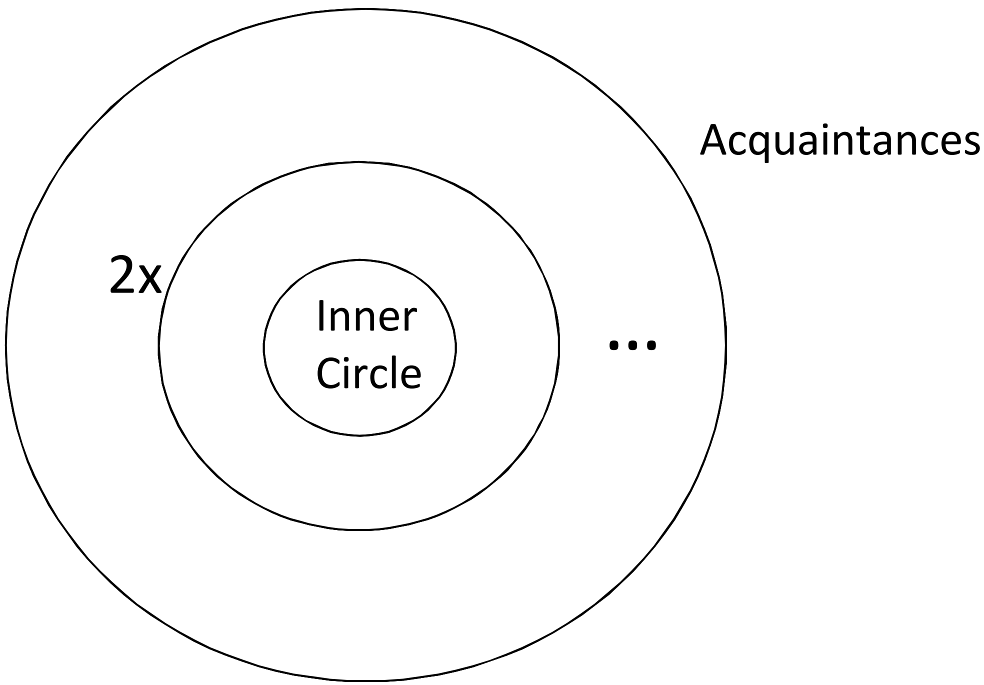
We don't know publicly available social networks with all these

- Closest: friendster

Given exemplars, could generate more instances with a network generator like BTER.

Varying Strength of Ties

- People “know” about 1500 others by face/name
- Hierarchy of strength

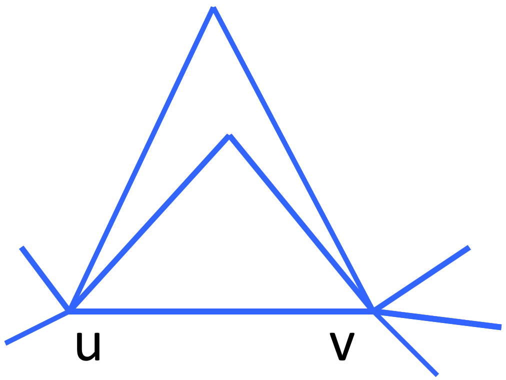


R. Dunbar, Social cognition on the internet: testing constraints on social network Size, Philosophical Transactions of the Royal Society B, Biological Sciences,367(1599):2192-2201, 2012

Edge strength

- A notion somewhat like Easley and Kleinberg 2010, and Berry et al., 2011

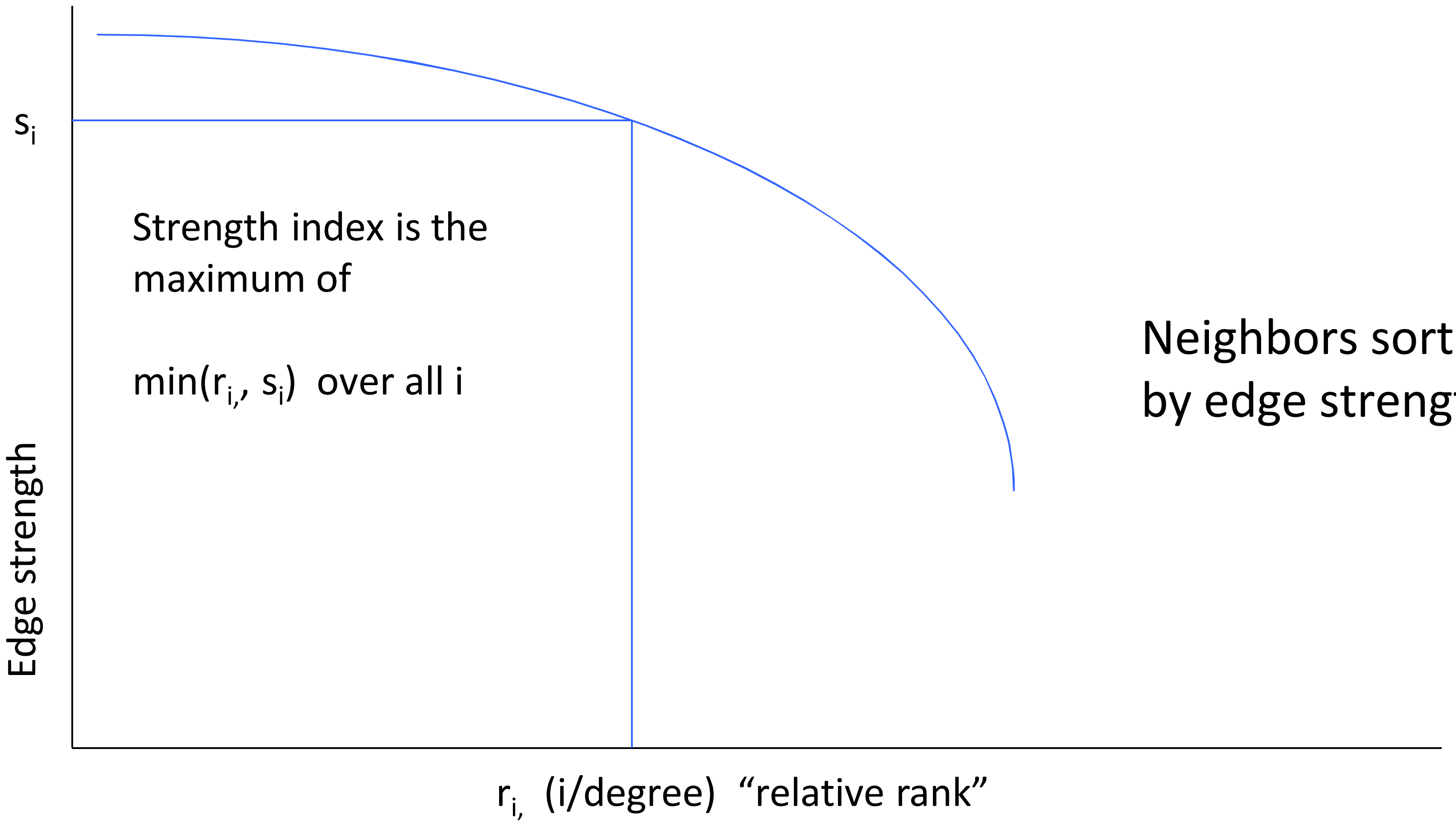
$$s(u,v) = \frac{2 * \#\ \text{triangles on}(u,v)}{d_u + d_v - 2}$$



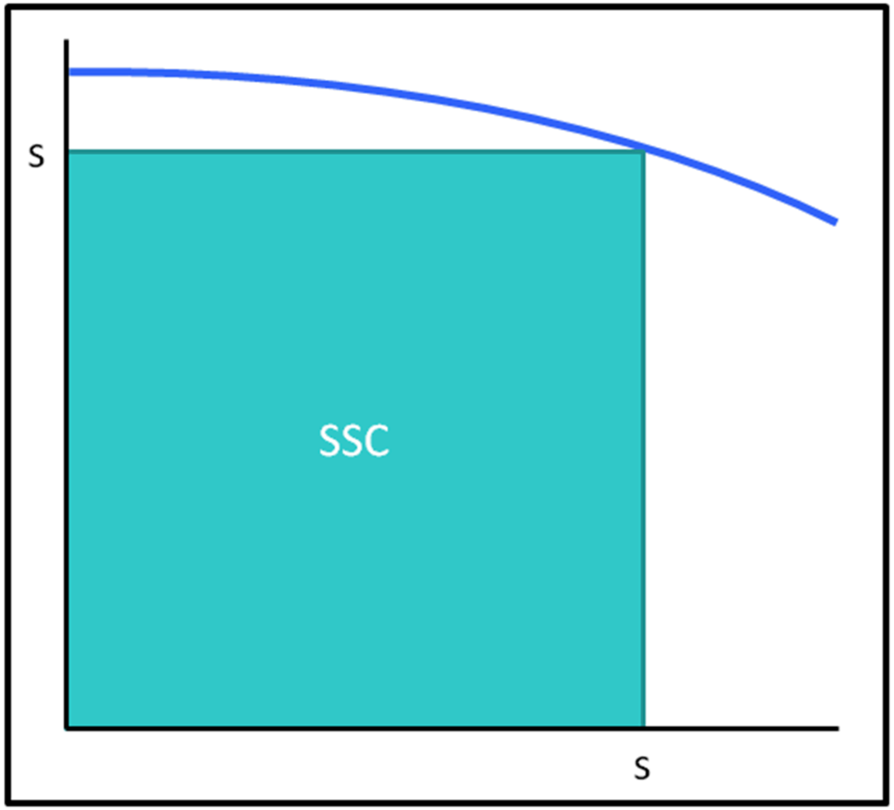
$$s(u,v) = \frac{2*2}{5+6-2} = \frac{4}{9}$$

- Idea: Total strength has a constant bound
  - Edge strength a continuum, not just strong/weak

"strength-index" for Nodes (like H-index)



Strength-Index Property



SSC: "Symmetric Strength Component"

Suppose strength-index = s;

Dunbar-like constant = D,
S = Prefix sum of strengths<=s

Then: $D \geq S \geq s^2 * \text{ degree}$

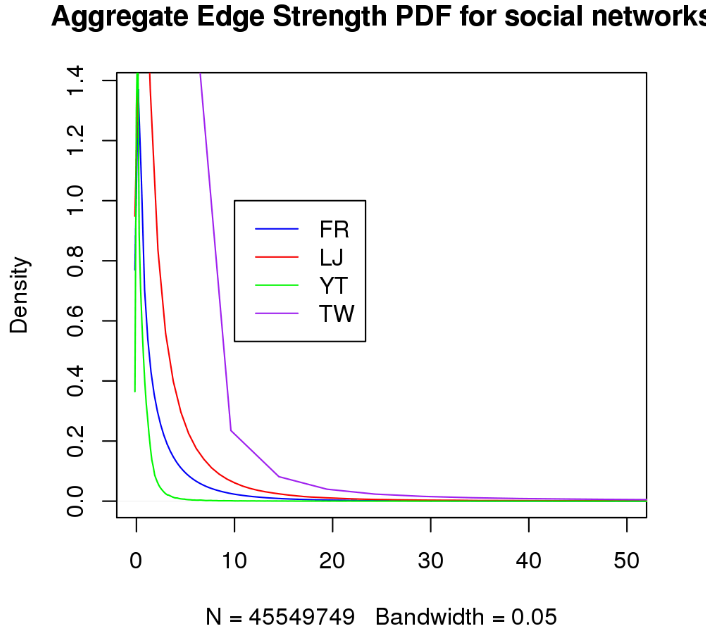
$$s \leq \sqrt{\frac{D}{d}}$$

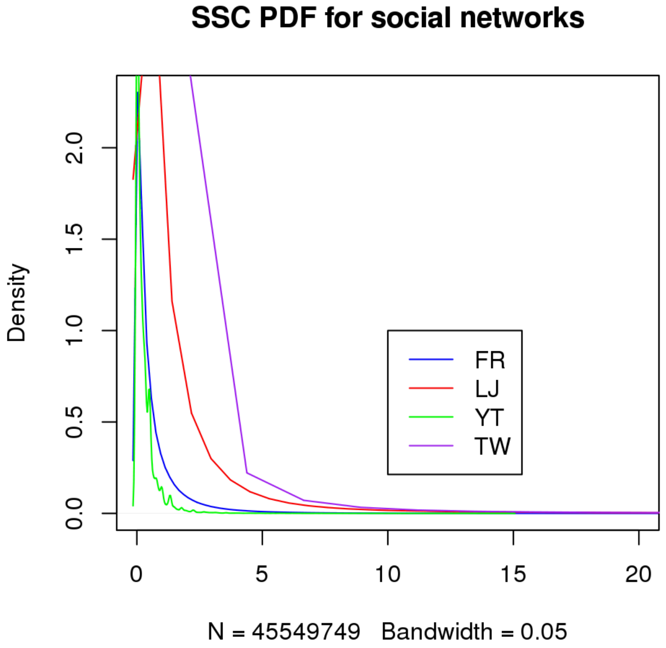
s = s-index
D = Dunbar-like constant
d = degree

Most important edges are free from tail effects

SSC and total strength distributions

SSC and total strength S are empirically bounded by small constants



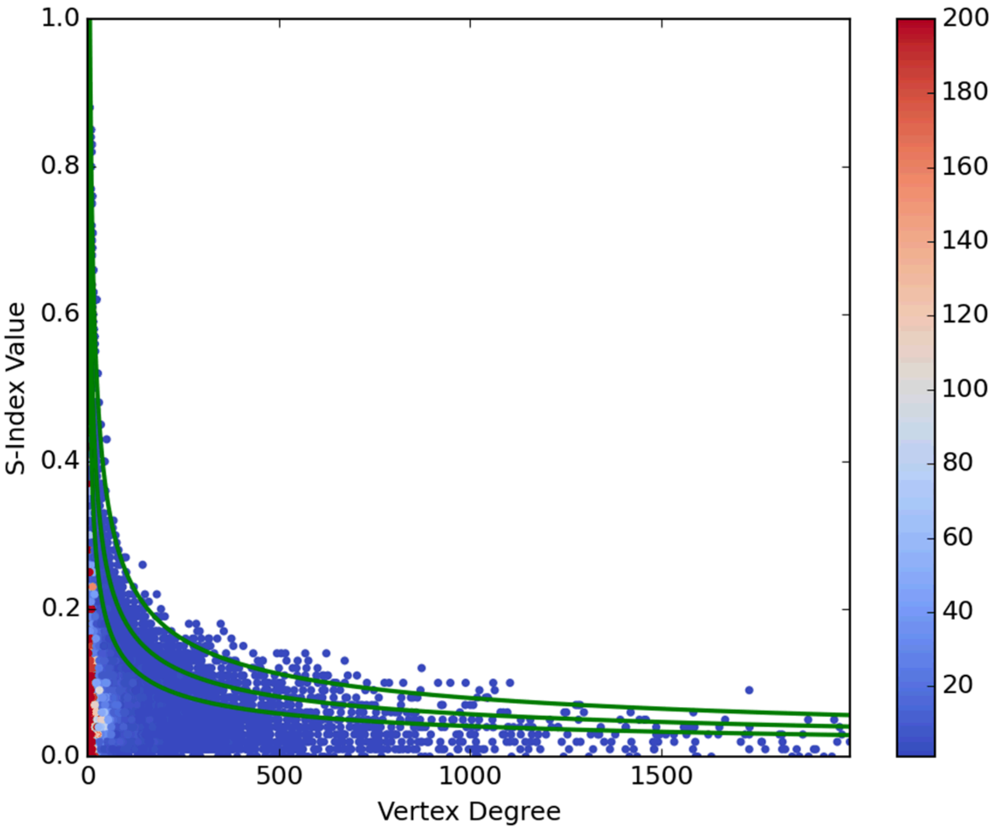


Cleaning Non-Human Nodes

- We assume $s \leq \sqrt{\frac{D}{d}}$ for entirely-human vertices
- Constant D will depend on the network
- Remove nodes with s above this curve (or edges connecting violators)
- Selecting D
  - Compute average SSC average μ and standard deviation σ
  - D = μ + kσ for user-defined parameter k
- Nodes above the line for a given k are kσ violators

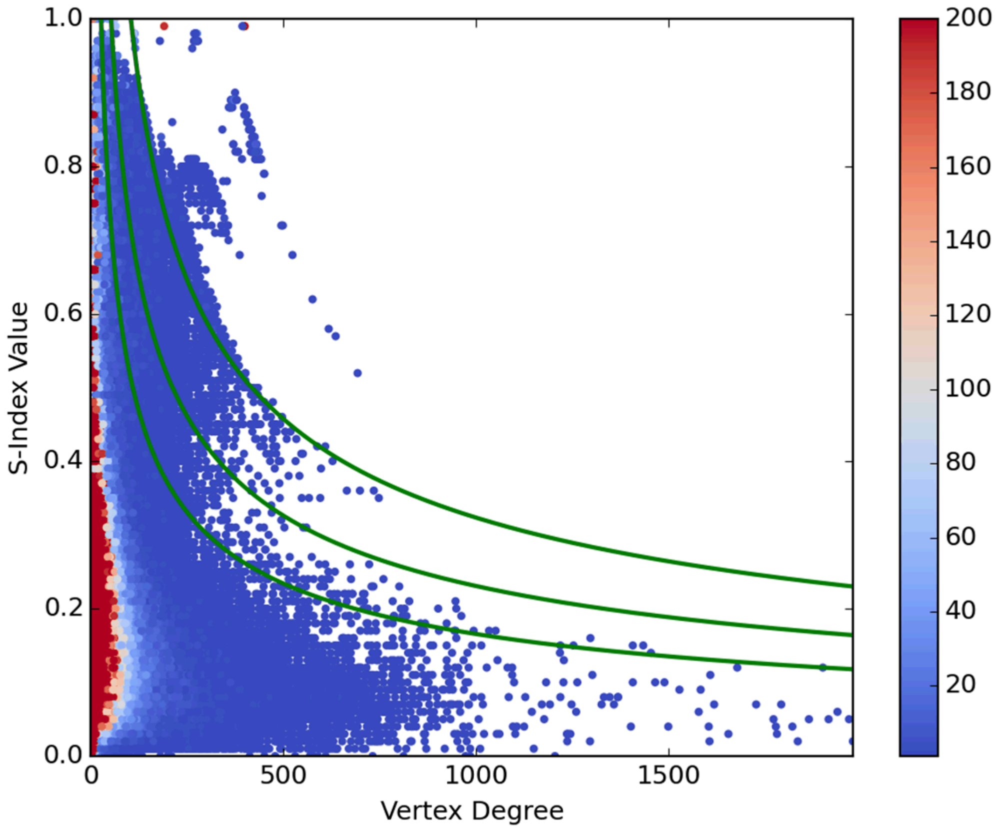
YouTube Heat Map

- Before cleaning. k=3,6,12



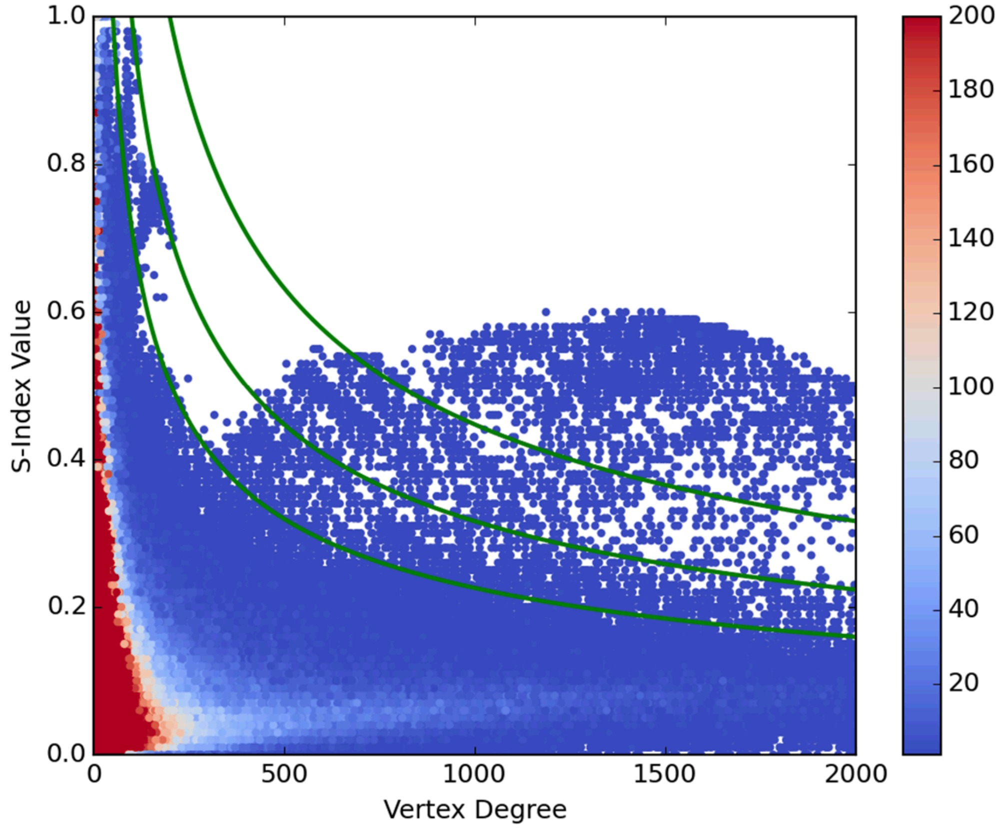
LiveJournal Heat Map

- Before cleaning. k=3,6,12



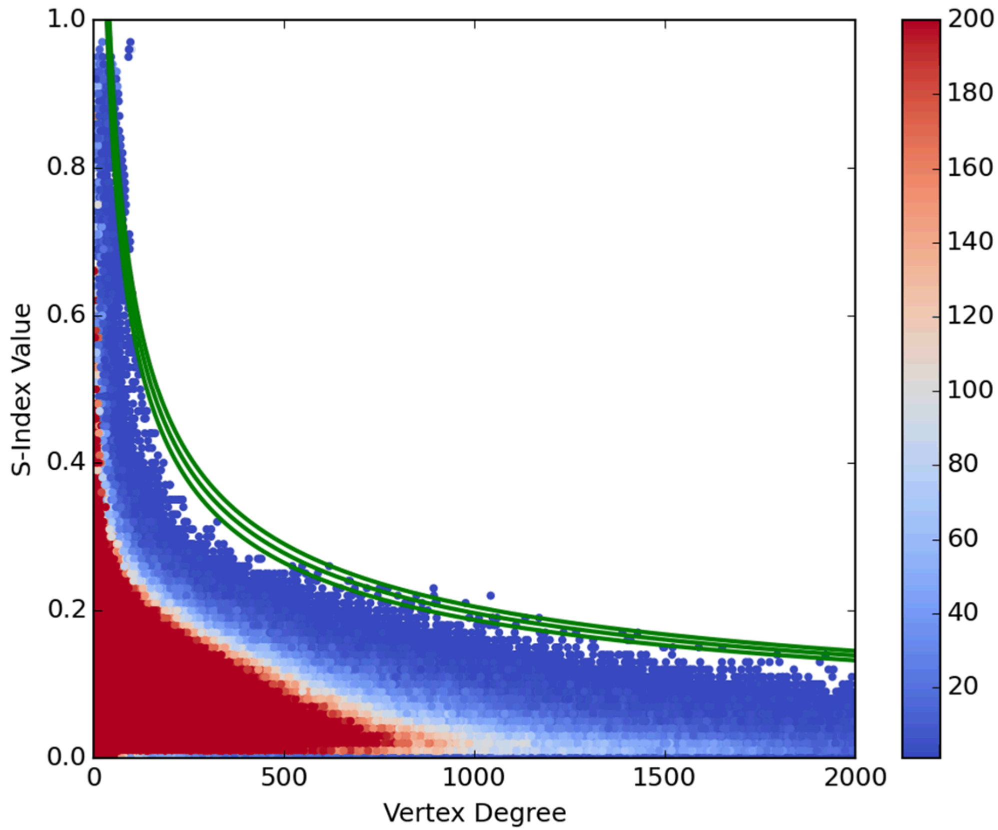
Twitter Heat Map

- Before cleaning. k=3,6,12



Friendster

- Before cleaning. k=3,6,12. Already clean!



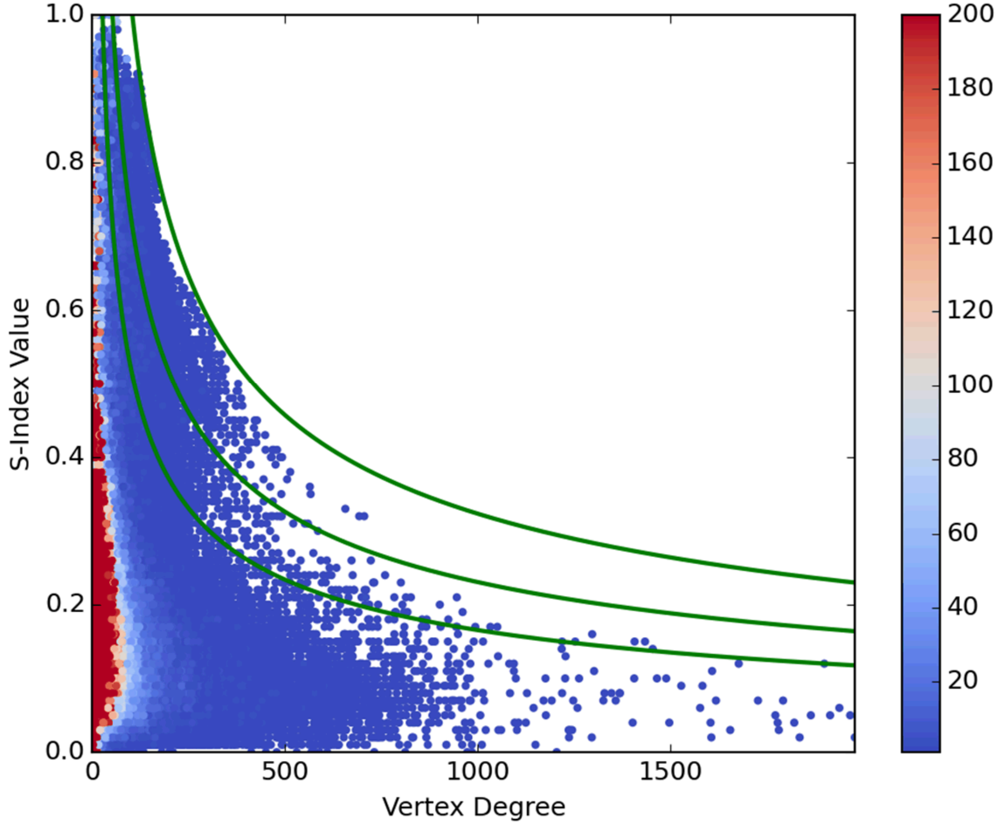
Cleaning

- Sometimes small number of vertices have a large fraction of edges

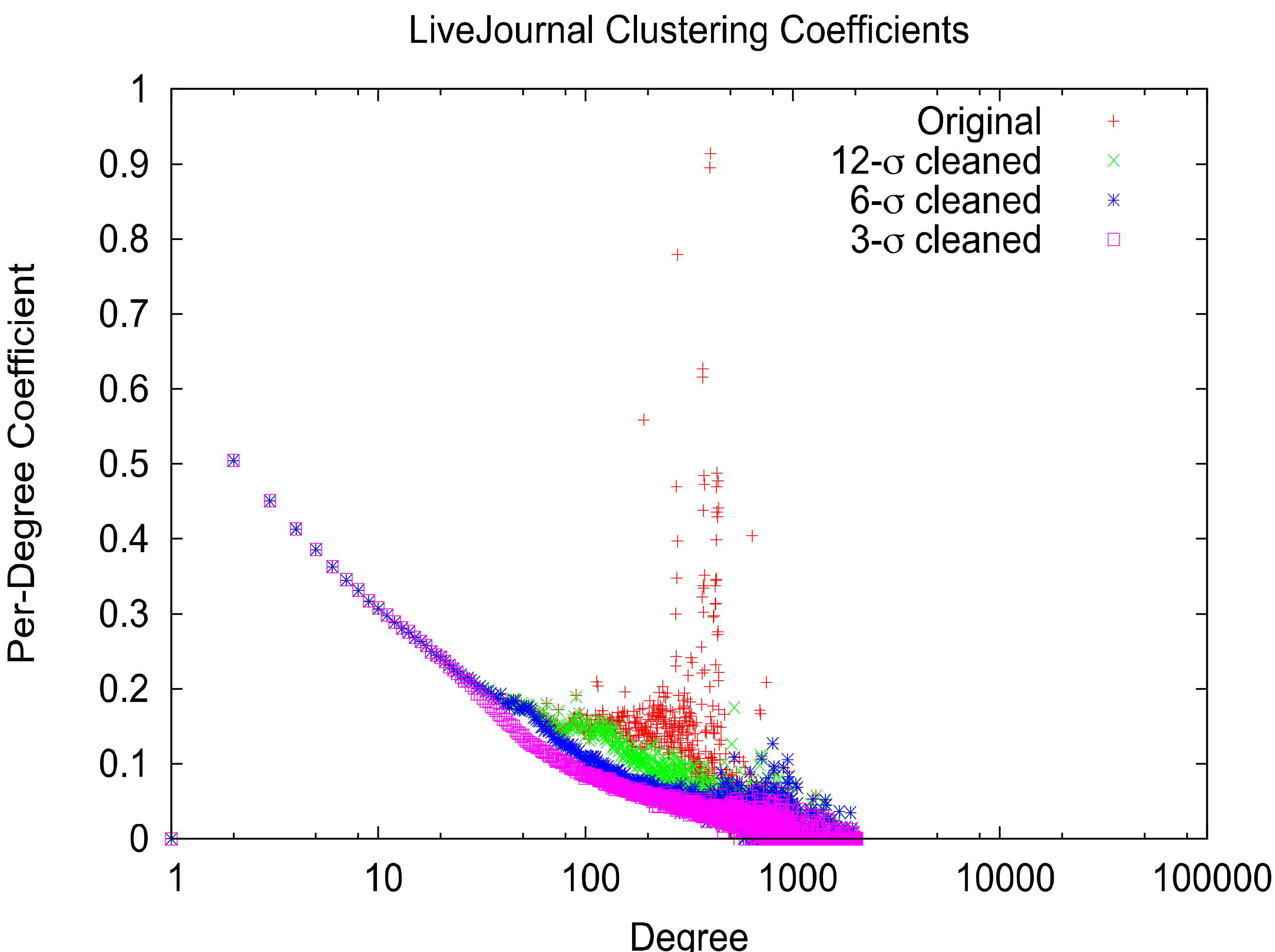
Network	percentage of vertices removed	percentage of edges removed
com-youtube($12\bar{\sigma}$)	0.01%	2.5%
com-youtube($6\bar{\sigma}$)	0.11%	10.76%
com-youtube($3\bar{\sigma}$)	1.18%	**32%**
ljournal-2008($12\bar{\sigma}$)	0.05%	1.57%
ljournal-2008($6\bar{\sigma}$)	0.14%	3.13%
ljournal-2008($3\bar{\sigma}$)	0.36%	5.38%
twitter-2010($12\bar{\sigma}$)	0.02%	**26.4%**
twitter-2010($6\bar{\sigma}$)	0.046%	**34.3%**
twitter-2010($3\bar{\sigma}$)	0.048%	**34.7%**

Cleaned LiveJournal

- k=12

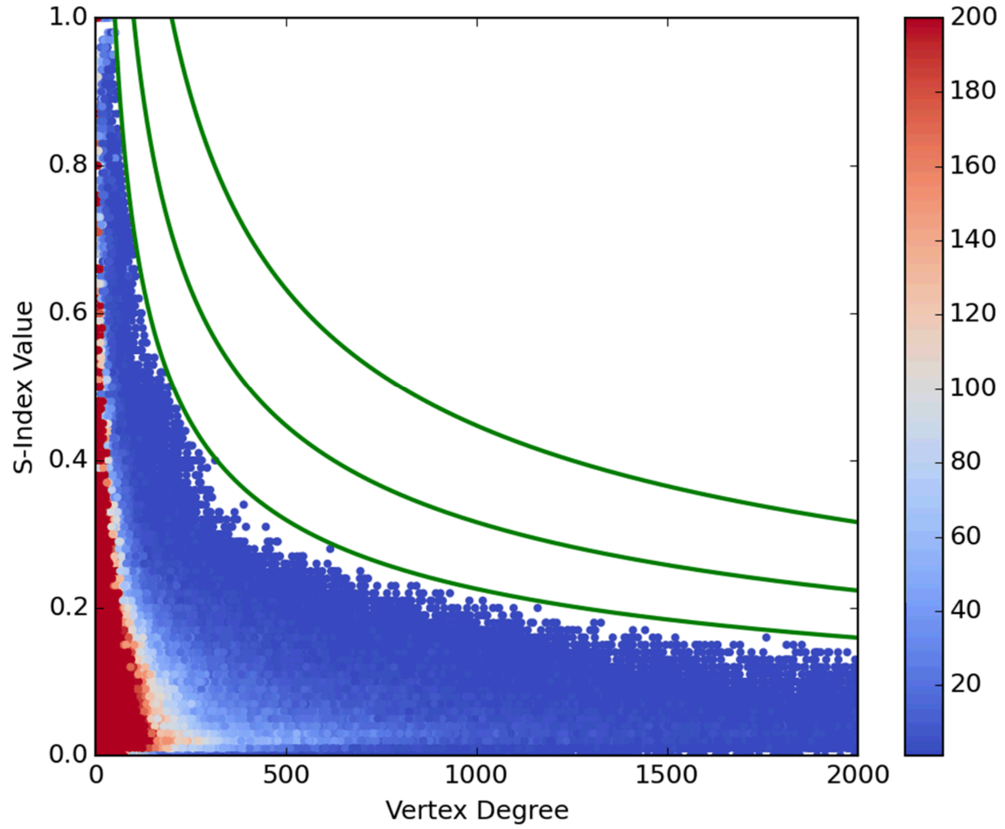


LiveJournal: Cleaned Clustering Coefficents



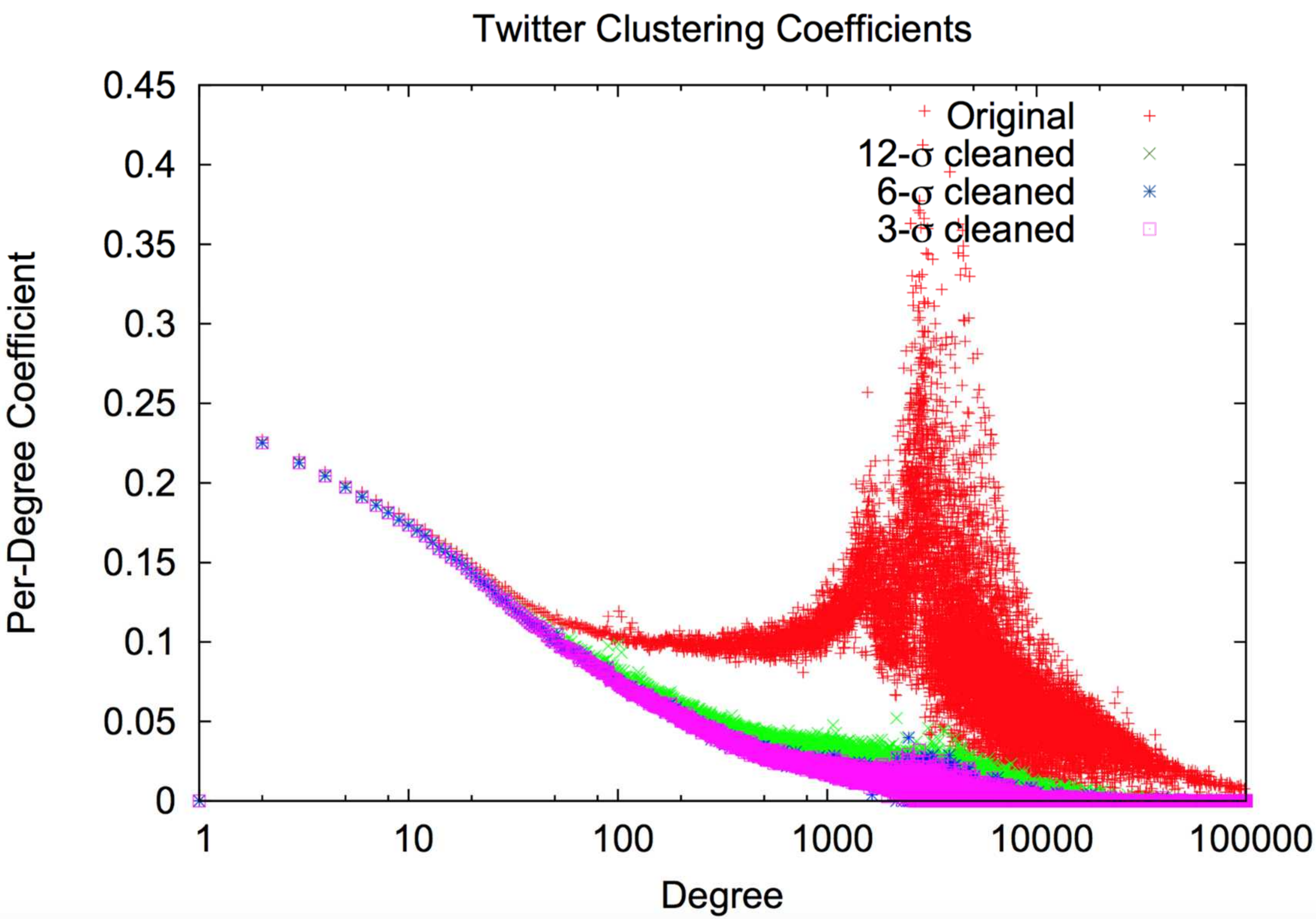
Cleaned Twitter

- k=3



Twitter

•



Twitter Clustering Coefficients
Per-Degree Coefficient
Degree
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1
10
100
1000
10000
100000
Original
12-σ cleaned
6-σ cleaned
3-σ cleaned

Validation Goal

Show empirically that we are not

"*throwing out the baby with the bath water*"

Working on it……

Computing and Social Networks

- Sandia joint work with Indiana U. described the main challenges for High-Performance Computing (HPC) and these graphs/networks

ANDREW LUMSDAINE et al, *Parallel Process. Lett.* **17**, 5 (2007). DOI: http://dx.doi.org/10.1142/S0129626407002843

CHALLENGES IN PARALLEL GRAPH PROCESSING

ANDREW LUMSDAINE

Indiana University, Bloomington, Indiana 47401, USA

DOUGLAS GREGOR

Indiana University, Bloomington, Indiana 47401, USA

BRUCE HENDRICKSON

Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

JONATHAN BERRY

Sandia National Laboratories, Albuquerque, New Mexico 87185, USA

Received: December 2006
Revised: January 2007

- Has influenced HPC, cloud, multicore graph computation

Summary

- Sandians have made contributions to social network analysis recently
- There's more related work on the horizon
- Main points of contact:
  - NM: Cindy Phillips, Jon Berry
  - CA: Tammy Kolda, Ali Pinar