SAND2015-11167C

Finding Non-Human Nodes in Social Networks

Jon Berry (Sandia National Laboratories)
Aaron Kearns (U. New Mexico)
Cynthia A. Phillips (Sandia National Laboratories)
Jared Saia (U. New Mexico)

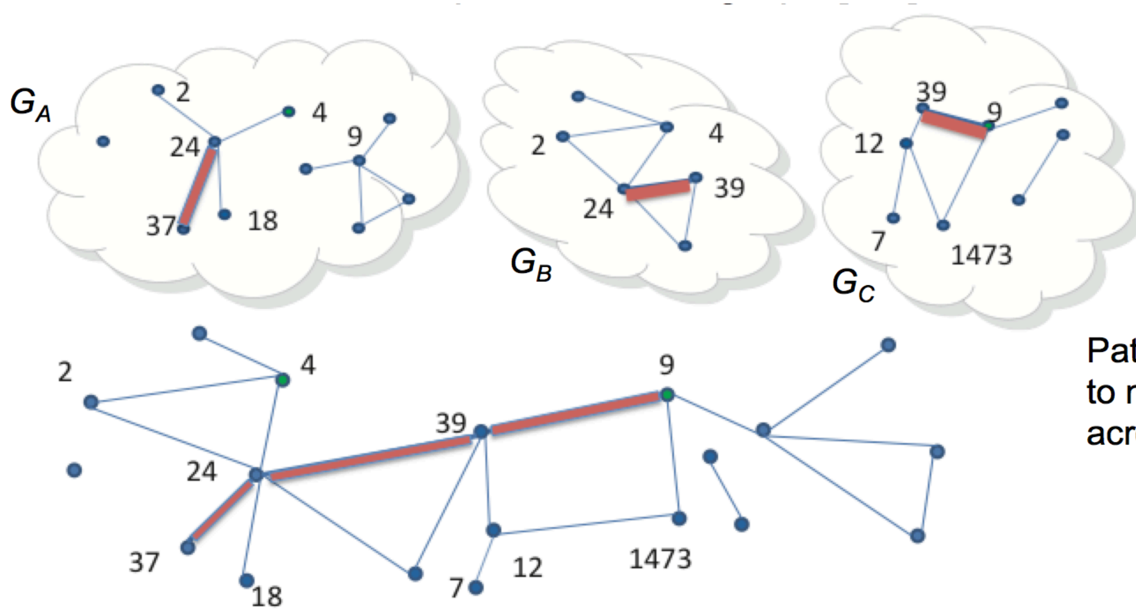
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

A 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

Sandia National Laboratories

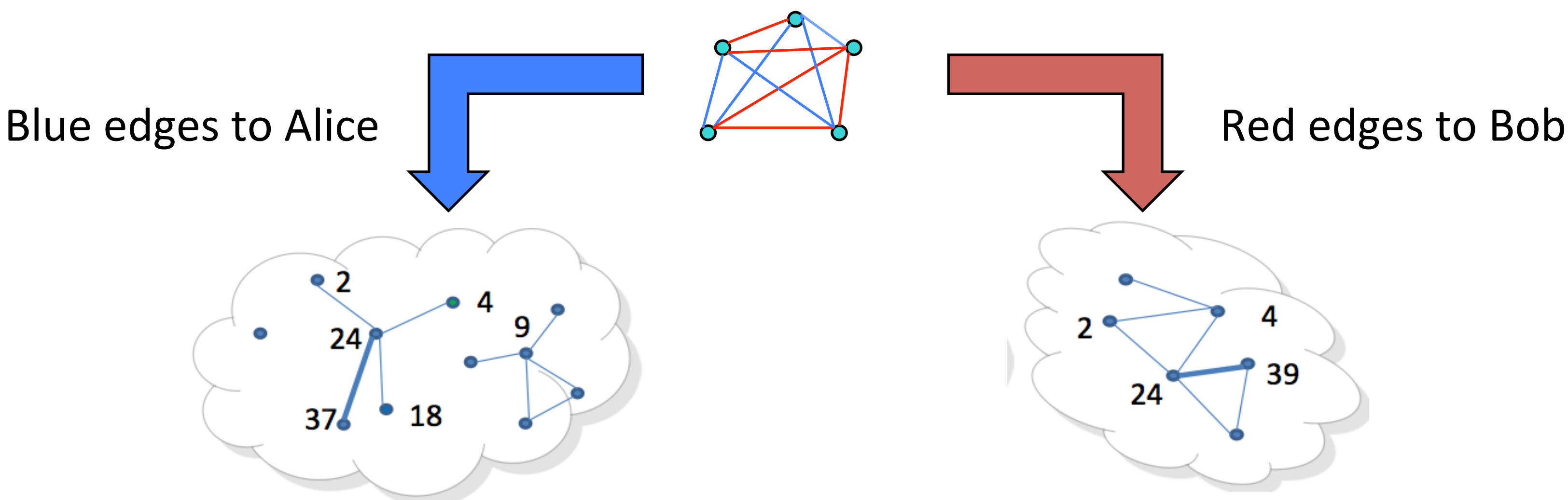
Previous Talk

- Algorithms for s-t connectivity in both models
  - Low communication, O($\log^2 n$) bits. Requires social network structure (giant component)
  - Low trust.
    - Alice gets no information beyond answer in honest-but-curious model.
    - Doesn't even reveal node names.
- Paper appeared in IPDPS 2015: "Cooperative computing for autonomous data centers"

Sandia National Laboratories

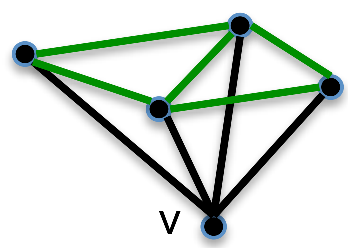
The Planted Clique Problem

- 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?
- For Erdos-Renyi, native is log n, can find $\sqrt{n/e}$ (Deshpande and Montanari, Alon, Krivelevich, Sudakov)



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 $O\left(\frac{1}{d^2}\right)$



Clustering coefficient (CC) of v =
Fraction of related neighbors.

$$\frac{\text{\# triangles on v}}{\text{\# wedges on v}} = \frac{5}{6}$$

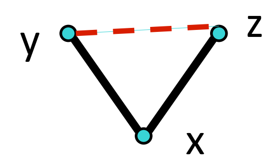
$c_{avg}(d)$ = Average CC over nodes of degree d.
Global CC = average over all nodes v

Please (for now) hold off on protests about what one sees in practice

Clustering Coefficient Justification

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 coincidental closures
  - experimental evidence: Kossinets, Watts 2006

Clustering Coefficient Justification

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
- Reasons for triadic closure all reduced as strength decreases

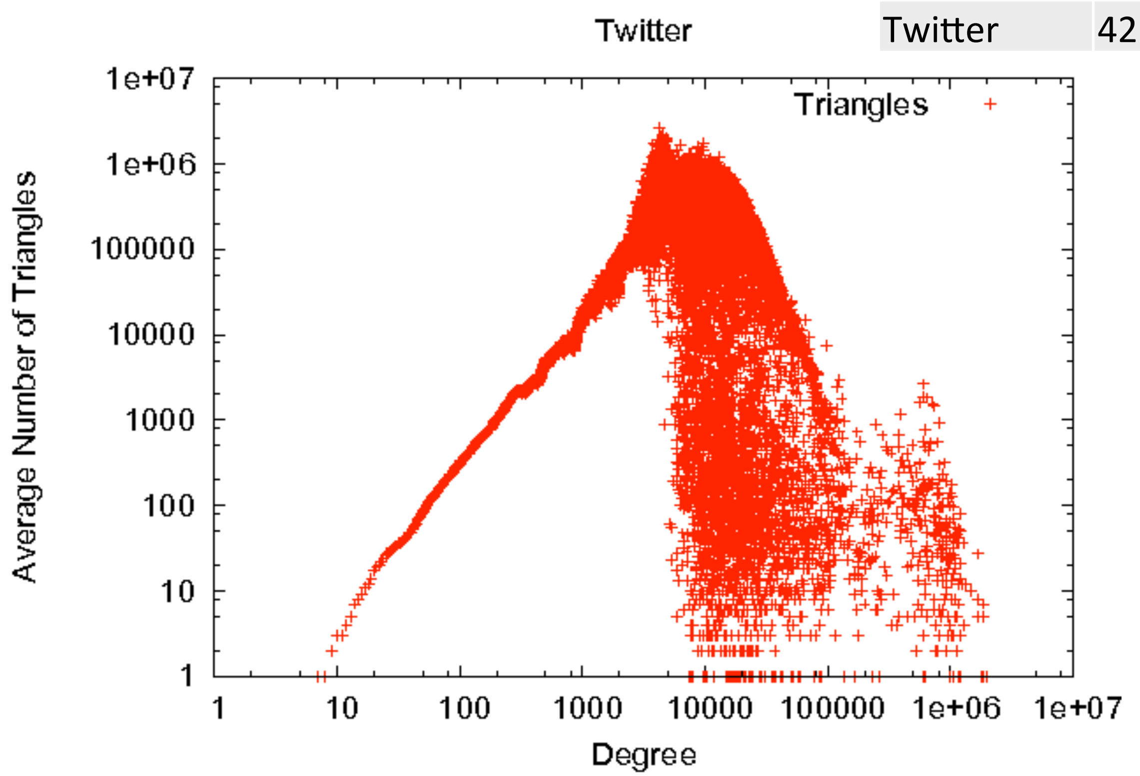
- Assumption implied on average whp by Kolda et al (SISC), where ξ fit from global CC: $c_{\text{avg}}(d) = c_{\max} \exp(-(d-1) \cdot \xi)$

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

Real Social Networks

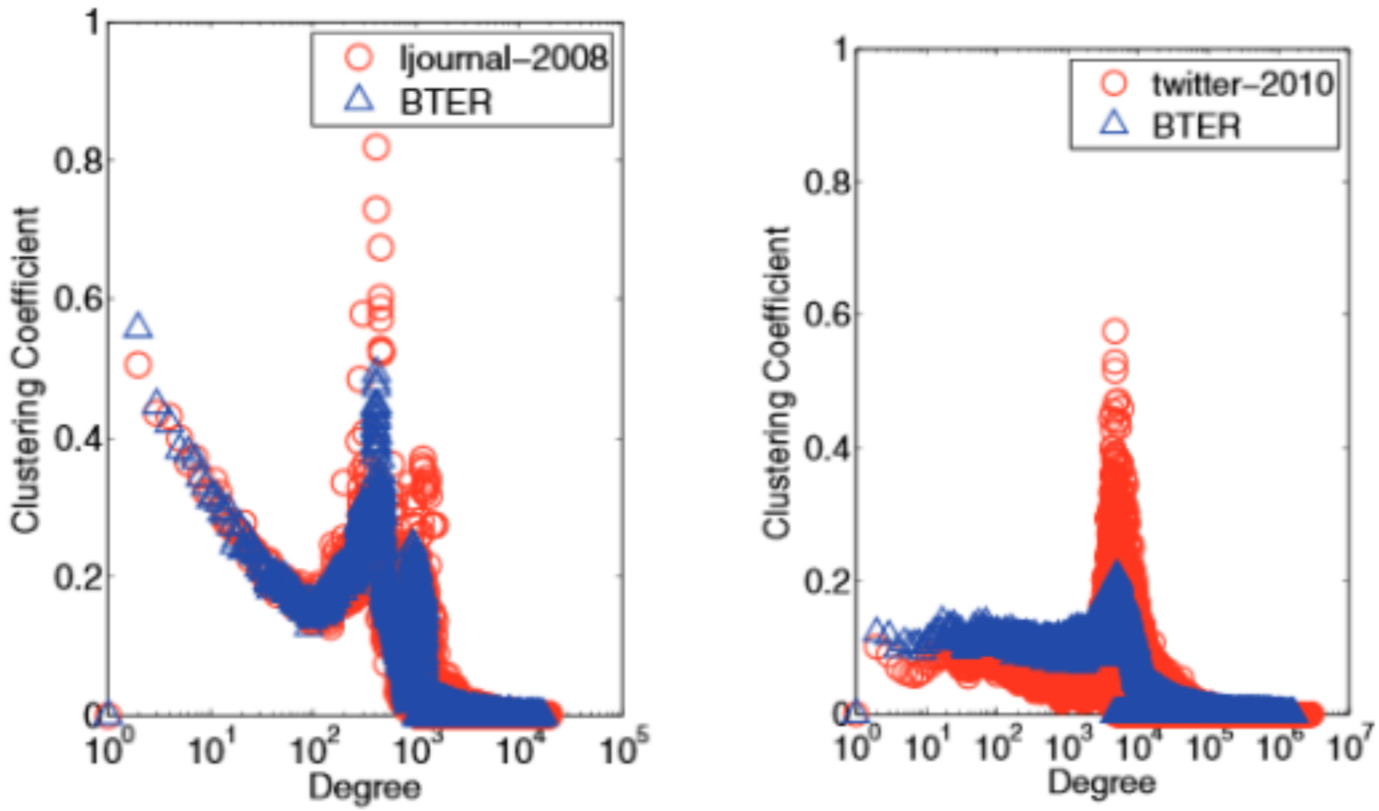
- Problem: Social networks don't obey our clustering coefficient assumption

Name	# nodes	#edges
Youtube	1M	3M
Orkut	3M	117M
LiveJournal	4M	35M
Twitter	42M	1.5B



Sandia National Laboratories

Clustering Coefficient "Rhino Horn"



Sandia National Laboratories

Human vs Automated

- Networks like Twitter contain a vast amount of non-human behavior
  - You can buy 500 followers for $5 US
- For our intended applications, the network owners (law-enforcement agencies) will have human-only networks
  - Networks are not public where entities can sign up
  - No cleaning problem
- We have no real data from law enforcement

Human vs Automated

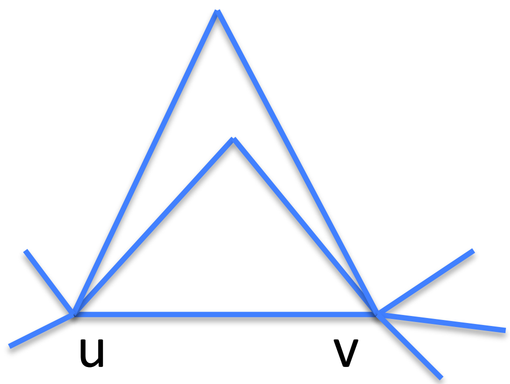
Goal: Clean (enough) non-human behavior to test our algorithms

- An idea: Real human relationships require attention
  - Attention can be divided
  - Total attention, time of day, etc, is limited

Edge strength

- A notion somewhat like the one we used for wCNM

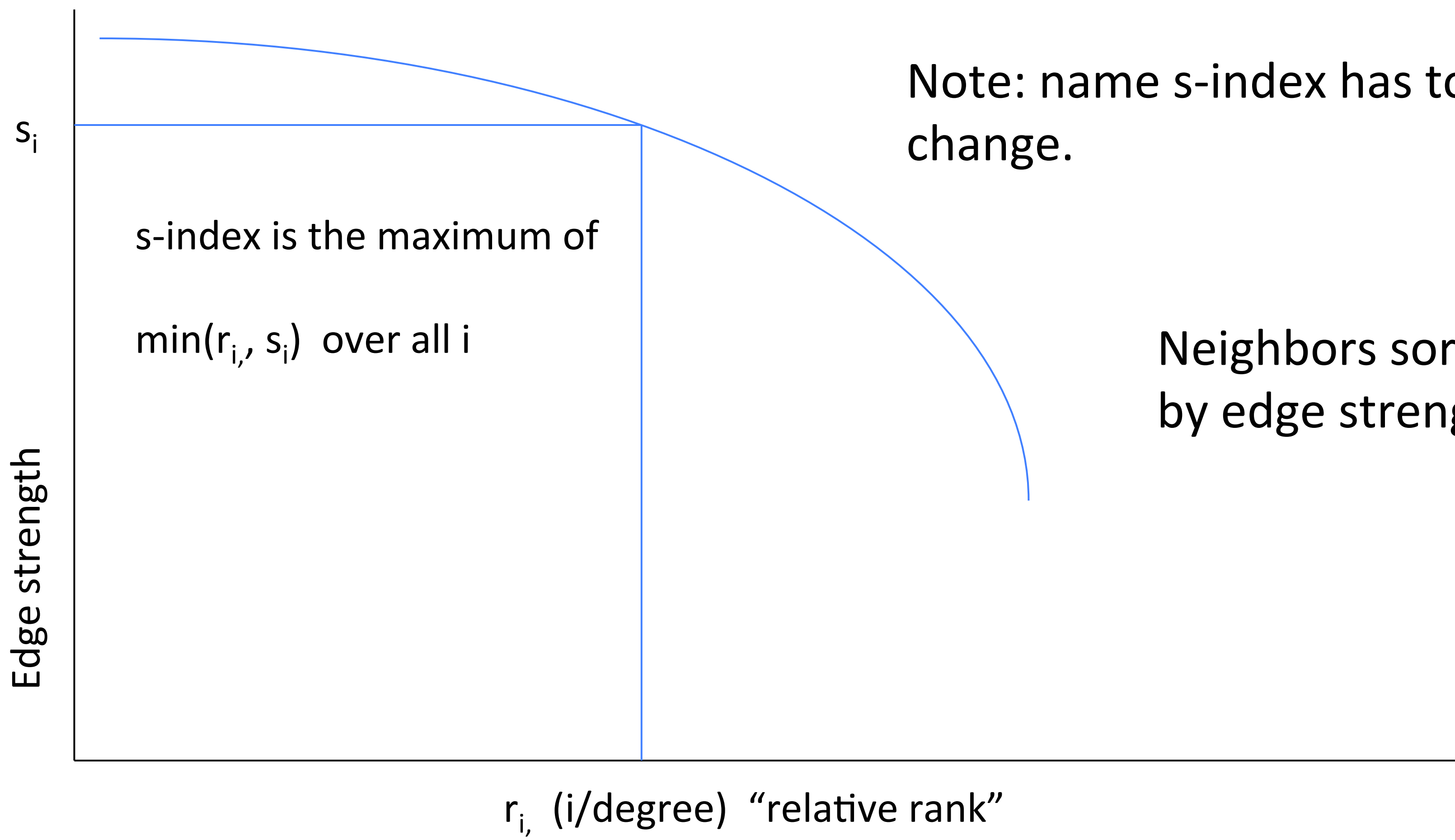
$$s(u,v) = \frac{2 * \#\ \text{triangles on}(u,v)}{d_u + d_v - 2}$$



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

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

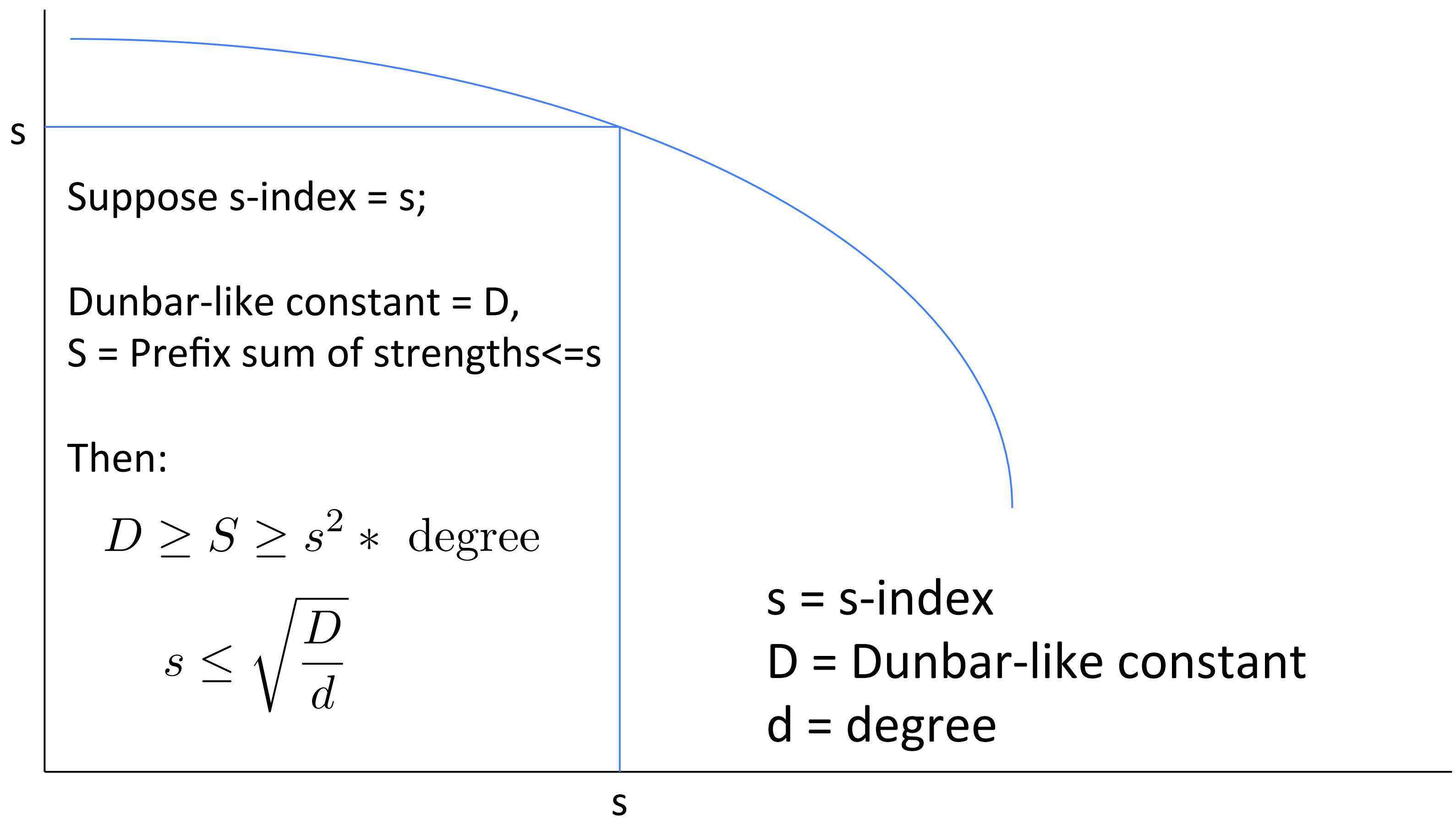
s-index



Note: name s-index has to change.

Neighbors sorted by edge strength

s-index



Suppose s-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

s-index vs degree plots

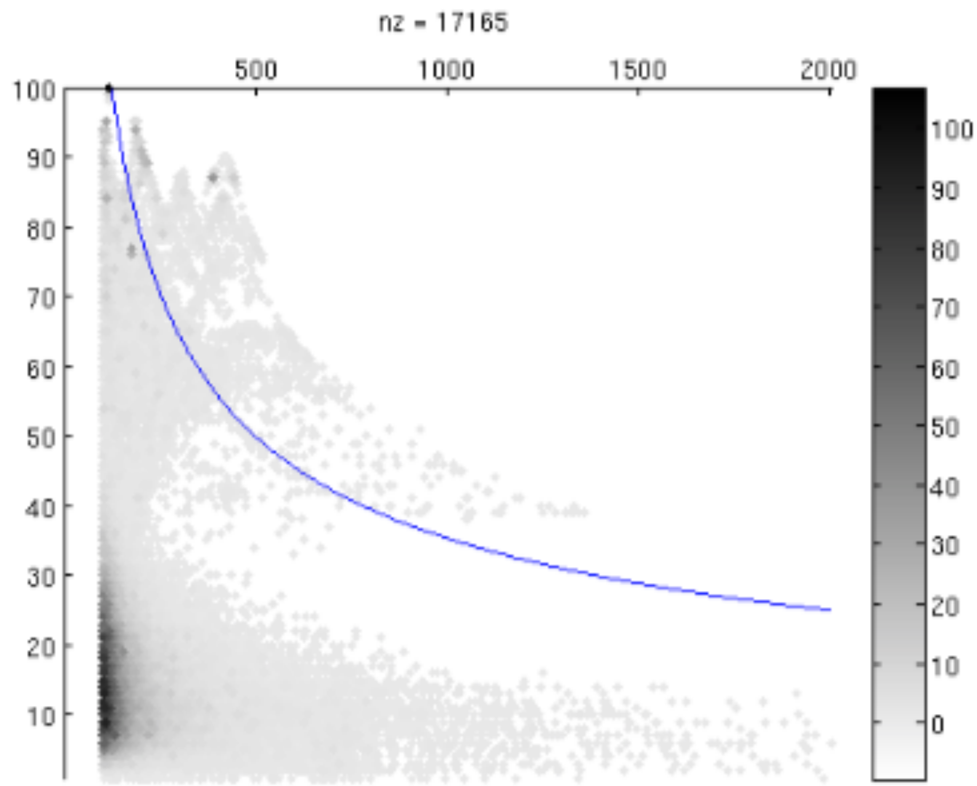


Figure 1: LiveJournal (original: no removal of non-reciprocating edges)

Sandia National Laboratories

s-index vs degree plots

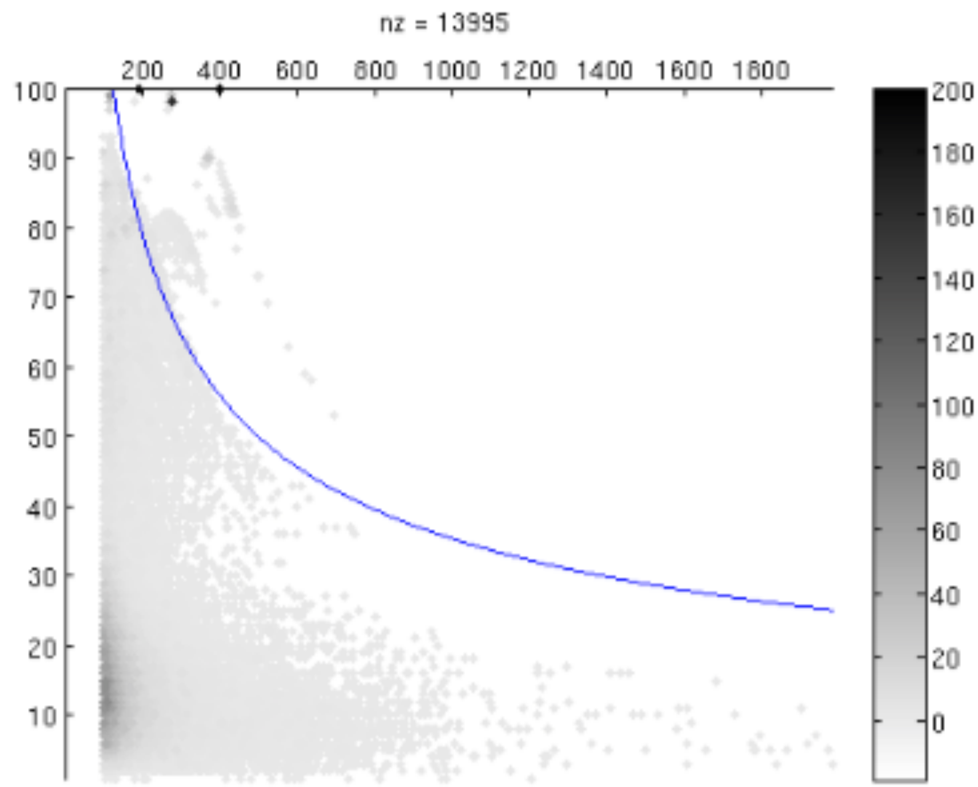


Figure 2: LiveJournal, non-reciprocating edges removed

s-index vs degree plots

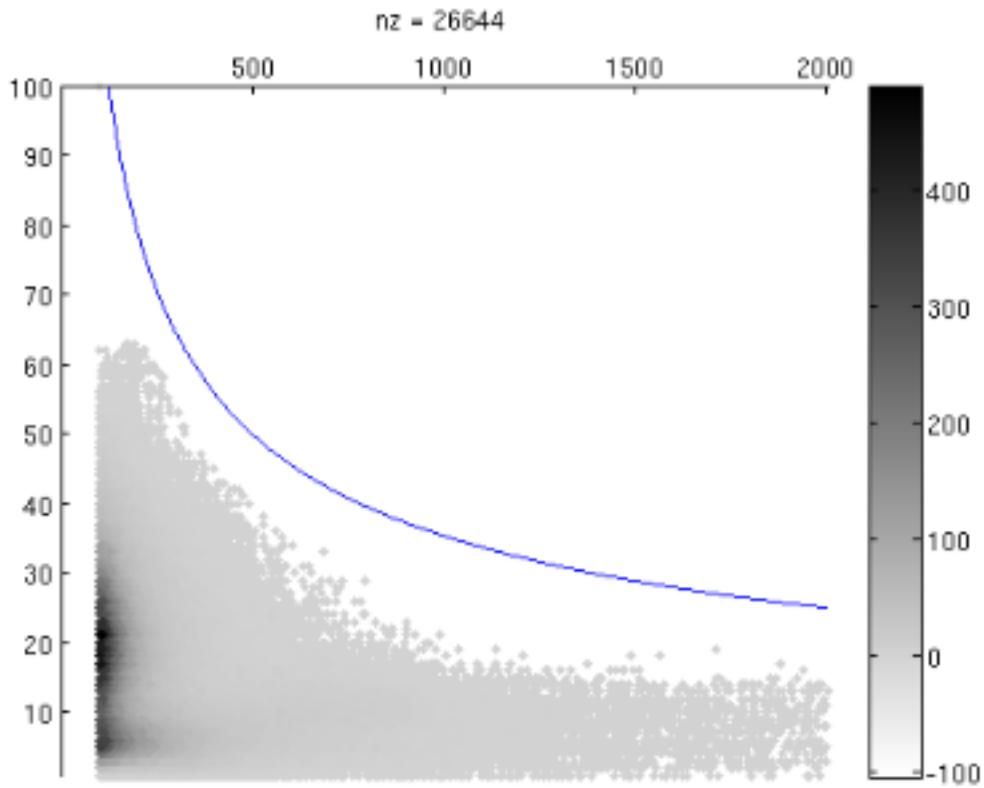


Figure 3: Orkut (original: no removal of non-reciprocating edges)

s-index vs degree plots

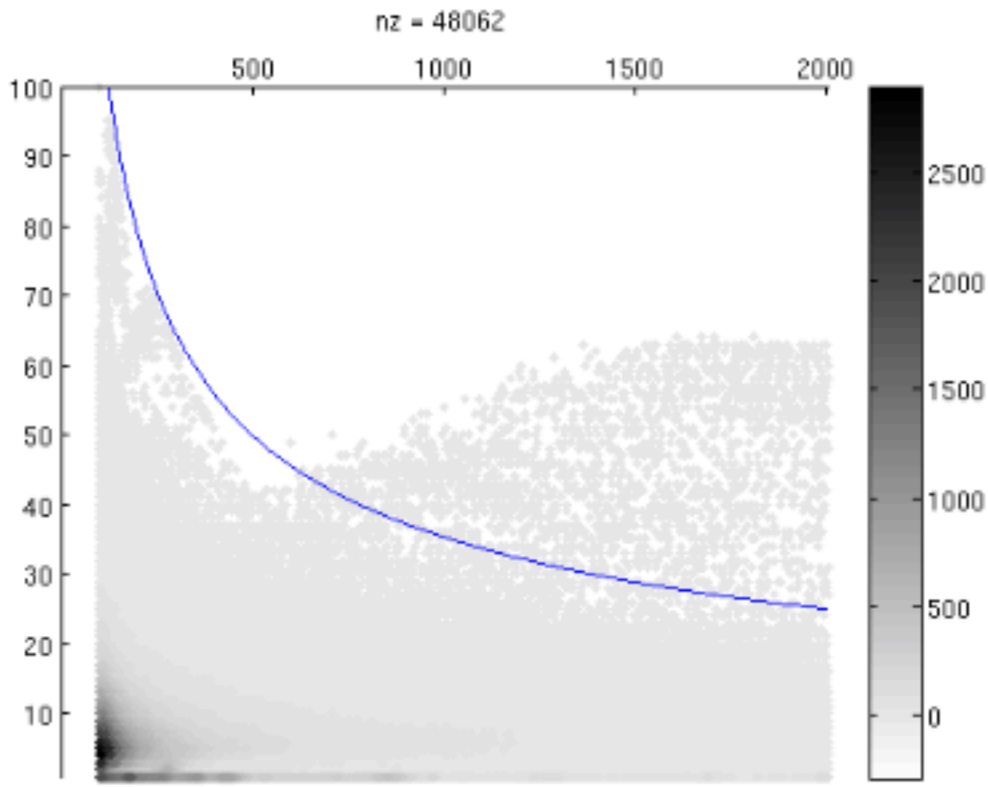


Figure 4: Twitter-2010 (original: no removal of non-reciprocating edges)

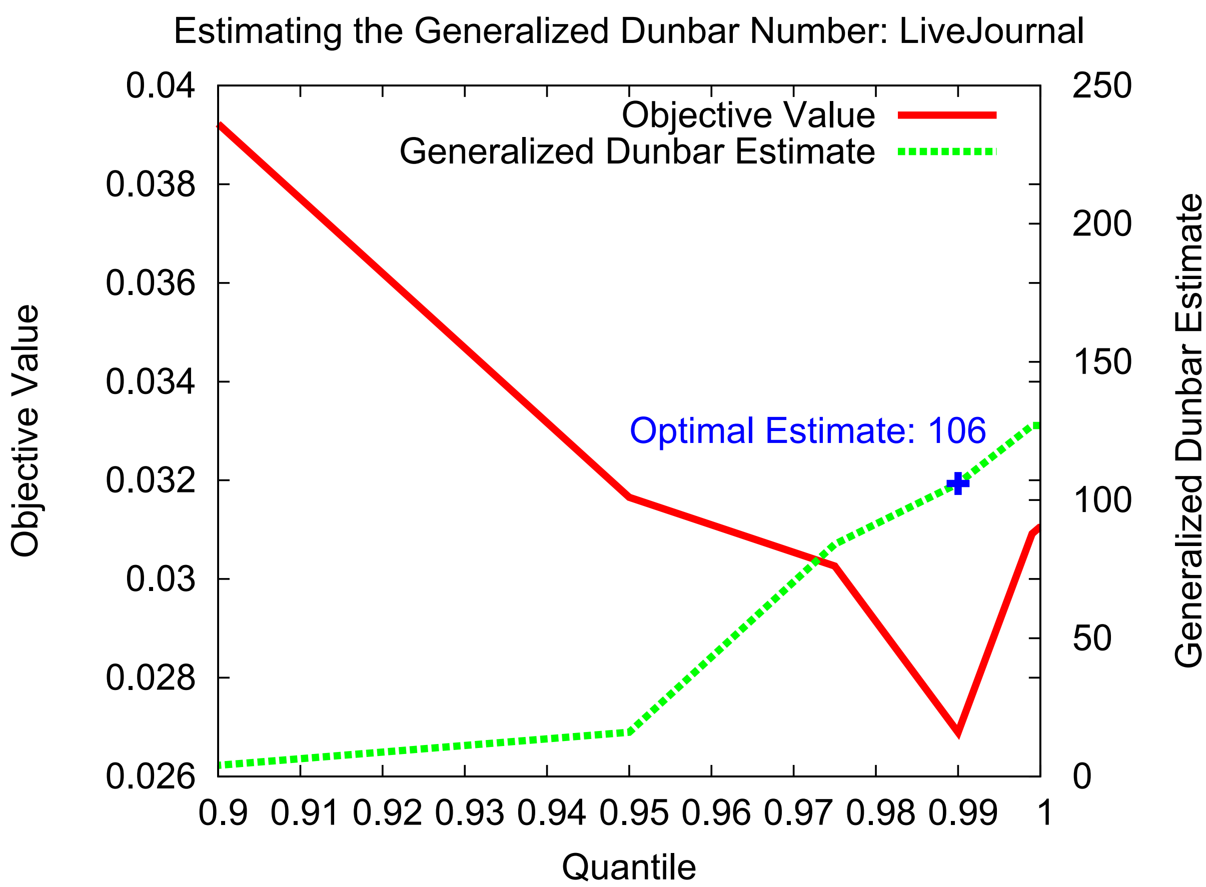
Finding D

- The previous plots used an “eyeballed” value of D
- We would like to calculate D in some statistically reasonable way
- Suggestion from Alyson Wilson:

$$\sum_d \left[\left(\sqrt{\frac{D}{d}} - y(q) \right)^2 * prop(d) \right]$$

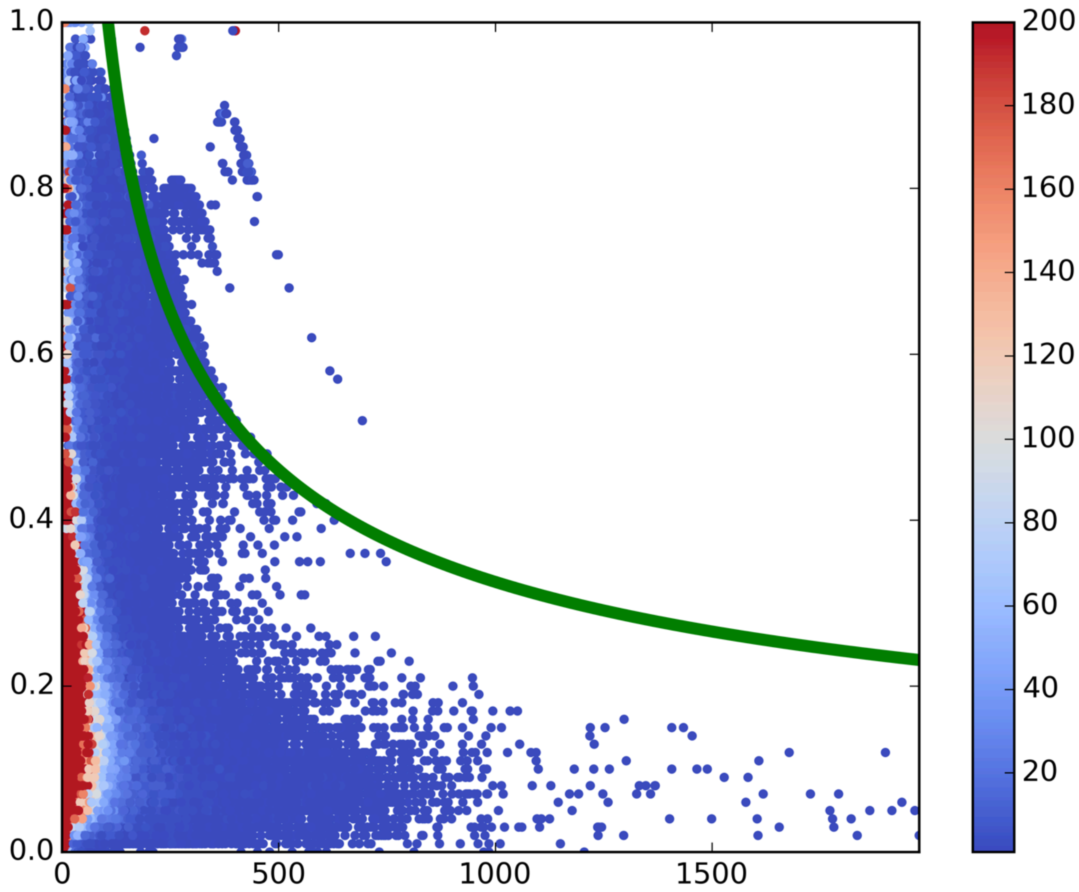
- Pick a quantile q (99% or 95%, etc)
- y(q) is the value at that quantile
- prop(d) is the proportion of nodes with degree d
- Skewed degree distribution requires exponential binning

Live Journal



Sandia National Laboratories

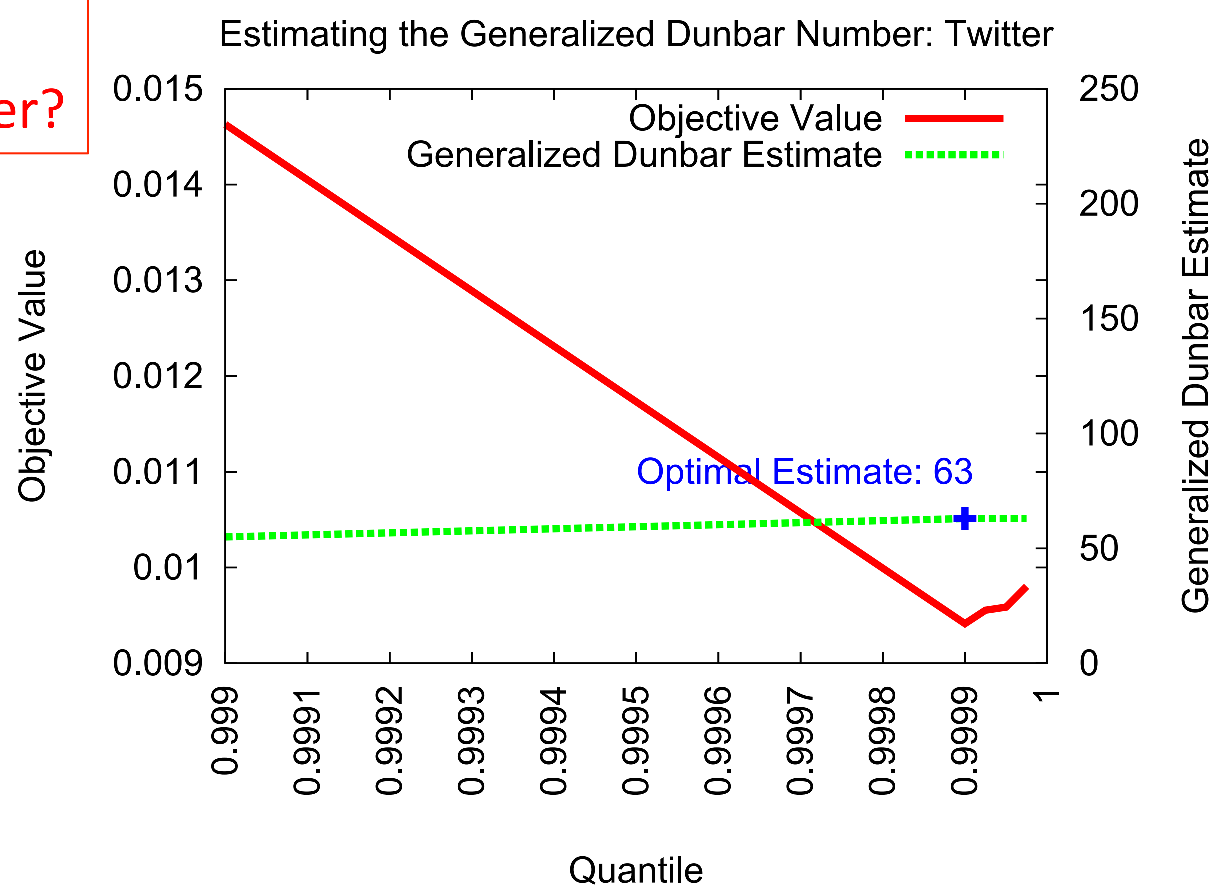
Live Journal



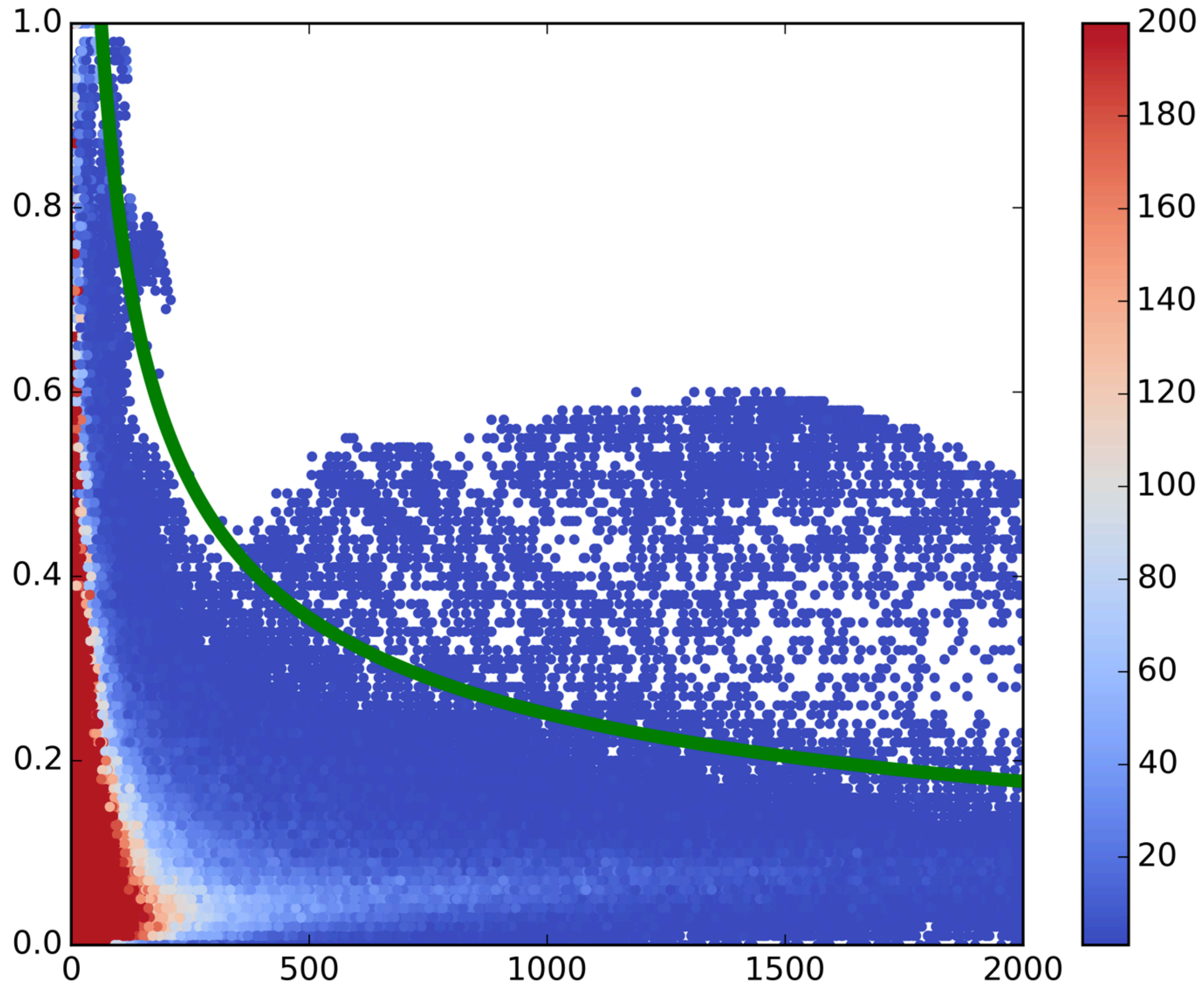
Sandia National Laboratories

Twitter

Why
Is D= 60
For Twitter?

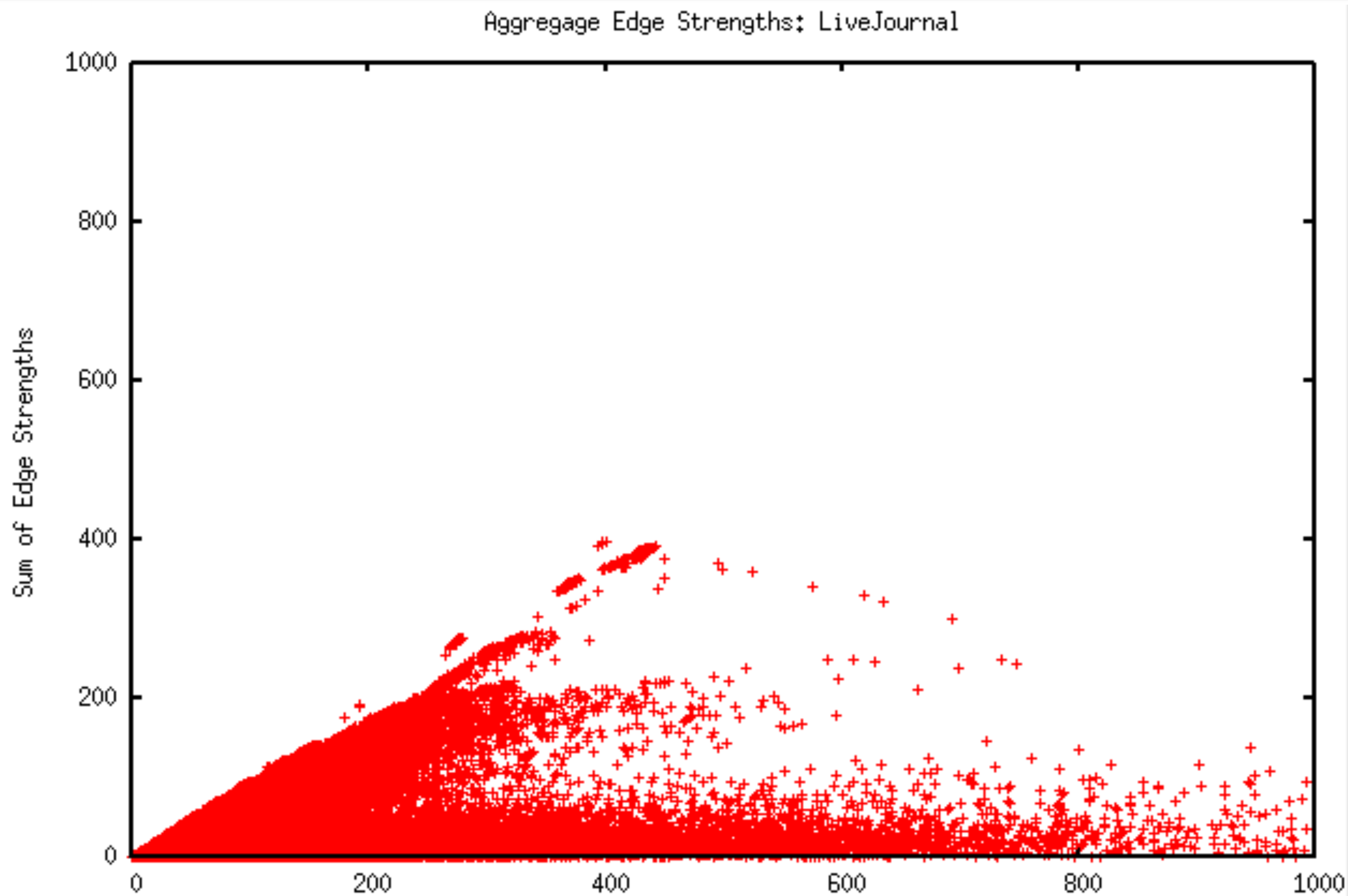


Twitter

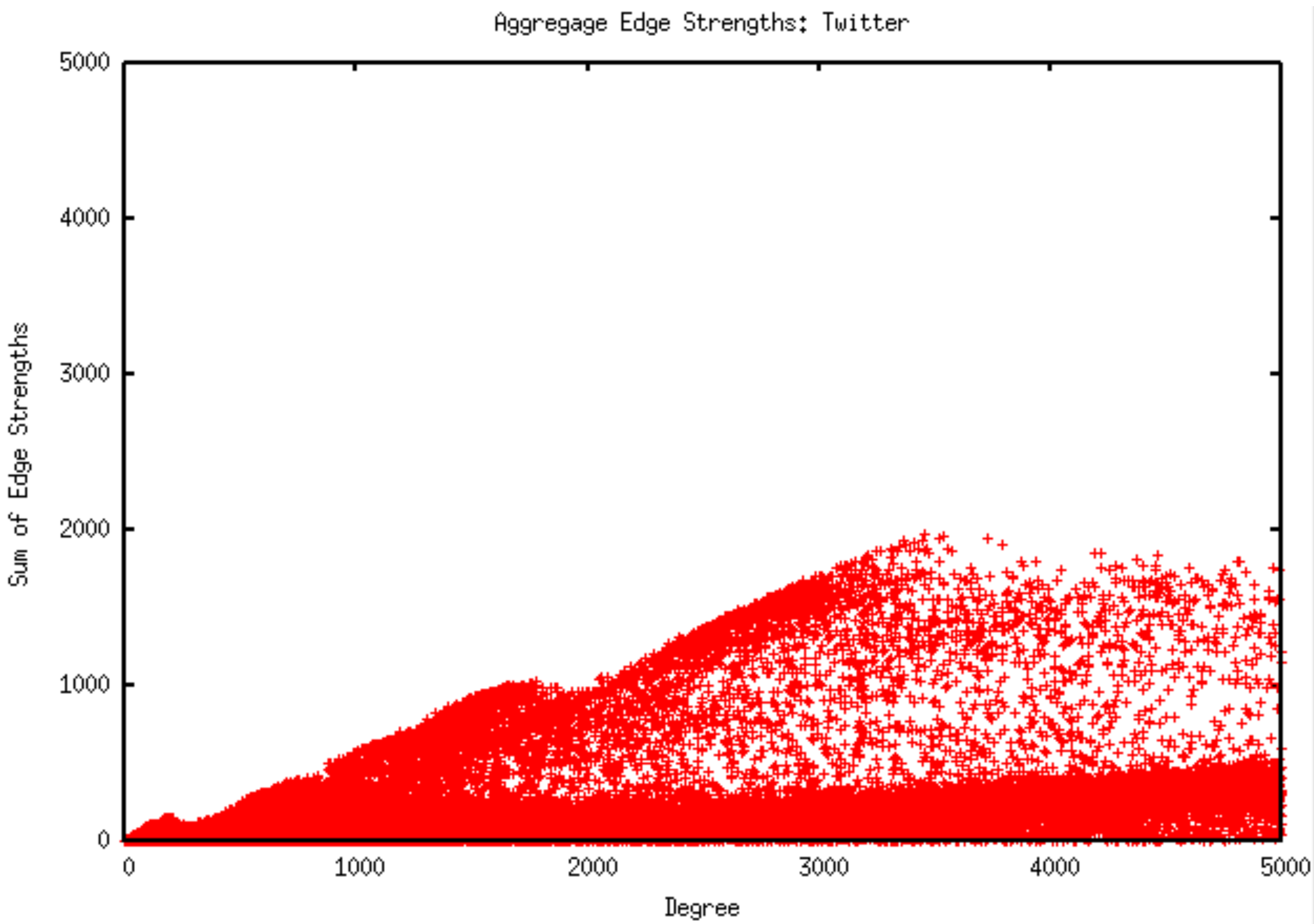


1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
200
180
160
140
120
100
80
60
40
20

Straight Strength - LiveJournal



Straight Strength - Twitter



Sandia National Laboratories

Summary

- An example where social network structure enables more efficient algorithms in theory and practice.
- Positive results in a model that captures constraints on cooperating autonomous data centers.
- A possible tool for cleaning non-human behavior from some social networks.

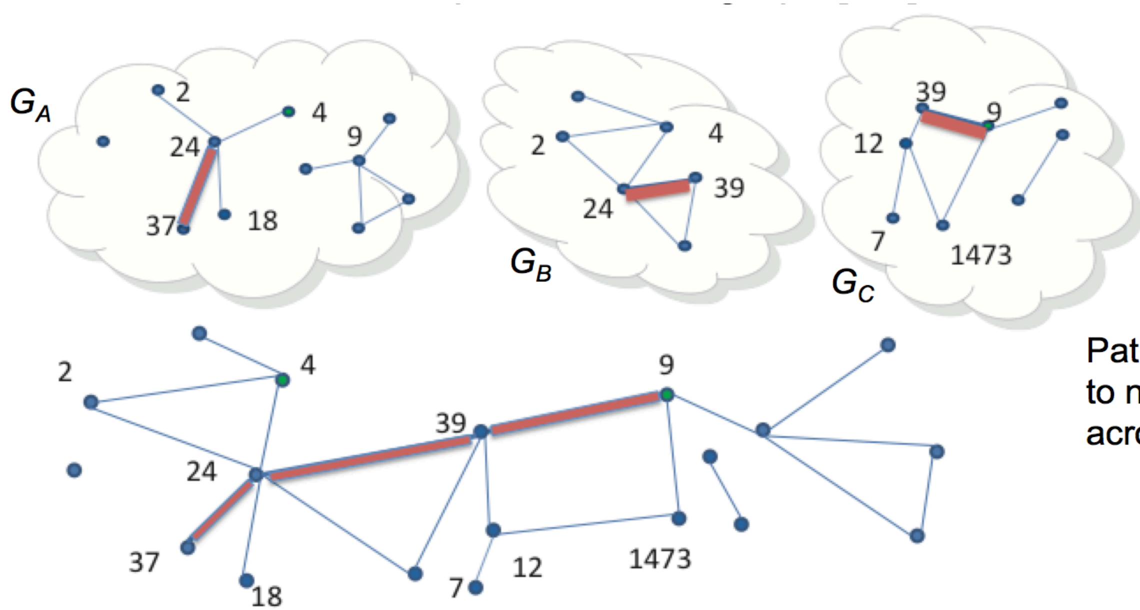
Backup Slides

Another Limited Sharing 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 (union G_c…).
Alice gets no information beyond answer in honest-but-curious model.



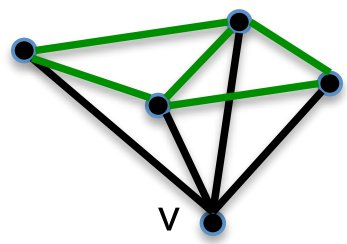
s-t connectivity
For 3 centers

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

Sandia National Laboratories

Maximum Triangle Density Subgraph (MTDS)

- Algorithmic tool
- Find subgraph that maximizes $\frac{\text{\# triangles in subgraph}}{\text{\# vertices in subgraph}}$



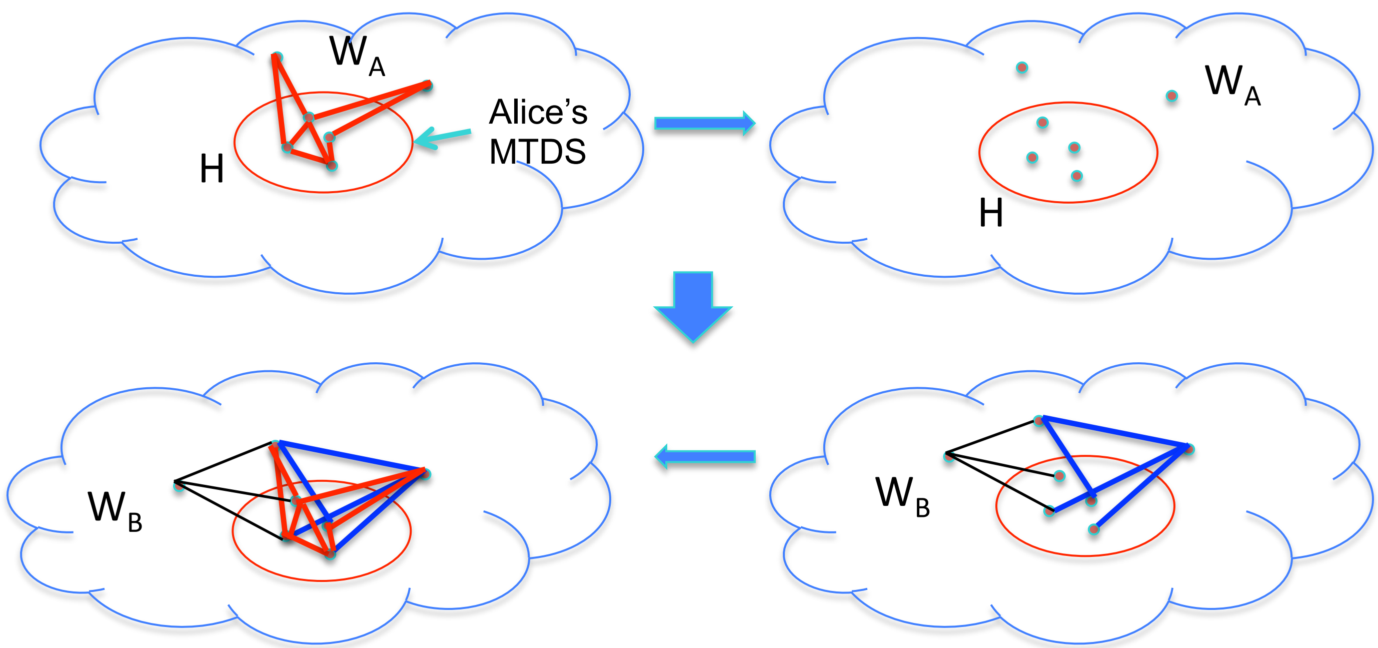
Triangle density = 7/5

- Solve in polynomial time via linear programming
  - Adjustment to Charikar's LP for maximum edge-density subgraph
- Greedy 3-approximation (from Charikar's 2-approx for edge density)

- **Theorem**: Alice's (WLOG) MTDS contains only nodes involved in the planted clique S
- **Theorem**: Whp any nodes not in S have O(1) edges into S.

Algorithm

1. Alice finds max triangle-density subgraph H and nodes (W_A) adjacent to at least half of H. Sends to H and W_A to Bob.
2. Bob finds nodes (W_B) adjacent to at least half of H and sends all induced edges (between V(H), W_A and/or W_B)
3. Alice finds clique (polynomial-time since O(log n))



Sandia National Laboratories