SAND2018-3529PE

Making Social-Network Data Sets More Human to Aid National-Security Graph Analytics

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

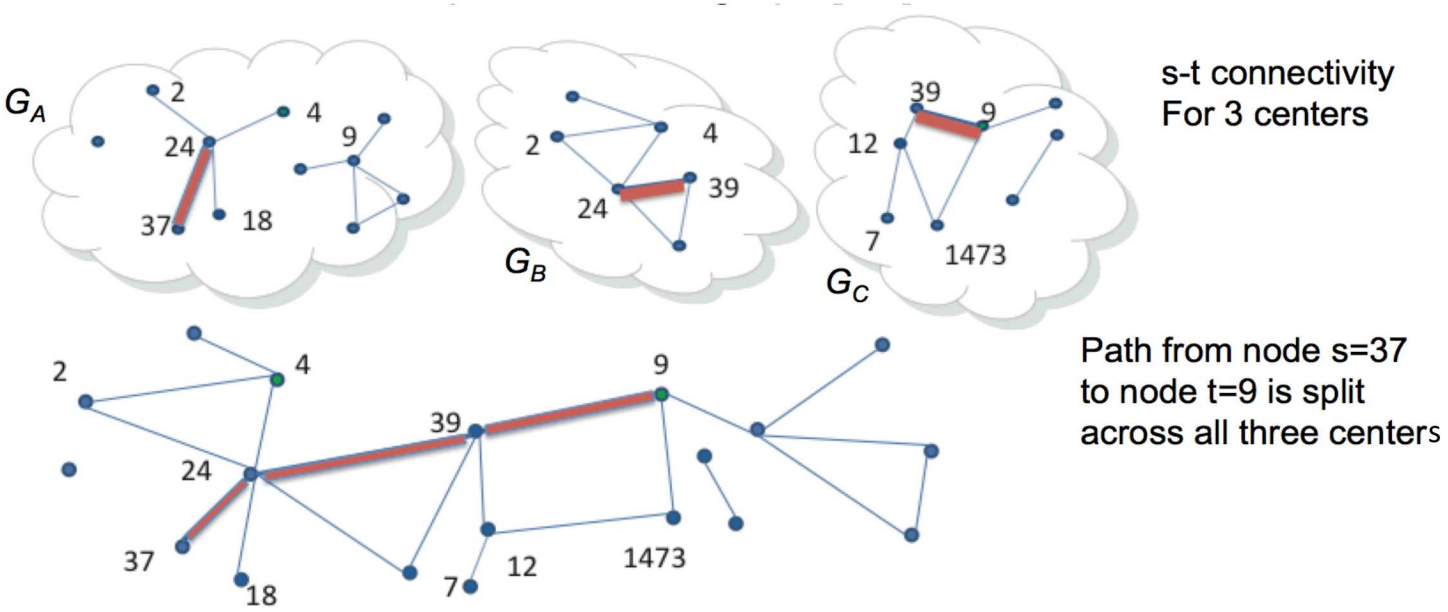
Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Origins: Distributed Graph Analytics

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

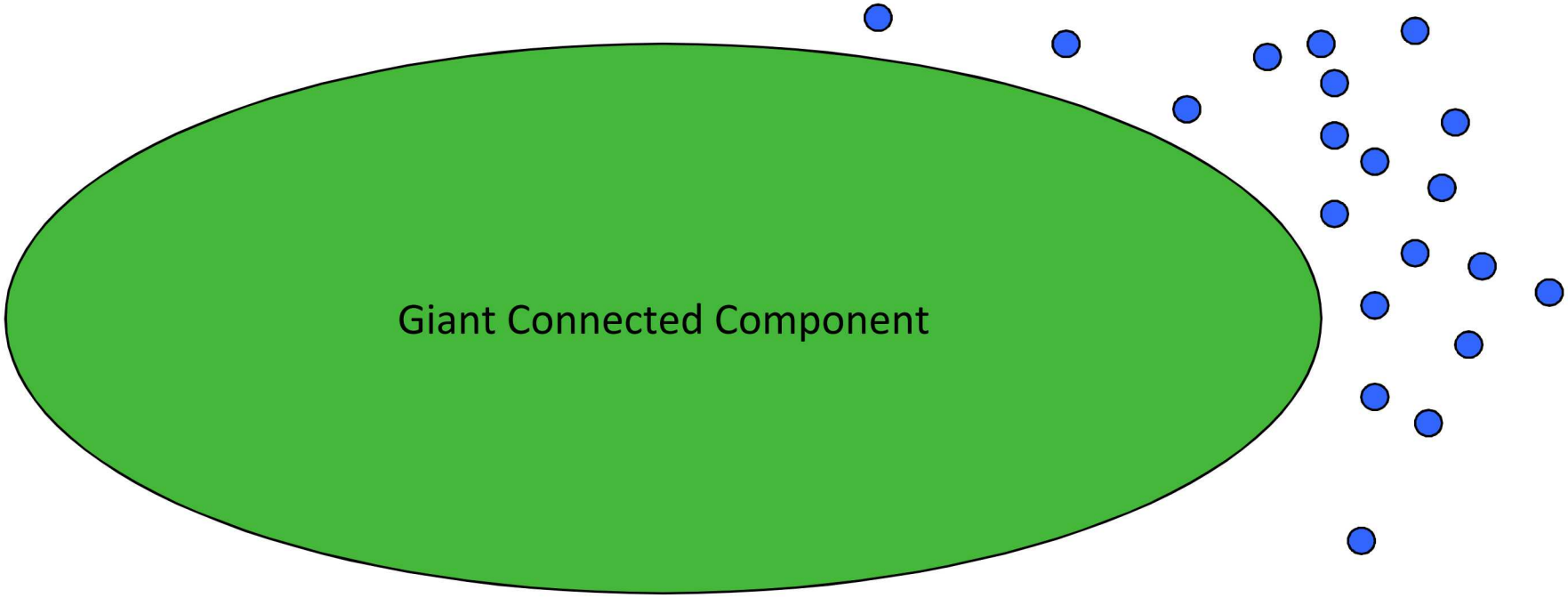
Goal: Cooperate to compute algorithms over G_A union G_B with limited sharing: $O(\log^k n)$ total communication for size n graphs, constant k

Motivation: National security: "connect the dots" for counterterrorism



Exploiting Graph Structure

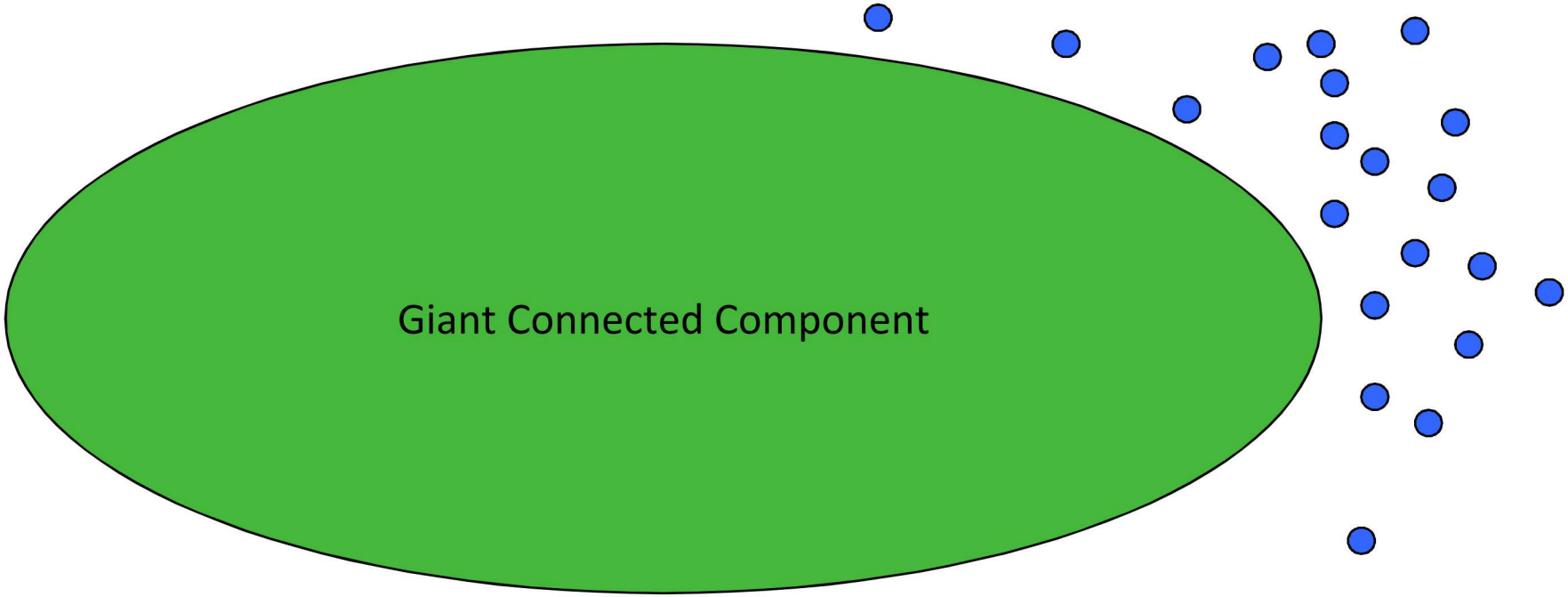
- Nodes are people, so exploit structure of social networks



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

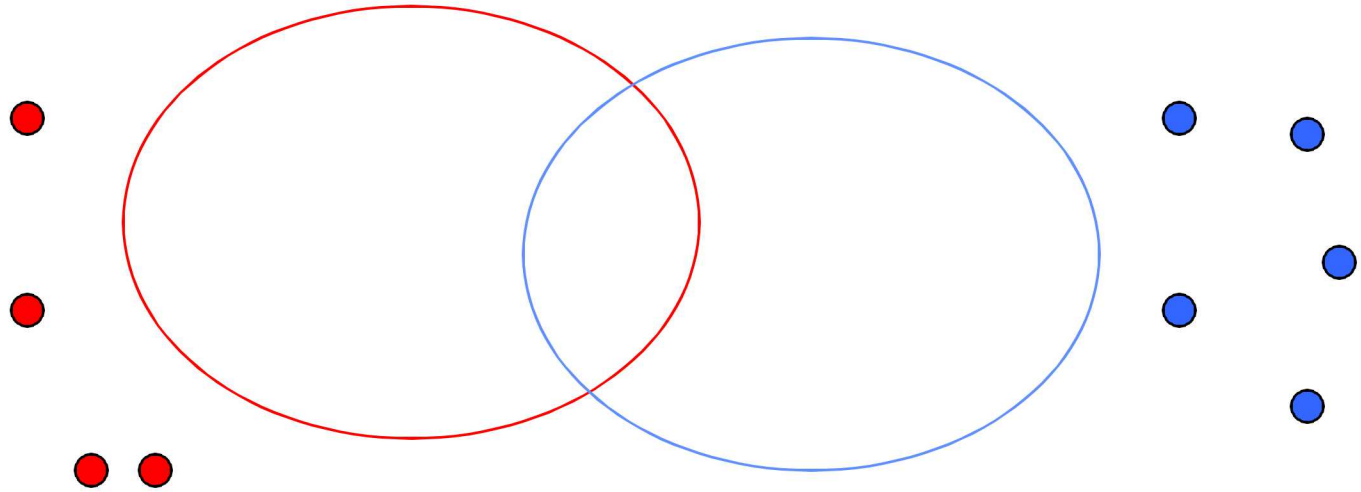
Social Network Structure

- 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+ε and 3.47



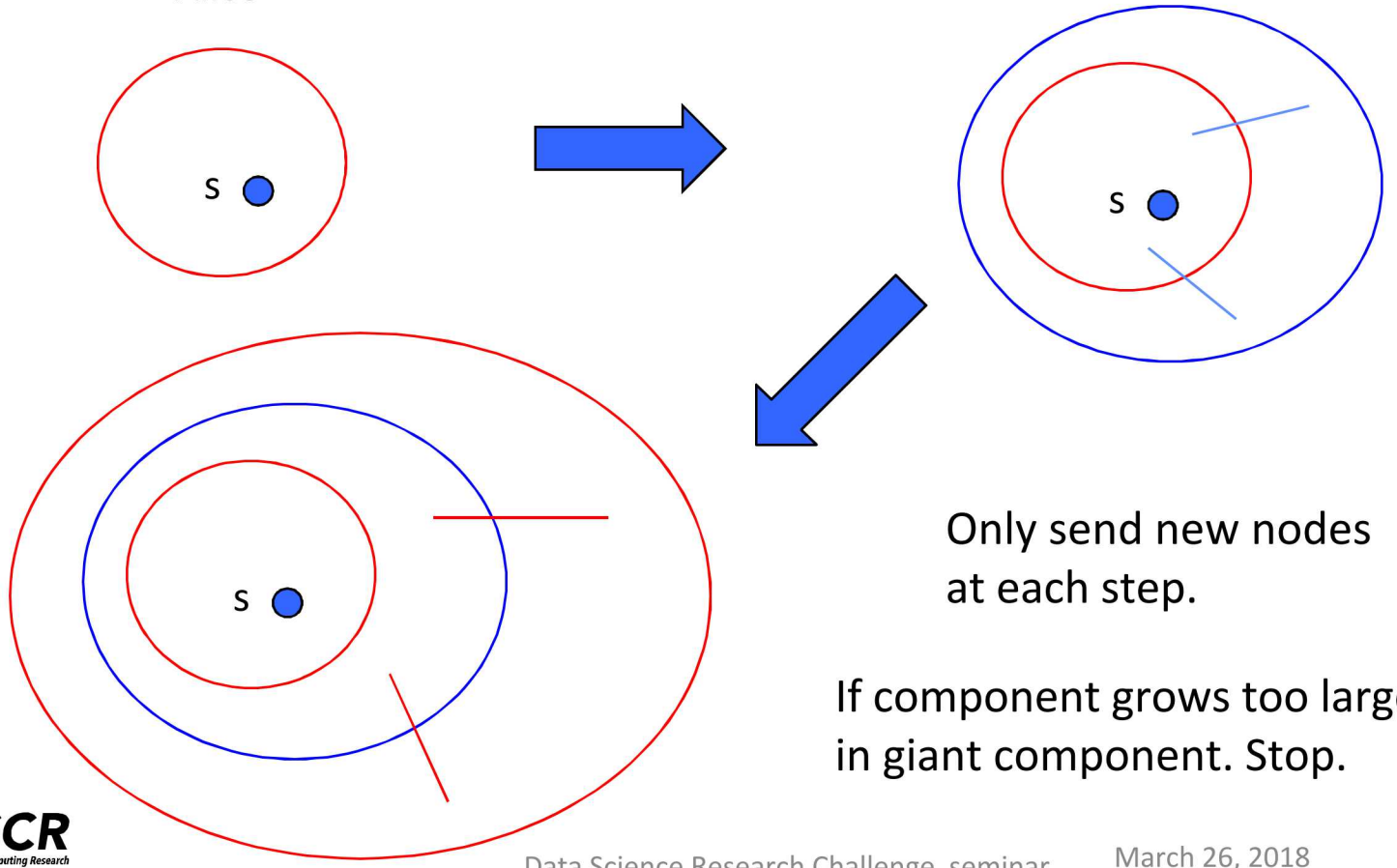
Assumptions

- 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)

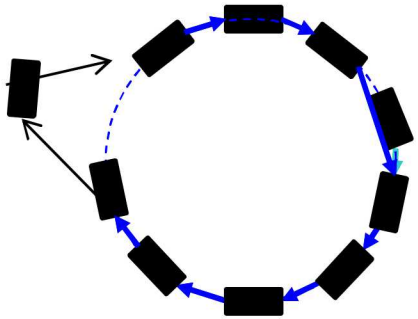


Only send new nodes at each step.

If component grows too large, in giant component. Stop.

Exploiting Graph Structure

- $O(\log^2 n)$-bit communication for s-t connectivity
  - Exploits giant component structure
  - Overcomes polynomial lower bounds for general graphs
    - $\Omega(n \log n)$ lower bound for general graphs (Hajnal, Maass, Turàn)
- Easy extension to multiple data centers



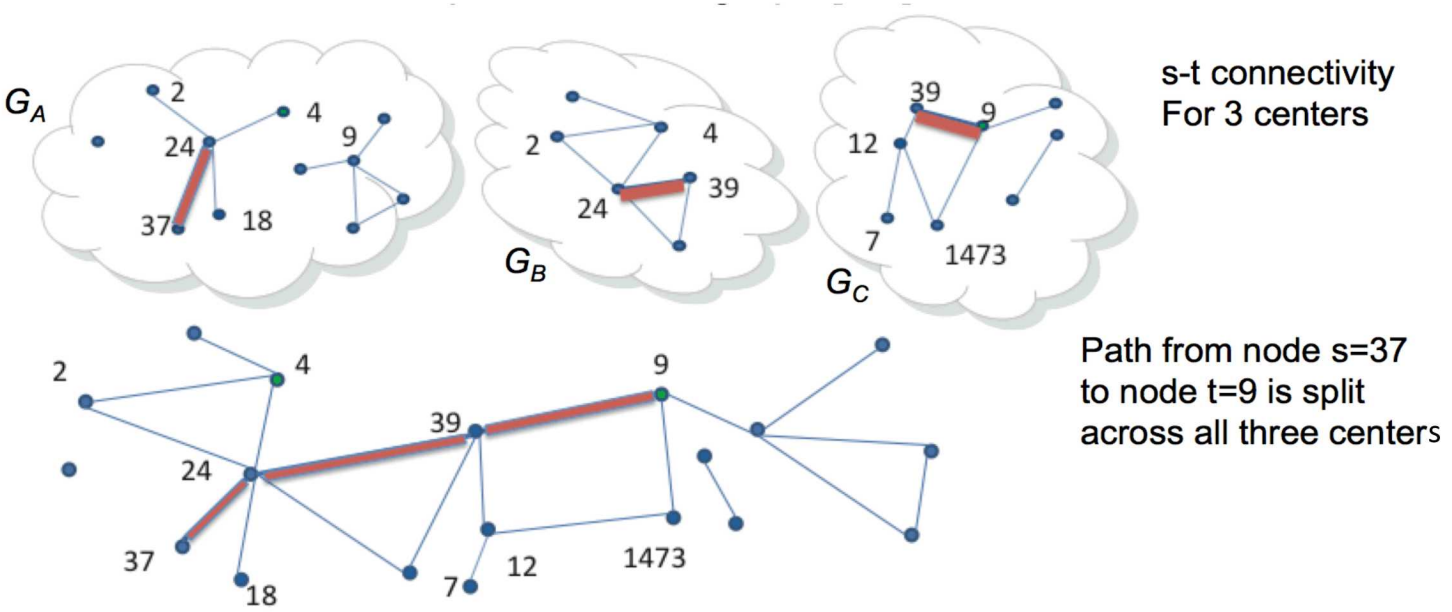
J. Berry, M. Collins, Aaron Kearns, C. Phillips, J. Saia, R. Smith, "Cooperative computing for autonomous data centers," Proceedings of the IEEE International Parallel and Distributed Processing Symposium, May 2015.

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

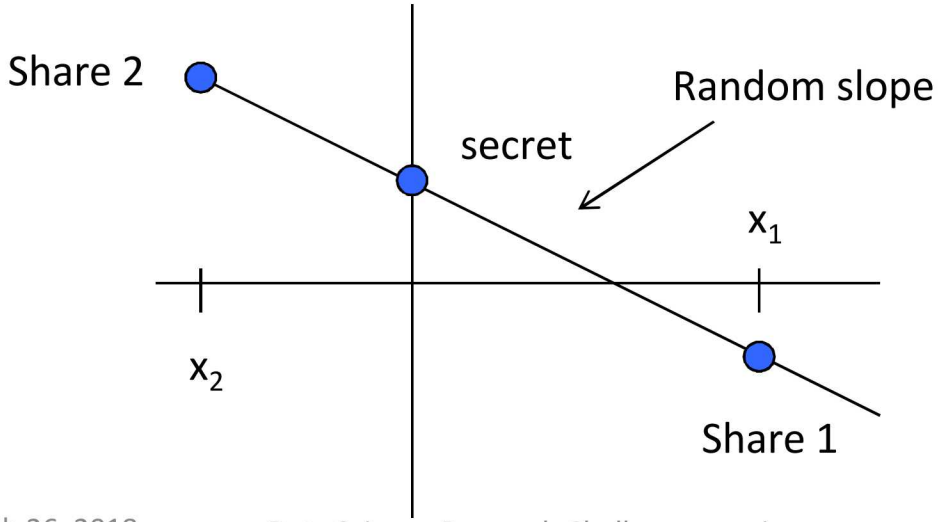


Secure Multiparty Computation Version

- Alice and Bob can determine that a path connects s and t without revealing anything about: the path, nodes seen by either party
- Similar to a model used by Brickell and Shmatikov
  - They assume known node names (shared customer lists)
- Secure multiparty computation
  - Usually many parties, small data (circuits, oblivious RAM)
    - Millionaire's problem
    - Beet farmers
  - We have small number of parties, large data

Tool #1

- Secret sharing
  - Secrets are in a finite field
  - Use a polynomial of degree d to encode a value, d+1 shares
    - All shares reveal secret, d reveals nothing
    - Solution is y intercept, secrets are polynomials at other x
- Key: Given a share of x (called $[x]_i$) and a share of y (called $[y]_i$), can get a share of the sum by adding shares: $[x+y]_i = x_i + y_i$



Tool #2: Secure MUX

$$\text{MUX}(c, a, b) = \begin{cases} a, & c \neq 0, \\ b, & \text{otherwise.} \end{cases}$$

- Need to be able to securely compute shares of MUX(c,a,b), given shares of a,b,c
- Information-theoretically secure protocols if at least 3 centers (Ben-or, Goldwasser, Wigderson)
- For 2 centers need Yao's garbled circuits (crytographic)
- This is expensive, requires communication

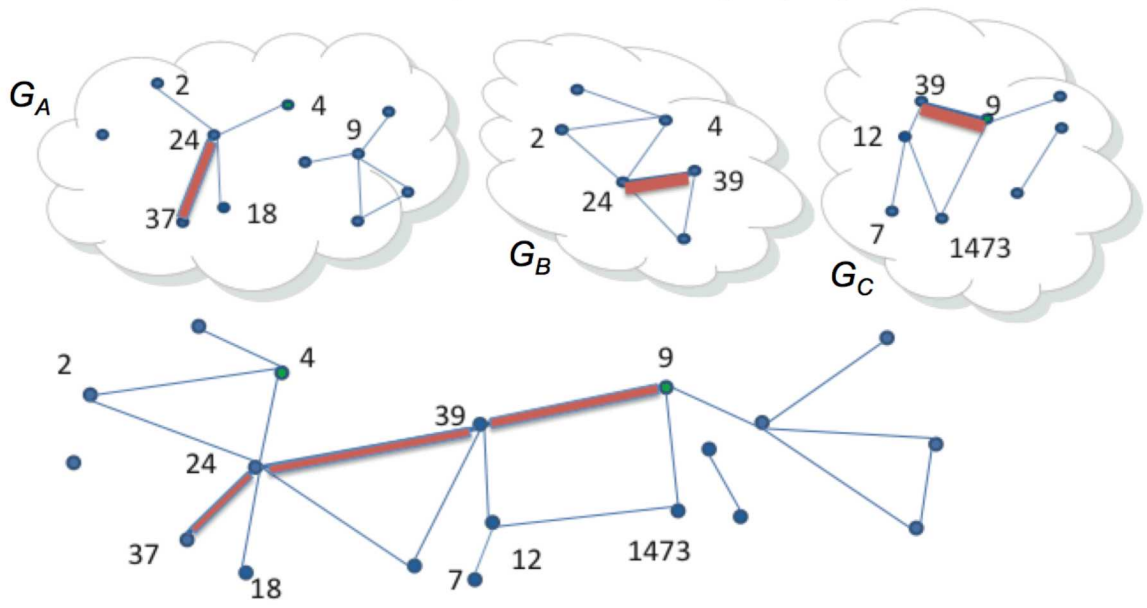
Algorithm Overview

- Alice (Bob) computes connected components on her (his) graph
- Secret share component names for each node (both Bob and Alice)
- Secret-shared shell expansion from s
- For each node compute secret-shared binary variable:
  - P(v) is 1 if node v in same component as s, else 0
- In end reveal P(t) by combining secret shares

- Can do this with hidden names except for s and t.

Computing in Secret-Shared World

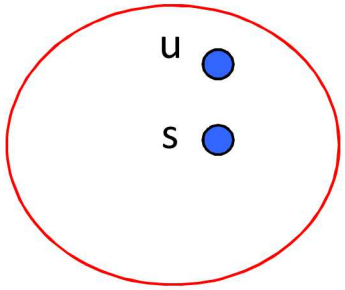
- Each data center has its part of the secret label for every node in every center
  - "Normal" computing on everything, but computing on shares
- Addition, subtraction, multiplication by a constant are local
- Comparison, multiplication of shares requires communication to all



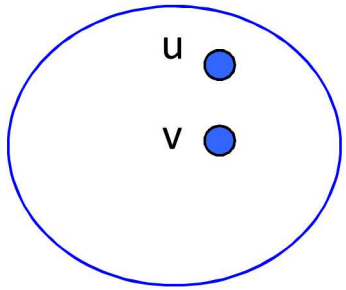
Propagating Connectivity Information

- P_v is a binary variable set to 0 iff there exists a node u such that $x_u=x_s$ and $y_u=y_v$.

Alice's variables
Are x_u, x_s



Alice



Bob

Bob's variables
Are y_u, y_v

Algorithm 1 OddStep

1: $P_v = 1$
2: **for** node u **do**
3: $P_v \leftarrow \text{MUX}((x_s - x_u + y_u - y_v), P_v, 0)$
4: **end for**

Pick labels so no Trivial zeros

Secure Multiparty Computation

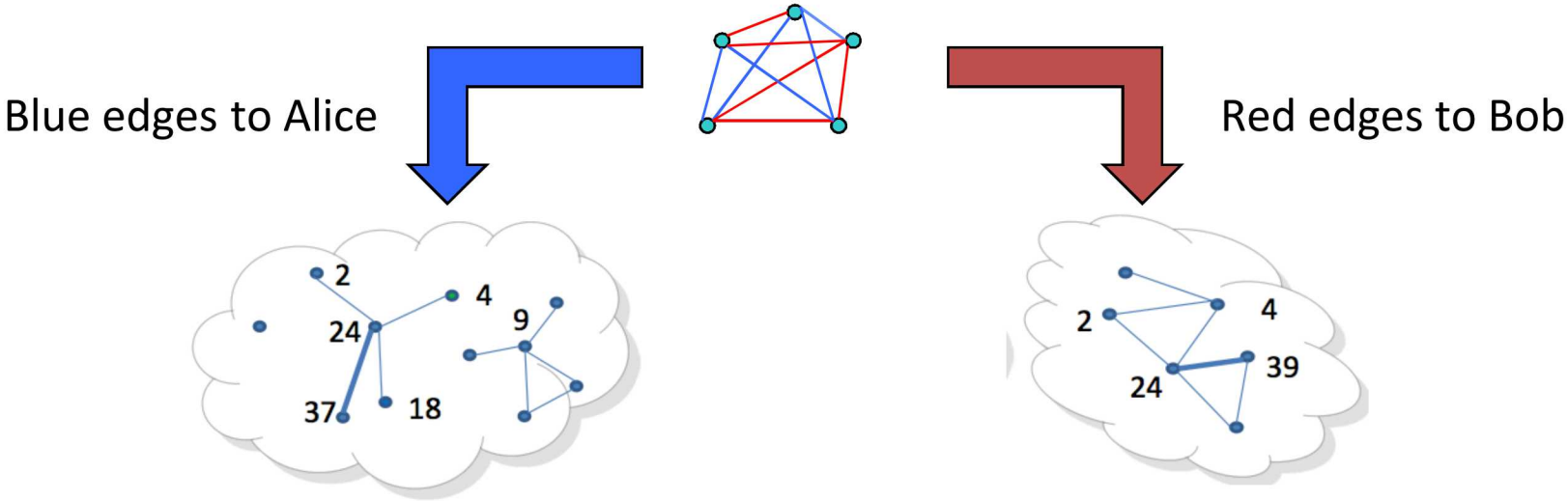
- Every step of shell expansion requires comparing everything
  - Don't know what shares represent
  - Values propagate (secretly), creating connected set
  - Expensive loops for large graphs and expensive computations
- Can reduce # of expansions if only care about short paths

- Considerable effort now to create faster secure multiparty computation
  - Motivated by cryptocurrency
  - New techniques: SNARKS (zero-knowledge proofs), P2P/secure data structures (ENIGMA company from MIT)

Exploiting Social Network Structure

- Next step: planted clique (dense subgraph anomaly detection)
  - Structural conjectures based on evolutionary psychology
  - Provably correct algorithm

- Experimental validation on some real networks failed!



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 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

*A. Horowitz and D. Horowitz, “Watch for fakes on social media”, The Costco Connection, 2014

Some Test Network Desired Properties

- Nodes are humans
- Edges plausibly represent a social bond
  - Even better if the relationship requires time/effort
- Large size (millions/billions of nodes/edges)
- Network is reasonably complete
  - Not an ego-network

Not too many publicly available social networks have all these.

Human vs Automated

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

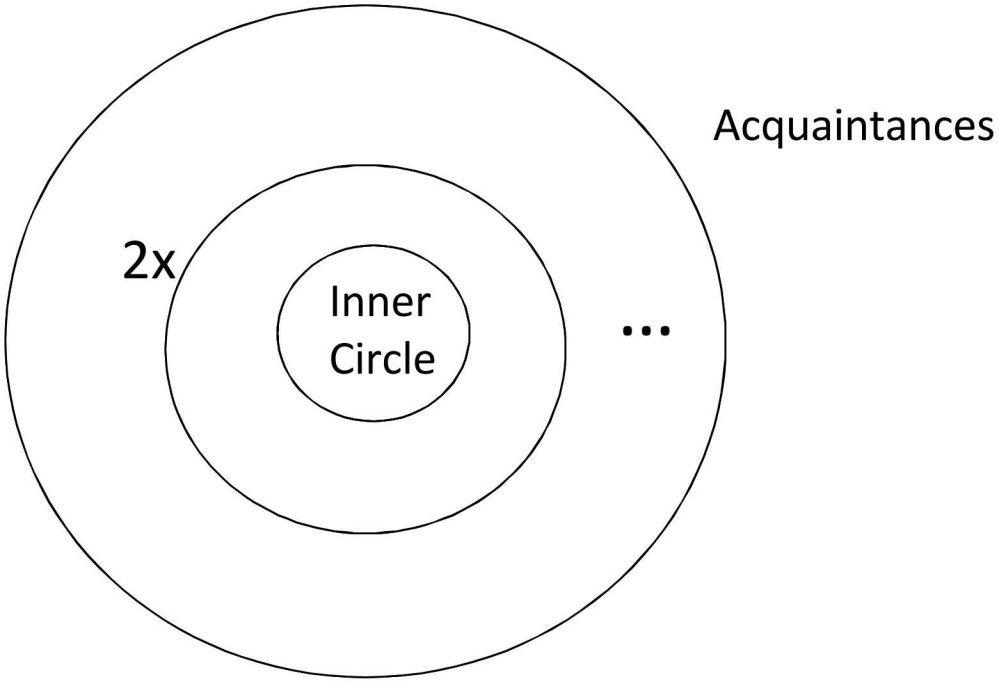
- Limitation: we have only topology
- An idea: Real human relationships require attention
  - Attention can be divided
  - Total attention, time of day, etc, is limited
- Nodes that show too many "strong" connections may not be human.
  - This includes humans, such as celebrities, who have a group of others manage their social media accounts.

- We'll give a method, then consider
  - Is it (plausibly) correct?
  - Should we care?

Varying Strength of Ties

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

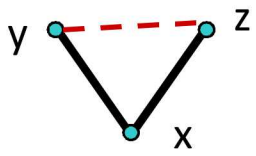
R. Dunbar, Social cognition on the internet: testing constraints on social network Size, Philosophical Transactions of the Royal Society B, Biological Sciences,367(1599):2192-2201, 2012



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

Triangle Significance

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



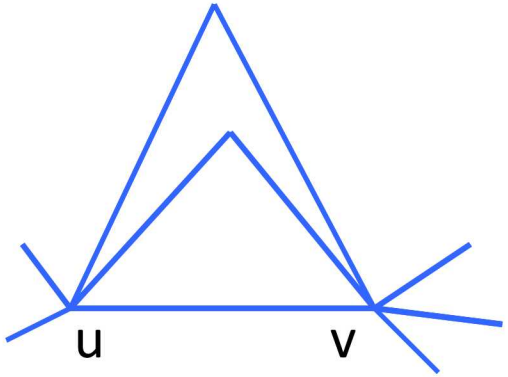
  - Converse of strong triadic closure: not (both edges strong) implies coincidental closures
    - experimental evidence: Kossinets, Watts 2006

"Communities have triangles"

Edge strength

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

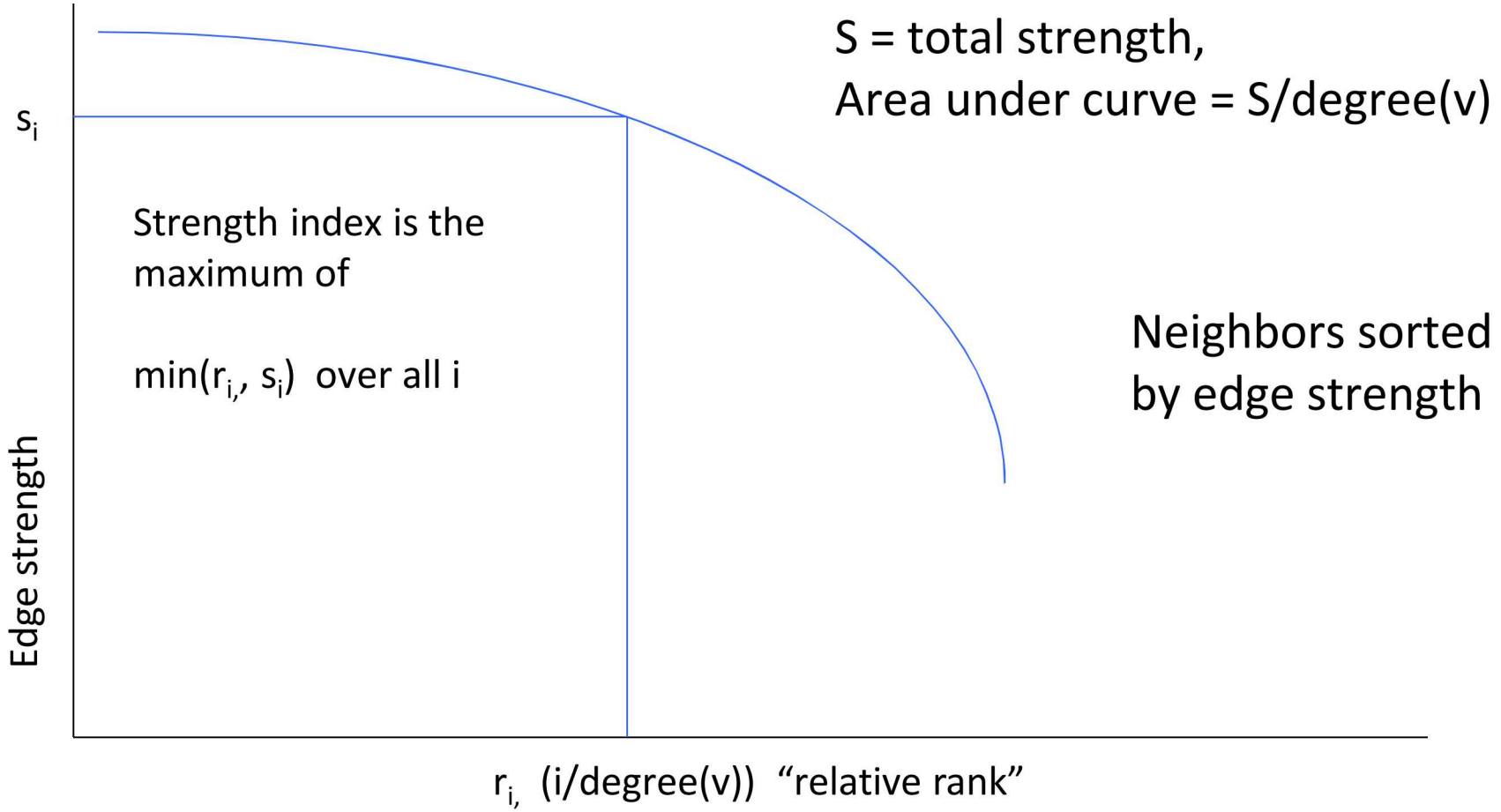
$$s(u,v) = \frac{2 * \#\text{ triangles on}(u,v)}{d_u + d_v - 2}$$



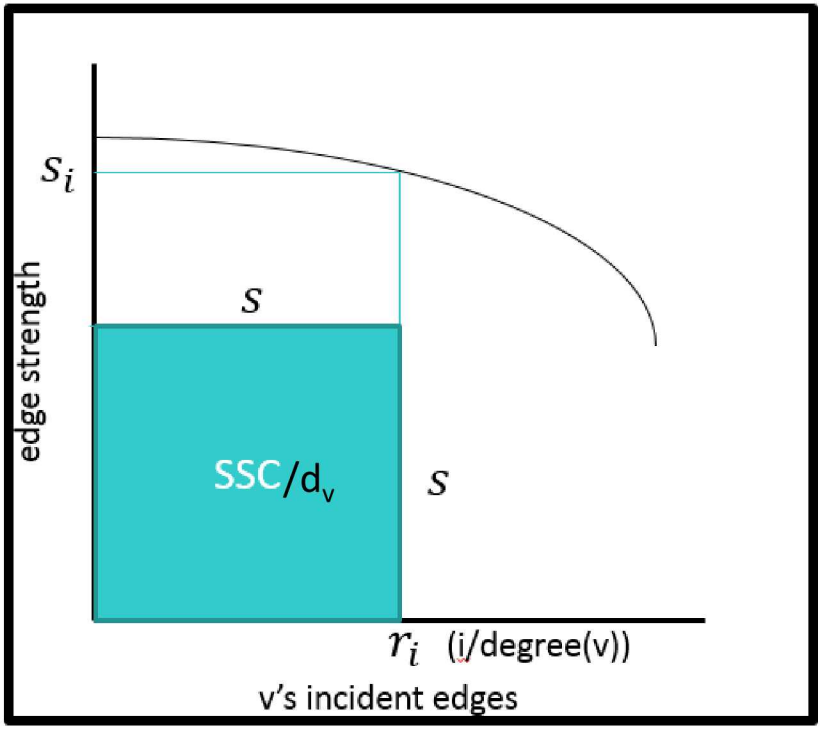
$$s(u,v) = \frac{2*2}{5+6-2} = \frac{4}{9}$$

- Assumption: Total strength of edges on a vertex has a constant bound D_G (network-dependent)
  - Edge strength a continuum, not just strong/weak

Strength-index for a vertex



Strength-Index Property



SSC = "Symmetric strength component"

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

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

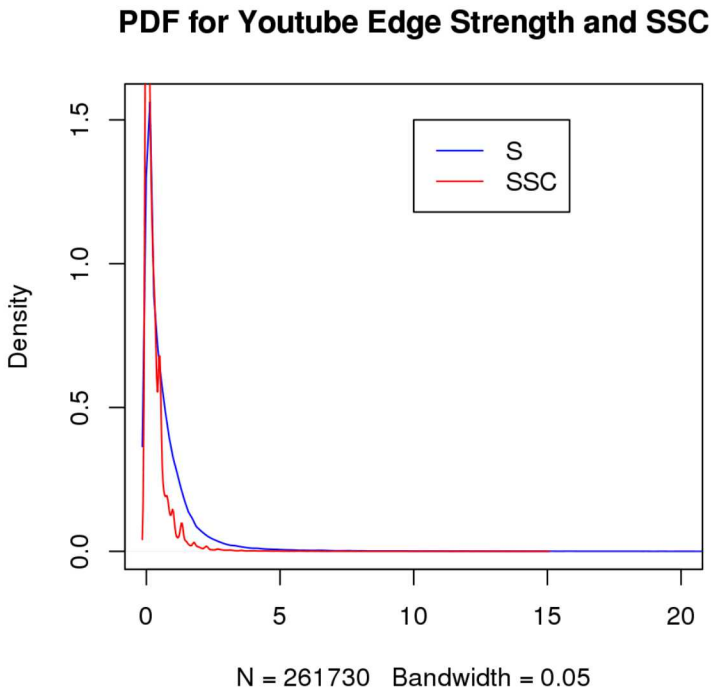
$$s \leq \sqrt{\frac{D}{d}}$$

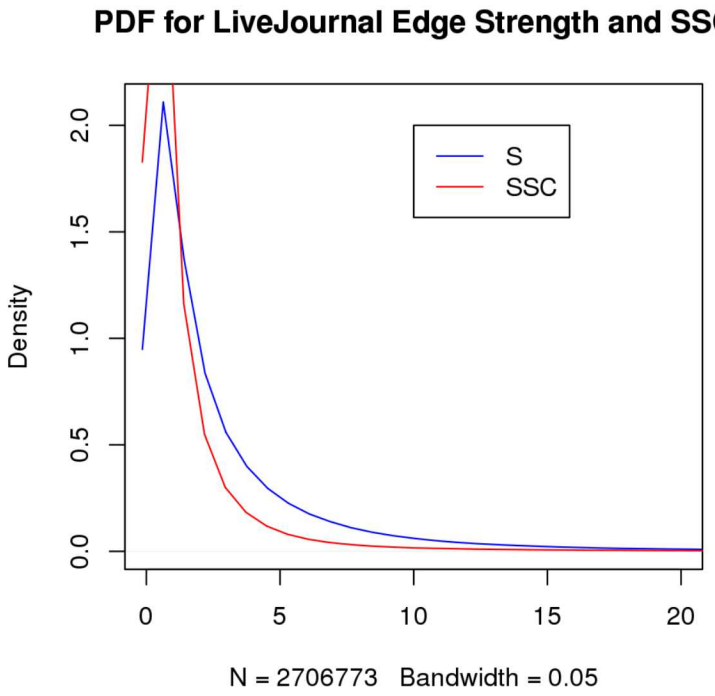
s = s-index
D = Dunbar-like constant
d = degree
SSC = $s^2 d_v$

Most important edges
Free from tail effects

SSC and total strength distributions

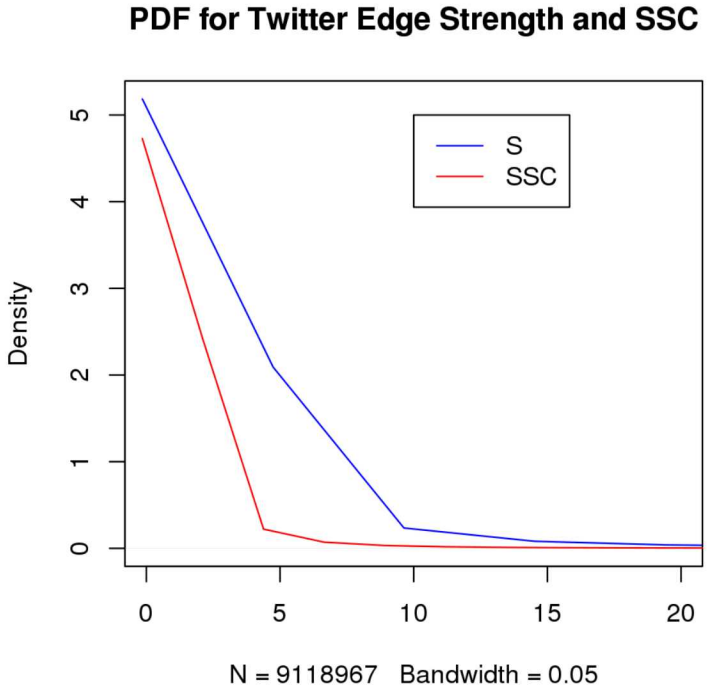
- SSC and total strength S seem to be (mostly) bounded by constant
- SSC seems to (mostly) be a good approximation to S

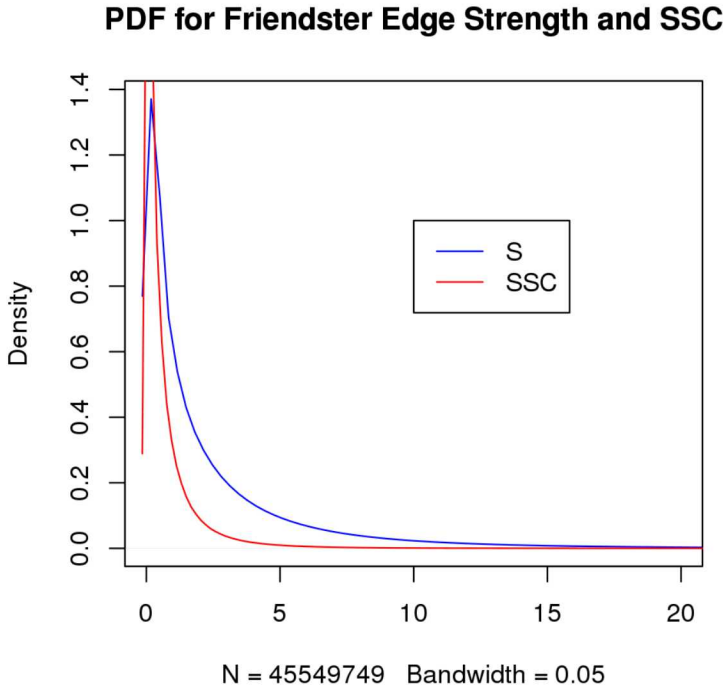




More Distributions

- Larger social networks





Cleaning Non-Human Nodes

- We assume $s \leq \sqrt{\frac{D}{d}}$ for all/most vertices
- Constant D will depend on the network
- Remove edges between nodes with s above this curve
- Selecting D
  - Compute average SSC average μ and standard deviation σ
  - D = μ + kσ for user-defined parameter k
- We use k=3

- We use only reciprocated edges

Why not remove whole vertex?

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

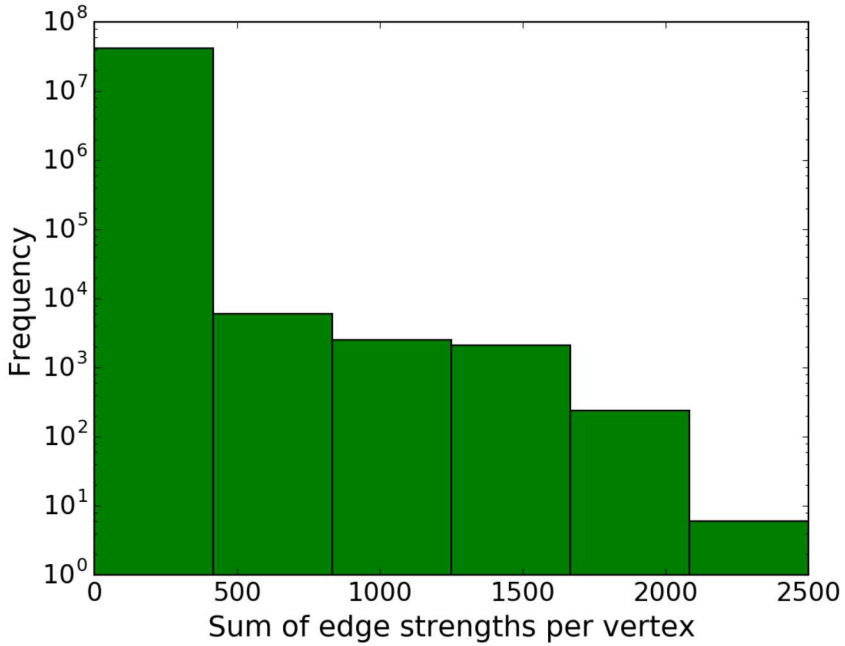
Network	percentage of vertices removed	percentage of edges removed
com-youtube($12\bar{\sigma}$)	0.01%	2.5%
com-youtube($6\bar{\sigma}$)	0.11%	10.76%
com-youtube($3\bar{\sigma}$)	1.18%	**32%**
ljournal-2008($12\bar{\sigma}$)	0.05%	1.57%
ljournal-2008($6\bar{\sigma}$)	0.14%	3.13%
ljournal-2008($3\bar{\sigma}$)	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%**

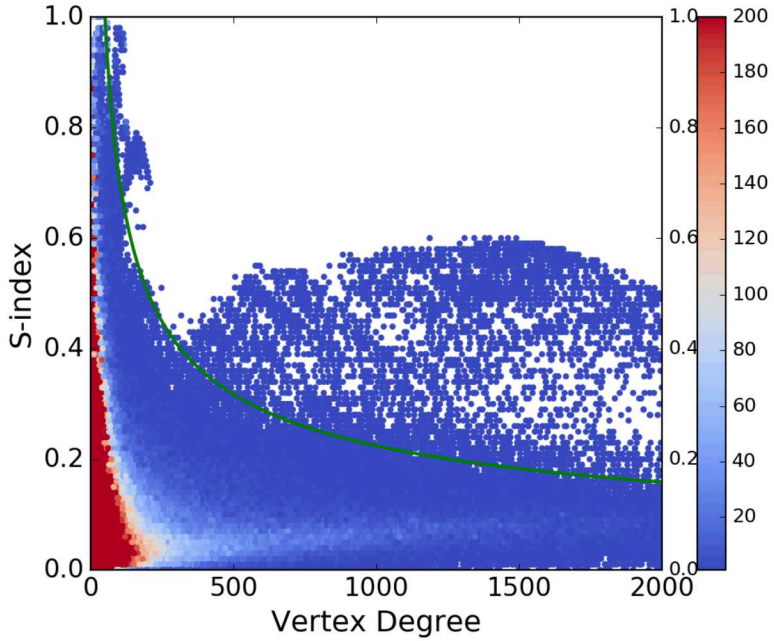
Why not remove whole vertex?

- Twitter is one social network where we can look up accounts
- Initial validation:
  - Strange connectivity:
    - A musician from a late-night show
    - A frisbee golf company (in New Jersey?)
    - Filmchair
    - Another unrelated Canadian company
    - Etc
- Conjecture: They paid a company to manage their Twitter accounts and the company connected them all

Twitter

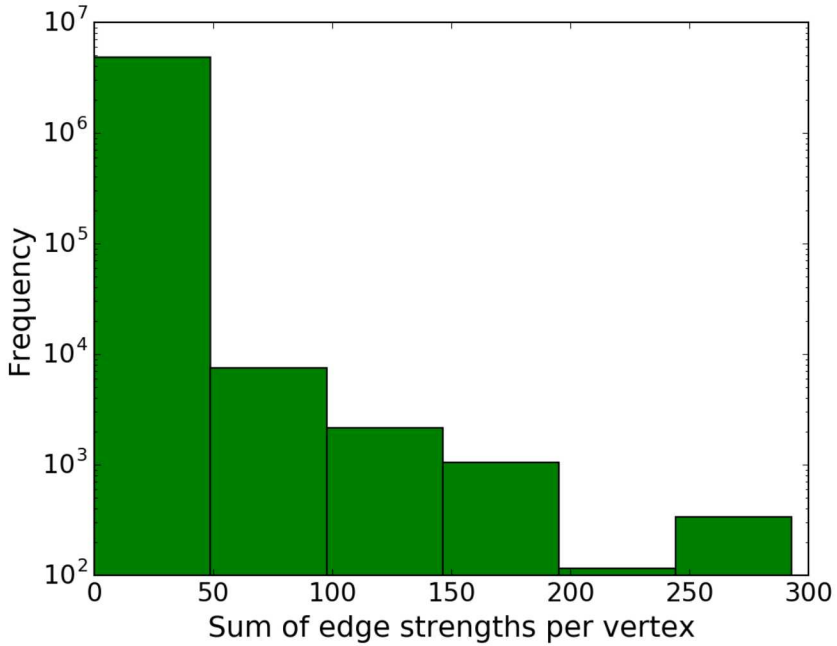
- 41.5M nodes, 266M reciprocated edges, D_G = 50

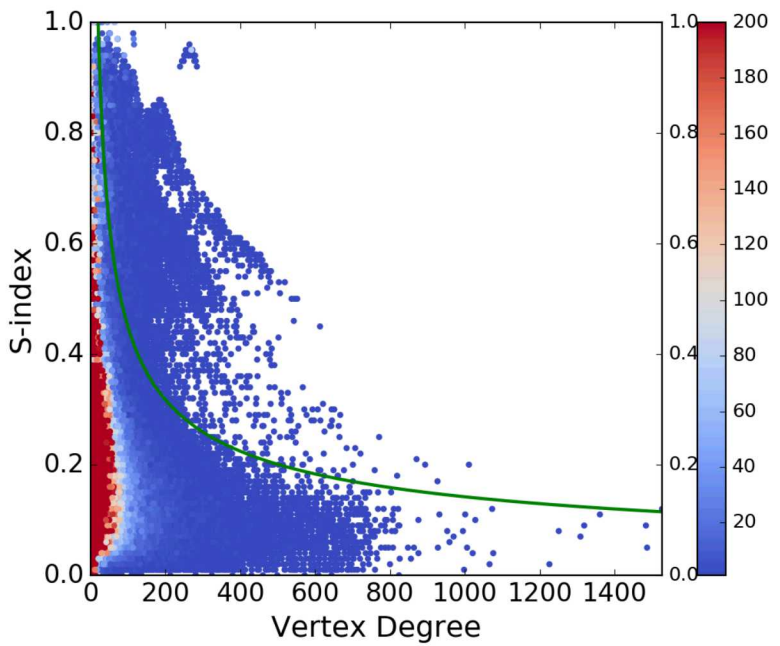




LiveJournal

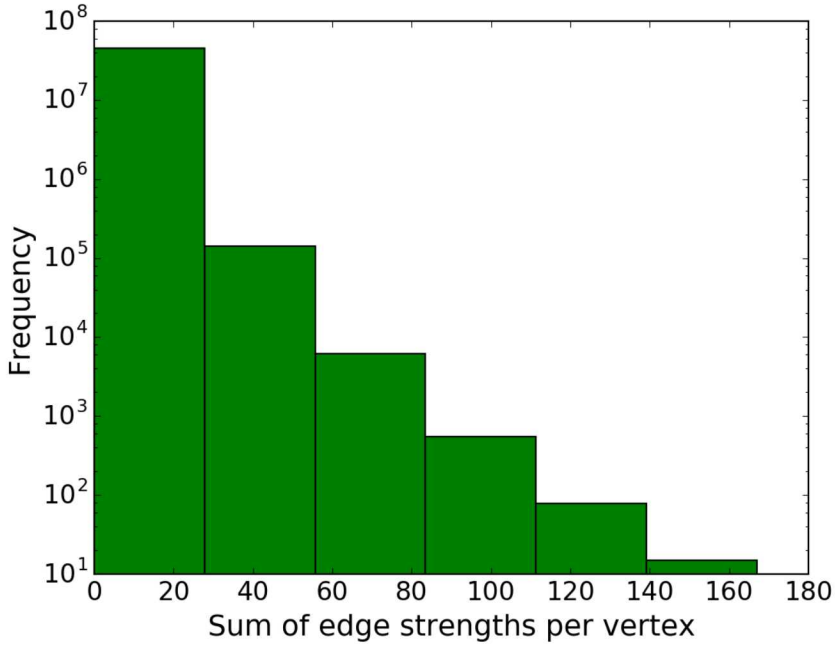
- 4.8M nodes, 25.6M reciprocated edges, D_G=20

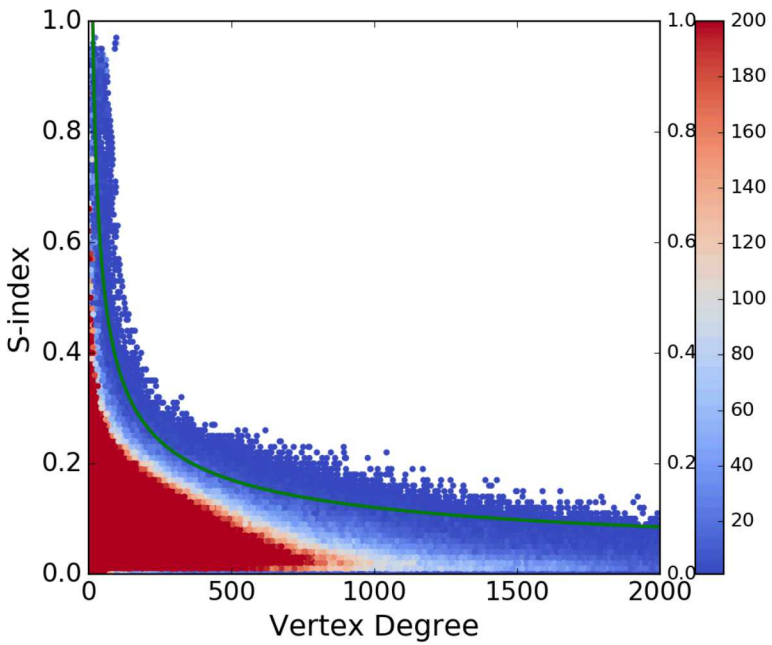




Friendster

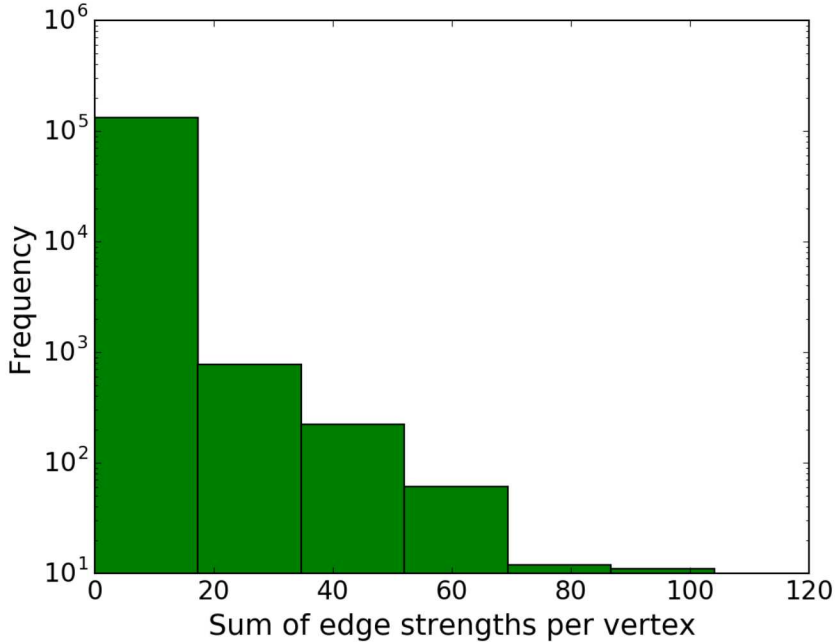
- 65.6M nodes, 1.8B reciprocated edges, D_G=14

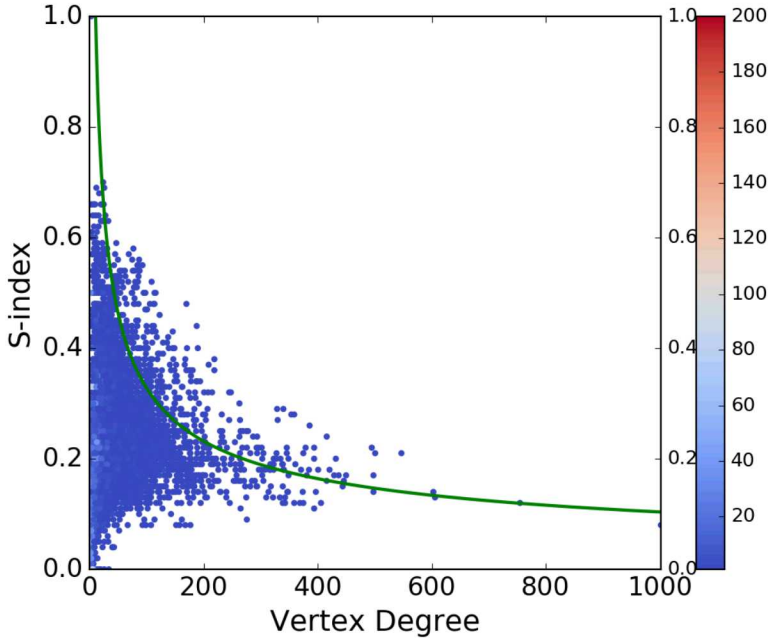




Ca-AstroPh (citation)

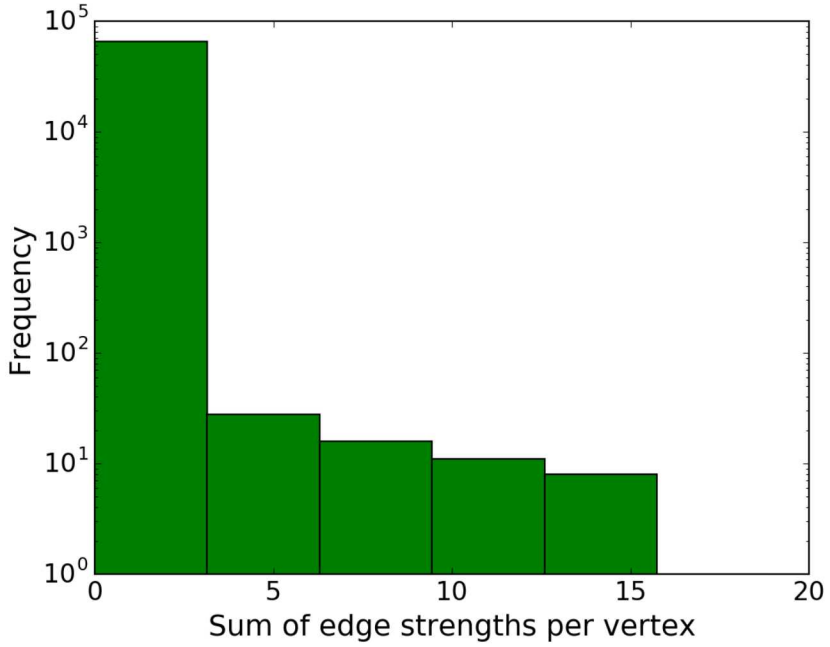
- 133K nodes, 198 reciprocated edges, D_G=11

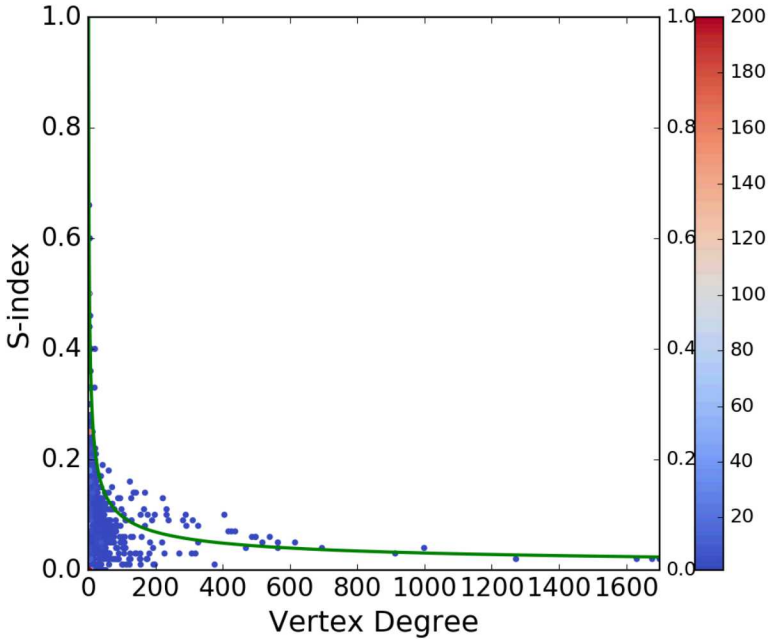




Caida (web)

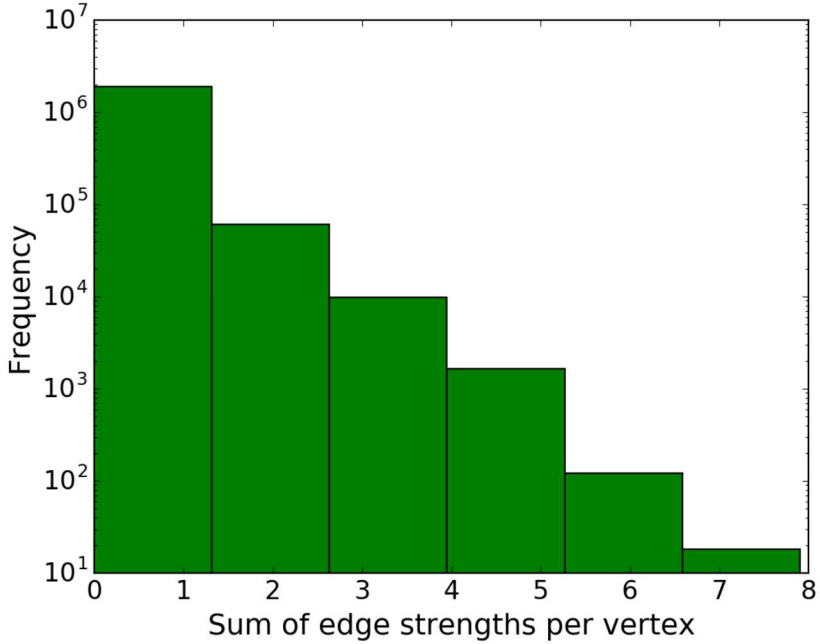
- 26K nodes, 53K reciprocated edges, DG=0.9

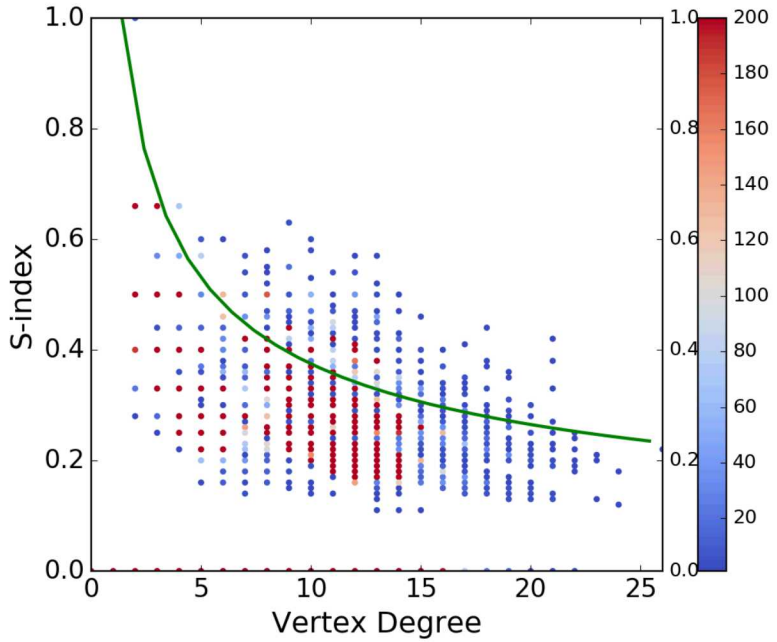




CA-RoadNet

- 2M nodes, 5.5M reciprocated edges, D_G=1.3



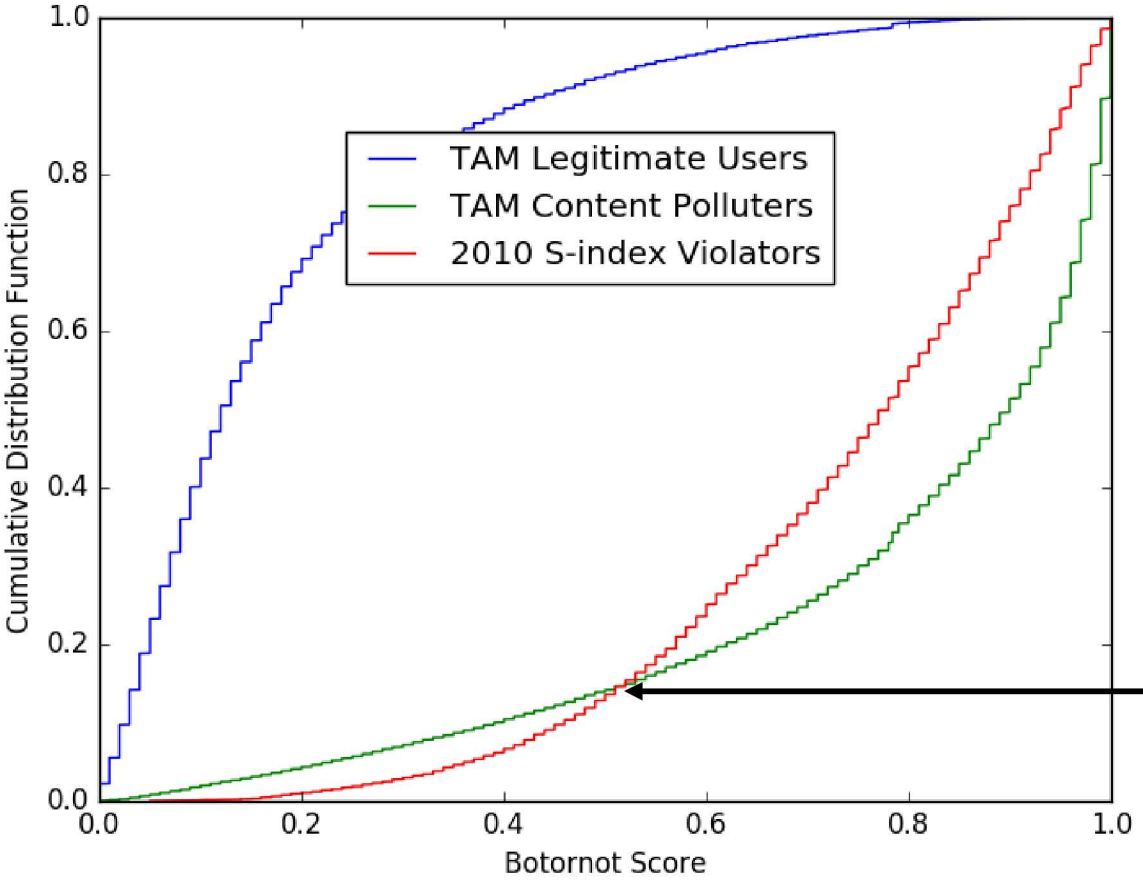


Heuristic Support 1

- Twitter has an API to look up users
  - O(10) account look ups/min
  - O(1) follower list look up/min
- First test "Bot-ness"
- Compare to
  - Texas A&M hand-labeled set of Twitter nodes (30K)
    - Human inspection
  - BotOrNot Scores (https://truthy.indiana.edu/botornot/)
    - BotOrNot uses account features
  - Our s-index violators came only from topology

Question: do our strength-index violators and the Texas A&M (TAM) Ground truth nodes have similar Botornot score distributions?

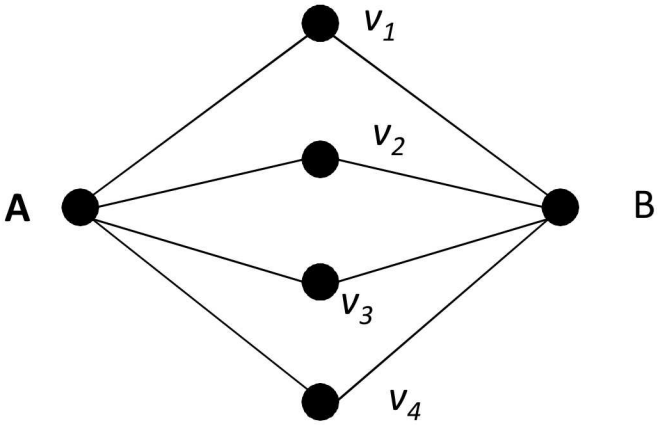
Bot-ness Results



~90% have BotOrNot Scores > 0.5

Support 2: Order of Following

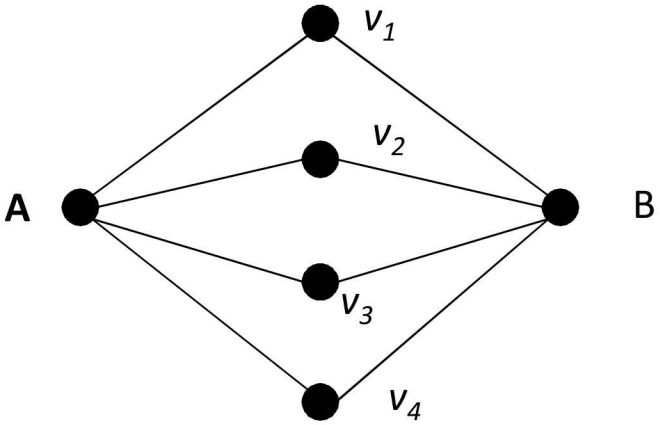
- An automated system might add followers in a given order
  - Adding whole botnet
  - Adding a new paying customer (add them to end of list)
- Consider order of adding shared neighbors



A's order v_1, v_2, v_3, v_4
B's order v_2, v_3, v_1, v_4

Support 2: Order of Following

- Consider order of adding shared neighbors
- Longest common subsequence of 2 length-n permutations
  - If added intentionally in order (automated) expect $\Theta(n)$
  - Random is $2\sqrt{n}$
    - Expect human to be more random



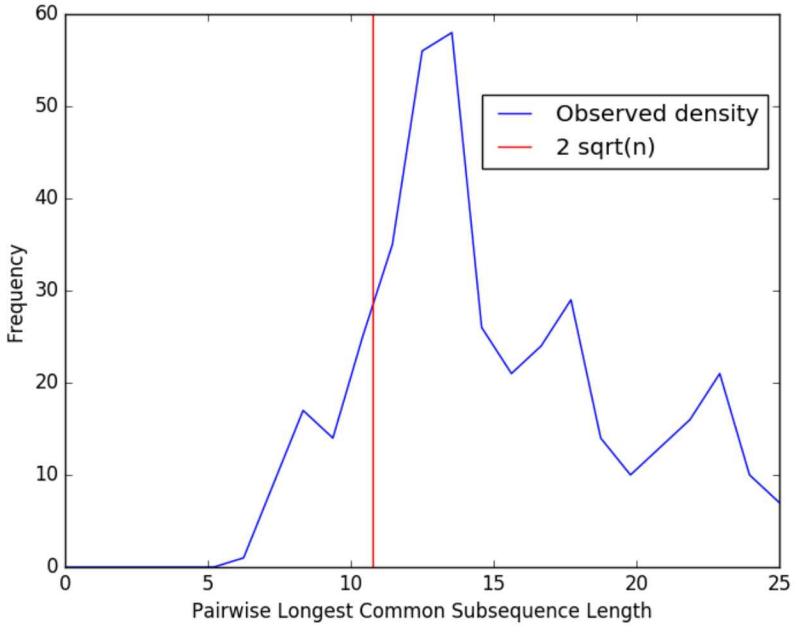
A's order v_1, v_2, v_3, v_4

B's order v_2, v_3, v_1, v_4

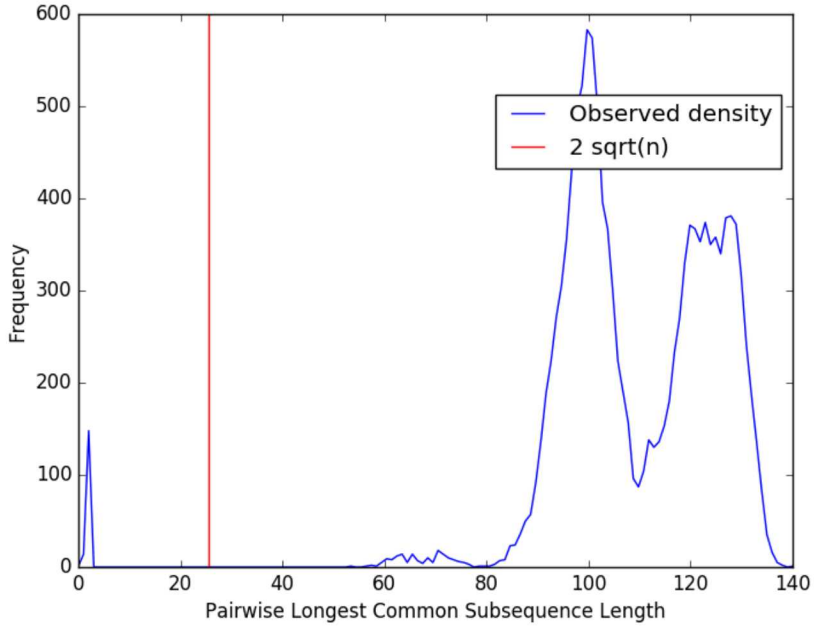
LCS is v_2, v_3, v_4

Order-of-Following Results

- Violators: largest clique 318 (in 2010), now 164
- Largest clique in non-violators was a small weather bot network
- Second-largest clique in non-violators 53 (in 2010), now 29



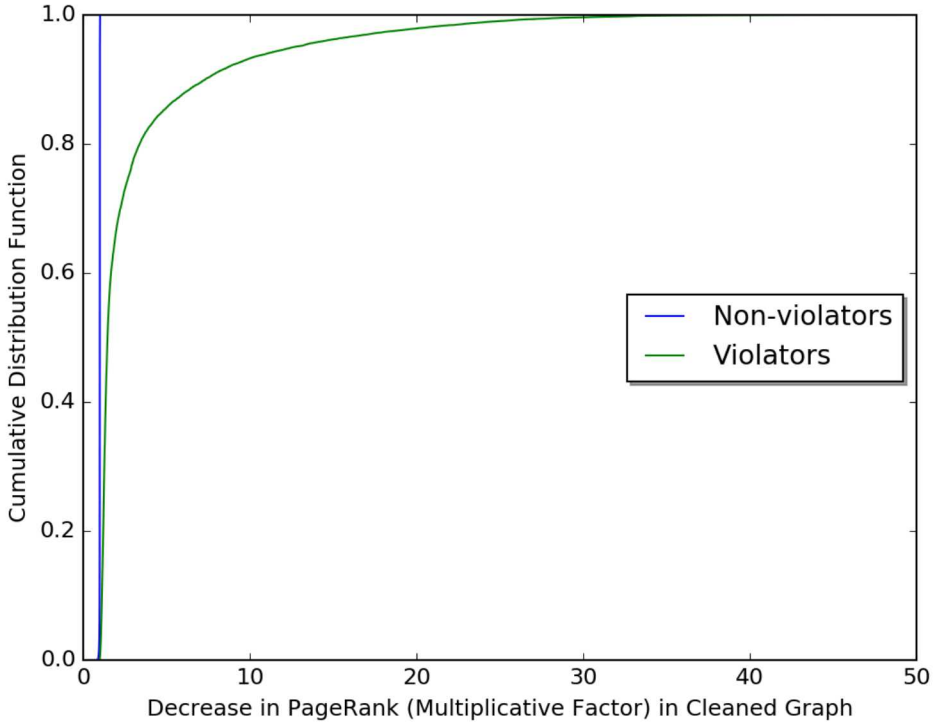
Non-violators



Violators

Consequences: PageRank

- People with access to real content on more social networks will need to further validate
- Does the cleaning matter?
- Yes for an algorithm like PageRank
  - 45% of violators have 2x decrease in cleaned graph
  - 16% decrease 5x

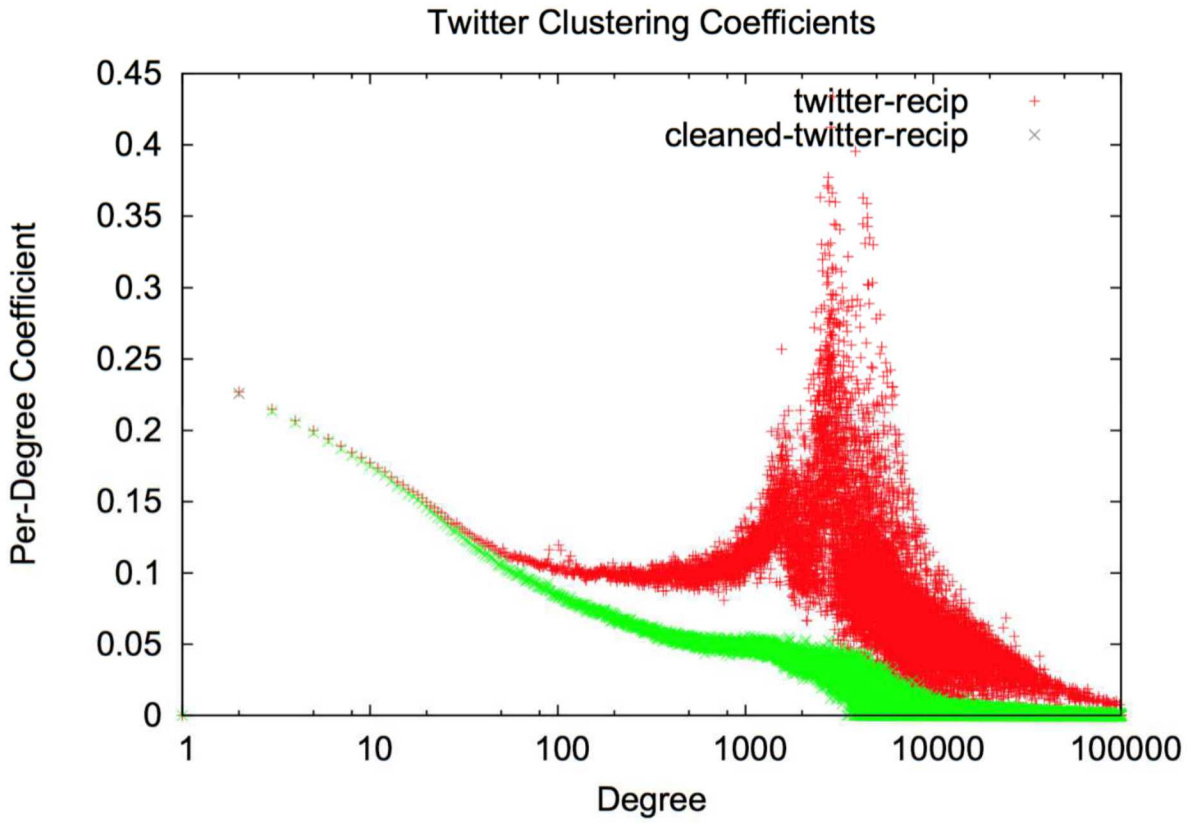


Clustering Coefficients

Fraction of wedges that close to a triangle

Consequences: Clustering Coefficients

- Clustering coefficients are a structural property
- The graph generator BTER uses only degree distribution and per-degree clustering coefficients



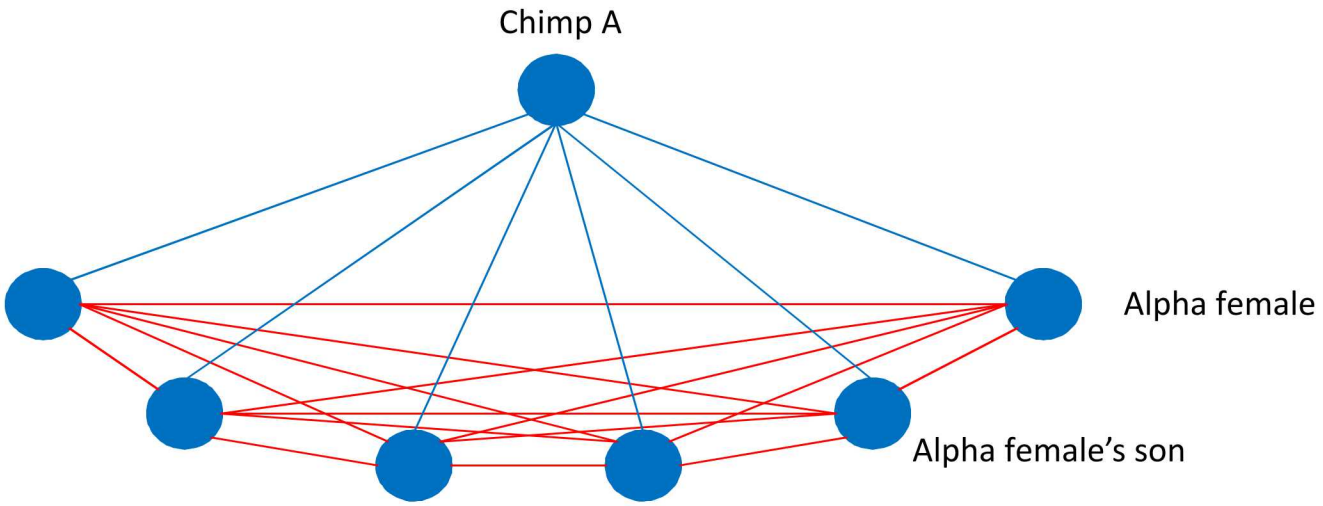
A New Concept

- We have assumed that each network has it's own "Dunbar constant" D

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

- Is there some hope for something more universal?

Dunbar's Number in a Nutshell



- Chimp A needs to worry about all of the inter-neighbor (red) edges in its ego network
- With *d* neighbors, there are $\binom{d}{2}$ such edges. We call this A's "cognitive load"
- Extrapolating to human brain size, if *d* > *150*, the load would be "too much"

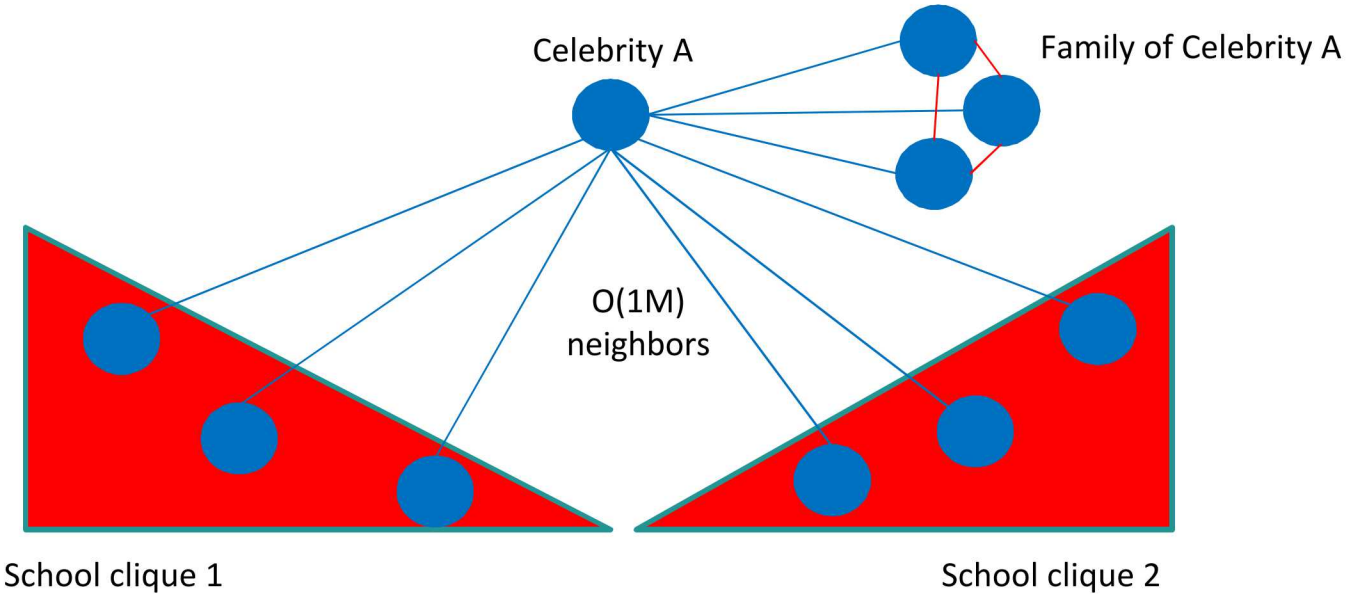
Support for Dunbar's Number

From Dunbar

- 150 estimated size of a Neolithic farming village
- 200 estimated upper bound on the number of academics in a discipline's sub-specialization
- 150 basic unit size of professional armies in Roman antiquity and modern armies
- Hutterites (Morovia) split communities at 150, since above that require police (https://www.theguardian.com/science/2011/apr/25/few-people-dunbars-number)
- W.L Gore and Associates (Gore-Tex) split employee units at 150 by trial and error, since more led to social problems [Malcolm Gladwell, The Tipping Point]

There are, of course, disagreements: social insects, size to survive without agriculture, larger modern estimated ties.

Dunbar's Number Applied to Human Celebrities



- Using Dunbar's model, the celebrity's aggregate cognitive load far exceeds Dunbar's limit
- The celebrity's family probably is a legitimate community
- The school cliques don't really impose a cognitive load on the celebrity, ***who doesn't care***
- But the intra-clique edges ***are indeed important*** to the students in the clique

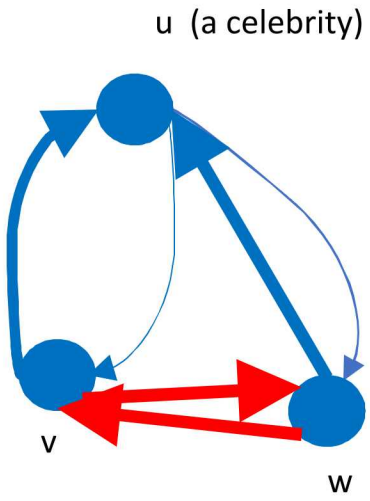
A Modeling Solution: "Directional Edge Weights

Shown:

- u doesn't care about relations with v or w
- v and w feel strongly about u
- v and w feel strongly about each other

Not shown:

- *How much does u have to care about the relationship between v and w?*
  - *That is: what contributes to u's "cognitive load"*

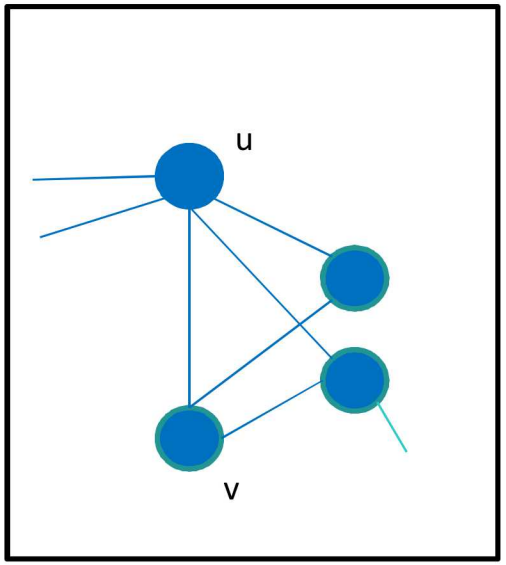


v and w: two school friends

A Formalization of Directional Edge Weights

d_u	the degree of u
Δ_u	the number of triangles incident on u
$\Delta_{(u,v)}$	the number of triangles on edge (u, v)
$w_u(u, v)$	the weight of edge (u, v) from vertex u's point of view

$$w_u(u,v) = \frac{1 + \Delta_{(u,v)}}{d_u}$$



(the proportion of edges incident to vertex u that are in triangles on edge (u,v))

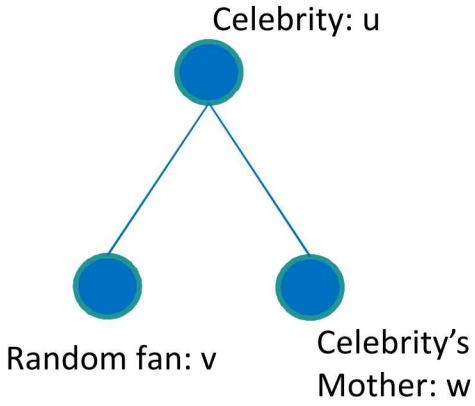
$$w_u(u,v) = \frac{1+2}{5} = 0.6$$
$$w_v(u,v) = \frac{1+2}{3} = 1.0$$

Modeling the "Cognitive Load" Imposed by a Relationship

Fundamental modeling assumption:

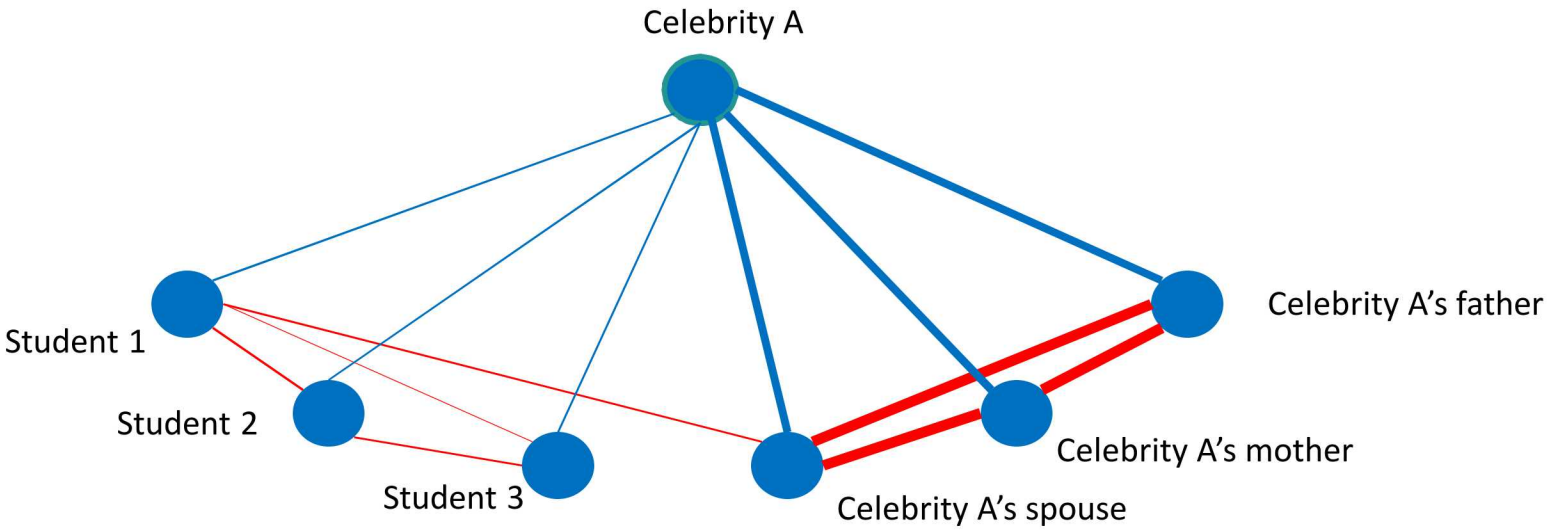
The "cognitive load" imposed by a relationship (v,w) upon a vertex u is the product of the directional edge weights (u,v) and (u,w).

- If either of these directional weights is low, the cognitive load is low.
- We express this as another edge weight:



$$w_u(v,w) = w_u(u,v)w_u(u,w)$$

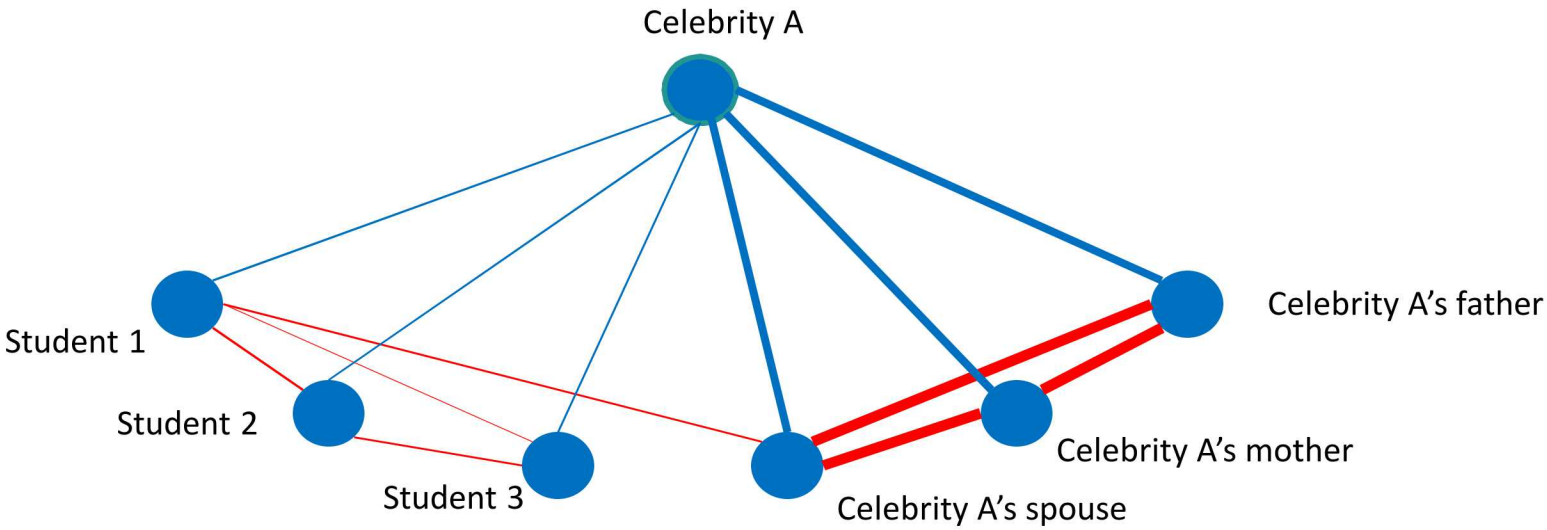
A Generalization of Dunbar's Limit



- Family relationships impose a cognitive load on Celebrity A
- Inter-fan relationships impose near-zero load
- We add up these fractional loads rather than Dunbar's original integer sum

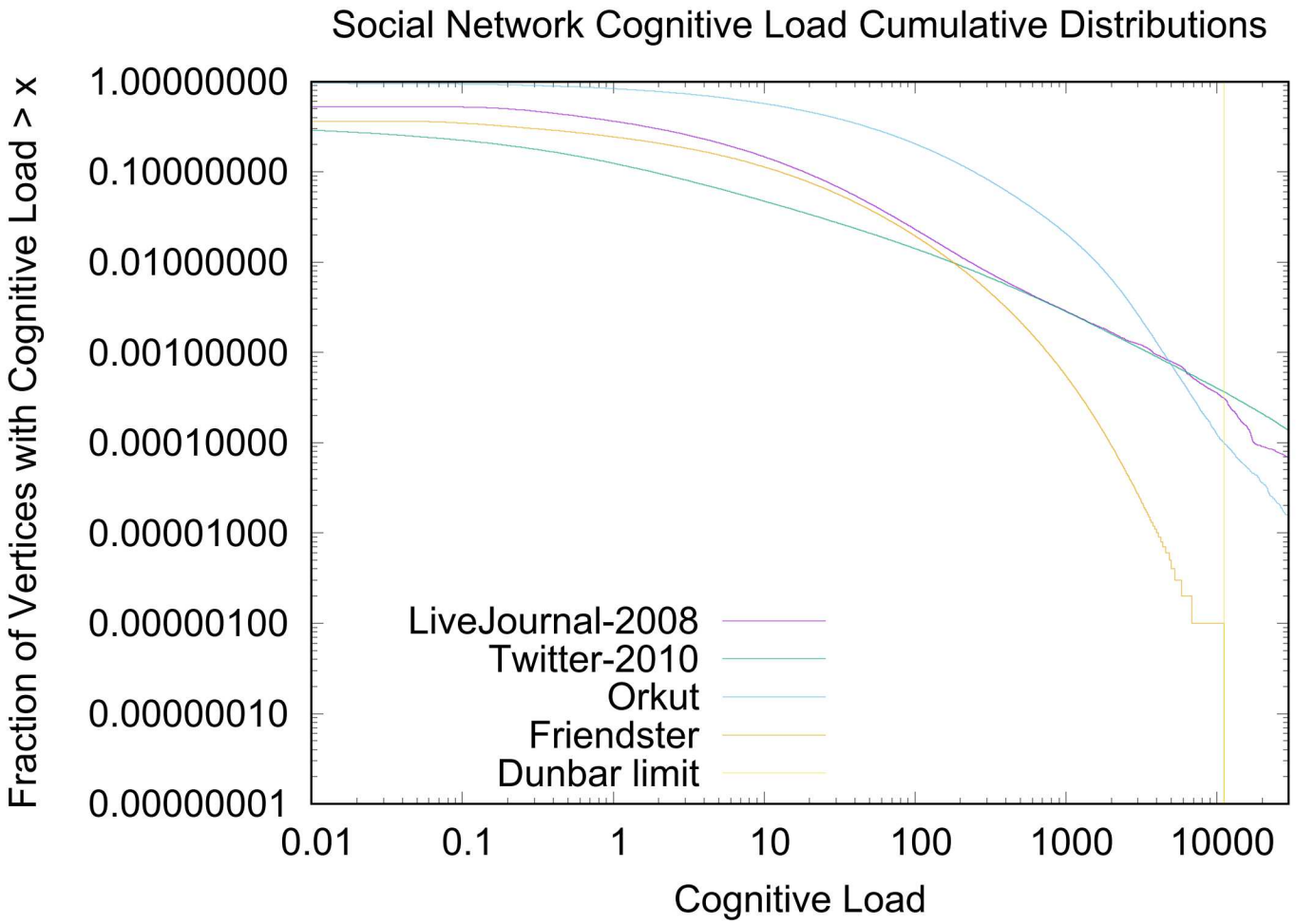
if $w_A(v, w) \approx 0.0$, it matters little to A whether edge (v, w) exists

Conjecture: Cognitive Loads are Still Bound by Dunbar's Number

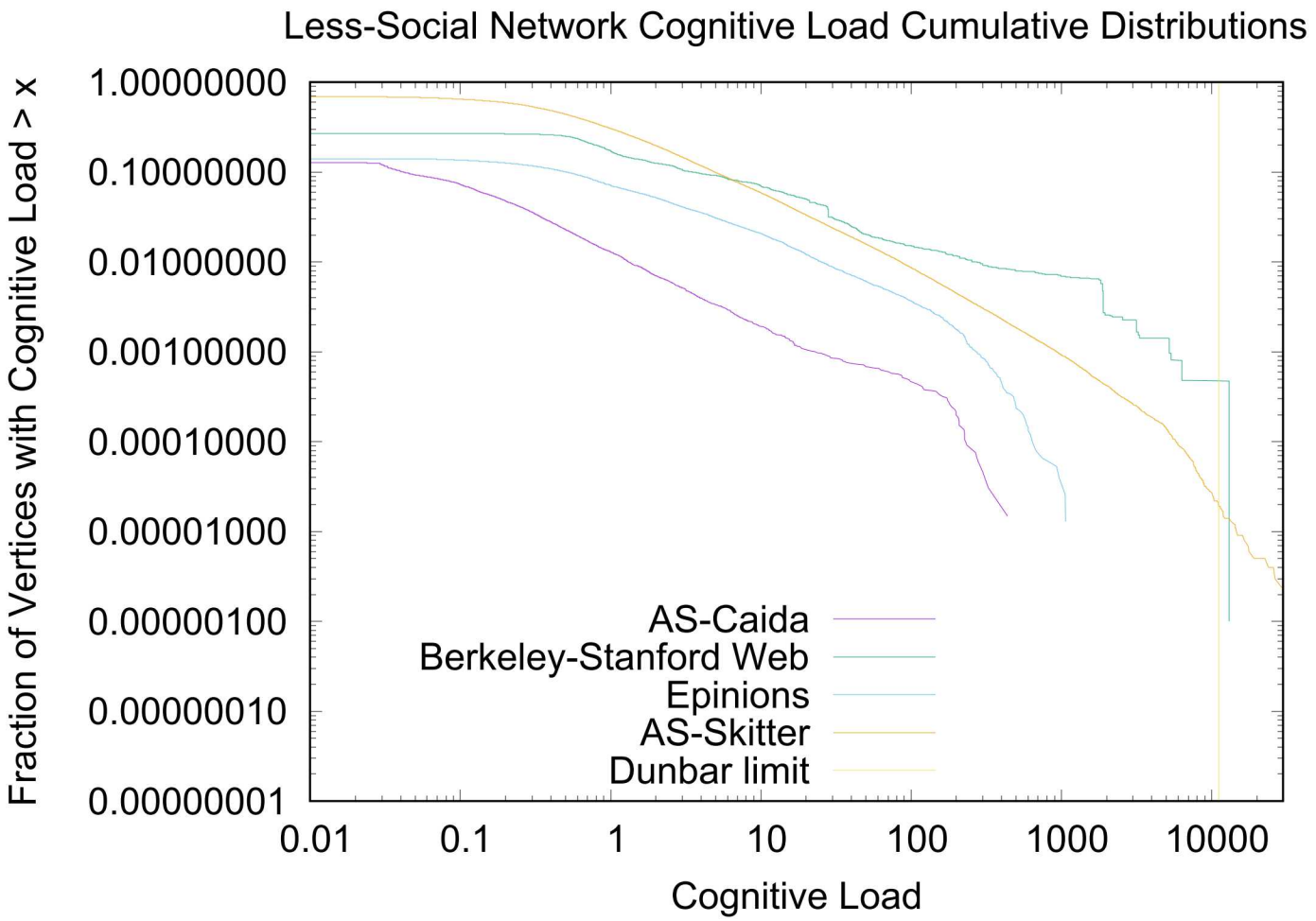


- It's the same human brain loaded down by these relationships
- We have given a formal way to down-weight unimportant relationships

Some results



Some Results



Summary

- A possible tool for cleaning some non-human behavior from some social networks.
  - conservative
- Social network structure enables more efficient algorithms in theory and practice, but requires human-only networks.
- This seems to be different from bot detection methods
  - Bad edges on non-bot nodes
- We won't be able to validate the other networks
- Theory implications are wide open
- We will release data sets and code with a publication/arXiv.