Multivariate Heavy Tails and Large Networks

D. Towsley

Collaborators within MURI: R. Davis, W. Gong,
S. Resnick, G. Samorodnitsky, Z. Zhang

Networks with MVHT distributions

- MVHT distributions ubiquitous in networks
  - in-degree, out-degree, reciprocated degree, labels, aggregate weights, …

Q: How to model, generate, estimate, classify, learn network structures?

Q: What effect does MVHT distributions have on answers to above questions?

Outline

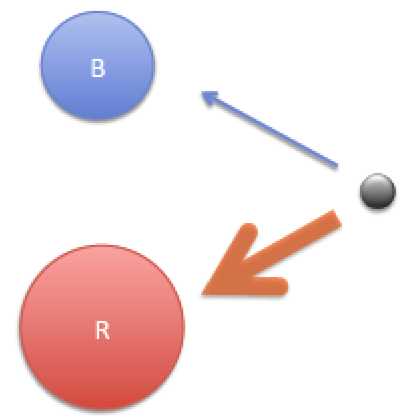
- competition in growing networks
- statistical inference in large networks
- reciprocity in large networks

Network growth

Cumulative advantage (CA)

- "rich gets richer"
- wealth (edge) attaches to nodes in proportion to function f of their wealth (degrees)

  (wealth accumulates in prop
- linear cumulative advantage (LCA) generates power laws



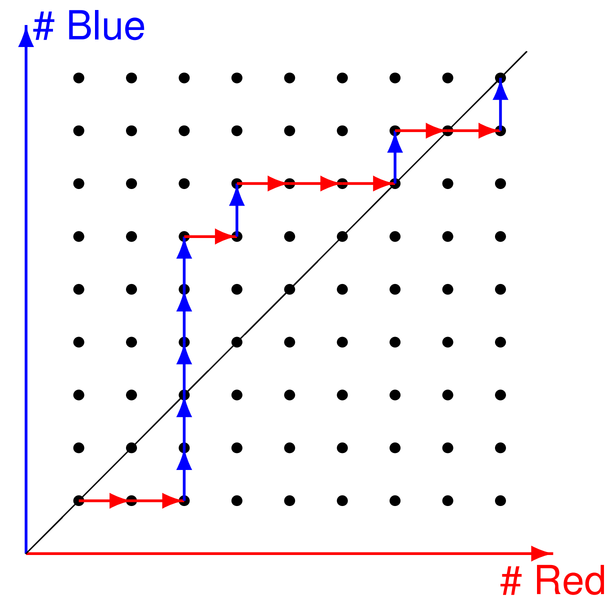
1. developed efficient algorithms to generate networks with $10^6 - 10^7$ nodes
   - studied network structure for different CA functions
2. studied competition under LCA

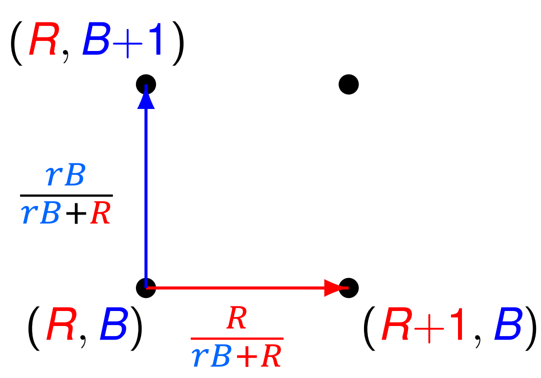
Competition under cumulative advantage (to appear J. Stat Mech.)

- duration of competition?
  - time taken for winner to emerge
- intensity of competition?
  - total # changes in leadership
- impact of inherent fitness?
- effect of cumulative function f?

Model under LCA

- two competitors
- state (R,B) in 2D lattice
  - each time, R or B increase by 1
  - transition rule, relative fitness $r \geq 1$ for B
  - generalized Pólya's urn model



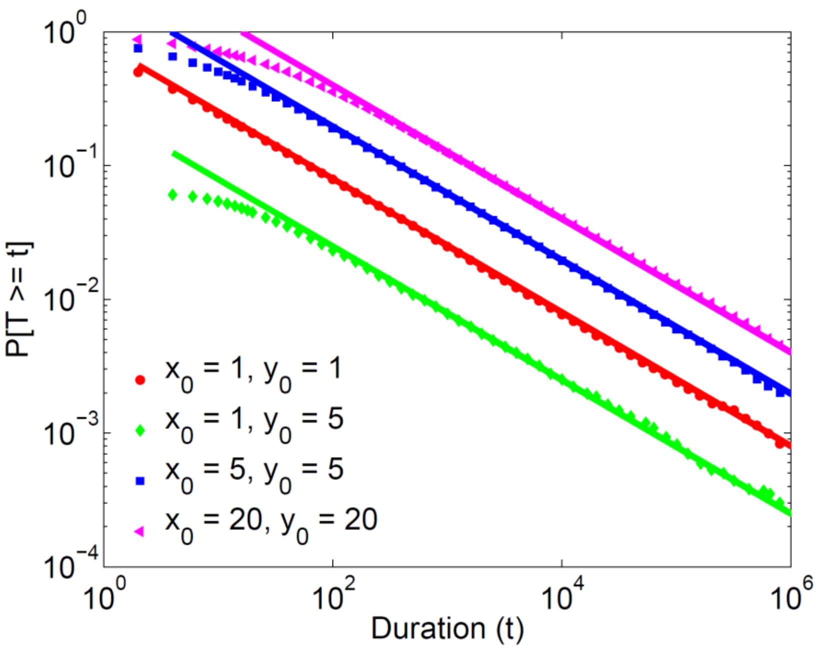


Equal fitness, $r = 1$

- derived joint PMF for duration, intensity
- duration, intensity both exhibit heavy tails

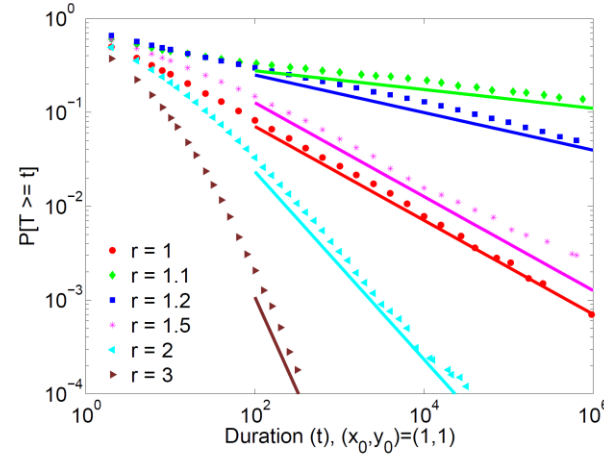
$$P(duration > t) \propto t^{-1/2}$$

$$P(intensity > n) \propto n^{-1}$$

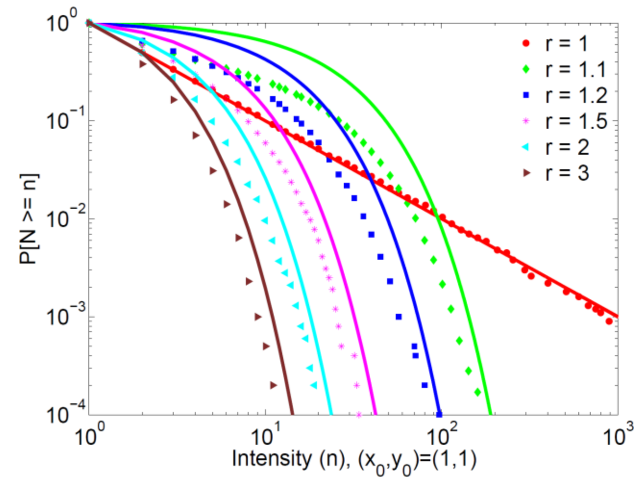


Different fitness, $r > 1$

- intensity exponentially tailed
- duration heavy tailed
- $P(duration > t) = \Omega\left(t^{-(r-1)b_0}\right)$
  - discontinuity at $r = 1$
  - ***tail much heavier*** at $r = 1 + \varepsilon$ than $r = 1$



Competition becomes less intense, but much longer



Nonlinear cumulative advantage

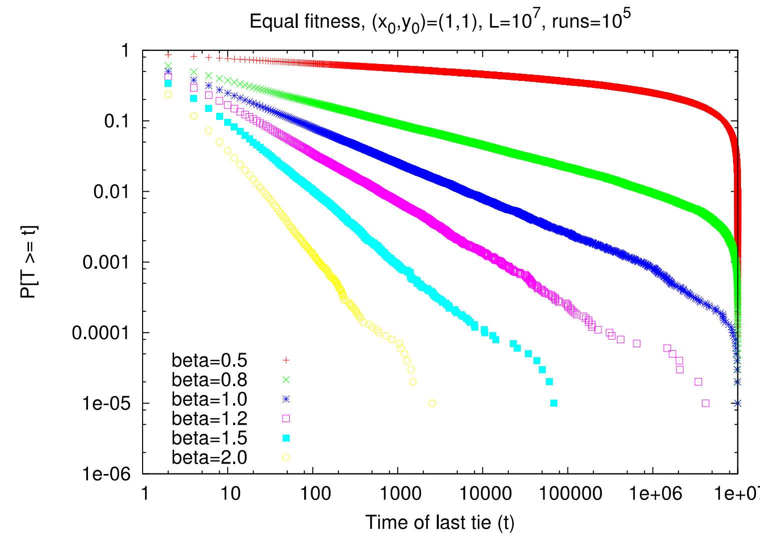
- model:

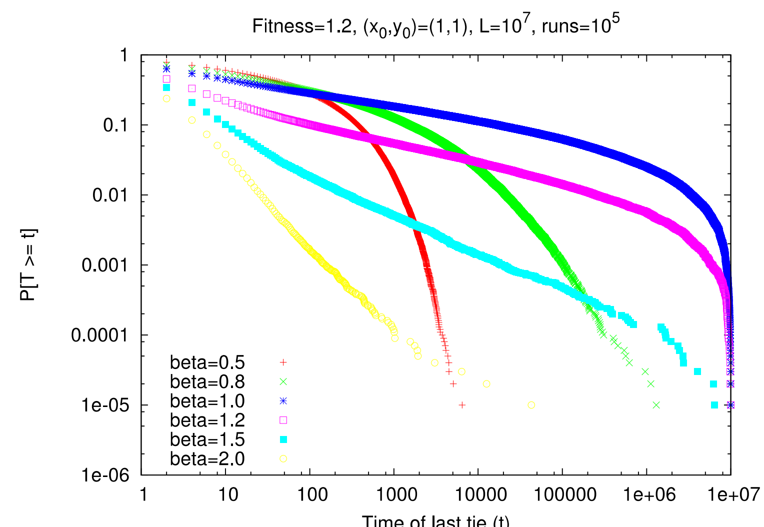
$$\frac{(rB)^\beta}{(rB)^\beta + R^\beta}, \quad \frac{R^\beta}{(rB)^\beta + R^\beta}$$

- equal fitness: power-law for $\beta > 1/2$

$$\lim_{t\to\infty} \frac{\log \mathbb{P}(\text{dur.} > t)}{\log t} = \left(\frac{1}{2} - \beta\right)^+$$

- conjecture: different fitness: power law for $\beta > 1$, light tailed for $1/2 < \beta < 1$





Summary

- (semi-)complete analysis of two party competition
- joint heavy tail distribution of duration/competition for equally fit parties
- surprising phase transition when one party becomes “slightly” more fit
  - intensity has lighter (exponential) tail
  - duration has heavier tail

Questions:

- 3+ parties
- parameter estimation
- non-linear CA rules

Inferring graph characteristics using random walks

Murai, Ribeiro, Towsley

Estimating joint degree distribution in directed graphs: random walks (RWs)

- networks extremely large, $10^6 - 10^7$ nodes
- sampling methods desirable/necessary
- random walk based methods standard for undirected networks

Q: adapt to directed graphs?

- transform digraph to undirected graph, degree distr.

$$\pi(l) = P(\text{degree} = l)$$

- collect samples using RW

$$s_1, s_2, \dots, s_n, \qquad\qquad s_k = (i_k, o_k)$$

- generate asymptotically unbiased estimates of
$$\pi_{i,j} = P(\text{indegree} = i, \text{outdegree} = j)$$

Pros/Cons RW-based sampling

Pros

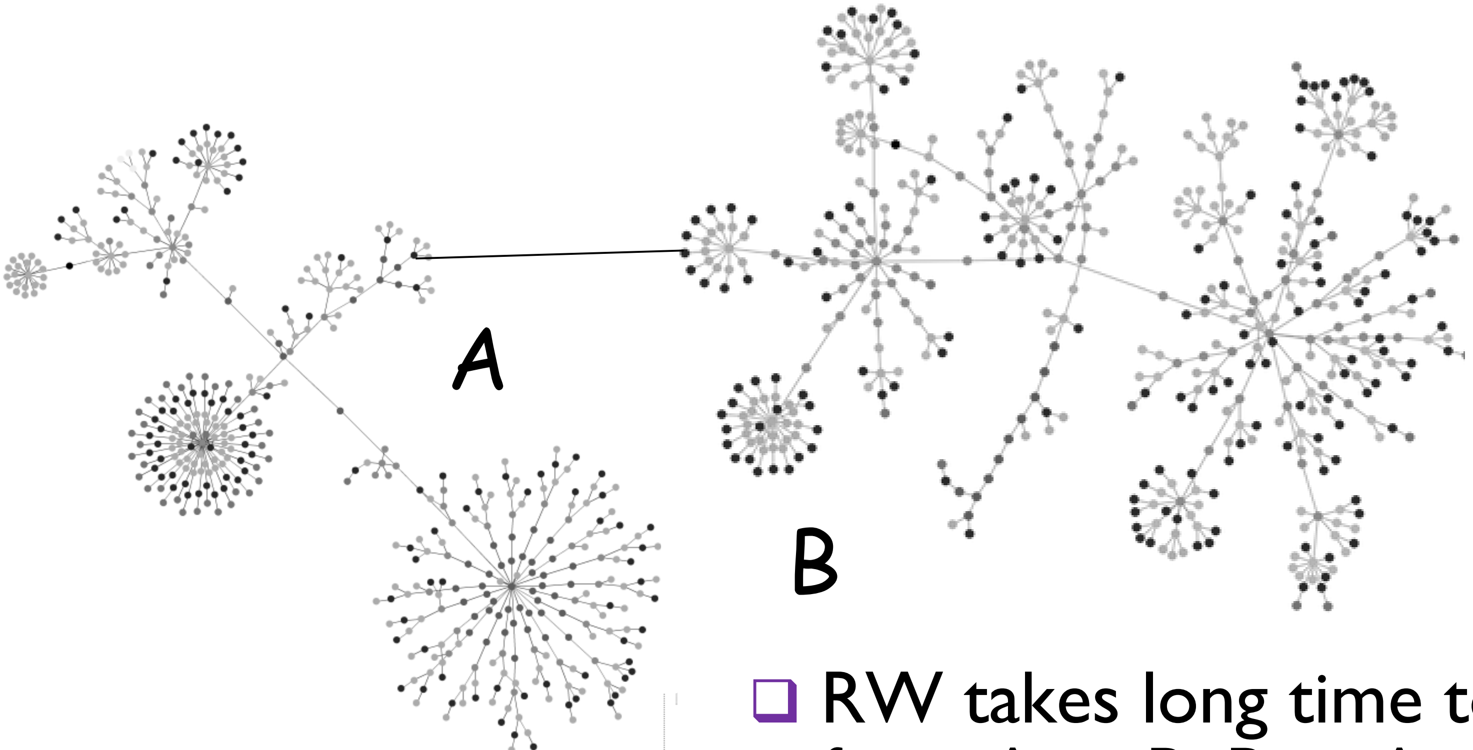
- samples in proportion to degree
  - good for heavy-tailed degree distributions (analytical, empirical)
- inexpensive
  - neighbors visible in many networks

(Twitter, Facebook, …)

Cons

- mixing times, dependence among samples
  - in contrast to uniform node sampling

Loosely connected components



Combine advantages of uniform vnode & RWs?

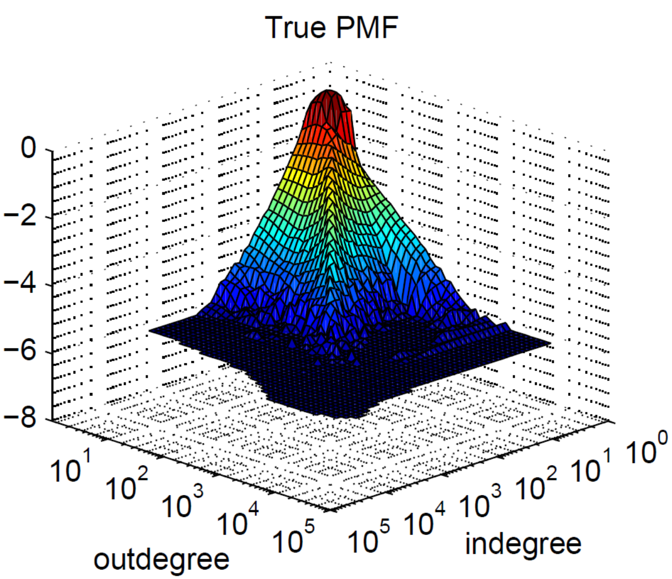
- RW takes long time to get from A to B; B to A
  - inexpensive
- uniform samples both A and B subgraphs
  - expensive

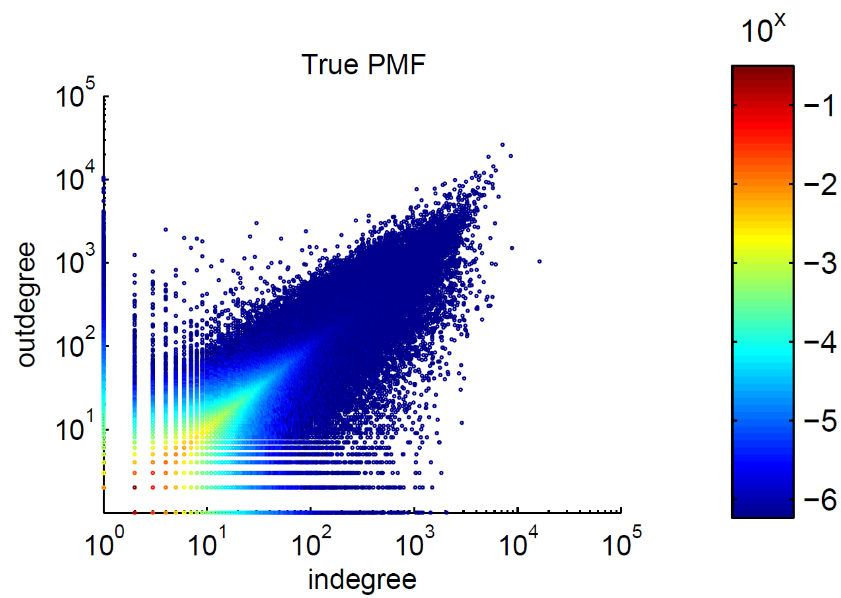
Frontier sampling

- multiple coupled RWs - Frontier sampling
  treat as virtual random walker
- mixing time decreases with number of walkers
- estimate combines initial uniform samples + RW samples
  - asymptotically unbiased

Flickr network: example

- 1.8M nodes
- 23M edges
- marginal heavy tails, strongly dependent





True CCDF

True CCDF

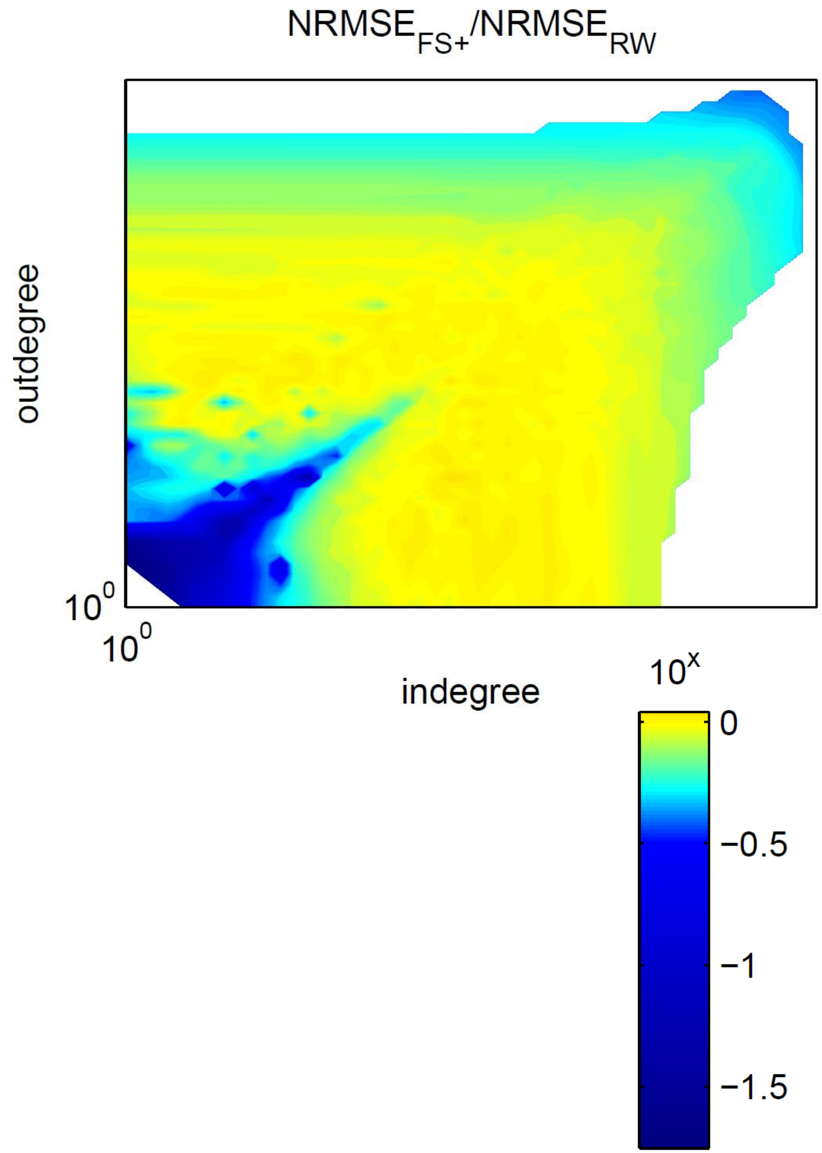
1(

Flickr, $B = 0.1$, FS vs RW

- FS, RW comparable for tail
- FS exhibits low error for head

Why?

- largest component 70% of network
- many small components
  $\rightarrow$ more low degree nodes



Summary

- frontier sampling superior to other RW-based sampling methods
- well suited to networks with heavy tailed degree distributions
- extend to other network inferencing problems
- promote to network/data scientists

Missing:

- better theoretical foundation

Reciprocity in directed networks

B. Jiang, D. Towsley (UMass),
Z. Zhang (U.Minn)

Presented at KDD 2015

Motivation

- reciprocity measures fraction of reciprocal edges

$$\text{reciprocity} = \frac{2}{3}$$

- important characteristic of directed networks
  - invites interpretation as network organizational principle,
  - e.g. reciprocal or anti-reciprocal

- nontrivial reciprocity observed in many real networks

	Google+	Swedish Wiki	Spanish Wiki
observed	34%	21%	15%
random	0	0	0
structural max	47%	28%	36%
ratio	73%	75%	42.5%

- how to interpret these numbers?
- most real social networks are reciprocal
- informative to compare with maximum reciprocity

Degree bi-sequence

- degree bi-sequence (d^+, d^-) of digraph
  - out-degree sequence: $d^+ = (d_1^+, \ldots, d_n^+)$
  - in-degree sequence: $d^- = (d_1^-, \ldots, d_n^-)$

- graphic bi-sequence: realizable by digraph

Maximum reciprocity problem

Given graphic bi-sequence (d^+, d^-)

maximize: reciprocity of G

subject to: G has degree bi-sequence (d^+, d^-)

- Max # reciprocal edges $\rho(d^+, d^-)$ upper bounded by

  $$\rho(d^+, d^-) \leq \beta(d^+, d^-) = \sum_i \min\{ d_1^+, d_1^+\}$$

- characterized for different random graph models exhibiting MVHT

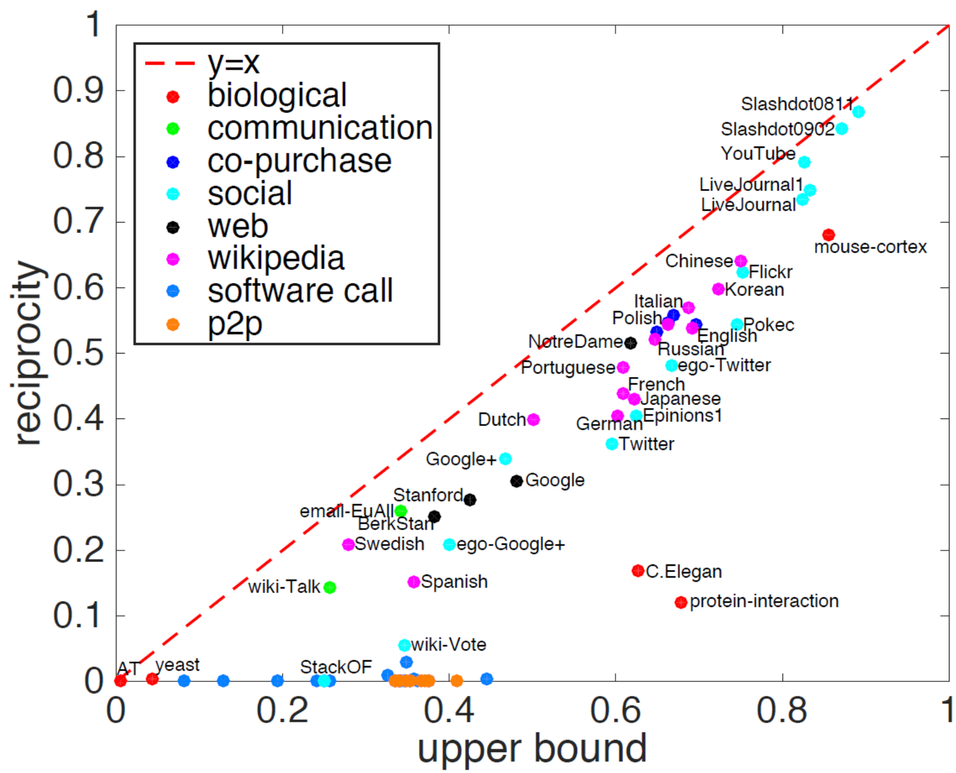
Empirical study

Datasets

- major directed social networks
- directed networks of other categories

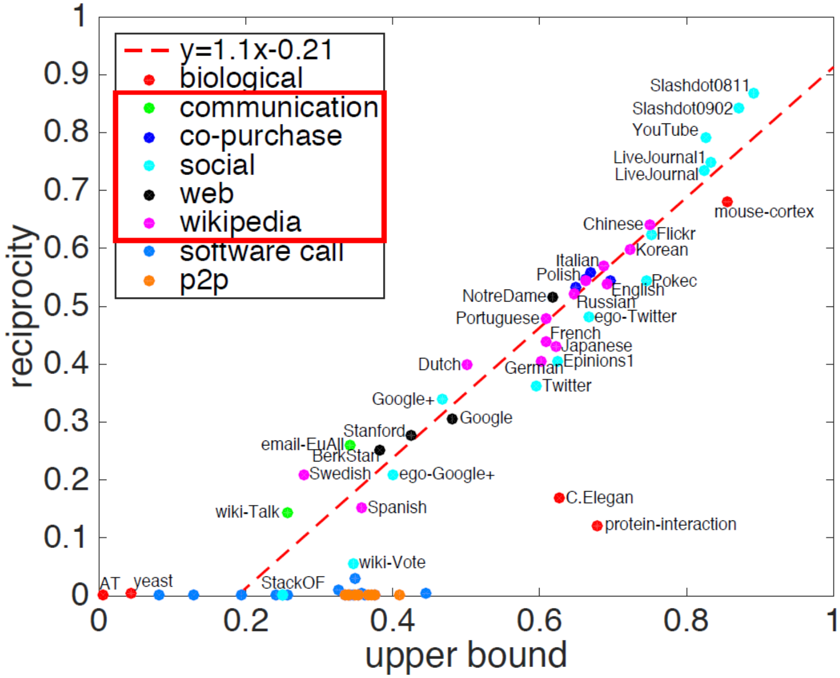
Reciprocity varies widely

- P2P: 0
- Slashdot: 90%
- high for social & Wiki
- low for P2P & software call

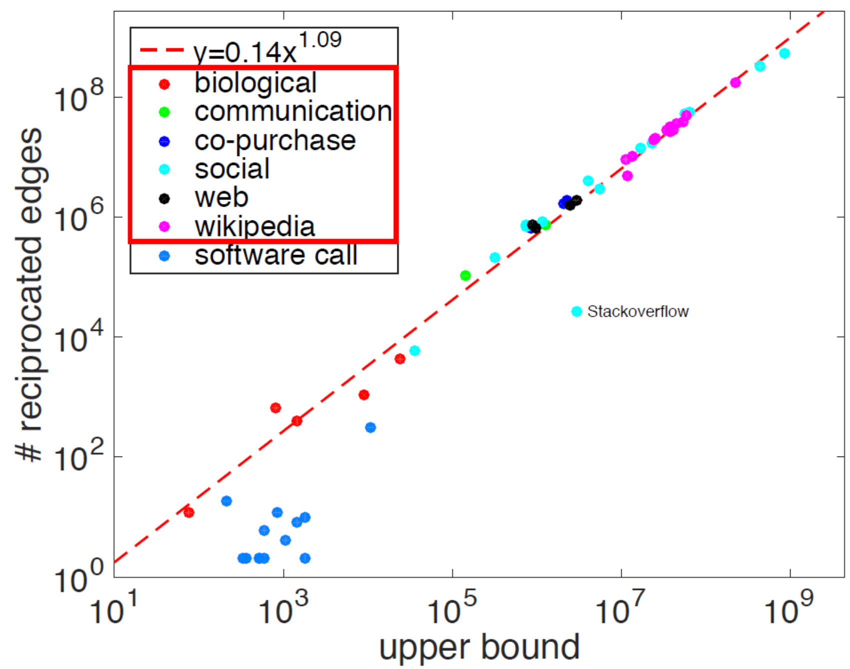


Strong linear relationship

Reciprocity

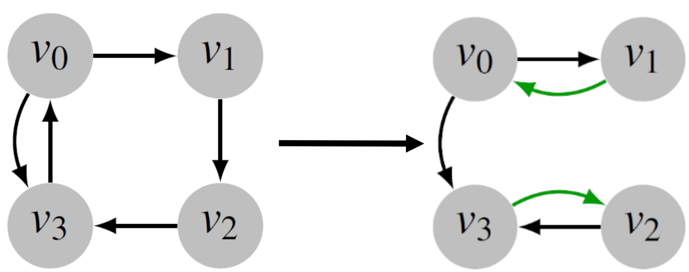


reciprocated edges

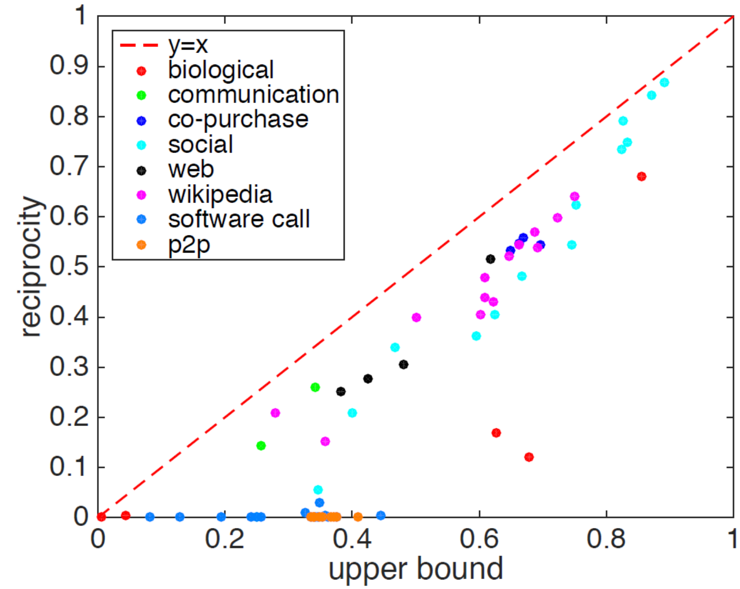


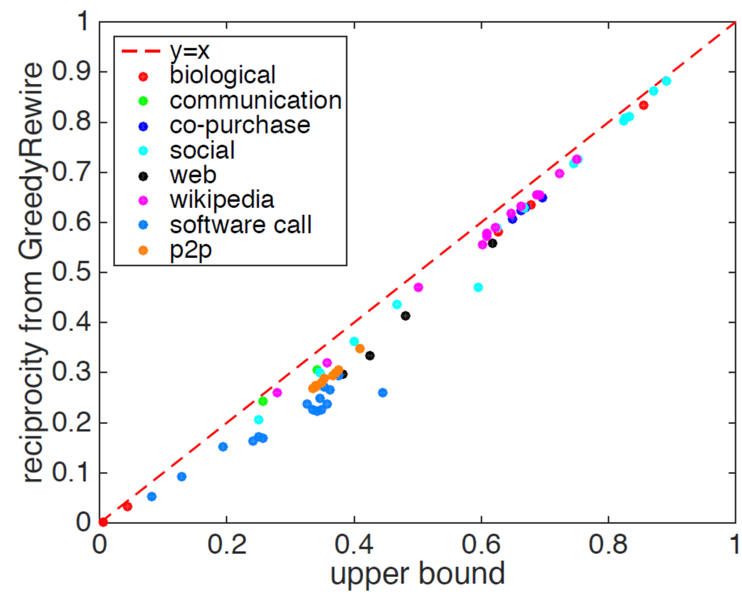
Tighter upper bound

- identified 4 node suboptimal motifs
- developed rules to increase reciprocity



- suboptimal 3-paths major source of loss in reciprocity





Future plans

- network exploration as multi-armed bandit problem
  - potential terrorists
  - donors to political parties
  - rewards exhibit MVHT behavior – how to exploit?
- principled network characterization
  - clustering, leveraging observed empirical MVHT behavior