A New Sampling Approach for Estimating Features of Large Networks

Gennady Samorodnitsky[1], Richard A. Davis[2]
Zhi-Li Zhang[3] and Jingjing Zou[2]

1. School of Operations Research and Information Engineering, Cornell University;
2. Department of Statistics, Columbia University;
3. Department of Computer Science and Engineering, University of Minnesota

Goals

- Estimate extreme characteristics in a very large network
  - The emphasis is on nodes, not edges.
  - Extremes of both in-degree and out-degree of interest.
  - Joint extremes of in-degree and out-degree also of interest.
- Challenges:
  - Hard to find and sample rare nodes in a very large network.
  - How do obtain approximately unbiased estimators?

The paradigm

- Assume power-like tails of in-degree and out-degree.
- Use multivariate extreme value theory.
- Use extreme value estimators.
  - The estimators benefit from larger sample sizes.
  - Devise sampling to achieve that.

The network

- Webgraph (network of webpages) from the Google programming contest (2002)
- A directed network
- 875,713 nodes (web pages), 5,105,039 edges (links)

Existing Sampling Approaches

- Single random walk: expensive, only represents a cluster.

Existing Sampling Approaches

- Single random walk: expensive, only represents a cluster.
- Multiple random walks: explore multiple clusters, but not specifically extreme nodes.

Existing Sampling Approaches

- Single random walk: expensive, only represents a cluster.
- Multiple random walks: explore multiple clusters, but not specifically extreme nodes.
- Frontier Sampling (Ribeiro and Towsley, 2010)
  - Start with multiple initial nodes
  - Each time pursue the most promising lead.

Existing Sampling Approaches

- Single random walk: expensive, only represents a cluster.
- Multiple random walks: explore multiple clusters, but not specifically extreme nodes.
- Frontier Sampling (Ribeiro and Towsley, 2010)
  - Start with multiple initial nodes
  - Each time pursue the most promising lead.
- In all cases: an attempt of uniform sampling of edge.
  - Requires adjustement to weight nodes equally.

Our Strategy

- We concentrate on “promising” nodes.
- Initial choice of nodes is random, as in other approaches.
- Check the neighbours of the chosen nodes.
- Build paths by discarding “non-promising nodes”.
- Adaptively decide on the depth of the search.

Tasks

- Nodes need to be weighted by the likelihood of being seen.
- That likelihood needs to be estimated.
- Not straightforward, since only outgoing edges are easily seen.
- Grouping nodes into equivalence classes helps.

Promising nodes

- "In": in-degree larger than the 95% quantile
- "Out": out-degree larger than the 95% quantile
- "Both": both in- and out-degrees larger than 95% quantiles
- "None": Neither

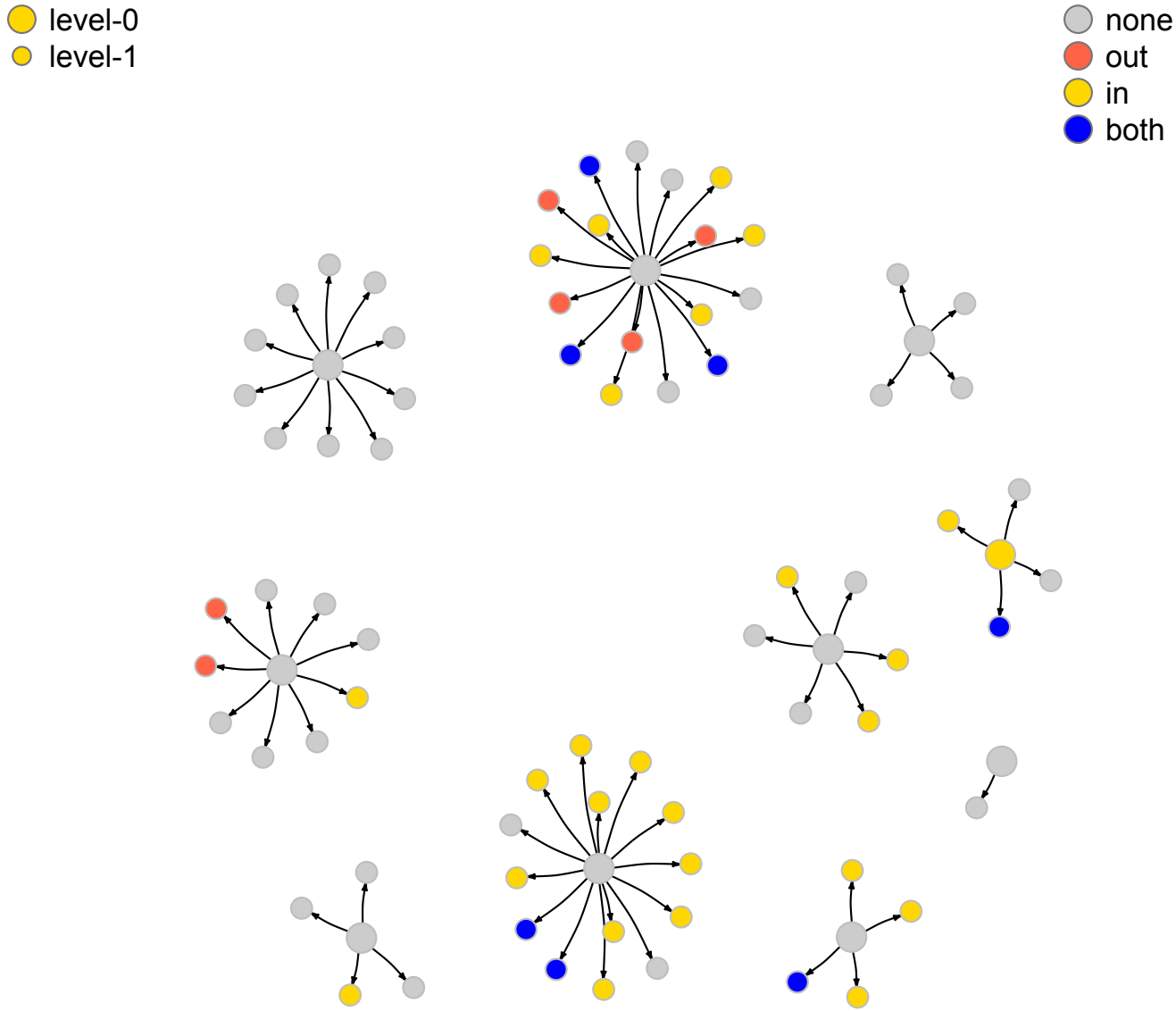
A small test

- Select at random 10 initial nodes.
- Distribution of node types in the initial selection

both	in	none	out
0.0	0.1	0.9	0.0

- The distribution of types in their neighbors?

Observed types

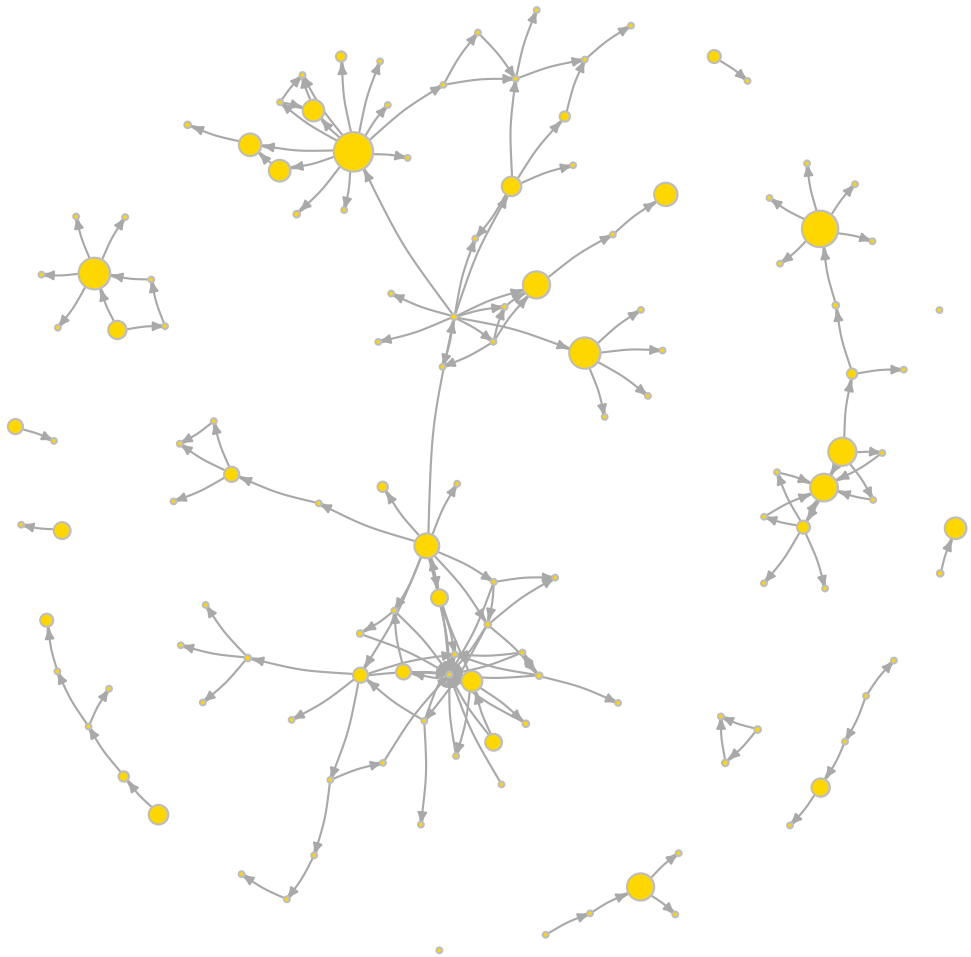


Observed types

Distribution of types among neighbors of the initial nodes:

both	in	none	out
0.0959	0.3425	0.4795	0.0822

The resulting sample of nodes



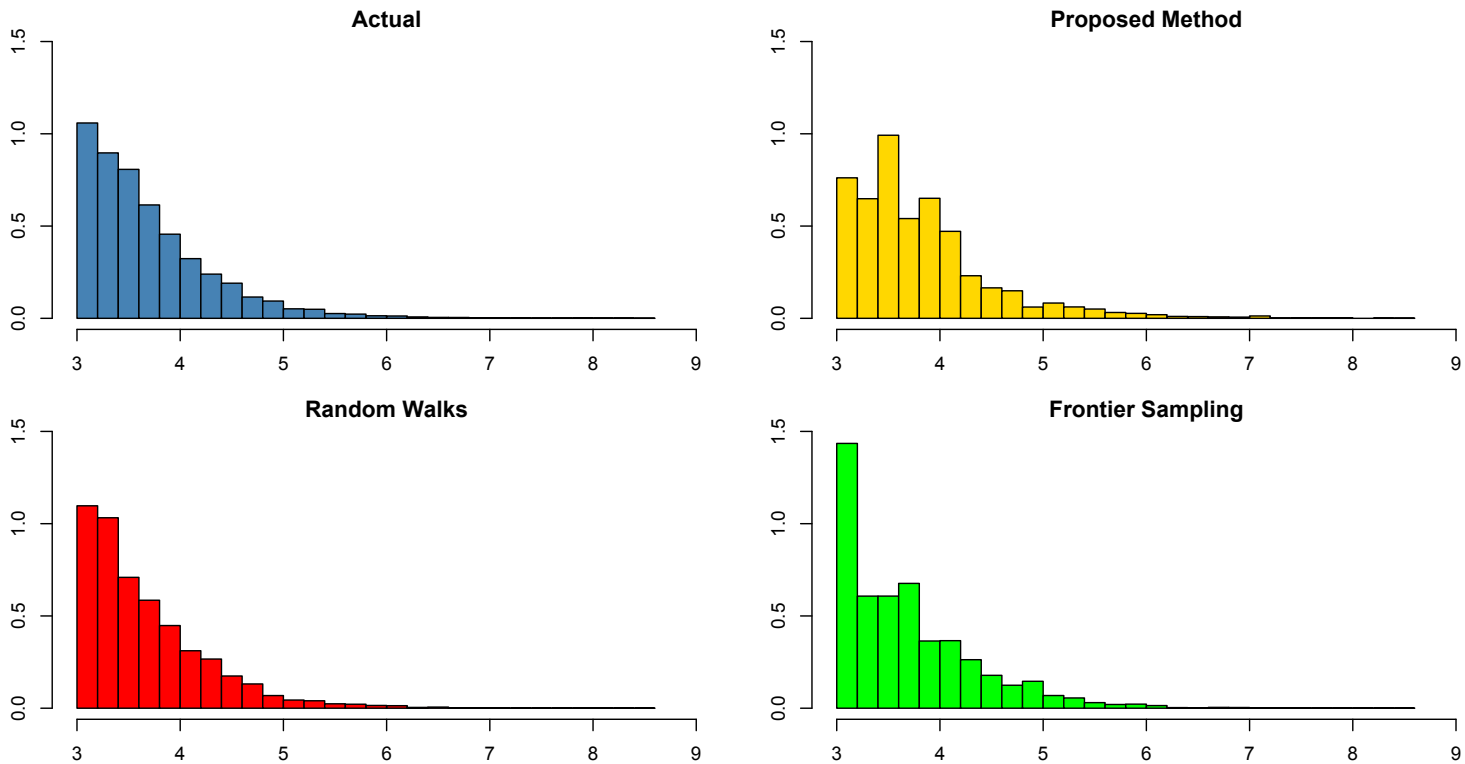
Weight Adjustment

- Achieve approximate lack of bias by using weighted averages of sampled nodes
- Desired weight: $w_i = 1/P(n_i \in S)$
- $P(n_i \in S) \propto$ no. of nodes "leading" to n_i
- The latter cannot be completely observed
- Use observed values and linear regression

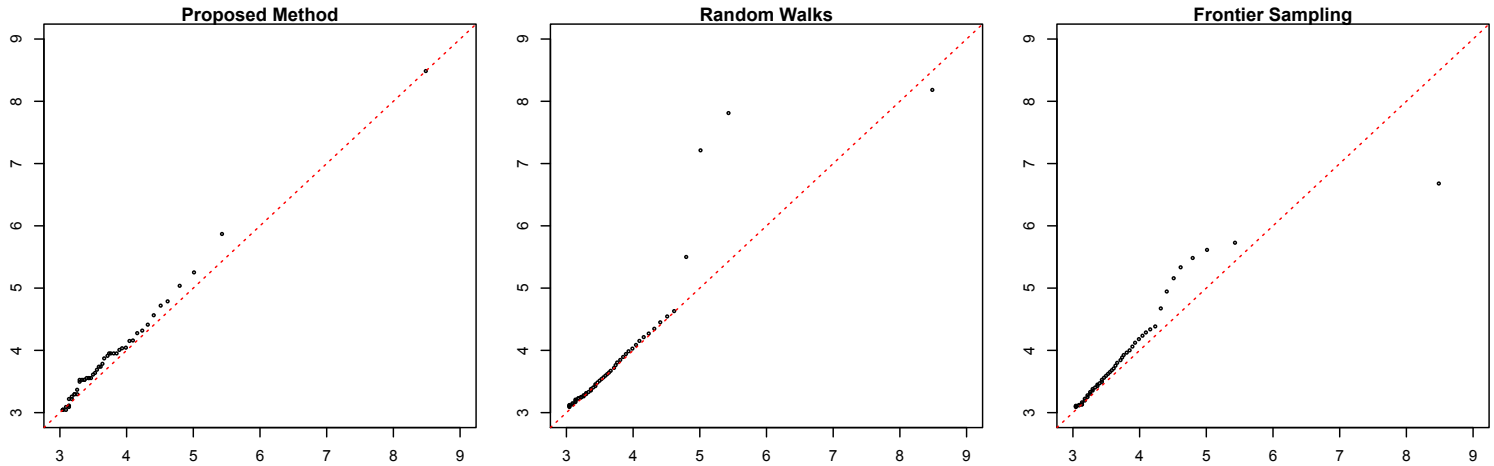
Estimation Results: Distribution of In-degree

- We are interested in the top 5% of the nodes
- Start from 20 nodes in our method
- For benchmarking: 200 initial nodes for Multiple Random Walks (RW) and Frontier Sampling (FS)

In-degree histograms, top 5% (Log-scale)



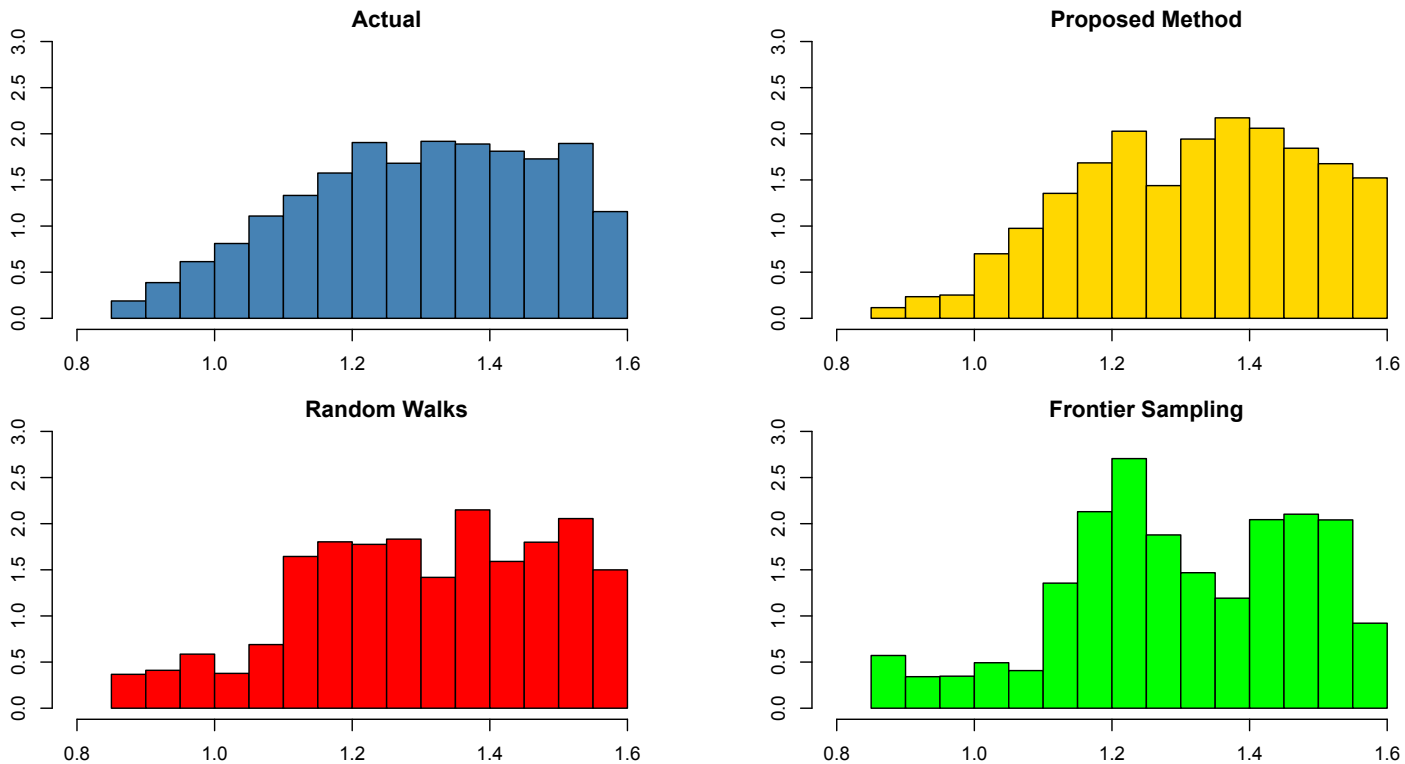
Q-Q Plots of In-degree, top 5%



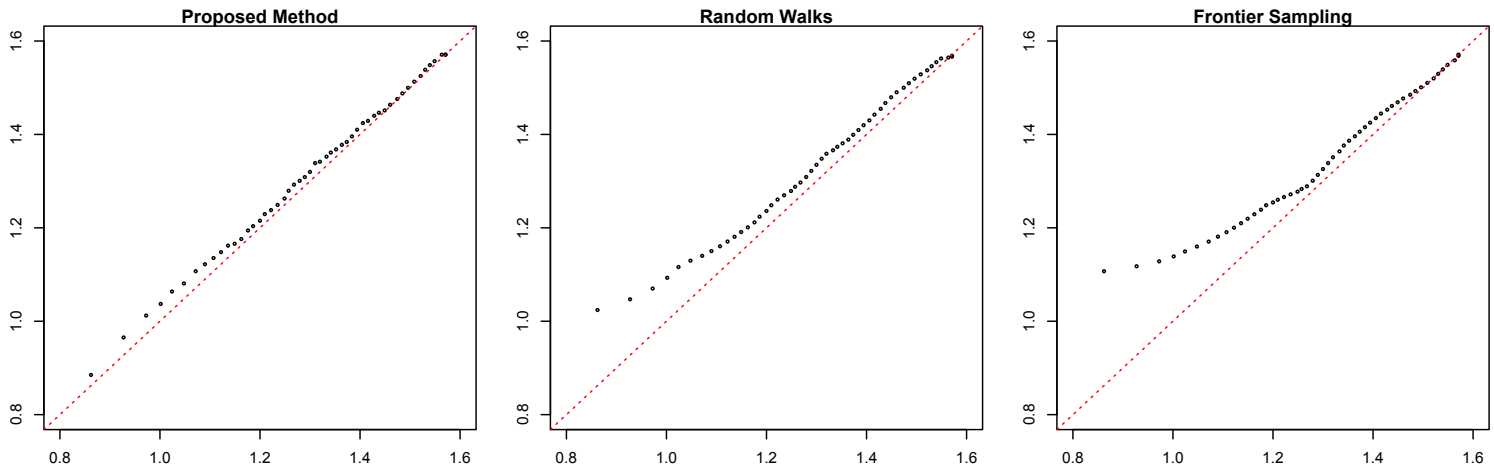
Estimation Results: Extrems of the Joint Distribution of In- and Out-Degrees

- Measured by the angle: $\arctan(\text{In}_k/\text{Out}_k)$
- Start from 200 initial nodes for all methods for benchmarking.

Histograms of the Angle



Q-Q Plots of the Angle



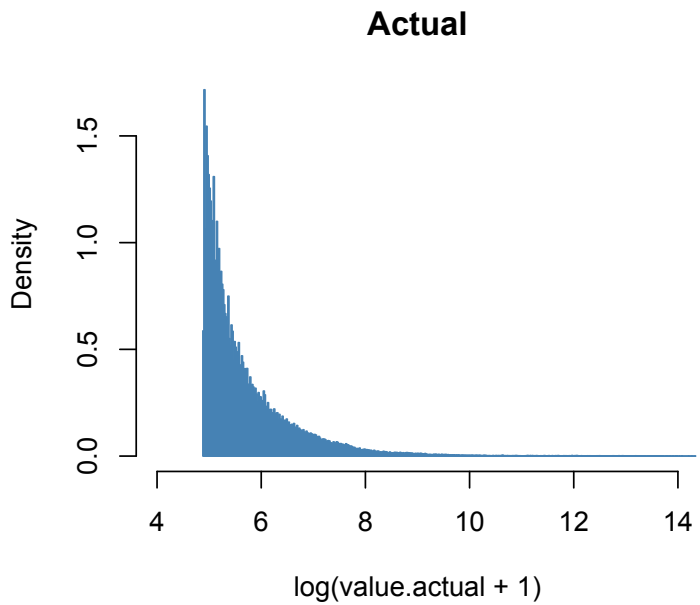
Google+ Data

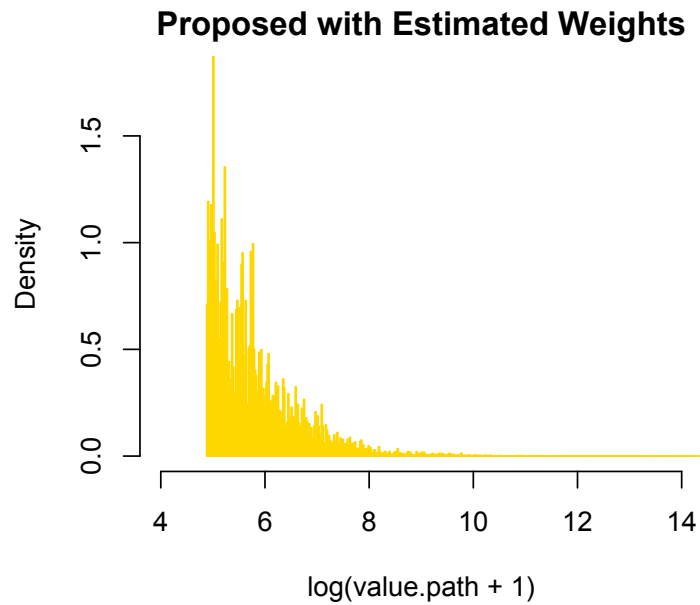
- A snapshot of the social network taken on Oct, 2012
- 76,438,791 nodes
- 1,442,504,499 edges

Our procedure

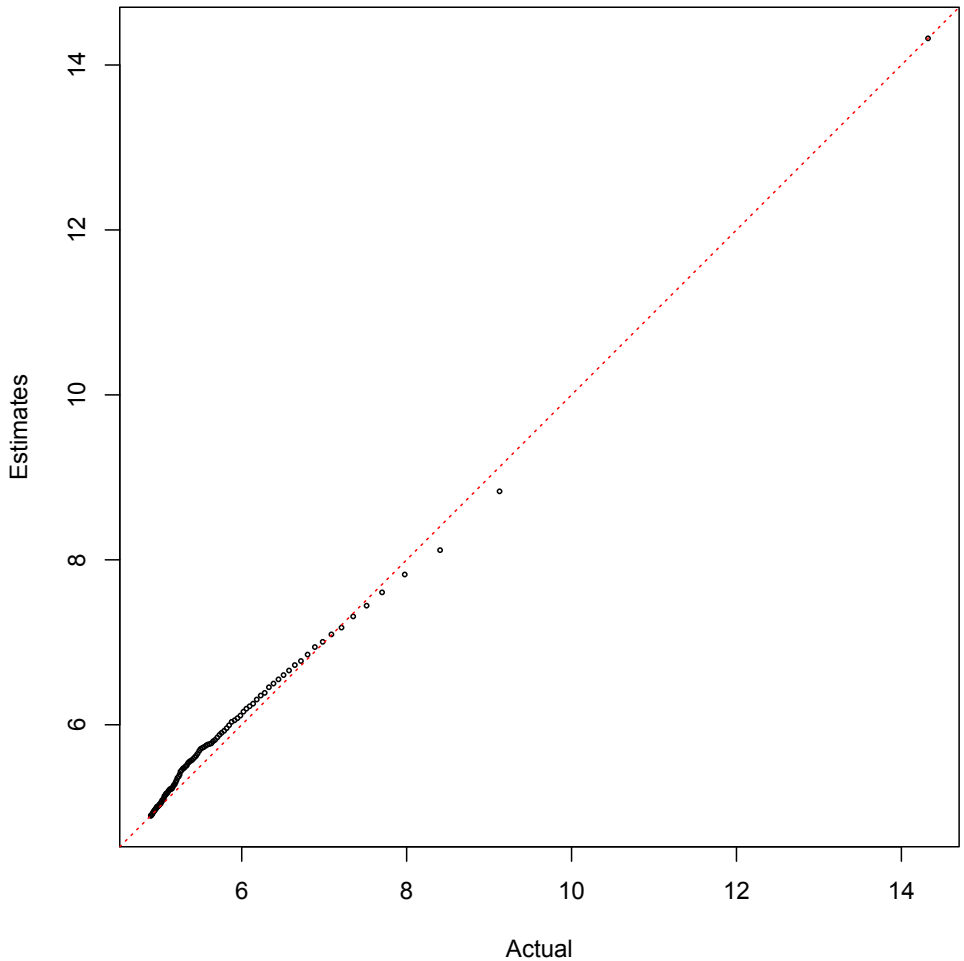
- "Promising" nodes: in the top 1%
- Start with 200 random initial nodes.
- Stopping rule: 3d generation of the initial nodes.
- Overall 17854 "promising" nodes sampled (for in-degree estimation).

Distribution of In-degree (Log-scale)

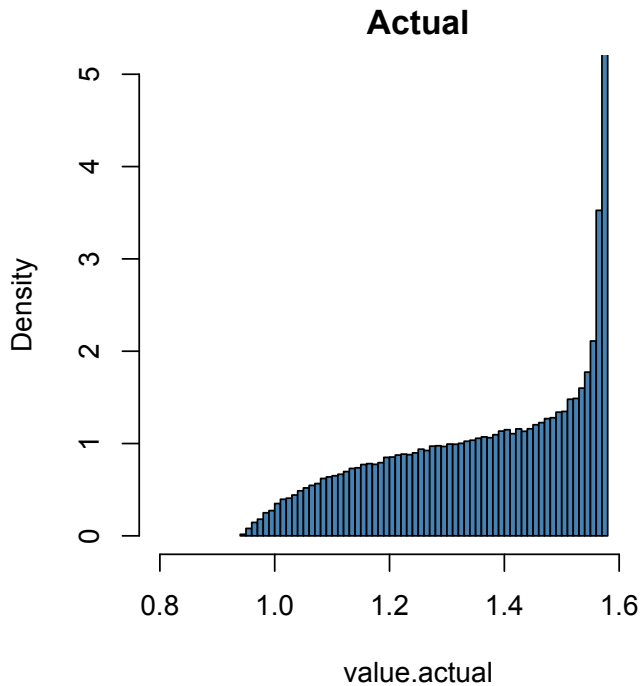


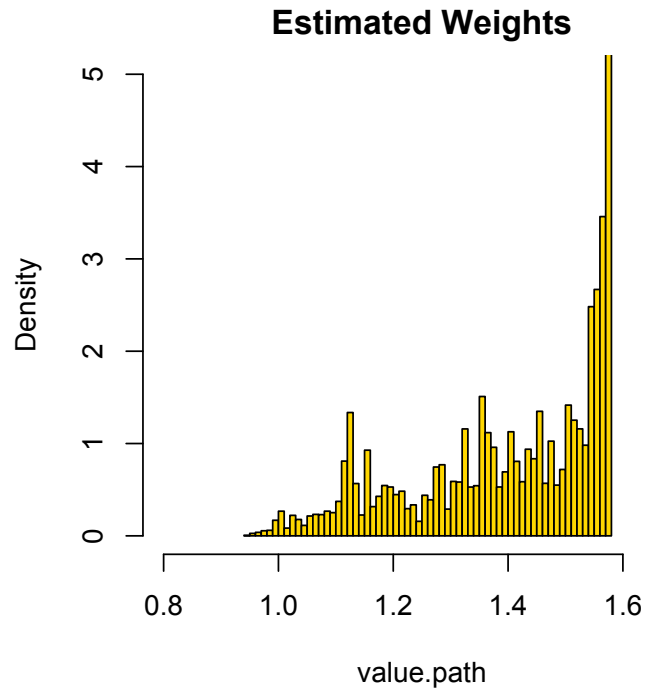


Q-Q Plot Of In-degree

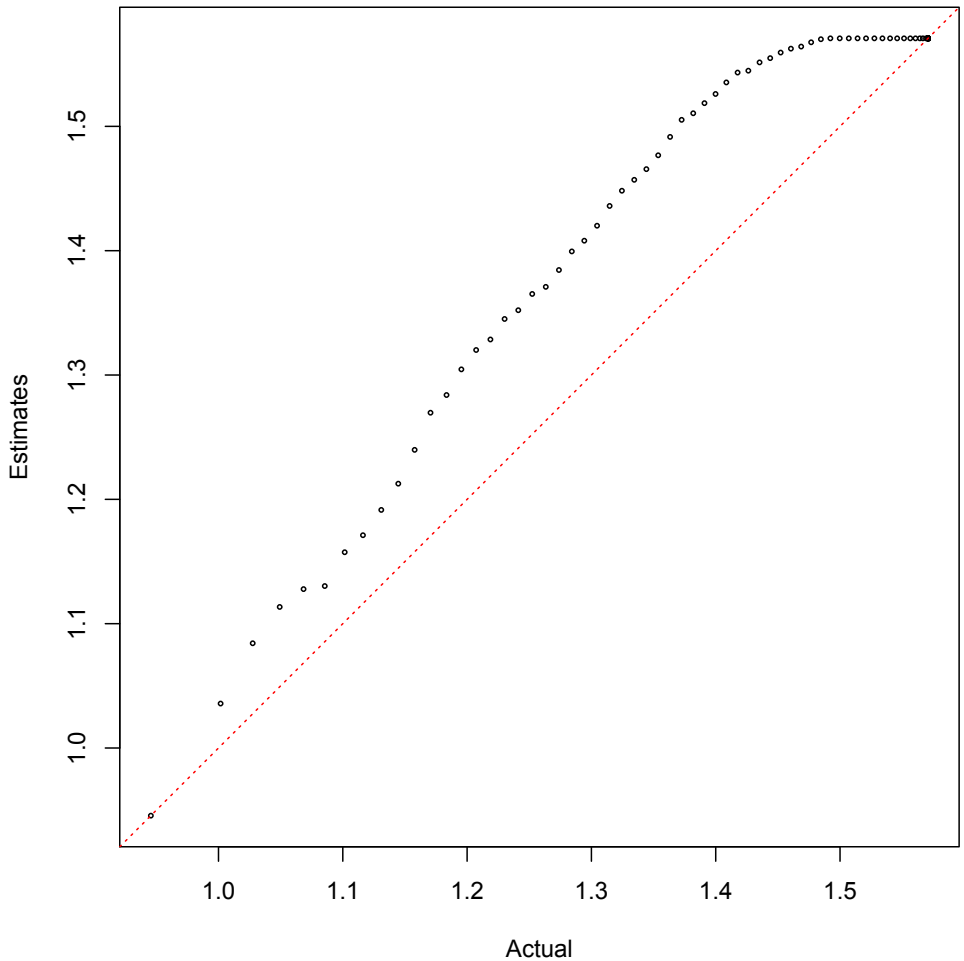


Histograms of the Angle





Q-Q Plot of the Angle



Computations: Webpages, In-Degree

- Proposed method (20 initial nodes): 1-3s for sampling, 1-2s for weight estimation (parallel computing)
- Multiple Random Walks (200 initial nodes): 3-10s for sampling
- Frontier Sampling (200 initial nodes): $>$ 5min for sampling

Computations: Webpages, Joint Distribution

- Proposed method (200 initial nodes): 1-3s for sampling, 1-3s for weight estimation (parallel computing)
- Multiple Random Walks (200 initial nodes): 3-10s for sampling
- Frontier Sampling (200 initial nodes): $>$ 5min for sampling