Sampling and Estimating Behaviors of Target Nodes in Networks

Jingjing Zou (Joint work with Richard A. Davis, Gennady Samorodnitsky, Zhi-Li Zhang)

April 15, 2016

The Network Data

- Usually recorded by edges
- In-degree: number of nodes to a specific node
- Out-degree: number of nodes from a specific node
- Interested in tail behavior

Data

- Webgraph from the Google programming contest (2002)
- Directed
- 875,713 nodes
- 5,105,039 edges

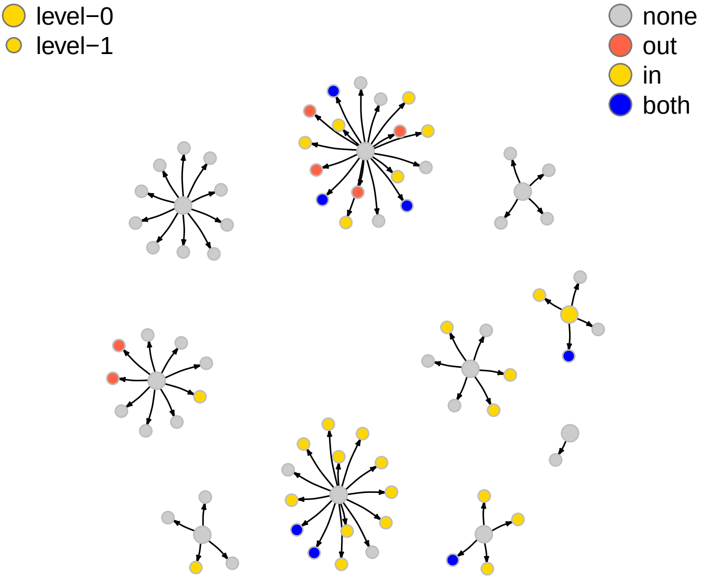
Node Types

- “In”: in-degree larger than 95% quantile (of interest here)
- “Out”: out-degree larger than 95% quantile
- “Both”: both in- and out-degree larger than 95% quantile
- “None”: Neither

Distribution of node types in initial selection:

```
##
## both   in none  out
##  0.0  0.1  0.9  0.0
```

Initial Selections



Distribution of types in neighbors of initial selection:

```
##
##   both      in    none     out
## 0.0959 0.3425 0.4795 0.0822
```

Goals

- Aim to study tail behavior of the network
- Sample nodes with extreme characteristics efficiently
- Construct unbiased estimators with sampled nodes

Strategies to Sample Target Nodes

- Single random walk: expensive, not representative with disjoint clusters
- Multiple random walks: able to explore multiple clusters
- Frontier Sampling (Ribeiro and Towsley, 2010)
- Uniform sampling of edges: use in-degree to adjust for bias

Our Strategy

- Use knowledge of distribution of neighbors' types
- Importance sampling / change of measure
- Construct estimators with weight adjustments

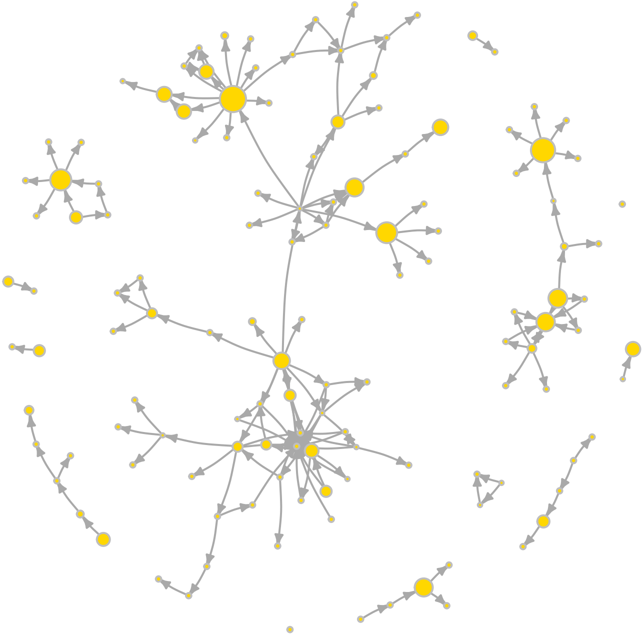
Our Strategy

- Step 0: randomly sample K nodes from the network
- Step 1: select neighbors of the K initial nodes
- Step 2: keep only the target (yellow) nodes
- Step 3: collect sample by following only paths of target (yellow) nodes

Final Selection

- Coarsening nodes connected in both directions to equivalence classes
- Nodes in the same equivalence class have the same weight (actual and estimated)

Final Selection



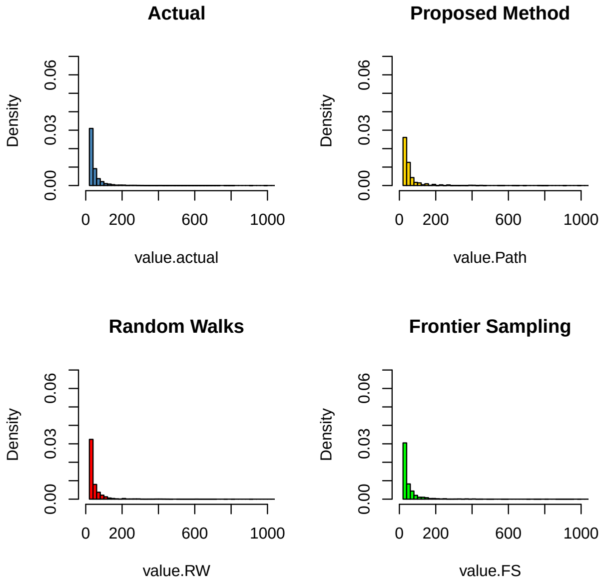
Estimators with Weight Adjustments

- Construct unbiased estimators using weighted averages of sampled nodes
- $w_i = 1/P(n_i \in S)$
- $P(n_i \in S) \propto$ no. of nodes leading to n_i
- Number of nodes leading to n_i cannot be completely observed
- Use observed values (proportional to the actual)

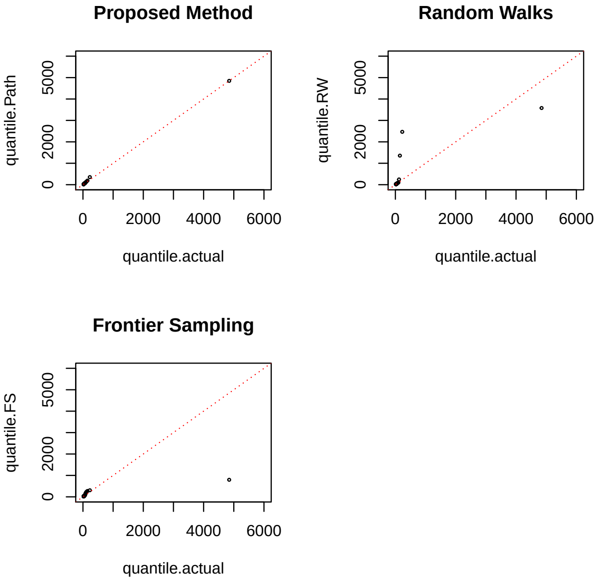
Estimation Results: Distribution of In-degree

- Start from 20 nodes in our method
- 200 initial nodes for Multiple Random Walks (RW) and Frontier Sampling (FS)

Distribution of In-degree



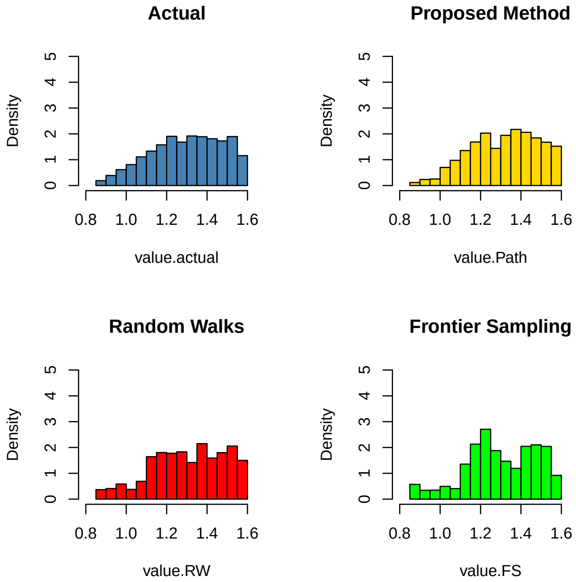
Q-Q Plots of Indegree



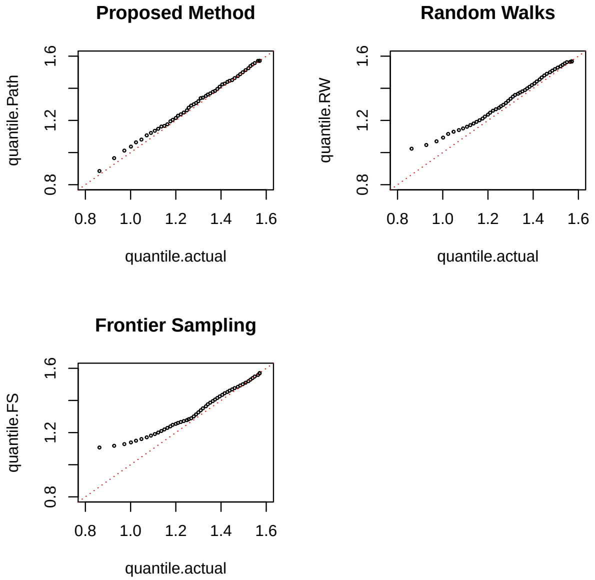
Estimation Results: Joint Distribution of In- and Out-Degrees

- Measured through $\arctan(\mathrm{In}_k/\mathrm{Out}_k)$
- Start from 200 initial nodes for all methods in comparison

Histograms of Angles



Q-Q Plots of Angles



Discussion on Computational Efficiency

- Cost of our method: choose cut-off, weight calculations
- Parallel computing

Camparison of Computing Time: Marginal Distribution of 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

Camparison of Computing Time: 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