569: Sampling Graphs (2013)

Jianguo Lu

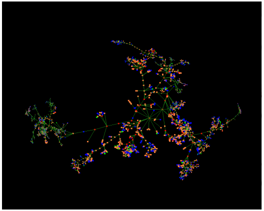
University of Windsor

November 18, 2013

Outline

Why sampling

Applications DeepWeb OSN SouceCode SemanticWeb

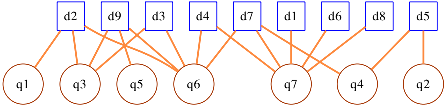


Properties Size Distribution Ranking Community Diameter CluterteringCoefficient

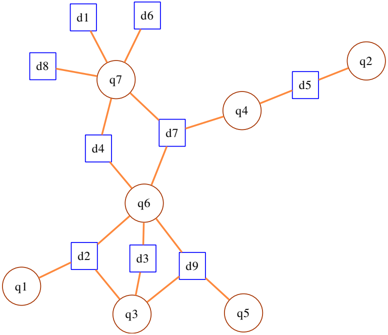
We need to estimate the properties for two reasons:

- Data in its entirety is not available (e.g., Facebook), or without central control (e.g., WWW), or evolving.
- Data is big. Quadratic algorithms are not feasible.

Deep web graph model



A: bipartite graph



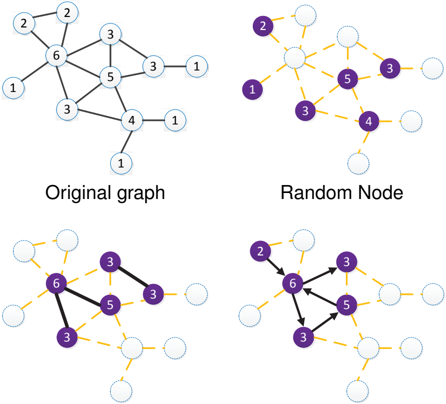
B: same graph as (A) in spring model layout

Figure: Hidden data source as a bipartite graph

What to sample for

- data size;
- distributions;
- clustering coefficient;
- communities;
- influential bloggers (degree, pageRank, Katz centralities, etc.)
- Outliers (spammers, zombies, inflated followers)

How to sample



Original graph | Random Node | Random Edge | Random Walk

- Sample by random node;
- Sample by random edge;
- Sample by random walk;
- Combinations and modifications. e.g., random walk with uniform random restart.

What is different from traditional sampling

- Most of the networks are scale-free. Degrees have very large variance. Uniform random sampling does not work.
- Precise sampling: quantities are digitalized, making the sampling process precise. e.g., know the exact degree, and can choose uniformly at random from the neighbouring nodes. Only possible for ONLINE social networks not real social networks.
- Access interface: provide interface APIs, many options. Can design new sampling schemes using APIs.

Size estimation

Applications

- Size of web, search engines
- Size of Online Social Network (Twitter, Weibo)
- Number of bugs in programs
- ...

Capture-Recapture Method

The estimator often used

When all the elements have equal probability of being sampled,

$$N = \frac{n_1 n_2}{d} \tag{1}$$

where n_1 and n_2 are the number of samples in two capture occasions, d is the duplicates.

- Lawrence and C. Giles. Searching the world wide web. Science, 280(5360):98-100, 1998.
- A. Broder and et al. Estimating corpus size via queries. In CIKM, pages 594-603. ACM, 2006.
- L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In WWW, pages 597-606. ACM, 2011.
- Petersson et al., Capture–recapture in software inspections after 10 years research—-theory, evaluation and application, Journal of Systems and Software, 2004.

Lots of research on obtaining uniform random sampling, using methods such as Metropolis-Hasting Random Walk

- Bar-Yossef et al. Random sampling from a search engine's index, JACM 2008.

Multiple Capture-Recapture

Equal sampling probability

$$N = \frac{n^2}{2C} \tag{2}$$

where n is total number of sampled elements, C is the number of collisions

Unequal sampling probability

$$N = \frac{n^2}{2C}\Gamma \tag{3}$$

where Γ is the normalized variance of the degrees of the graph

How large is Γ?

Γ in various datasets

Graph	$N(\times 10^3)$	γ or $\sqrt{\Gamma - 1}$	$\Phi(\times 10^{-5})$
WikiTalk [?]	2,388	26.32	2,700
BerkStan [?]	654	14.51	5.3
EmailEu [?]	224	13.66	13
Stanford [?]	255	11.51	5.8
Skitter [?]	1,694	10.46	56
Youtube [?]	1,134	9.64	440
NotreDame [?]	325	6.40	9.4
Gowalla [?]	196	5.54	1,200
Epinion [?]	75	4.02	610
Google [?]	855	4.00	62
Slashdot [?]	82	3.35	1,900
Facebook [?]	2,937	3.14	590
Flickr [?]	105	2.64	68
IMDB [?]	374	2.05	130
DBLP [?]	511	1.61	560
Amazon [?]	410	1.27	98
Gnutella [?]	62	1.21	9,100
CitePatents [?]	3,764	1.20	1,100

Table: Statistics of the 18 real-world graphs, sorted in descending order of the coefficient of degree variation γ. Φ is the conductance.

For Twitter data, $\Gamma \approx 1300$.

Outline

Bias Correction

Theorem

The relative bias of $\widehat{N}$ can be approximated by the reciprocal of $E(C)$, i.e.,

$$RB \approx \frac{1}{E(C)} \tag{4}$$

Jianguo Lu, Dingding Li, Bias Correction in Small Sample from Big Data, TKDE, 2013.

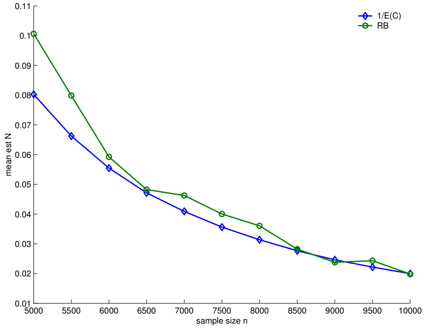


Figure: RB and 1/E(C) against sample sizes in simulation study. It shows that $\widehat{N}$ is biased upwards, and the relative bias can be approximated by the reciprocal of E(C).

Our bias corrected estimators

$$N = \frac{n^2}{2(C+1)} \quad (5)$$

$$N = \frac{n^2}{2(C+1)}\Gamma \quad (6)$$

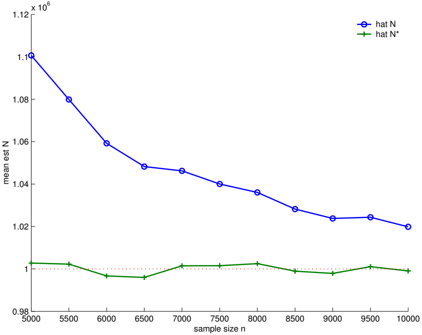
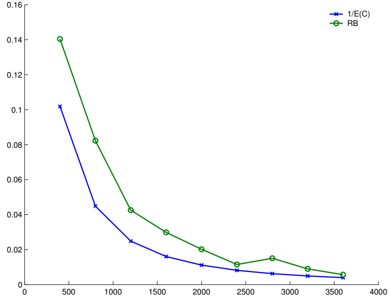


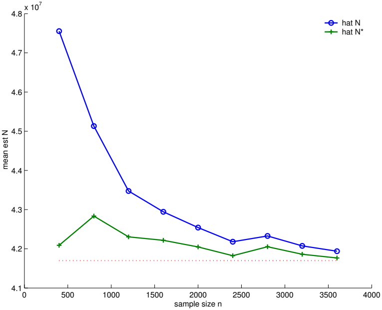
Figure: $\widehat{N}$ and $\widehat{N_S}$ over 10^4 runs for various sample size in simulation study. Red dotted line is the true value.

Twitter data

N=41million.



Expected bias



Corrected estimation

Outline

Uniform Random Sampling not Recommended

Lemma (Variance of $\widehat{N}_N$)

The estimated variance of RN estimator $\widehat{N}_N$ is

$$\widehat{var}(\widehat{N}_N) \approx \frac{N^2}{E(C)} \approx \frac{2N^3}{n^2} \tag{7}$$

Lemma (Variance of $\widehat{N}_E$)

The estimated variance of RE estimator $\widehat{N}_E$ is

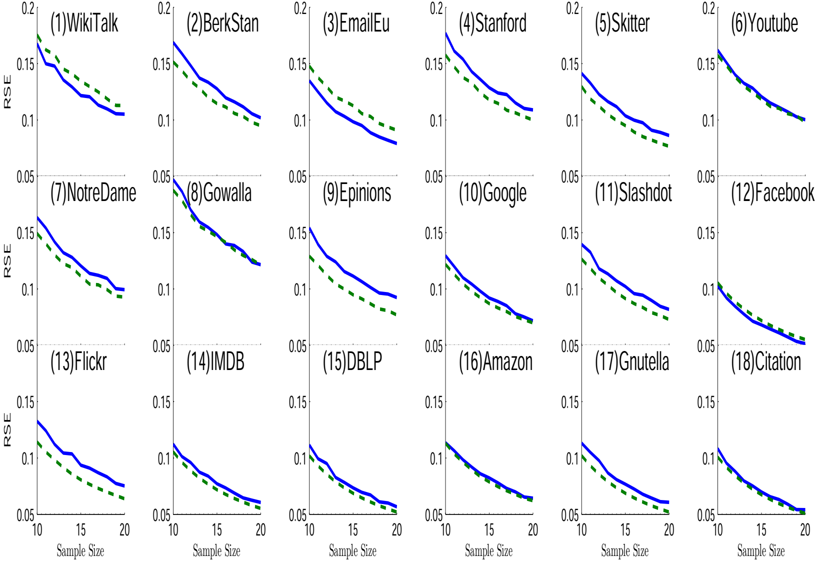
$$\widehat{var}(\widehat{N}_E) \approx \frac{2N^3}{n^2\Gamma}\left(1 + \frac{n\Gamma CV^2(\Gamma)}{2N}\right), \tag{8}$$

where $CV(\Gamma)$ is the coefficient of variation of Γ.

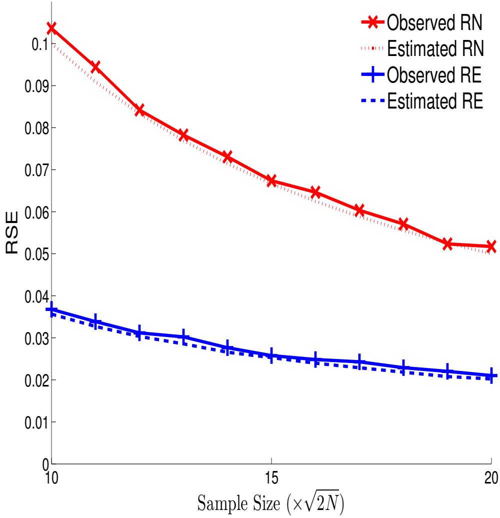
Theorem (RN vs. RE)

To achieve the same variance of $\widehat{N}_E$, $\widehat{N}_N$ needs to use at most $\sqrt{\Gamma}$ times more samples.

Estimated vs. observed



RN and RE Sampling on Facebook Data



Comparison of RN, RE, and RW Sampling

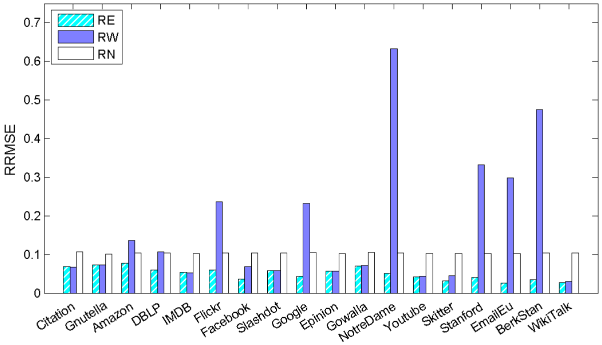


Figure: Comparison of three sampling methods. The sample size $n = \sqrt{2NC}$ where $\sqrt{C} = 10$. It shows that for RN sampling (red solid bars), the relative standard error is equal to $1/\sqrt{C} = 0.1$ across all the datasets. RE sampling is consistently smaller than RN sampling.RW sampling can approximate RE sampling for some datasets. For NotreDame etc. that have low conductance, RW is grossly wrong.

Why RW Varies

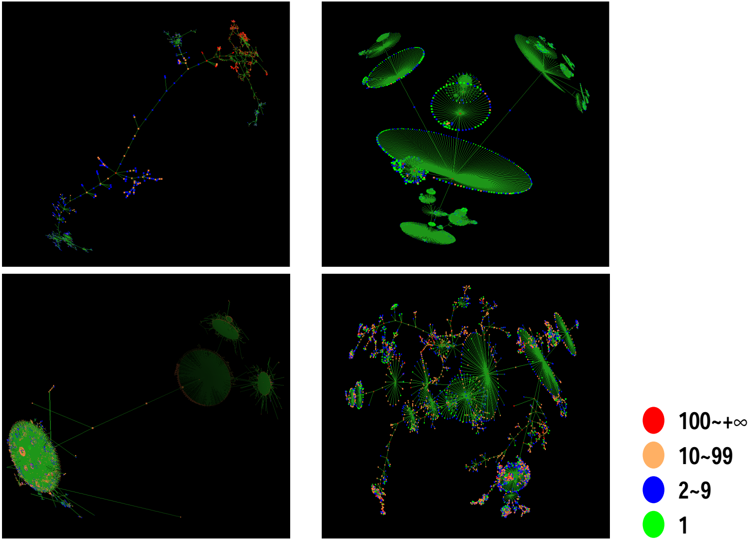


Figure: Subgraphs obtained by RW sampling from Flickr, EmailEu, Stanford and Youtube. Each subgraph contains 60,000 nodes. Node colour represents its degree in the original graph. Green=1; Blue=2 $\sim$ 9; Orange= 10$\sim$99; Red=100$\sim$ ∞.

Sampling for average degree

- Average degree is an important metrics in any network
- In and out average degrees in Weibo are different.
- Naive method–arithmetic sample mean
- Problem– Variance is too large because of the power law
- Solution– Use PPS (RE) sampling and harmonic mean estimator
- On Twitter, PPS sampling can be hundreds of times better

Average Degree Estimation

$$\widehat{\langle d \rangle}_{RN} = \frac{1}{n} \sum_{i=1}^{n} d_{xi} \tag{9}$$

$$\widehat{\langle d \rangle}_{RE} = \widehat{\langle d \rangle}_{RW} = n \left[\sum_{i=1}^{n} \frac{1}{d_{x_i}} \right]^{-1} \tag{10}$$

Theorem

Suppose the degrees follow Zipf's law with exponent one, i.e., $d_i = \frac{A}{\alpha+i}$. The variance of the random node estimator is

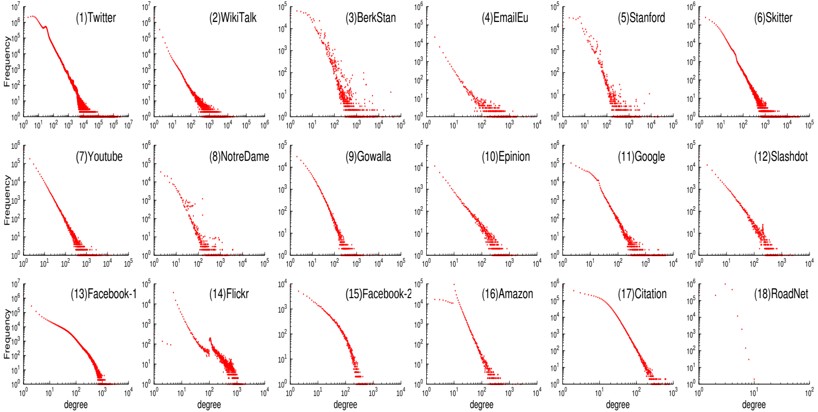
$$var(\widehat{\langle d \rangle}_{RN}) \approx \frac{\langle d \rangle^2}{n} \left(N \left[(\alpha+1) \ln^2 \frac{N+\alpha}{1+\alpha} \right]^{-1} - 1 \right). \tag{11}$$

$$var(\widehat{\langle d \rangle}_{RE}) \approx \frac{\langle d \rangle^2}{n} \left(\frac{1}{2} \ln \frac{N+\alpha}{1+\alpha} - 1 \right). \tag{12}$$

Main conclusion

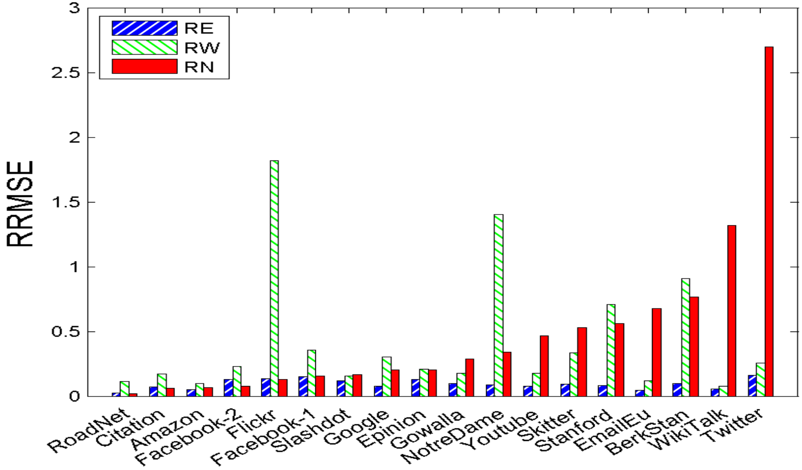
RE is better than RN in many cases; RW depends on graph conductance.

Degree distribution

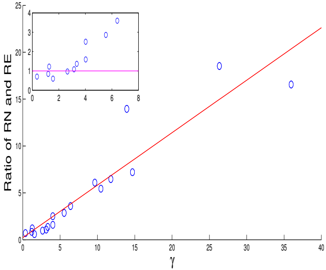


- Plots are sorted in decreasing order of coefficient of variation γ.
- Most of them follow power-law, yet they are very different

Comparison of Three Sampling methods



Comparison of RN and RE



Comparison of RN, RE, and RW Samplings

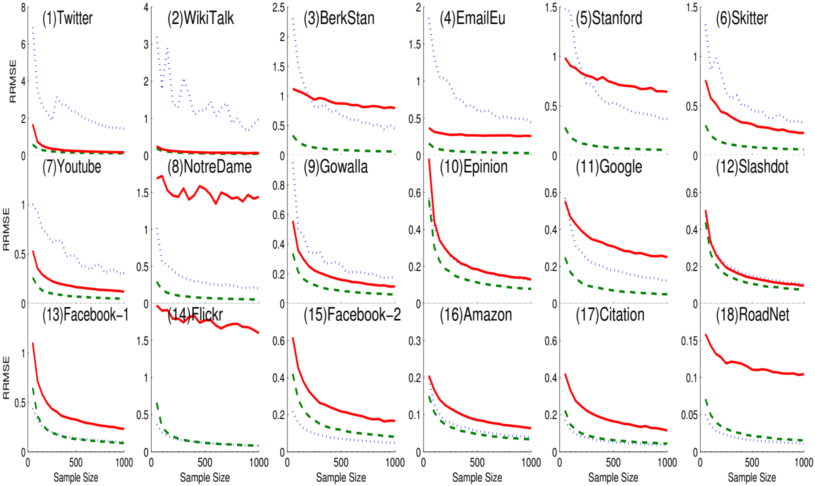


Figure: **RRMSEs of RN, RE, and RW samplings as a function of sample size for 18 graphs. The dotted, dashed, and solid lines are for RN(. . .), RE(− −), and RW(–) samplings respectively. It shows that in most cases the sample size does not change the relative positions of the sampling methods. The exceptions are the web graphs 3 and 5 where RW**

Sample distribution

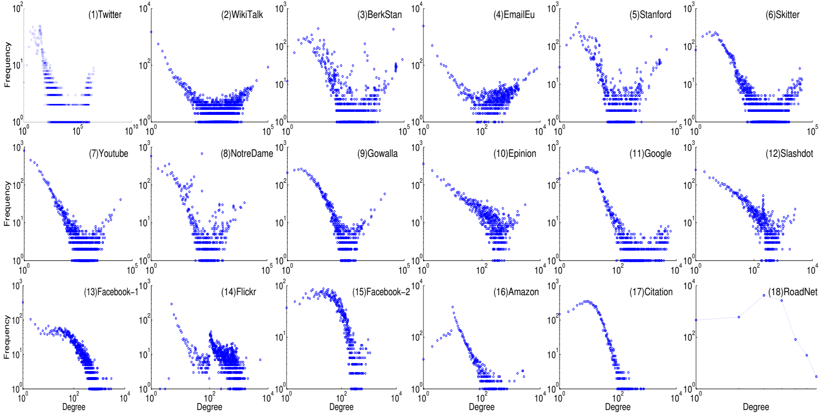


Figure: **The degree distributions of the samples obtained from RE (Random Edge) samplings. n=8,000. The log-log plots in the first two rows exhibit a “V" shape, where the sampled small nodes resemble the distribution of the original graph, while the sampled large nodes have a tail pointing upwards. These plots in the first two rows indicate that both small and large nodes are well represented in the sample. The plots in the last row indicate that the sample distribution is similar to the original distribution, therefore the RRMSE of RE sampling is similar to that of RN sampling.**

Graph conductance and RW sampling

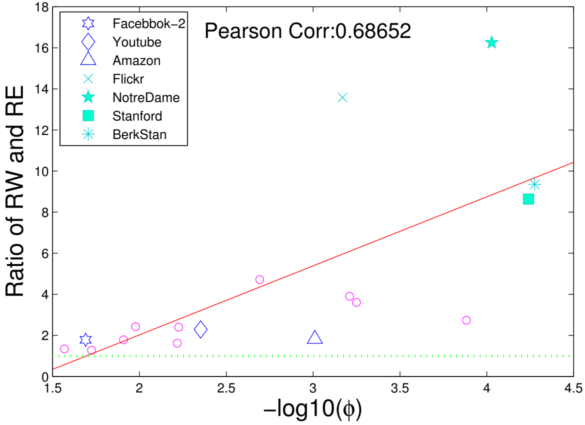


Figure: **Standard error ratio between RW and RE vs. graph conductance Φ for 18 datasets. Sample size is 400.**

Network Structure

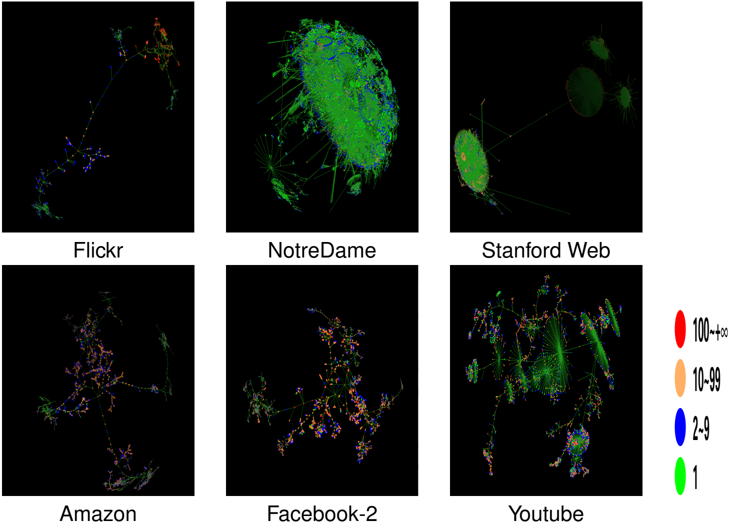


Figure: **Random walks on six networks. Flickr, NotreDame and Stanford have loosely connected components while Amazon, Facebook and Youtube are well enmeshed. Each random walk contains** 6×10^4 **nodes except NotreDame which has** 15×10^4 **nodes. Node colour indicates the degree of the node. Green=1; Blue=2**$\sim$**9; Yellow=10**$\sim$**99; Red=100+.**

Conductance

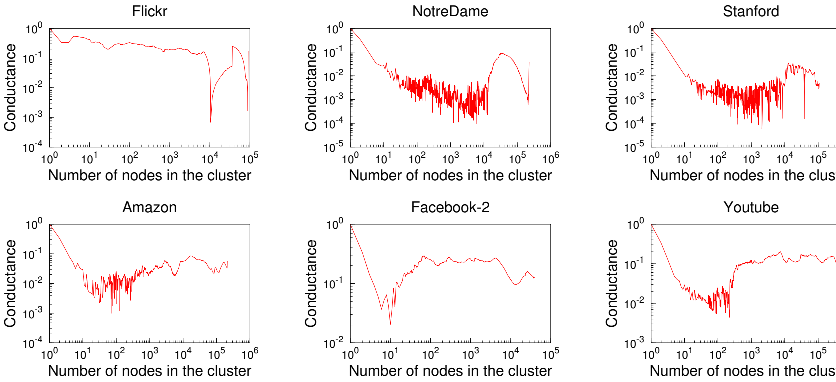


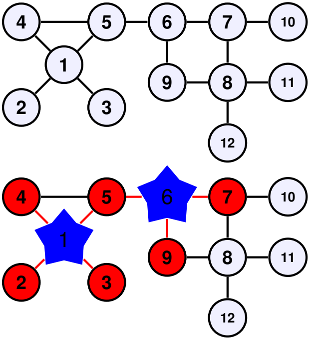
Figure: **Conductance $\Phi(S)$ over $|S|$, the size of the the components, for six networks. Plots are drawn using SNAP API described in [?].**

Weibo

- Very important
- We can access only partial information
- What is the global picture?
  - Size
  - Distribution
  - Most influential
  - Overall topology (e.g. clustering coefficient)
  - Message diffusion, Critical nodes
  - Communities
  - ...

Star Sampling

- Select nodes uniformly at random (e.g., nodes 1 and 6);
- Take all the neighbours as sample (nodes 2,3,4,5,5,7,9);
- It approximate PPS (probability proportional to size) sampling;
- More efficient than random walk by taking all the neighbours instead a random one;
- We sampled around one million Weibo stars;



Degrees and Messages

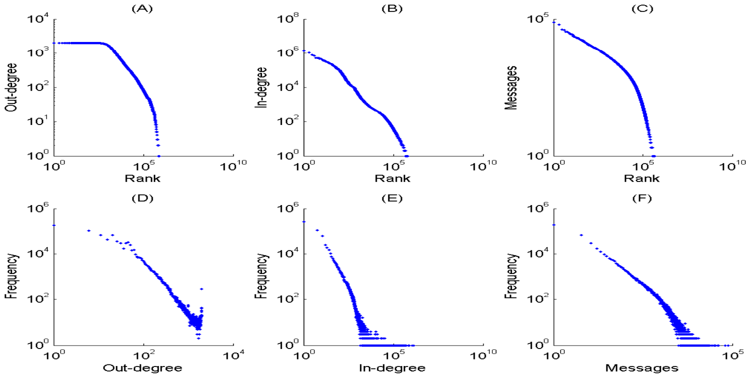


Figure: **Estimated out-degree, in-degree, and message distributions of Weibo.**

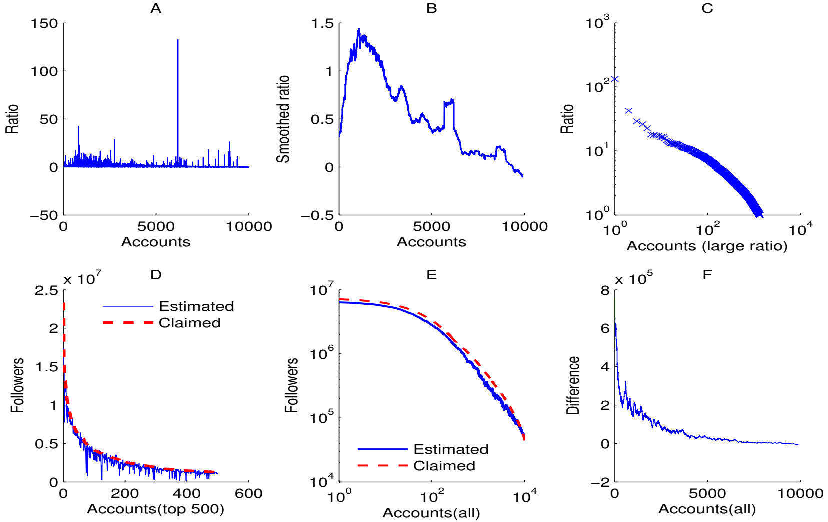
Average in-degree and out-degree as 32.10 (CI 31.91, 32.29) and 54.39 (CI 49.02, 59.76), respectively.

Estimation of followers

	f_i	d_i	$\widehat{\langle d \rangle}_i$	Difference	Ratio
1	85016	23,335,290	16,859,105	6,476,185	0.38
2	75243	15,945,306	14,921,069	1,024,237	0.06
3	71417	15,247,604	14,162,354	1,085,250	0.07
4	37914	13,394,620	7,518,539	5,876,081	0.78
5	61962	13,278,161	12,287,380	990,781	0.08
6	63308	13,153,177	12,554,298	598,879	0.04
7	59969	12,990,041	11,892,158	1,097,883	0.09
8	57100	12,604,270	11,323,220	1,281,050	0.11
9	59406	12,097,122	11,780,512	316,610	0.02
10	54264	12,003,137	10,760,827	1,242,310	0.11

Table: Estimation for the top 10 Weibo accounts. f_i: capture frequency of account i; d_i: claimed in-degree or number of followers; $\widehat{\langle d \rangle}_i$: estimated number of followers; $Ratio = (d_i - \widehat{\langle d \rangle}_i)/\widehat{\langle d \rangle}_i$.

Estimated Followers



Estimated vs. Claimed Followers

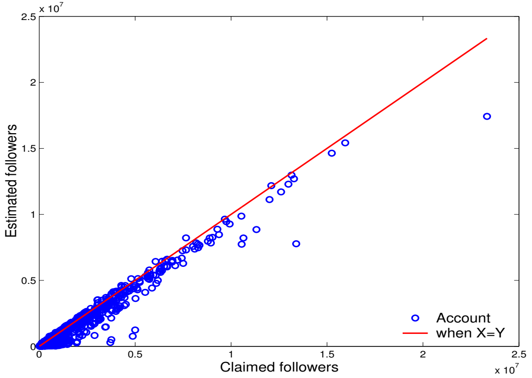


Figure: **Estimated followers vs. claimed followers in log-log scale. The Pearson correlation coefficient is 0.9797.**

Relative standard deviation of the estimator is

$$RSD(\hat{d}_i) = 1/\sqrt{f_i}. \tag{13}$$

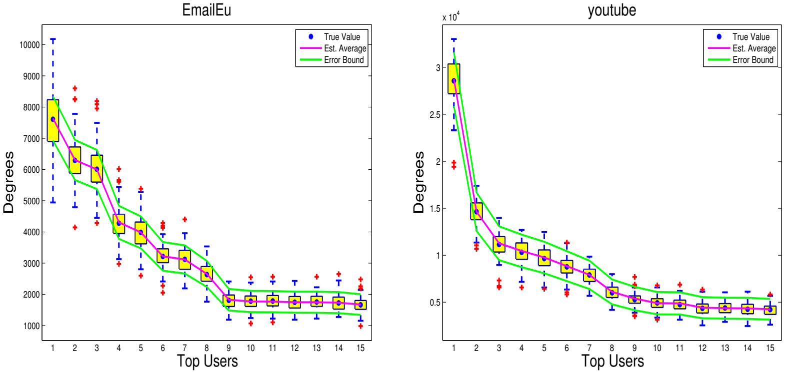


Figure: Sampling accuracy on existing data