Detect Inflated Follower Numbers in OSN Using Star Sampling

Hao Wang, Jianguo Lu
School of Computer Science, University of Windsor
401 Sunset Avenue, Windsor, Ontario N9B 3P4. Canada
E-mail: {wang115o, jlu}@uwindsor.ca.

Abstract—**The properties of online social networks (OSNs) are of great interests to the general public as well as IT professionals. Often the raw data are not available and the summary released by the service providers are sketchy. Thus sampling is needed to reveal the hidden properties of the underlying data. While uniform random sampling is often preferred, some properties such as the top bloggers need to be obtained using PPS (probability proportional to size) sampling. Although PPS sampling can be approximated using simple random walk, it is not efficient because only one sample is taken in every step. This paper introduces an efficient sampling method, called star sampling, that takes all the neighbours as valid samples. It is more efficient than random walk sampling by a factor of the average degrees. We derive the estimator and its variance, and verify the result using six large real-networks locally where the ground-truth are known and the estimations can be evaluated.**

Then we apply our method on Weibo, the Chinese version of Twitter, whose properties are rarely studied albeit its enormous size and influence. Along with other conventional metrics such as size and degree distributions, we demonstrate that star sampling can identify ten thousand top bloggers efficiently. In general, the estimated follower number is consistent with the claimed number, but there are cases where the follower numbers are inflated by a factor up to 132.

Index Terms—**Online social network, sampling, graph sampling, Weibo.**

I. INTRODUCTION

The properties of online social networks (OSNs) are of interests to a variety of stakeholders, including the general public as well as IT professionals [3]. Often the raw data are not available and the summary released by the service providers are sketchy. OSNs are so large that exhaustive exploration of the network is infeasible. Instead, we can only obtain a small portion of the network and estimate the properties of the network using some sample datasets.

There are many studies on sampling methods for OSNs. Two of the basic sampling methods are uniform random sampling and PPS (probability proportional to size) sampling. Uniform random sampling is the norm of the practice, the method opted for whenever possible, also the method not easy to implement in many applications. In OSN studies, it is often realized by uniform ID sampling [6] [7], or Metropolis-Hasting random walk [7].

Some properties, such as the top bloggers, are innately not suitable for uniform random sampling, especially for scale-free networks where most of the bloggers have a small number of followers [2]. It was widely accepted [8], as well as demonstrated in this paper, that most OSNs are scale-free networks. Uniform random sampling gives each blogger an equal probability of being sampled, meaning that top bloggers have no more chance of being sampled than other people. Consequently, the sample is mostly comprised of small accounts. Most top bloggers are not even sampled at once, let alone to study their properties.

Therefore, there is a need to use PPS sampling to sample the large microblog accounts more often. PPS sampling is hard to implement directly using existing OSN access methods. Most OSNs support random walk sampling, which can approximate PPS sampling in the sense that the sampling probability of a node(account) is proportional to its degree asymptotically [14]. It is not efficient in that in every random walk step, only one random sample is obtained from all the neighbours of the current node. This problem becomes more acute in OSN sampling where each step involves remote access to the API through internet, and sometimes there are daily quotas for the total number of accesses allowed. When one API call retrieves all the neighbouring nodes (followees), it is too costly to select only one of them and discard the others.

Thus, we propose star sampling that is an efficient approximation to PPS sampling. It selects random nodes first using ID sampling that is enabled by several OSNs, including Weibo OSN being studied. Then, for each random node we select all its neighbours connected by outgoing links, as if expanding the node to a star. In this way, the sampling process is faster by a factor of the average degree of the nodes, and it can be run in parallel. Yet, it is a kind of PPS sampling as we will demonstrate in this paper. This method also avoids other difficulties in random walk sampling, such as dead-ends, infinite loops, and isolated components [14].

Before applying star sampling on Weibo OSN, we first verify it on six networks whose ground-truth values are known. We compare the empirical average with the true value, and the empirical variance with the theoretical predication. All six datasets support our method very well. Based on this result, we apply our method to explore a variety of properties of Weibo, including degree distributions and follower numbers. Although most of the accounts conforms to our estimation, there are outliers whose claimed followers are much higher than our predication.

In the following, we will first introduce the background knowledge and related work. Then we use ID sampling to obtain uniform random samples. From the uniform random nodes, we apply star sampling to samples whose capture probability is proportional to its size. From these samples, we estimate their followers and reveal the discrepancy.

II. Background and Related Work

A. OSN Access Methods

An efficient sampling method needs to fully utilize the access interfaces provided by the OSN service provider. There are several approaches to accessing OSN data, including:

- By probing account IDs: In some microblog sites such as Weibo and Twitter, microblogger's account can be accessed using http request such as www.weibo.com/1234567890, where the number is the account ID. Because every account can be accessed using an ID, and the ID space is not very large (a 10 digit number for Weibo), uniform random accounts can be found by generating a random number within the ID space. This method is used to obtain uniform random nodes (accounts) from Facebook [7], Youtube [23] and Weibo [6] .
- By crawling using web API: Most OSNs provide programmable web APIs, typically supporting programmers to navigate in the network, such as getting the outgoing and in-coming links. New blogger data can be obtained by following the links provided in the current account.
- By crawling HTML pages and screen scraping: Instead of using more organized web APIs, OSN data can be also directly extracted from its HTML pages. By following the hyperlinks imbedded inside the web pages, we can find the neighbours of the current blogger.
- By sending queries: Most OSNs provide searchable interfaces, either by providing an API or an HTML form. In either way, we can send queries and retrieve matched pages.

We use ID probing and web API calls in combination. ID probing is used to get uniform random samples, while web API is used to get all the outgoing links of those random bloggers.

B. Graph sampling

An OSN can be modelled as a graph, where an account(or a blogger) is a node, and nodes are connected by following relationship. In general, a graph can be sampled by random node, random edge, and random walk. Their comparative studies are conducted in [12] [21] [13].

1) Uniform Random Node Sampling: In this sampling method, each node is sampled with equal probability. It can be realized by selecting the nodes directly, as in random ID sampling, or by following the links using certain strategies such as Metropolis-Hasting random walk [16] [9] [7].

2) Random edge sampling: In random edge sampling each edge is selected with equal probability. Consequently, each node is selected with probability proportional to its degree. Thus it is a PPS sampling that we want to perform. However, random edge sampling is not easy to realize in many cases, and is often approximated by random walk sampling.

3) Random walk sampling: Simple random walk sampling selects the next node from one of its neighbours with equal probability. The variations come when we decide how many nodes to select in the next step, how to choose the next step (with same probability or different probability depending on some measurement, and what we can do when the walk is stuck in a dead end and loop.

C. Weibo and other OSN sampling

There are very few papers depicting the landscape of Weibo OSN despite its enormous size and influence. Very recently, [6] uses uniform samples to estimate the properties of Weibo OSN. They report a wide range of estimated properties such as the number of accounts, active accounts according to messaging information, and geographic distributions. Due to the limitation of uniform random sampling, they are not able to find out the degree distribution, neither the number of followers of top bloggers. We apply both uniform random sampling and PPS sampling, thereby obtain more interesting results.

Other OSNs are extensively studied, including Facebook [7] [22] [5] [18] and Twitter [1] [8] [11] [15]. Compared to this groups of work, we are not aware of the star sampling method proposed in this paper, neither the study on properties of top bloggers.

III. Uniform ID Sampling and Size Estimation

Suppose that the set of possible ID is $\{1, 2, \ldots, U\}$. Among them there are N number of valid IDs, and $U - N$ number of invalid IDs. Our target is to obtain a uniform random sample.

The ID sampling process (Algorithm 1) can be explained as follows: A random number is generated within the range of $1, \ldots, U$, and is tested whether it is valid by sending the ID to the web site. Overall n number of tests are made, among them $v = |V|$ number of tests are valid. Note that the random numbers can have duplicates and they are included in the counting. The number of accounts can be estimated by

$$\hat{N} = \frac{v}{n} U, \tag{1}$$

whose approximate relative standard deviation (RSD) is

$$RSD(\widehat{N}) = \sqrt{1/v}. \tag{2}$$

In the case of Weibo each user account ID is a 10-digit number, i.e., $U = 10^{10}$. Although some accounts have account names, they are still have a 10-digit ID that is accessible by our method. As Equation 2 shows, the success of ID sampling hinges on the value of v, which in turn is decided by the ration of N/U. If the universe U were

Algorithm 1: Uniform ID sampling

```
Input: ID range 1..U, sample size $n$;
Output: Valid IDs $V$.
$V$=empty sequence;
$i = 0$;
while $i < n$ do
    $i++$;
    generate a random number $id$ within 1..U;
    if $id$ is a valid account then
        add $id$ into $V$ ;
    end
end
```

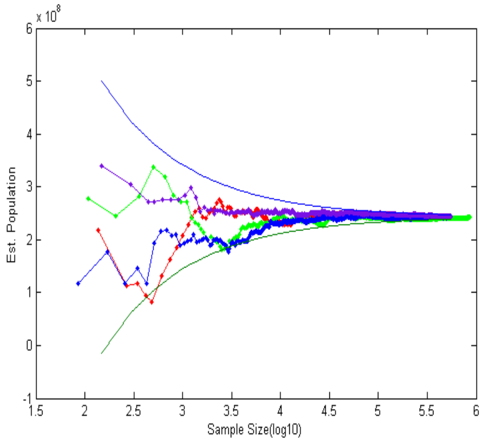

Fig. 1. **Estimated number of accounts against sample size. The estimation stabilizes when only 20,000 random IDs are tested.**

very large (say, by allowing for arbitrary length of letters), most of the randomly generated IDs would be invalid ones. Such low ratio will render the ID sampling infeasible. Fortunately, in the case of Weibo, the ratio is rather large and the probability of success is $21104/848969 \approx 0.025$. To expedite the process, the DNS resolution is done once and cached for later use.

We run ID sampling in December 2011. Figure 1 shows four independent sampling processes, along with the projected error bound derived from Equation 2. Each process has a sample size around 500,000. Overall, the estimation of the total number of accounts is 243 million (95% Confidence interval is between 238 million and 247 million).

Our results coincide with the size estimation reported in [6], where 269 millions of account are projected in January 2012. Within one month, we observe an increase about 9 % of user accounts. Another observation we have is that relatively small number of samples are needed to reach an accurate estimation of the user account number.

A. Degree and message distributions

Using the same ID sampling algorithm, we obtain further 1,184,964 uniform IDs. This time we do not record the times the ID probing fails. Instead, we focus on the valid IDs by downloading its degree and message information to study their distributions.

It is reported that a uniform random sample can reflect the distributions of the original data [21]. Figure 2 shows the distributions of the in-degree, out-degree, and messages. All are in log-log plot since they have long tails. Each data is plotted in two ways: The degree-rank plots in the first row focus on the top nodes, while the frequency-degree plots on the second row focus on the nodes with small degrees. In a degree-rank plot, all the nodes are sorted according to its degree in increasing order, then a rank is assigned to each node. In frequency-degree plot, the occurrence frequency of a degree is plotted against the degree.

Figure 2 (A) shows that the out-degree has a limit around two thousand. Its corresponding frequency-degree plot in subplot (D) shows that there are more than 10^5 nodes that have only one outgoing edge among the one million sampled nodes.

In-degrees are closer to power-law distribution, similar to most other networks such as Twitter [11], Facebook [7], and the Web graph [4]. Subplot (B) shows an almost straight line with exponent one, similar to that of Twitter data [11]. This plot also demonstrates that uniform random sampling can only reveals the shape of the follower distribution, not the details of top bloggers. For instance, there are only two of the sampled accounts that have follower number greater than one million. Using star sampling discussed in the next section, we found that there are 691 millionaires who have more than one million followers. Subplot (E) is the corresponding frequency-degree plot that shows most of the bloggers have small number of followers. For instance, in the sampled nodes, there are more than 10^5 number of nodes/bloggers who have only one follower.

Subplots (C) and (F) describe the message distribution among the samples. Subplot (C) is the rank-message plot, describing that the number of messages decreases quickly. The top sampled blogger sends close to 10^5 number of messages. Overall, the curve fit better with Mandelbrot law [19].

The standard deviations are 103.2708 and 2916.8887 for in-degrees and out-degrees, respectively. From the uniform random samples we can estimate the average in-degree and out-degree as 32.10 (CI 31.91, 32.29) and 54.39 (CI 49.02, 59.76), respectively. Surprisingly, the average in-degree is markably larger than the average out-degree. Such inconsistency can be caused by several factors. One may be the inflated follower number as suggested in the next section.

IV. Star sampling and Follower Number Estimation

Fake OSN followers has become a multimillion dollar business. In Twitter, zombi followers are sold in large quantities ranging from thousands to millions [20]. There are robots to generate zombies to follow designated bloggers, and there are also tools to detect the percentage of

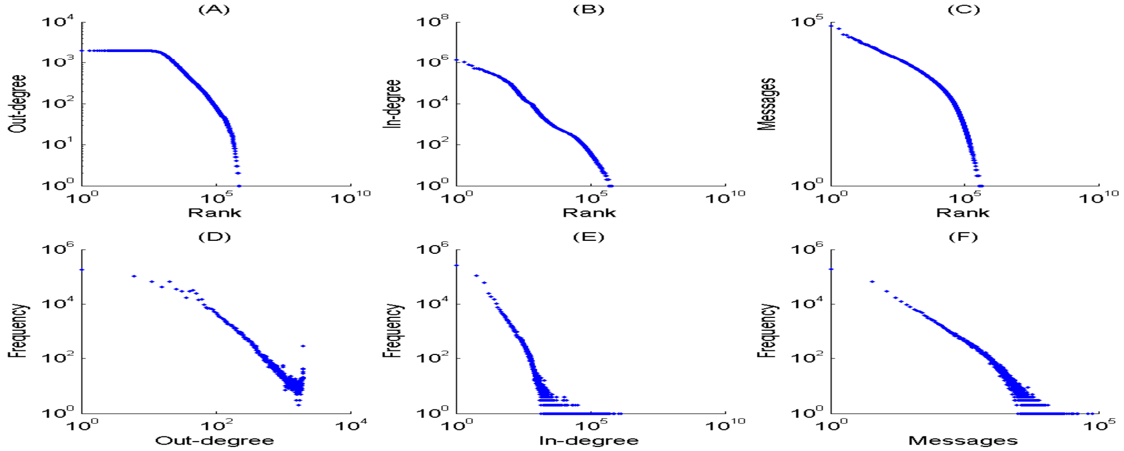

Fig. 2. **Estimated out-degree, in-degree, and message distributions of Weibo.**

zombi followers[1]. This paper addresses another type of fake follower number, the follower number that is artificially inflated regardless of the zombies.

A. *Star sampling*

To find the follower number of the top bloggers, it is no longer effective to use uniform random sampling, where all the nodes, most of them are small ones with few connections, have equal probability of being sampled. To have top bloggers sampled more often, large nodes should have high probability of being sampled. Therefore, we opt for PPS sampling where nodes are sampled with probability proportional to their degrees. There are several choices to run PPS sampling, such as random edge and random walk samplings. Random edge sampling is not easy to implement in Weibo sampling, while random walk can approximate PPS especially in OSNs where the mixing time is small. However, random walk is not efficient in that in every step only one random node is selected among all the neighbours.

Since we already have the uniform random IDs, and every remote API call gets back all the neighbours, we utilize all the neighbouring nodes by employing the following star sampling as described in Algorithm 2: select a set of uniform random nodes, and expand each node as a star that contains all the neighbours. Put all the nodes in the neighbours into the sample. Compared with the random walk where only one node is taken, this method speeds up by a factor of the average out-degree $\langle d \rangle$. In Weibo dataset, it is 32 times more efficient. Next, we need to establish the accuracy of the estimation.

Suppose that in the directed graph G, there are N number of nodes labeled as $1, 2, \ldots, N$, whose in-degrees are denoted by d_i for $i \in \{1, 2, \ldots, N\}$. Let E denote the number of edges, which is the sum of all the in-degrees (or out-degrees), i.e., $E = \sum_{i=1}^{N} d_i$. If we ignore the structure of the graph, we can view it as a sequence of occurrences of node IDs. Node i has d_i incoming links, so it will occur in the sequence d_i times. Obviously the sequence is of length E.

Algorithm 2: Star sampling

```
Input: Valid ID set $V$, sample size $n$;
Output: Sequence of sample IDs $S$, including
        duplicates.
$i = 0$;
$S$ is empty;
while $i < n$ do
    $x$ = an ID selected from $V$ uniformly at random;
    $(x_1, x_2, \ldots, x_k)$ = all the neighbours of $x$;
    Add nodes $(x_1, x_2, \ldots, x_k)$ to $S$;
    i=i+k;
end
```

If we select an occurrence from the sequence uniformly at random, each occurrence is selected with equal probability $1/E$, and node i has the probability $p_i = d_i/E$ being selected. The number of times node i is selected after n sample nodes are taken can be described by the binomial distribution $B(n, p_i)$. Thus, the expected number of captures of node i is $E(f_i) = np_i$. Or, given observation of f_i, p_i can be estimated by:

$$\hat{p}_i = \frac{f_i}{n}, \tag{3}$$

where f_i is the number of times node i is sampled.

Star sampling resembles the above sampling process, except that it selects multiple nodes at once instead of one node at a time. In other words, it is sampling without replacement, and results in hypergeometric distribution instead of binomial distribution. It is well known that binomial distribution can approximate hypergeometric distribution well if $n \ll N$. In our Weibo OSN application, out-degrees (or equivalently the star size) has a up limit 5000, well below the population size N in the order of 10^8.

With the knowledge of p_i, d_i, the number of followers

[1] For instance in http://www.socialbakers.com/twitter/fakefollowercheck/

of node i is estimated by

$$\widehat{d_i} = \hat{p}_i\; E = \frac{f_i E}{n} = \frac{f_i}{n}\widehat{N\langle d\rangle}^{out}. \tag{4}$$

Because of the binomial distribution, the variance of f_i is

$$var(f_i) = np_i(1-p_i) \approx np_i. \tag{5}$$

The approximation is valid because p_i is very small in our scenario. The variance of the estimator is

$$var(\widehat{d_i}) = var(f_i)E^2/n^2 = f_i E^2/n^2. \tag{6}$$

Hence the relative standard deviation is

$$RSD(\widehat{d_i}) = 1/\sqrt{f_i}. \tag{7}$$

Equation 7 gives the guideline to select the sample size so that satisfactory estimation can be obtained. For instance, if we want the 95% confidence interval to be within $\widehat{d_i} \pm 0.2\widehat{d_i} \approx \widehat{d_i} \pm 1.96 \times \sqrt{1/f_i}\widehat{d_i}$, f_i needs to be greater than 100. We use this guideline to design our experiments.

B. Pilot study on local datasets

Our estimator and its variance are deduced based on the binomial distribution for sampling with replacement where each edge is selected one at a time. That edge is put back, and can be sampled again in the next sampling occasion. In our star sampling, a set of edges in a star are sampled simultaneously without replacement–there is no chance that an edge within the same star can be sampled twice when that star is selected. This sampling process will result in hypergeometric distribution. When the size of star is much less than the total population, which is true in our case, it is known that binomial distribution can approximate the hypergeometric distribution very well.

To validate our assumption, we carried out a pilot experiment on local datasets. Since the ground truth values are known, the estimator and the variance can be evaluated. Our local datasets are six networks whose statistics are summarized in Table I. Star sampling are applied to each network. We evaluate the estimation performance on the top 15 nodes for each network, by comparing their empirical variance with the theoretical variance, and empirical average with the true value as demonstrated in Figure 3.

The 95% error bounds are calculated from Equation 7, the box plots are obtained from 100 repetition of the experiments. The average of 100 estimations fit well with the true value across all the networks and all the top 15 nodes. This indicates that star sampling is indeed unbiased. In addition, most of the estimations fall within the estimated error bound, demonstrating that star sampling can approximate PPS sampling. The sample size is controlled so that each of the 15-th node can be sampled at least 50 times. Depending on the degree of the 15-th node and the overall degree distribution, varying sample size is needed for each network (50K for WikiTalk, 200K for Skitter, 80K for Youtube, 80K for NotreDame, 200K for Stanford and 40K for EmailEU).

Network	# Nodes	γ	$\langle d\rangle$	Max degree
WikiTalk[13]	2,394,385	26.34	3.89	100,029
EmailEu[13]	265,009	13.93	2.75	7,636
Stanford[13]	281,903	11.79	14.14	38,625
Skitter[13]	1,696,415	10.46	13.08	35,455
Youtube[17]	1,138,499	9.65	5.25	28,754
NotreDame[13]	325,729	6.40	5.25	10,721

TABLE I
STATISTICS OF THE 6 REAL-WORLD GRAPHS, SORTED IN DESCENDING ORDER OF THE COEFFICIENT OF DEGREE VARIATION $\gamma = variance/\langle d\rangle^2$.

	f_i	d_i	$\widehat{d_i}$	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 II
ESTIMATION FOR THE TOP 10 WEIBO ACCOUNTS. f_i: CAPTURE FREQUENCY OF ACCOUNT i; d_i: CLAIMED IN-DEGREE OR NUMBER OF FOLLOWERS; $\widehat{d_i}$: ESTIMATED NUMBER OF FOLLOWERS; $Ratio = (d_i - \widehat{d_i})/\widehat{d_i}$.

C. Results for Weibo data

During October 2011 and January 2012, we selected 1,184,964 number of uniform random nodes. On average, each random node has 32.08 outgoing links. We expand each uniform random node as a star, and collect all the nodes pointed by the outgoing edges as the sample. Overall 38,019,277 number of sample nodes are collected, including duplicates. The largest account has claimed 23,335,290 number of followers, and is captured 85,016 times. According to Equation 7 we reckon that around 100 of captures are required to produce meaningful estimation. Thus, we take only the top 10,000 accounts. The lowest has 16,038 followers and is captured 65 times. Some relevant data are available at http://cs.uwindsor.ca/~jlu/weibo.

The estimation is consistent with the claimed number for many accounts. Let ratio denote the relative inflation rate, i.e.,

$$ratio = (d_i - \widehat{d_i})/\widehat{d_i}, \tag{8}$$

where d_i is the claimed number of followers (in-degree), and $\widehat{d_i}$ is the estimated number of followers. Table II listed the estimations for the top 10 accounts in Weibo.

For the claimed top 10,000 accounts, there are in total 6,069 accounts whose ratio is between -0.2 and 0.2, 52 is smaller than -0.2, and 3,930 is larger than 0.2. The minimal ratio value is -0.482, while the highest is 132.8413. Overall the total number of claimed followers is 23% more than the estimated followers. The claimed follower numbers are taken at the end of the experiment, while the whole

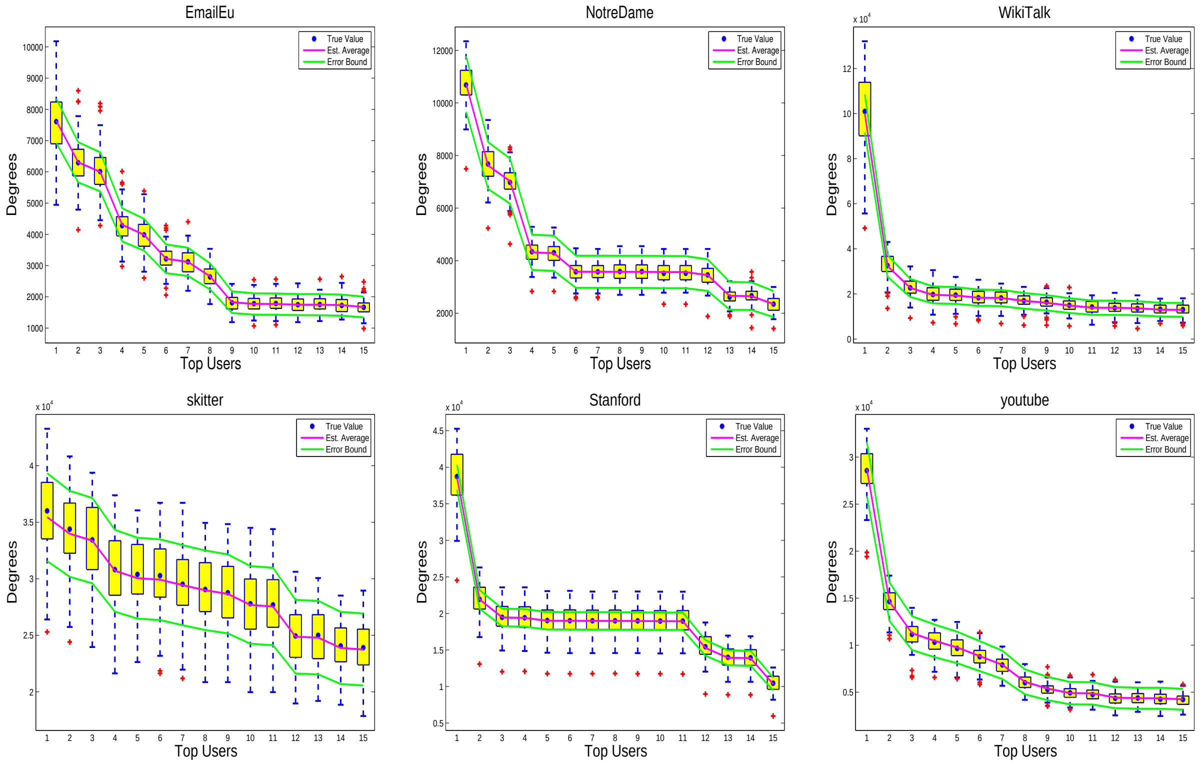

Fig. 3. **Degree estimation of six networks using star sampling. Boxplots are obtained from 100 repeated experiments.**

sampling process spans a few months. Given the dynamic nature of the network, especially the fast increasing number of new accounts and followers, it is understandable that overall the claimed number is higher than the estimated number.

However, there are many accounts with very high inflation rate. The inflation rate for all 10,000 accounts are plotted in Figure 4 (A), where the accounts are sorted by their claimed follower numbers in decreasing order. Figure 4 (B) is the corresponding smoothed ration plot using moving average with 500 window size. Those two plots show that there are higher inflation rate for the accounts ranked between 1,000 to 2,000. In general, the inflation rate drops for smaller accounts. Altogether, there are 194 accounts whose ratios are greater than five (1,342 accounts greater than one), a very large discrepancy that is hard to explain. We plot those 1,342 accounts in log-log scale in Figure 4 (C). Interestingly enough, the inflation rate also follows power law.

Figures 4 (D) depicts the comparison between the estimated and claimed follower numbers for the top 500 accounts. There are spikes pointing downwards, indicting the accounts having very low estimation, or high inflation rate. Figure 4(E) gives an overall picture of all the $10^4/$ accounts, both lines are smoothed using window size 100 using log-log plot. Figure 4(F) is the smoothed difference (claimed -estimated) for all the accounts. The smoothing window size is also 100.

Lastly, we draw the correlations between the claimed and estimated followers in Figure 5. The inflated number does not change greatly the overall landscape of the rankings of the accounts. The estimation is closely related to the claimed number as evidenced by the high Pearson correlation coefficient 0.9797. However, it is also clear from the plot that some accounts deviate a lot from the estimations.

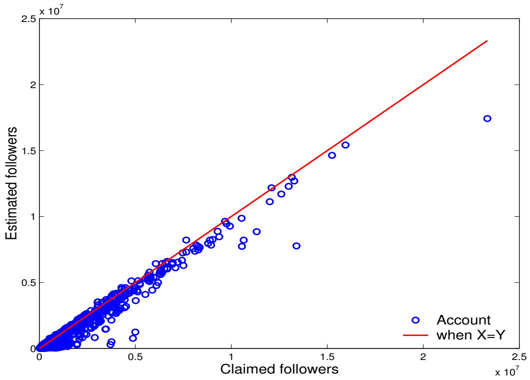

Fig. 5. **Estimated followers vs. claimed followers in log-log scale. The Pearson correlation coefficient is 0.9797.**

From these analyses, it seems that some accounts have their follower numbers artificially inflated, while most of the accounts, especially the smaller ones, have the follower number consistent with our estimation.

V. Discussions and Conclusions

This paper proposes the star sampling to estimate properties of online social networks. Some network properties

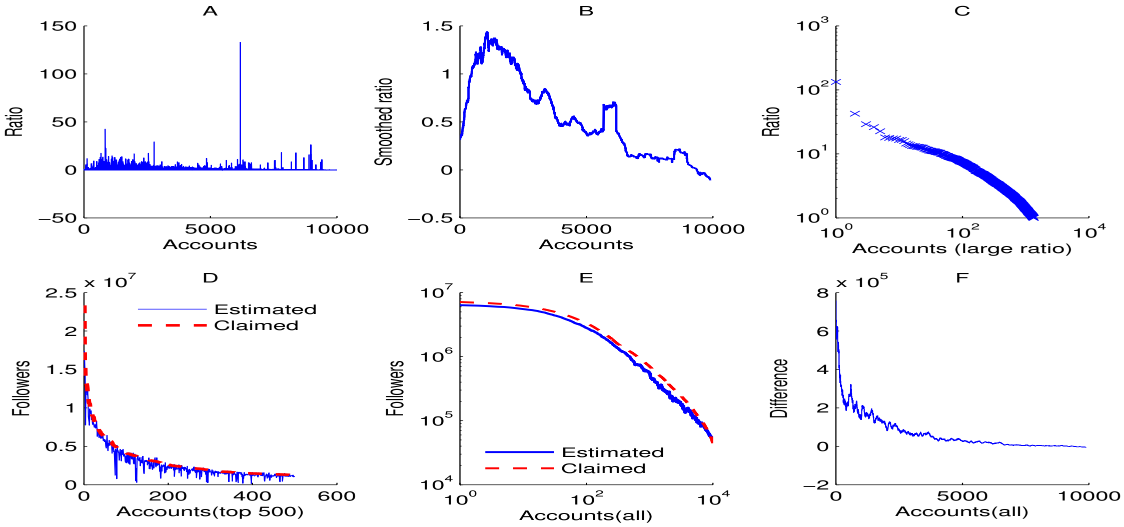

Fig. 4. **Weibo followers estimation for the top 10^4 accounts. Panel A: inflation ratio over 10^4 top accounts. Panel B: the smoothed version of A. Panel C: All the accounts whose inflation ratio is higher than one. Panel D: top 500 accounts. Panel E: comparison of top 10^4 accounts, smoothed. Panel F: difference between the claimed and estimated followers. Smoothed.**

prefer PPS sampling, which is not easy to carry out for most online social networks. Random walk can approximate PPS sampling, but it collects only one random node in its neighbours. Star sampling improves the performance of random walk sampling by a factor of $\langle d \rangle$, the average degree of the network. We demonstrate on six local datasets that star sampling approximates PPS sampling very well.

We then applied the star sampling to explore Weibo, the Chinese version of Twitter that has 243 million accounts in 2011. We find that Weibo is a power-law network, and has similar degree distribution as Twitter. In particular, we find that there are some accounts whose claimed follower numbers are much higher than our estimations.

We will apply the star sampling to discover other network properties, such as community structure of top bloggers.

VI. Acknowledgement

This research is supported by NSERC (Natural Sciences and Engineering Research Council of Canada).

References

[1] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, pages 835–844. ACM, 2007.

[2] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] R. Bond and et al. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[5] S. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Crawling facebook for social network analysis purposes. *Arxiv preprint arXiv:1105.6307*, 2011.

[6] K.-w. Fu and M. Chau. Reality check for the chinese microblog space: a random sampling approach. *PLOS ONE*, 8(3):e58356, 2013.

[7] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. A walk in facebook: Uniform sampling of users in online social networks. *Arxiv preprint arXiv:0906.0060*, 2009.

[8] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. 2008.

[9] C. Hubler, H. Kriegel, K. Borgwardt, and Z. Ghahramani. Metropolis algorithms for representative subgraph sampling. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 283–292. IEEE, 2008.

[10] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[11] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600. ACM, 2010.

[12] S. Lee, P. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73(1):016102, 2006.

[13] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *SIGKDD*, pages 631–636. ACM, 2006.

[14] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.

[15] J. Lu and D. Li. Bias correction in small sample from big data. *TKDE, IEEE Transactions of Knowledge and Data Engineering, in Press*, 2013.

[16] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[17] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *SIGCOMM*, pages 29–42. ACM, 2007.

[18] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.

[19] M. Montemurro. Beyond the zipf–mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3):567–578, 2001.

[20] N. Perlroth. Fake twitter followers become multimillion-dollar business. 2013.

[21] M. Stumpf and C. Wiuf. Sampling properties of random graphs: the degree distribution. *Physical Review E*, 72(3):036118, 2005.

[22] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August 2009.

[23] J. Zhou, Y. Li, V. Adhikari, and Z. Zhang. Counting youtube videos via random prefix sampling. In *SIGCOMM*, pages 371–380. ACM, 2011.