Sampling online social networks by random walk

Jianguo Lu[1], Dingding Li [2]

[1] School of Computer Science, University of Windsor
[2] Department of Economics, University of Windsor
Email: {jlu, dli}@uwindsor.ca
401 Sunset Avenue, Windsor, Ontario N9B 3P4. Canada

ABSTRACT

This paper proposes to use simple random walk, a sampling method supported by most online social networks (OSN), to estimate a variety of properties of large OSNs. We show that due to the scale-free nature of OSNs the estimators derived from random walk sampling scheme are much better than uniform random sampling, even when uniform random samples are available disregarding the notorious high cost of obtaining the random samples. The paper first proposes to use harmonic mean to estimate the average degree of OSNs. The accurate estimation of the average degree leads to the discovery of other properties, such as the population size, the heterogeneity of the degrees, the number of friends of friends, the threshold value for messages to reach a large component, and Gini coefficient of the population. The method is validated in complete Twitter data dated in 2009 that contains 42 million nodes and 1.5 billion edges.

Keywords

OSN, Online Social Network, Hansen-Hurwitz, Estimator, Scale free network, Harmonic mean

1. INTRODUCTION

The properties of online social networks are of great interests to general public as well as IT professionals. Yet the raw data are usually not available to the public and the summary released by the service providers is sketchy. Thus sampling is needed to reveal the hidden properties or structure of the underlying data [5, 20, 13].

For instance, we may want to learn the average number of friends in a network, or the average degree of a graph. One obvious but often impractical method is to select randomly a set of users $\{U_1, U_2 \ldots, U_n\}$, count their degrees $\{d_1, \ldots, d_n\}$ for each user, and calculate the sample mean as the estimate of the population mean:

$$\widehat{d}_{SM} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad (1)$$

The sample mean estimator $\widehat{d}_{SM}$ is an unbiased estimator of the population, if the users can be selected randomly with equal probability. Unfortunately this is not the case in most practice. When micro bloggers are selected, they are often not picked randomly due to the limited access methods provided by OSN sites. Rather, more popular bloggers tend to have a higher probability of being sampled if users are crawled by following the links.

There are studies on the sampling methods for OSN [5, 20] and in related areas such as social networks [22, 26], graphs [13, 25], web URLs [8], and search engine index and deep web [1, 17, 16]. The typical underlying techniques include Metropolis Hasting Random Walk (MHRW) [18] for uniform sampling and Random Walk (RW) [14] for unequal probability sampling.

A random walk on graph follows one of the links with an equal probability among all the links. A blogger with more followers will have higher probability of being sampled. It is well known that the asymptotic probability of a node being sampled is proportional to its degree [14]. Therefore, the sample mean tends to overestimate the population average degree.

MHRW is reported rather good at obtaining a random sample of random networks. However, in the sampling process many nodes are retrieved, examined, and rejected. The cost is rather high especially for OSN where the node retrieval needs network traffic and usually there are quota for daily accesses.

Even when uniform random samples are obtained, the sample mean estimator has a high variance because the degree distribution of OSNs usually follows power law. Many nodes have small degrees, while some nodes may have very large degree. The inclusion/exclusion of a super large node in a sample will make the estimates diverge.

When uniform random samples are hard to obtain, it is rather common to use PPS (Probability Proportional to Size) sampling and Hansen-Hurwitz related estimators [7]. In particular, the harmonic mean instead of the arithmetic mean of the sample can be used as the estimator of the average degree of OSN:

$$\widehat{d}_H = n \left[\sum_{i=1}^{n} \frac{1}{d_i} \right]^{-1} \tag{2}$$

Here the subscript H indicates that it is the harmonic mean, and that it can be derived from the traditional Hansen-Hurwitz estimator as described in the next section. For this estimator the sample is obtained by simple random walk, resulting in the node selection probability proportional to its degree. This estimator was first derived and studied in depth by Salganik et al. [22] to estimate the properties of hidden population such as drug-addicts. In that setting the true values are unknown, the assumptions such as sampling probability are flimsy, thus the veracity of the estimator is impossible to evaluate.

In the context of OSN, Kurant et al. [11, 5, 6] studied various sampling methods, including random walk, to discover network properties such distribution of node degrees. [5] studied the sampling of Facebook, in particular the Re-Weighted Random Walk that can be also traced back to Hansen-Hurwitz estimator. [11] mentioned harmonic mean estimator, but fell short of the analysis and comparison of the estimator.

Rasti et al. [21] studied re-weighted random walk sampling in peer-to peer networks. Both [5] and [21] compare their methods with Metropolis-Hasting random walk, not uniform random samples. The comparison to uniform random samples was conducted in [10] for the estimation of population size not average degree.

This is the first paper to show that in a real large network the harmonic mean estimator is much better than sample mean estimator in uniform random samples, even ignoring the cost of obtaining the uniform samples. In practice as demonstrated in Twitter network, the sample size can be thousands times smaller than uniform random samples to achieve similar accuracy. In theory, the improvement can be unlimited with the growth of the network size.

The *contributions* of this paper are 1) the properties of the estimator (bias and variance) are analyzed and empirically verified in a large real network; 2) the advantage over uniform random sample is analyzed and compared. In particular we found that in Twitter data the estimator is much better–it has a very small bias, and the variance is orders of magnitude smaller than the sample mean estimator; 3) the cause is identified as the heterogeneity of the data induced by the scale-free nature of the network. Coefficient of variation is proposed to quantify the heterogeneity; 4) the accurate estimation of the average degree can lead to the discovery of a string of other network properties such as the network size, the heterogeneity of the degrees, the threshold value for message diffusion, and the inequality of the friends in the network.

We want to emphasize that our method is not limited to the estimation of direct connections between users in OSN. The average degree can be the average number of friends in the case of Facebook or Linkedin, or average followers and followees in Twitter and Weibo networks. In addition to such explicit graph where edges represent the following (or friend) relationships, in OSNs there are implicitly derived graphs where an edge exists if two nodes share messages, groups, etc.., resulting in message networks and group networks. In a message network, two persons are linked if they shared a message. In group network, two persons are connected if they belong to the same group. Thus, the degree can represent the direct connections to friends, the number of message reposts on the network, or the number of groups people are associated with.

2. ESTIMATORS

2.1 Sample mean estimator

Suppose that in the population there are N number of users. Each user has a property $Y_i, i \in \{1, 2, \dots, N\}$, which can be age, number of friends, or number of messages etc..

Let the population total is $\tau = \sum_{i=1}^{N} Y_i$, and population mean is $\overline{Y} = \tau/N$. Our task is to estimate $\overline{Y}$ using a sample. In particular, this paper focuses on the degree property, i.e., estimating the average degree $\overline{d}$ using a sample $\{d_1, d_2, \dots, d_n\}$.

If a uniform random sample $Y_1, \dots, Y_n$ is obtained, the sample mean is an unbiased estimator as defined below:

$$\widehat{Y}_{SM} = \frac{1}{n} \sum_{i=1}^{n} Y_i \tag{3}$$

When Y_i is the degree of node i, i.e., $Y_i = d_i$, the above equation becomes the sample mean estimator for degrees:

$$\widehat{d}_{SM} = \frac{1}{n} \sum_{i=1}^{n} d_i \tag{4}$$

The variance of the estimator $\widehat{d}_{SM}$ is [24]

$$var(\widehat{d}_{SM}) = \frac{N-n}{N} \frac{\sigma^2}{n} \tag{5}$$

where σ^2 is the population variance for degrees that can be calculated by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} d_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} d_i\right)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N} d_i^2 - \overline{d}^2 \qquad (6)$$

where $\overline{d}$ is the arithmetic mean of all the degrees in the total population.

The estimated variance of the estimator $\widehat{d}_{SM}$ is

$$\widehat{var}(\widehat{d}_{SM}) = \frac{N-n}{N}\frac{s^2}{n} \qquad (7)$$

where s^2 is the sample variance of $d_1, d_2, \ldots, d_n$.

The problem with this sample mean estimator is that uniform sample is not easy to obtain. Moreover, the population variance σ^2, and consequently the estimator variance, are large due to the scale-free nature of the network. The degree distribution in online social networks follows power law or Zipf law. That is, if we rank all the nodes according to their degrees in decreasing order $(d_1, d_2, \ldots, d_N)$, then

$$d_i = \frac{A}{i^\alpha}, \qquad (8)$$

where A and α are constants. α is called the exponent or slope that is typically around one in various scale-free networks [1].

With such degree distribution the population variance is very large, leading to large variance of the sample mean estimator. Suppose that $\alpha = 1$, which is typical for many scale free networks [19] including Twitter network [12]. σ^2 can be approximated as below by combining Equations 8 and 6:

$$\sigma^2 = E(X^2) - E^2(X)$$
$$= \left(\frac{E(X^2)}{E^2(X)} - 1\right)E^2(X)$$
$$= \left(\frac{N\sum_{i=1}^{N} d_i^2}{(\sum_{i=1}^{N} d_i)^2} - 1\right)\overline{d}^2$$
$$= \left(\frac{N\sum_{i=1}^{N} i^{-2}}{(\sum_{i=1}^{N} i^{-1})^2} - 1\right)\overline{d}^2$$
$$\approx \left(\frac{N}{ln^2 N} - 1\right)\overline{d}^2 \qquad (9)$$

It shows that the variance does not converge when the network size N grows to the limit.

2.2 Harmonic mean estimator

When sampling probability is not equal for each unit, a common approach is to use Hansen-Hurwitz estimators. One of them is to estimate the population total [24]:

$$\widehat{\tau}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{p_i}, \qquad (10)$$

where p_i is the selection probability of unit i, $\tau = \sum_{i=1}^{N} Y_i$ is the population total, and $\sum_{i=1}^{N} p_i = 1$. Selection probability of unit i is the probability it is selected in one draw of the sample elements. Note that Hansen-Hurwitz estimator is used when sampling with replacements, i.e., a unit can be sampled multiple times just the same as in random walk sampling.

When $Y_i = 1$ for all $i \in \{1, 2, \ldots, N\}$, the above estimator is reduced to another version of Hansen-Hurwitz estimator that estimates the total number of nodes $N = \sum_{i=1}^{N} Y_i$:

$$\widehat{N}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i} \qquad (11)$$

In our OSN case, samples are often obtained by random walk. It is well known that random walk obtains a biased sample. Asymptotically the probability of a user being visited in a random walk is proportional to its degree, i.e., in the case of random walk,

$$p_i = \frac{d_i}{\sum_{j=1}^{N} d_j} = \frac{d_i}{\tau} \qquad (12)$$

Therefore an estimator for degree mean $\widehat{d_H}$ can be derived from the unbiased Hansen-Hurwitz estimator for N as follows:

$$\widehat{d_H} = \frac{\tau}{\widehat{N}_{HH}}$$
$$= \tau\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\tau}{d_i}\right]^{-1}$$
$$= n\left[\sum_{i=1}^{n}\frac{1}{d_i}\right]^{-1} \qquad (13)$$

The estimator for the arithmetic mean degree turns out to be the harmonic mean of the degrees in the sample. Salganik et al [22] gave a similar derivation using the ratio of two estimators in the setting of respondent driven sampling.

Although $\widehat{N}_{HH}$ is an unbiased estimator, its inverse may not be unbiased. Cochran [3] showed that the bias is on the order of $1/n$. Since the sample size n in social network sampling is rather large in general, the bias is negligible.

[1] Note that there are two ways to describe the property of power law, one using the Zipfian approach as used here, the other is the frequency of the degrees that is equivalent to Zipfian approach except that the exponent is greater by one.

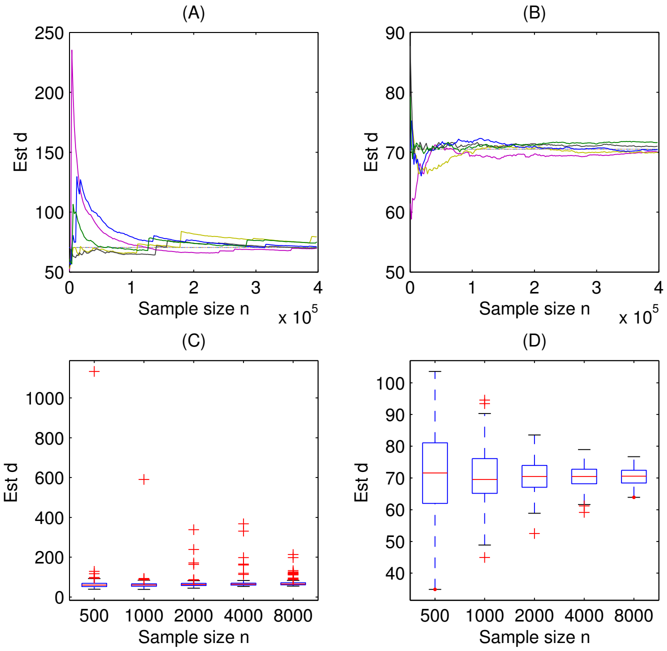

Figure 1: Comparison of $\widehat{d}_{SM}$ in UR (Uniform Random) sampling and $\widehat{d}_H$ in RW (Random Walk) sampling. Panels A (for UR) and B(for RW) show that the estimation fluctuates with the increase of sample size. Panels C (for UR) and D (for RW) show the box plots of 100 estimations for sample sizes ranging between 500 and 8000.

The variance of $\widehat{N}_{HH}$ is

$$var(\widehat{N}_{HH}) = \frac{1}{n}\sum_{i=1}^{N} p_i(1/p_i - N)^2 \qquad (14)$$

It can be estimated from a sample using

$$\widehat{var}(\widehat{N}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(1/p_i - N)^2 \qquad (15)$$

Using Delta method the variance of estimator d_H is

$$\widehat{var}(\widehat{d_H}) = \frac{s_v^2}{\overline{v}^4 n} \qquad (16)$$

where $v_i = 1/d_i$, $\overline{v}$ and s_v^2 are the sample mean and variance of v_i's. This equation will be used in calculating the error bound in Figure 2.

3. EXPERIMENTS

3.1 Data

The estimator is verified on the Twitter network data that are provided by Kwak et al. [12], characterizing the complete Twitter network as of July 2009. The data contain about 1.47 billion edges and 41.7 million nodes or users, occupying around 20 gigabytes hard drive space. Since they are too large to fit into the memory of commodity computers, we index them using Lucene, a popular index engine. Then the random walk and uniform random sampling are performed on the index that are stored in hard drive. Since our method is better to be used in undirected graph, we remove the direction in Twitter data.

Table 1: Empirical bias and standard error of the two estimators over 100 runs for various sample size n.

n	Bias		Standard error	
	UR	RW	UR	RW
500	2.6295	1.1444	108.1054	12.0539
1000	-4.1512	0.1016	53.8785	8.7383
2000	-0.8226	-0.0320	36.2923	5.6482
4000	4.0328	-0.2842	45.0989	4.1571
8000	2.1037	-0.0674	25.1908	2.7238

3.2 Results

Two estimators, $\widehat{d}_{SM}$ in Equation 1 and $\widehat{d}_H$ in Equation 13, are tested on the data for five different sample sizes 500, 1000, 2000, 4000, and 8000. For each sample size 100 samples are selected using uniform random sampling and random walk sampling respectively. Their bias and standard errors are tabulated in Table 1.

It shows that indeed $\widehat{d}_H$ has a very small bias as expected. What is striking is that its standard error is much smaller than $\widehat{d}_{SM}$.

We use Figure 1 to explain the result further. Panels C is the box plot for $\widehat{d}_{SM}$ using uniform sampling. It shows that the estimation fluctuates very much, can even go as high as 1000 when n=500, where the true mean is 70.5. The big variance problem is ameliorated slightly but remains large with the growth of the sample size.

On the other hand the box plot for $\widehat{d}_H$ in Panel D shows much smaller variance.

We also run five large samples, each with size 4 × 10^5, as depicted in panels A (for UR) and B (for RW). Note that in the case of uniform random sampling, the estimate jumps from time to time even when the sample size is rather large.

Figure 2 shows four estimations bounded by the 95% confidence interval calculated by Equation 16.

3.3 Discussions

This paper shows that the biased sampling is much better than uniform sampling for the estimation of average degrees. In the past, people try to obtain uniform samples whenever possible, and resort to biased

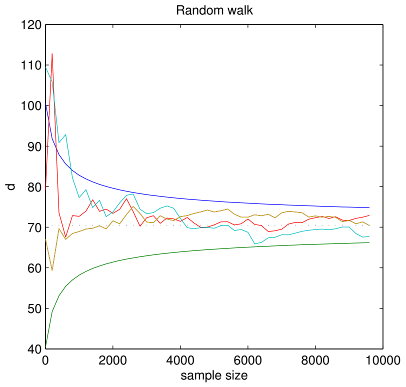

Figure 2: 95% confidence interval and four RW (Random Walk) estimation processes using $\widehat{d}_H$ estimator. The error bound is drawn from Equation 16.

sampling such as PPS (Proportional To Size) sampling only when uniform sampling is impossible [22] or costly. The results of this paper suggest that in the context of online social networks, random walk sampling instead of uniform sampling should be used, even when uniform random samples are readily accessible.

It is easy to understand that the variance of uniform random estimator $\widehat{d}_{SM}$ is large because online social networks are mostly scale-free as shown in Equation 9. The smaller variance of $\widehat{d}_H$ can be explained below.

Let d^W be the random variable for the degrees sampled by random walk. First we draw its empirical distribution and its comparison with uniformly sampled degrees in Figure 3. Uniform random (UR) samples resemble the distribution of the total population [23] that obeys power law with exponent around one. On the other hand, in random walk (RW) sampling scheme d^W has a flatter starting section and a drooping tail, which can be approximated by the Mandelbrot law:

$$d_i^W = \frac{B}{(a+i)^b} \tag{17}$$

where b is the exponent, B is a normalization constant, a is a constant that corresponds to the position where the curve droops down.

Let

$$\overline{v} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_i^W}. \tag{18}$$

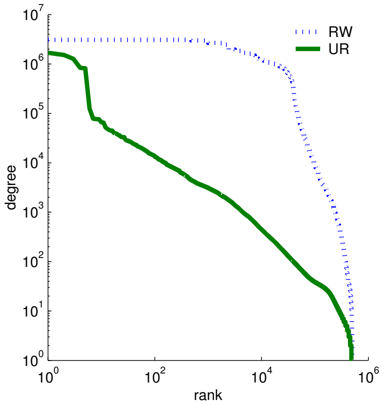

Figure 3: The degree distributions of the samples obtained from UR (Uniform Random) and RW (Random Walk) samplings. n=500,000. The nodes, including the ones being repeatedly sampled, are ranked in decreasing order of their degrees, and drawn with degrees against their ranks.

The variance of the reciprocal of the variable is

$$\begin{aligned}
var(1/d^W) &= \left(\frac{n\sum_{i=1}^{n}(i+a)^{2b}}{\left(\sum_{i=1}^{n}(i+a)^b\right)^2} - 1\right)\overline{v}^2 \\
&= \left(n\sum_{i=1}^{n}(i+a)^{2b}\left(\sum_{i=1}^{n}(i+a)^b\right)^{-2} - 1\right)\overline{v}^2 \\
&\approx \left(n\left[\frac{1}{2b+1}n^{2b+1}\right]\left[\frac{1}{b+1}n^{b+1}\right]^{-2} - 1\right)\overline{v}^2 \\
&\approx \left(\frac{(b+1)^2}{2b+1} - 1\right)\overline{v}^2
\end{aligned}$$

Thus $var(1/d^W)$ is a constant that does not grow with the population size as σ^2 does.

4. IMPLICATIONS

Average degree plays a pivotal role in discovering other properties of a large network. Its accurate estimation has a ramification on a string of other hidden properties of large networks. One immediate result is the total number of edges in the graph when user size is known. However, the more profound consequence is that we can discover the heterogeneity, CV (Coefficient of Variation), of the entire network with a small sample using average degree. The discovery of CV will in turn deduce other properties such as the total number of users, the inequality of degrees (friends of friends and Gini coefficient).

4.1 Estimate heterogeneity

$\overline{d}$ can be used to estimate CV, Coefficient of Vari-

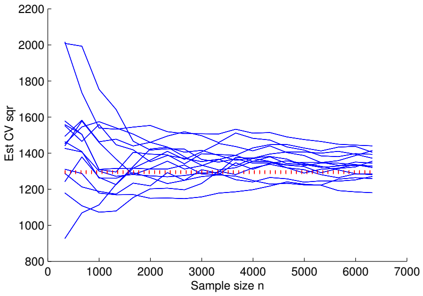

Figure 4: 15 Estimation processes of γ^2 in Twitter data using Equation 20. The red dotted line is the true value.

ation (denoted as γ), that is an important metric to measure the heterogeneity of degree distribution. It is defined as the standard deviation normalized by the average degree: $\gamma^2 = \sigma^2/\overline{d}^2$. Expanding the definition for variance we have

$$\begin{aligned}\gamma^2 + 1 &= \frac{\overline{d^2} - \overline{d}^2}{\overline{d}^2} + 1\\ &= \frac{1}{N}\sum_{i=1}^{N} d_i^2 \left[\frac{1}{N}\sum_{i=1}^{N} d_i\right]^{-2}\\ &= N\sum_{i=1}^{N} d_i^2 \left[\sum_{i=1}^{N} d_i\right]^{-2}\end{aligned}$$

On the other hand the sample mean of the degrees obtained by random walk is

$$\begin{aligned}\overline{d^W} &= \frac{1}{n}\sum_{i=1}^{n} d_i^W\\ &= \sum_{i=1}^{N} p_i d_i\\ &= \frac{1}{N\overline{d}}\sum_{i=1}^{N} d_i^2 \qquad (19)\end{aligned}$$

Combining the two equations we derive the estimator for CV as follows:

$$\widehat{\gamma}^2 + 1 = \frac{\overline{d^W}}{\overline{d}}, \qquad (20)$$

where $\overline{d^W}$ is the sample mean of the degrees obtained by random walk, $\overline{d}$ can be estimated by the arithmetic mean of the same data. The convenience of the method is that only one random walk is needed. Figure 4 shows 15 estimates that converge quickly with the growth of the sample size.

4.2 Population estimation

Once γ^2 is available, it can be used to estimate the population size as follows, which is a special case of Eq 3.20 in [2]:

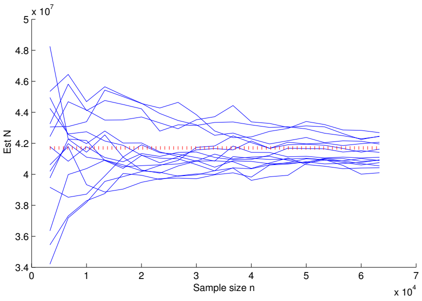

Figure 5: 15 estimation processes of twitter accounts N using Equation 21. Red dotted line is the true value.

$$\widehat{N} = (\gamma^2 + 1)\binom{n}{2}\frac{1}{C}, \qquad (21)$$

where n is the sample size, C is the number of collisions, and the sample is obtained by random walk [2]. In the area of capture-recapture research [2, 17, 15], it has been a perplexing problem for the population estimation of heterogeneous data whose capture probabilities are unequal, mainly due to the difficulty of estimating the heterogeneity. Now in the setting of OSN, the problem is solved thanks to the estimator $\widehat{d}_H$.

Because of the accurate predication of the heterogeneity of the data (γ^2), the estimation of population size is rather good as shown in Figure 5. Since this estimator hinges on collision times, extra caution should be taken to avoid spurious collisions caused by random walk. For instance if a node A is only connected to node B, a visit to A will cause node B visited twice. To avoid such loops, we take samples spaced every a few steps apart.

4.3 Other properties

4.3.1 Friends of friends

γ^2 can be also used to measure the ratio between the number of friends of your friends , and the number of your friends. As the saying goes, your friends have more friends than you do. To be more precise, your friends have $\gamma^2 + 1$ times more friends than you do.

The mean number of friends of friends is [4]

$$\sum_{i=1}^{N} d_i^2 / \sum_{i=1}^{N} d_i = \overline{d} + \sigma^2/\overline{d} \qquad (22)$$

[2] Here is a simple derivation for the estimator. The expected number of collisions is

$$E(C) = \binom{n}{2}\sum_{i=1}^{N} p_i^2 = \binom{n}{2}\frac{1}{\tau^2}\sum_{i=1}^{N} d_i^2 = \binom{n}{2}\frac{\gamma^2+1}{N}$$

The above equation shows that your friends have no less than the friends you have. Simple rearranging the equation results in:

$$\begin{aligned}\frac{\sum_{i=1}^{N} d_i^2/\sum_{i=1}^{N} d_i}{\overline{d}} &= 1+\sigma^2/\overline{d}^2 \\ &= 1+\gamma^2 \end{aligned} \tag{23}$$

In words, the equation says that on average your friends have $1+\gamma^2$ times more friends then you do. In a homogeneous network where everybody has the same number of connections, $\gamma = 0$, thus your friends have the same number of friends as you do. In twitter society, γ^2 is around 1000, thus your friends have a thousand times more friends than you do.

4.3.2 Message diffusion

Along the same line γ^2 can be used to quantify the diffusion of messages that is borrowed from epidemiology. In particular, it can be derived that the threshold for the occurrence of large component, or the occurrence of epidemics [9] (Eq 7.8) is

$$\pi = \frac{(\gamma^2+1)\overline{d}-2}{(\gamma^2+1)\overline{d}-1}, \tag{24}$$

where π is the proportion of the nodes that are immuned uniformly from the network.

4.3.3 Clustering Coefficient

Some structural network properties can be also derived using γ^2. For instance, one important network property is Clustering Coefficient, indicating the proportion whether your friend of friend is also your friend. It is hard to calculate directly for a large network, but can be estimated [19] (eq 13.47) by

$$\overline{d}\gamma^4/n. \tag{25}$$

4.3.4 Gini coefficient

Gini index is used to measure the inequality of wealth. It can also be used to measure the inequality of friendships in OSNs. Using $\overline{d}$ the Gini coefficient can be approximated by

$$\widehat{G} = \frac{1}{2n(n-1)\overline{d}} \sum_{i=1}^{n}\sum_{j=1}^{n} |d_i - d_j| \tag{26}$$

The classic problem of Gini coefficient estimation is that the mean is hard to obtain. Thanks to the estimation of average degree, in Twitter network, we find its Gini coefficient is around 0.70-0.82.

5. CONCLUSIONS

This paper proposes to use random walk to sample a network and use the harmonic mean to estimate the average degree. The empirical experiments show that the estimator is much better even than uniform random samples.

The method is very practical in that in thousands or even hundreds of steps of random walk we can learn the average degree of a large network containing tens of millions of nodes and billions of edges.

The method works well because of the scale-free nature of the underlying network where the variance tends to be very large, potentially unlimited when the network size becomes infinitely large. For such networks, we analytically showed that the harmonic mean estimator removed the large variance problem.

Therefore the estimator works not only for online social networks, but also any scale-free networks that are ubiquitous and more common than random networks. For instance, we also validated the estimator in document-term graph where document and terms are nodes, and they are connected if a document contains a term.

The method relies on the assumption that random walk produces samples whose selection probability is proportional to their degrees. Theoretically this is true only asymptotically. Therefore the samples before the mixing time should be thrown away. Our experiments show little difference whether or not to include the first batch of samples in the random walk.

The degree estimation is not only important by itself but also crucial for discovering other network properties. The success solution of average degree can lead to the discovery of the heterogeneity of the underlying data, the user and link size etc.

The method is not restricted to the degrees of the explicit networks where the edges are the friendship relations. Instead, the edges can be forged by other implicit relations, such as sharing the same message.

6. ACKNOWLEDGEMENTS

We thank the reviewers for their detailed comments, and the support from NSERC (Natural Sciences and Engineering Research Council of Canada) and SSHRC (Social Sciences and Humanities Research Council of Canada).

7. REFERENCES

[1] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. *Journal of the ACM (JACM)*, 55(5):24, 2008.

[2] A. Chao, S. Lee, and S. Jeng. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, pages 201–216, 1992.

[3] W. Cochran. *Sampling techniques.* Wiley-India, 2007.

[4] S. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, pages

1464–1477, 1991.

[5] 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.

[6] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1872–1892, 2011.

[7] M. Hansen and W. Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 1943.

[8] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. *Computer Networks*, 33(1-6):295–308, 2000.

[9] M. Jackson. *Social and economic networks.* Princeton Univ Pr, 2008.

[10] L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *Proceedings of the 20th international conference on World wide web*, pages 597–606. ACM, 2011.

[11] M. Kurant, A. Markopoulou, and P. Thiran. Towards unbiased bfs sampling. *Selected Areas in Communications, IEEE Journal on*, 29(9):1799–1809, 2011.

[12] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[13] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 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. Efficient estimation of the size of text deep web data source. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1485–1486, Napa Valley, California, USA, 2008. ACM.

[16] J. Lu. Ranking bias in deep web size estimation using capture recapture method. *Data & Knowledge Engineering*, 69(8):866–879, 2010.

[17] J. Lu and D. Li. Estimating deep web data source size by capture–recapture method. *Information retrieval*, 13(1):70–95, 2010.

[18] 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.

[19] M. Newman. *Networks: an introduction.* Oxford University Press, Inc., 2010.

[20] M. Papagelis, G. Das, and N. Koudas. Sampling online social networks. *Knowledge and Data Engineering, IEEE Transactions on*, (99):1–1, 2011.

[21] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *INFOCOM 2009, IEEE*, pages 2701–2705. IEEE, 2009.

[22] M. Salganik and D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1):193–240, 2004.

[23] M. Stumpf, C. Wiuf, and R. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4221, 2005.

[24] S. Thompson. *Sampling.* Wiley, 2012.

[25] T. Wang, Y. Chen, Z. Zhang, T. Xu, L. Jin, P. Hui, B. Deng, and X. Li. Understanding graph sampling algorithms for social network analysis. In *the 3rd ICDCS Workshop on Simplifying Complex Networks for Practitioners*, 2011.

[26] C. Wejnert and D. Heckathorn. Web-based network sampling. *Sociological Methods & Research*, 37(1):105–134, 2008.