Variance Reduction In Large Graph Sampling

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

Abstract

The norm of practice in estimating graph properties is to use uniform random node (RN) samples whenever possible. Many graphs are large and scale-free, inducing large degree variance and estimator variance. This paper shows that random edge (RE) sampling and the corresponding harmonic mean estimator for average degree can reduce the estimation variance significantly. First, we demonstrate that the degree variance, and consequently the variance of the RN estimator, can grow almost linearly with data size for typical scale-free graphs. Then we prove that the RE estimator has a variance bounded from above. Therefore, the variance ratio between RN and RE samplings can be very large for big data. The analytical result is supported by both simulation studies and 18 real networks. We observe that the variance reduction ratio can be more than a hundred for some real networks such as Twitter. Furthermore, we show that random walk (RW) sampling is always worse than RE sampling, and it can reduce the variance of RN method only when its performance is close to that of RE sampling.

Keywords: Uniform random sampling, Random walk, Graph sampling, Online Social Network, Scale-free network, Harmonic mean.

1. Introduction

The data on the Web and online social networks can be often viewed as graphs. The data in its entirety may not be available for various reasons. They can be distributed over many machines (e.g., the Web and P2P networks), hidden behind searchable interfaces (e.g., search engines and online social networks), scattered among a larger graph (e.g., various communities in online social networks). Regardless of the causes, a common challenge is to reveal the properties of such datasets when we do not own the entire data. In the past, extensive research was carried out to explore the profile of search engines [17] and other data collections [4, 6, 33]. Most of them focused on obtaining uniform random node (RN) samples, such as uniform random web pages from the Web [11] and search engines [2], and uniform random bloggers from online social networks [9]. Once uniform random samples are obtained, network properties, in particular the attributes of the nodes including average degree, could be estimated with statistical guarantee.

In many cases, RN sampling works only in theory. The majority of real world networks are scale-free [3], whose degree distributions follow a power law. Such scale-free networks often induce a large variance of the degrees. In theory, the variance does not exist when the exponent of the power law falls in certain range. In practice, the variance can be extremely high for very large networks. For instance, the coefficient of variation of the Twitter user network collected in 2009 [16] is as high as 35.95. To understand the impact of such a high coefficient of variation, let us have a quick calculation for the sample size needed to reach 20% accuracy for its average degree 70.51. More precisely, to make sure that the estimation is within the range of 70.51 ± 14.10 with 95 % confidence, the relative standard error RSE should be around 0.1, and the required sample size $n = 35.95^2 * /(0.1)^2 = 129,240$. In addition, uniform random samples are obtained with high cost because they are not provided directly by the data sources. Costly sampling methods, such as rejection sampling, have to be employed to obtain uniform samples. In the process, many samples are retrieved and rejected as invalid. The actual samples retrieved are many times larger than $129,240$, depending on the sampling methods allowed. Considering the network traffic involved and the daily quota imposed by the service provider, it is prohibitive to use uniform random sampling to obtain meaningful estimations. With increasingly more applications of big data analyses, there is an urgent need to find a method to reduce such a large variance.

Recent developments made empirical observations that simple random walk (RW) sampling or its extensions can improve degree estimator performance for P2P networks [30], Facebook user network [9], Twitter user network [22], and term-document bipartite graphs [39]. Similar empirical observations are made for node size estimation [13, 15]. We find that these observations are data dependent. Random walk can be much worse than uniform random sampling for other datasets, even when the graph is scale-free and the variance is very high as we will show in Section 4.3.

We find that it is random edge (RE) sampling, not RW sampling, that reduces the variance for graphs with large degree variation. In addition to this empirical observation, we explore the reason why RE outperforms RN sampling, and why RW does not. While it is easy to understand that uniform random sampling does not work well for scale-free networks, it was not clear whether RE sampling works better.

This paper shows that the variance of the RE estimator is bounded from above by a polynomial in the average degree and sample size. It implies that the performance of RE sampling does not deteriorate with the growth of degree variance of the graph, thereby it guarantees the superiority of RE sampling when degree variance is large. This result is particularly important for large graphs whose variance becomes larger compared with smaller data with the same distribution. Improvement ratios as high as 100 are observed on Twitter and other networks. Such a large gap has implications for both practitioners and researchers. Practitioners can greatly save the estimation effort and give a worst case error bound. Researchers can devise new sampling methods that approximate RE sampling when it is not directly supported by the data source under investigation. For instance, random walk with restart [1] can succeed because it is similar to RE sampling and exploits the large gap between RN and RE sampling.

The major contribution of the paper is our development of the upper bound of the variance of RE estimator. The result holds independent of degree distribution and graph topology. We verify the result using both simulated datasets and 18 real world networks. A direct consequence of the upper bound is the improvement ratio between RN and RE methods. To illustrate that the ratio can be very large, we first demonstrate that degree variance (consequently RN estimator variance) can be in the order of $O(N/\ln^2 N)$ under the assumption of the power law distribution. Now that RE variance is upper bounded, the improvement ratio tends to be infinite when data size goes infinitely large. Finally, we show that random walk sampling can approximate the performance of RE sampling only when the conductance of the graph is not very small, or, when the graph is well-enmeshed.

In the following sections, we first introduce the background of the research in Section 2, including the sampling methods and their corresponding estimators, the related work, and its applications. Then in Section 3 we derive the variances of RN and RE estimators. By giving the upper bound of the variance of the RE estimator, we quantify the performance ratio between RN and RE methods. In Section 4, we verify our result on 18 real networks, and demonstrate that the performance of RW sampling depends on both degree variance and graph conductance.

2. Background and Related Work

2.1. RN, RE, and RW Sampling

Given an undirected graph $G(V, E)$ where V is the set of nodes, and E the set of edges. Let $|V| = N$. Nodes are labeled as $1, 2, \ldots, N$, and their corresponding degrees are $d_1, d_2, \ldots, d_N$. The volume of the graph is $\tau = \sum_{i=1}^{N} d_i$, the average degree is $\langle d\rangle = \frac{1}{N}\sum_{i=1}^{N} d_i = \tau/N$. The variance σ^2 of the degrees in the population is defined as:

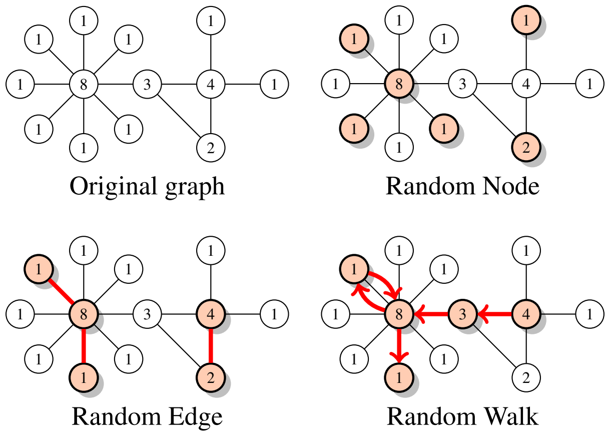
$$\sigma^2 = \langle d^2\rangle - \langle d\rangle^2, \tag{1}$$

where $\langle d^2\rangle = \sum_{i=1}^{N} d_i^2/N$ is the second moment, i.e., the arithmetic mean of the square of the degrees in the total population. The coefficient of variation (denoted as γ) is defined as the standard deviation, or the square root of the variance, normalized by the mean of the degrees:

$$\gamma^2 = \frac{\sigma^2}{\langle d\rangle^2} = \frac{\langle d^2\rangle}{\langle d\rangle^2} - 1. \tag{2}$$

Suppose a sample of n elements $(d_{x_1}, \ldots, d_{x_n})$ is taken from the population, where $x_i \in \{1, 2, \ldots, N\}$ for $i = 1, 2, \ldots, n$. Our task is to estimate the average degree $\langle d\rangle$ using the sample. Table 1 summarizes the notations used in this paper.

There are different ways to take the samples, notably by RN, RE, and RW samplings. In RN sampling, each node is sampled uniformly at random with replacement. In RE sampling, edges are selected with equal probability and two



Original graph

Random Node

Random Edge

Random Walk

Figure 1: A graph and three sampling methods to select six sample nodes. The three sampling methods are random node (RN), random edge (RE), and random walk (RW). Nodes can be sampled multiple times as shown in sub-figures for RE and RW samplings.

nodes incident to a random edge are collected. In this way, RE sampling is a kind of PPS (probability proportional to size) sampling in that each node is sampled with probability proportional to its degree. RW sampling selects the next node in the current neighbourhood uniformly at random. Its node selection probability is proportional to the degree asymptotically.

Different sampling methods require different estimators. The arithmetic mean is an unbiased estimator for RN sampling:

$$\widehat{\langle d\rangle}_{RN} = \frac{1}{n}\sum_{i=1}^{n} d_{x_i}, \tag{3}$$

In RE or PPS sampling, the arithmetic mean estimator tends to overestimate the average degree $\langle d\rangle$ by a factor of $(\gamma^2 + 1)$. Instead, the harmonic mean should be used for these samples:

$$\widehat{\langle d\rangle}_{RE} = n\left[\sum_{i=1}^{n}\frac{1}{d_{x_i}}\right]^{-1}. \tag{4}$$

We refer to [32] for the detailed derivation of this estimator. RW sampling assumes that the sampling probability is proportional to the degree, therefore the same estimator is used. What we are interested in this paper is the variance of the RE estimator, particularly the comparison with the variance of the RN estimator.

The sampling and estimation methods can be illustrated using Fig. 1. The average degree of the graph is 2. The sample degrees taken by RN, RE, and RW sampling methods are (1, 1, 1, 1, 2, 8), (1, 8, 1, 8, 2, 4), and (4, 3, 8, 1, 8, 1), respectively. The estimations for RN, RE, and RW samples are:

$$\begin{aligned}
\widehat{\langle d\rangle}_{RN} &= \frac{1+1+1+1+2+8}{6} = 2.5,\\
\widehat{\langle d\rangle}_{RE} &= \frac{6}{\frac{1}{1}+\frac{1}{8}+\frac{1}{1}+\frac{1}{8}+\frac{1}{2}+\frac{1}{4}} \approx 2,\\
\widehat{\langle d\rangle}_{RW} &= \frac{6}{\frac{1}{4}+\frac{1}{3}+\frac{1}{8}+\frac{1}{1}+\frac{1}{8}+\frac{1}{1}} \approx 2.11.
\end{aligned}$$

To develop intuition about the high variance of RN sampling, and the variance reduction enabled by RE sampling to be discussed in the next section, consider a pedagogical example depicted in Fig. 2. It is a star graph that has a large node connecting with every other node (degree=N-1), while all the remaining (N-1) nodes connect with the large node only (degree =1). Such a graph in a much larger scale is also found as a subgraph in the real NotreDame web graph as shown in Fig. 11. The average degree is $(N-1+N-1)/N \approx 2$, assuming $1/N \approx 0$. Most of the

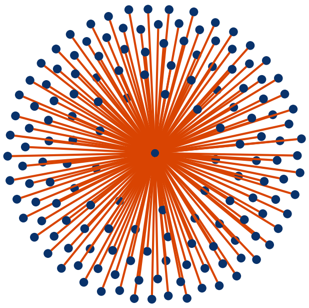


Figure 2: **An illustrative example that favours random edge (RE) sampling.**

Table 1: Summary of notations

Notation	Meaning	Properties
N	population size	
n	sample size	
d_i	degree of node i	
τ	volume of all the nodes	$\tau = \sum_{i=1}^{N} d_i = N\langle d\rangle$
d_{x_j}	degree of the j th sampled node	$x_j \in \{1, 2, \dots, N\}$
p_i	probability of node i being visited	$p_i = d_i/\tau$, $\sum_{i=1}^{N} p_i = 1$
$\langle d\rangle$	mean degree	$\langle d\rangle = \tau/N$
$\langle d^2\rangle$	mean of the squared degrees	$\langle d^2\rangle = \sum_{i=1}^{N} d_i^2/N$
σ^2	variance of the degrees	$\sigma^2 = \langle d^2\rangle - \langle d\rangle^2$
γ^2	coefficient of variation	$\gamma^2 = \sigma^2/\langle d\rangle^2 = \langle d^2\rangle/\langle d\rangle^2 - 1$
$\langle d^E\rangle$	asymptotic mean degree of RE sampling	$\langle d^E\rangle = \langle d^2\rangle/\langle d\rangle$

uniform random samples will include the small nodes only, even when the sample size is close to N. Thus most of the estimations will be 1, while occasionally there are very large estimations when the large node is sampled. When RE sampling is used, both small and large nodes are sampled, resulting in sampled degree sequence $(1, N-1, 1, N-1, \dots)$. For these sampled degrees, the sample mean is $N/2$, which over estimates grossly because a node is sampled with the probability proportional to its degree. Such samples need a different estimator, i.e., the harmonic mean instead of arithmetic mean. The harmonic mean of four sample degrees is $4/(1 + 1/(N - 1) + 1 + 1/(N - 1)) \approx 2$. This approximates the true value very well.

2.2. *Pertinent Work*

Graph sampling has been widely studied [19, 38], and finds its applications in online social networks [9, 29, 7, 31], real social networks [32, 40], web graphs [11], search engine indexes [2], and deep web data sources [21]. The norm of the practice is to use uniform random samples whenever possible. Quite often, uniform random sampling is not directly supported. Other methods, such as Metropolis Hasting Random Walk (MHRW) [24] and rejection sampling, are utilized to approximate uniform random sampling [2]. When uniform random samples are not available, numerous sampling methods are proposed, in particular RW [20] for unequal probability sampling.

Recently, there are empirical observations that RW sampling can outperform MHRW sampling [30, 9]. Although MHRW does produce uniform random samples, it incurs additional cost, and is not the same as the direct RN sampling. Therefore, it is easier to observe that RW can be better than MHRW sampling. Rasti et al. observed that random walk sampling can outperform MHRW in the context of peer-to-peer networks [30], Gjoka et al. showed that RW (called re-weighted random walk in their paper) and MHRW are comparable [9]. This paper claims that, it is RE sampling that is better than RN sampling for large and scale-free networks. RW sampling, on the other hand, is always inferior to RE sampling, and can be much worse than RE and RN sampling when the graph conductance is very small.

Our earlier work on the comparison between RW and RN samplings on the Twitter data [22] motivated the studies conducted in this paper. [22] found that, on the Twitter data, RW sampling is much better than RN sampling. [39] experimented RW sampling on bipartite graphs representing term-document relationship in search engines. On such bipartite graphs, the performance of RW sampling does not have an obvious advantage over RN sampling. Our further study on dozens of other datasets also generated mixed results. Thereby, this paper tries to answer the question as for when and why RW is better than RN sampling. We identified two orthogonal factors influencing the sampling method: the degree variation and the conductance. High degree variation will guarantee that RE sampling works well, and the lack of loosely connected components ensures that RW sampling can approximate RE sampling.

The harmonic mean estimator was first derived and studied in depth by Salganik et al. [32] to estimate the properties of hidden populations such as drug-addicts. The degree sampling of networks, which is the focus of this paper, has also received special attention. Stump et al. studied the sampling of degree distribution [35] for two sampling schemes, i.e., random sampling and the degree dependent sampling of the nodes. For average degree estimation, both [8] and [10] used uniform random sampling of the nodes. [8] discussed the lower bound of the estimation. Based on this result, [10] proposed a sampling scheme that put more weight on the nodes that have less probability of being sampled.

The impact of sampling methods on the discovery of graph properties has also been studied in [19, 36, 35, 18]. They cover a wide range of network properties, and focus on the properties of the derived sub-graph, instead of the estimation of the properties of the original graph. For instance, [19] investigated several network characteristics like the distribution of connected components. [18] showed that random node sampling performs better than random edge sampling in approximating the clustering coefficient of the graph.

2.3. Why Average Degree Estimation

Graph properties need to be estimated when the graph in its entirety is not available. This happens when the graph is distributed without central data deposit, such as the Web, or, when it is hidden behind searchable web interfaces, such as search engines, online social networks, and millions of textual corpora hidden behind HTML search boxes. In either case the direct calculation of the property is impossible. Next, we highlight the importance of average degree estimation, and practical implications of RE sampling.

Average degree is an important metric for any graph [14], and has many incarnations in the real world: when the graph is the Web, the average degree is the average number of in/out-links, which is an important property to characterize the Web [5]; when the graph is an online social network, the average degree is the average number of friends, messages and followers [9]; when the graph is a term-document network implemented by a search engine, the average degree is the average document size and average query matches [6] [2]; when the graph is an email network, the average degree is the average number of email contacts.

The applications can go beyond computer science to other disciplines such as finding the average degree (friends) of drug addicts [32]. Furthermore, average degree estimation can be generalized as the problem of estimating the expectation of a random variable. RN and RE samplings correspond to uniform sampling and PPS sampling, respectively. Thus our results can be extended to other scenarios without a graph representation. In this paper, we discuss the problem in the setting of graph so that the RW sampling can be compared within the same framework. Besides, graph representation gives us a tangible illustration and straightforward implementation for the sampling process.

What is more important is that average degree can be used to derive other properties such as the variance and the data size. The variance $\langle d^2 \rangle - \langle d \rangle^2$, or equivalently, $\gamma^2 = \langle d^2 \rangle / \langle d \rangle^2 - 1$, is dependent on average degree $\langle d \rangle$. More interestingly, it can be estimated using $\gamma^2 = \langle d^E \rangle / \langle d \rangle - 1$, where $\langle d^E \rangle$ is the average degree of the samples obtained by RE sampling. γ in turn can be used to estimate the number of nodes by $\widehat{N} = (\gamma^2 + 1)\frac{n^2}{2C}$ [23, 22], where n is the sample size, C is the number of collisions in the samples. 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 on average. To be more precise, your friends have $\gamma^2 + 1$ times more friends than you do. 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 [12] (Eq 7.8) is $\pi = \frac{(\gamma^2+1)\langle d \rangle - 2}{(\gamma^2+1)\langle d \rangle - 1}$, where π is the proportion of the nodes that are immuned uniformly from the network.

3. Variance Reduction Using RE Sampling

The performance of estimators can be evaluated in terms of bias and variance. The RN method is unbiased, and the RE method has a small bias that can be ignored compared with the variance when sample size is large [32]. Therefore, we focus on the comparison of the variances of the two methods.

The variance of the RN method depends on the variance of the degrees, which varies from data to data, and typically grows with the size of the data for a given degree distribution. On the other hand, we find that the variance of the RE method has an upper bound that does not depend on the degree variance. This upper bound guarantees the reduction of the variance of the RE method for very large data.

We develop the upper bound in three steps: first we use V_1, a value that is obtained from Taylor expansion, to approximate the RE estimator. We demonstrate that such approximation is accurate using simulated data and real networks of various topologies. V_1 contains the variance of the reciprocal of the sampled degrees, which is difficult to quantify and compare with. Therefore, V_2 is derived as an upper bound of V_1, by exploiting the fact that all the degrees are greater than one, and that most of the nodes are of very small degrees in scale-free networks. In a typical scale-free network with exponent one, we show that $V_2 \approx 2V_1$. In real networks, we observe that the ratio between V_2 and V_1 varies widely between 1.19 and 7.55. V_2 can be simplified further into V_3 when the average degree is large.

3.1. The Large Variance Problem of RN Sampling

The variance of the arithmetic mean estimator $\widehat{\langle d\rangle}_{RN}$ for RN sampling is:

$$var(\widehat{\langle d\rangle}_{RN}) = \frac{\sigma^2}{n} = \frac{\gamma^2\langle d\rangle^2}{n}, \tag{5}$$

where n is the sample size. Sometimes, it is easier to interpret the variance relative to its true value using RSE (relative standard error) as defined as below:

$$RSE(\widehat{\langle d\rangle}_{RN}) = \frac{\sqrt{var(\widehat{\langle d\rangle}_{RN})}}{\langle d\rangle} = \frac{\gamma}{\sqrt{n}}. \tag{6}$$

Although $\widehat{\langle d\rangle}_{RN}$ is an unbiased estimator, its variance can be very large for some scale-free networks. The degrees of most real life networks are close to Zipf's distribution [28], inducing a large variation of the degrees. However, it is hard to quantify the variance exactly because real data do not fit exactly the Zipf's law, and the exponent and cut-off value vary from data to data. Nonetheless, we can assume a distribution to gain some understanding of the variance.

When d_i follows the power law, $d_i = A/i^\alpha$, where A is a normalizing constant that satisfies

$$\sum_{i=1}^{N} d_i = A\sum_{i=1}^{N} 1/i^\alpha \approx A\zeta(\alpha) = N\langle d\rangle,$$

where $\zeta(.)$ is the Riemann-zeta function $\zeta(\alpha) \approx \sum_{i=1}^{N} 1/i^\alpha$ based on the assumption that N is a very large number. That is, $A = N\langle d\rangle/\zeta(\alpha)$. Note that this exponent α is for the degree-rank plot. The corresponding frequency-degree plot has slope $-(\alpha+1)$ [28]. Since the vast majority of networks have degree-frequency slope around -2 [27], in the following we derive the variance when the slope is exactly -2, i.e., $\alpha = 1$ in the degree-rank equation. Note that $\zeta(1) \approx \ln N$, and $\zeta(2) \approx 1.6$. Therefore,

$$\sum_{i=1}^{N} d_i = \sum_{i=1}^{N} \frac{A}{i} = A\zeta(1) \approx A\ln N, \tag{7}$$

$$\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} \frac{A^2}{i^2} = A^2\zeta(2) \approx 1.6A^2. \tag{8}$$

By the definition of variance, we can derive the variance of the degrees as below:

$$var(d) = \langle d^2\rangle - \langle d\rangle^2 = \frac{1}{N}\sum_{i=1}^{N} d_i^2 - \frac{1}{N^2}\left(\sum d_i\right)^2 = \frac{1.6A^2}{N} - \frac{A^2\ln^2 N}{N^2} = \langle d\rangle^2\left(\frac{1.6N}{\ln^2 N} - 1\right). \tag{9}$$

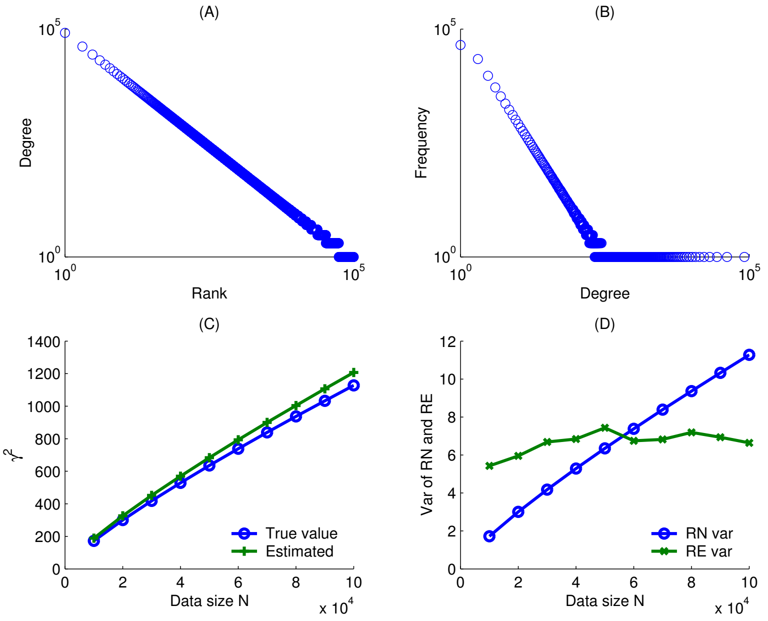


Figure 3: γ^2 **grows almost linearly with data size** N **when the degree distributions are the same. Panel(A) Degree-rank log-log plot when** $N = 10^5$**. (B) The frequency-degree plot of the same data. (C) Observed** γ^2 **against the data size** N**, along with the projected** γ^2 **in Eq. 10. (D) Variances RN and RE samplings when sample size n=100.**

Therefore, when exponent α of the Zipf's law is 1, the coefficient of variation is:

$$\gamma^2 = \frac{1.6N}{\ln^2 N} - 1. \tag{10}$$

The intuition of this equation is that the variance grows almost linearly with data size N, in the order of $O(N/\ln^2 N)$. The sample size n needs to be in the order of $O(N/\ln^2 N)$ so that satisfactory estimates can be obtained. When the data is very large, almost all the nodes need to be checked before an estimation can be made. That is equivalent to saying that the estimation is infeasible for very large scale-free graphs using uniform random sampling.

For instance, Twitter has $N \approx 5 \times 10^8$ users in 2012. If the degree distribution followed the power law as we assumed, its coefficient of variation would be $\gamma^2 = 1.6N/\ln^2 N = 2.0 \times 10^6$. According to Eq. 6, this means that we would need a sample size $n = \gamma^2/0.1^2 = 2.0 \times 10^8$ so that the RSE could be 0.1. To achieve the 95% confidence interval $5 \times 10^8 \pm 10^8$, the sample size is already in the same order of the total population. For the downloaded Twitter data in 2009 that contains 4.1×10^7 users, we find that the sample size needs to be 129,240 so that RSE is 0.1.

To verify our derivations, we generate 10 synthetic datasets with the same distribution ($\alpha = 1$) but different data size N ranging between 10^5 and 10^6. Panel (C) in Fig. 3 shows the observed γ^2 values along with the ones projected by Eq. 10. Clearly, γ^2 grows almost linearly with the data size, and the projection is rather accurate. Panels (A) and (B) plot the degree distributions when $N = 10^5$. Panel (A) is the degree-rank plot with slope -1, i.e., $\alpha = 1$. (B) is the frequency-degree plot for the same data. We can see that the slope of the frequency-degree plot is $-(\alpha + 1) = -2$. Panel (D) compares the estimator variance for RE and RN methods when the sample size is 100. It shows that, for the same data distribution, the performance of RE method remains almost constant. But compared with the increasing variance of RN method, RE sampling becomes better when $N > 50,000$, and the advantage becomes larger with the increase of data size.

3.2. Variance of RE Sampling

Given a set of degrees $\{d_{x_1}, d_{x_2}, \ldots, d_{x_n}\}$ obtained by RE sampling, recall that the harmonic mean estimator is:

$$\widehat{\langle d \rangle}_{RE} = n \left[\sum_{i=1}^{n} \frac{1}{d_{x_i}} \right]^{-1}. \tag{11}$$

Let random variables $v = 1/d_{x_i}$, and $V = \sum_1^n 1/d_{x_i}$. By the harmonic mean estimator, we have $E(v) = E(1/d_{x_i}) = 1/\langle d \rangle$, and $E(V) = n/\langle d \rangle$. Our first approximation to the variance is obtained by applying the Taylor expansion of $\widehat{\langle d \rangle}_{RE}$ around $E(V)$. The result is:

$$\widehat{\langle d \rangle}_{RE} = \frac{n}{V} = n \left(\frac{1}{E(V)} - \frac{V - E(V)}{E(V)^2} + \ldots \right) \tag{12}$$

By applying the variance on the first two terms of the Taylor expansion, we obtain an approximate variance of $\widehat{\langle d \rangle}_{RE}$ (denoted by V_1) as follows:

$$V_1 = \frac{n^2 var(V)}{E(V)^4} = \frac{\langle d \rangle^4 var(v)}{n}. \tag{13}$$

Next, we need to find a bound for $var(v)$. By the definition of variance,

$$var(v) = E(v^2) - (E(v))^2 = E\left(\frac{1}{d_{x_i}^2} \right) - \frac{1}{\langle d \rangle^2}. \tag{14}$$

Since d_{x_i} is obtained with probability $p_{x_i} = d_{x_i}/\tau$, where $\tau = \sum_{i=1}^N d_i$,

$$E\left(\frac{1}{d_{x_i}^2} \right) = \sum_{i=1}^{N} p_i \frac{1}{d_i^2} = \sum_{i=1}^{N} \frac{1}{\tau d_i} = \frac{1}{N \langle d \rangle} \sum_{i=1}^{N} \frac{1}{d_i}. \tag{15}$$

$\sum_{i=1}^N 1/d_i$ varies from data to data, but a safe upper bound is $N > \sum_{i=1}^N 1/d_i$ since every $d_i > 1$. In scale-free networks, $\sum_{i=1}^N 1/d_i$ is not far away from N, as we will show in the simulation study and in real networks. Therefore we derive an upper bound for $var(\widehat{\langle d \rangle}_{RE})$, which is called V_2 as defined below:

$$V_2 = \frac{\langle d \rangle^4}{n} \left(\frac{1}{\langle d \rangle} - \frac{1}{\langle d \rangle^2} \right). \tag{16}$$

When $\langle d \rangle$ is a large number, the second term $1/\langle d \rangle^2$ can be neglected, resulting in a simplified upper bound V_3,

$$V_3 = \frac{\langle d \rangle^4}{n} \frac{1}{\langle d \rangle} = \frac{\langle d \rangle^3}{n}. \tag{17}$$

In summary, $var(\widehat{\langle d \rangle}_{RE}) \approx V_1 < V_2 < V_3$. Thus, we derive the following theorem:

Theorem 1. *The upper bound for the variance of RE sampling is $\langle d \rangle^3/n$. Or,*

$$var(\widehat{\langle d \rangle}_{RE}) < \frac{\langle d \rangle^3}{n}. \tag{18}$$

We highlight two points regarding this result. First, the upper bound does not depend on the degree distribution or other topological characteristics like graph conductance. As long as the nodes are selected with probability proportional to their degrees, the result holds no matter whether it is a scale-free graph, or has tightly knit communities. On the other hand, the performance of RW sampling depends on graph conductance as we will explain in Section 4.3.

Second, the upper bound is surprisingly simple, involving only average degree and sample size. Unlike the variance of RN sampling that is associated with the degree variance, the variance of RE sampling is bounded from above by a constant determined by average degree. Recall that the variance for RN estimator is $var(\widehat{\langle d \rangle}_{RN}) = \frac{\langle d \rangle^2}{n} \gamma^2$. Comparing the variances for estimators $\widehat{\langle d \rangle}_{RN}$ and $\widehat{\langle d \rangle}_{RE}$, we have:

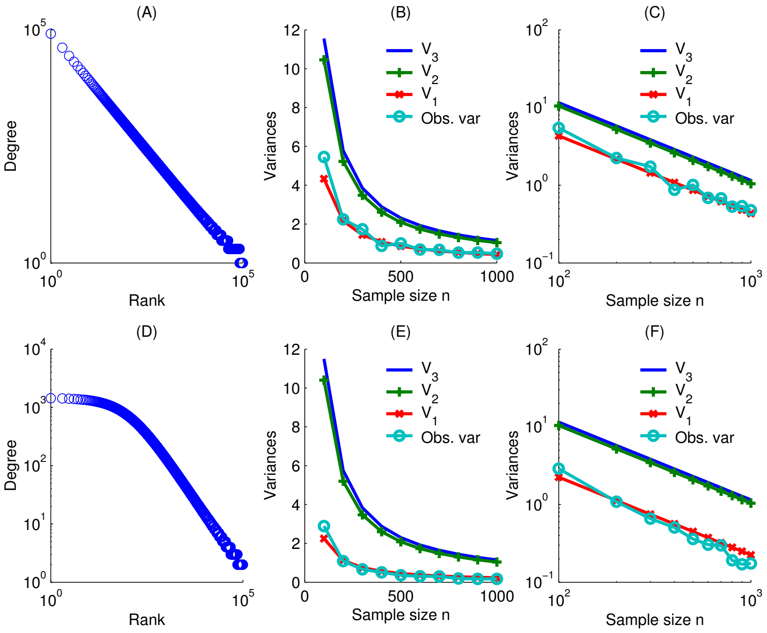


Figure 4: **The upper bound V_3 and the observed variances, along with the approximations V_1 and V_2. Sample sizes range between 100 and 1000. $N = 10^6, \alpha = 1, \langle d \rangle = 10$. Observed variance is obtained from 100 repetitions.**

Corollary 1. *RE sampling reduces the variance at least by a factor of $\gamma^2/\langle d \rangle$, i.e.,*

$$\frac{var(\widehat{\langle d \rangle}_{RN})}{var(\widehat{\langle d \rangle}_{RE})} > \frac{\gamma^2}{\langle d \rangle}. \tag{19}$$

3.3. Simulation Studies

To understand the relationship between the upper bound and the true variance, we demonstrate the variances using synthetic datasets whose degree distributions follows a power law $d_i = A/(\beta + i)$, where A is a normalizing constant $N\langle d \rangle / \sum_{i=1}^{N} 1/(\beta + i)$. This is called Zipf-Mandelbrot law [26] that can model the real data better than $d_i = A/i^\alpha$. When $\beta = 0$, it is reduced to the Zipf's law described in Section 3.1.

We experiment with two distributions, and report the results in Fig. 4. The first row is for the distribution with $\beta = 0$, and the second row has $\beta = 100$. For both distributions, we generate the degrees satisfying such distribution where the data size $N = 10^6$, and conduct RE sampling 100 times. From 100 repetitions, we record the variance of the RE estimator, along with its approximations V_1, V_2 and V_3. Panels (A) and (D) are the degree-rank plots for the data, giving a visual understanding of the distribution. Panels (B) and (C) are the plots for the variances (V_1, V_2, V_3, and observed variance) over various sample sizes ranging between 100 and 1000. Panels (C) and (D) are the corresponding log-log plots.

We make the following observations from the simulation study:

- **V_3 vs. V_2:** According to our analysis in the previous subsection, the ratio between V_3 and V_2 should be that of $1/\langle d \rangle$ and $1/\langle d \rangle - 1/\langle d \rangle^2$, which is $\langle d \rangle / (\langle d \rangle - 1)$. In this particular data, $\langle d \rangle = 10$, and we can see that the ratio between V_3 and V_2 are very close to 10/9 as expected.

- **V_2 vs. V_1:** The difference between V_1 and V_2 is determined by the difference of $\sum_{i=1}^{N} 1/d_i$ and N. When $\beta = 0$,

$$\sum_{i=1}^{N} 1/d_i = \frac{1}{A} \sum_{i=1}^{N} i = \frac{\ln N}{N\langle d \rangle} \frac{N(N-1)}{2} \approx \frac{N}{2}. \tag{20}$$

Table 2: Statistics of the 18 graphs, sorted in decreasing order of the coefficient of degree variation γ. Each graph has a citation indicating where the data is from.

Graph	γ	$\langle d \rangle$	# Nodes
Twitter [16]	35.95	70.51	41,652,230
WikiTalk[19]	26.32	3.90	2388953
BerkStan[19]	14.51	20.10	654782
EmailEu[19]	13.66	3.02	224832
Stanford[19]	11.51	15.21	255265
Skitter[19]	10.46	13.09	1694616
Youtube[25]	9.64	5.27	1134890
NotreDame[19]	6.40	6.69	325729
Gowalla[19]	5.54	9.67	196591
Epinion[19]	4.02	10.69	75877
Google[19]	4.00	10.03	855802
Slashdot[19]	3.35	12.27	82168
Facebook-1[41]	3.14	14.27	2937612
Flickr [19]	2.64	43.52	105720
Facebook-2 [37]	1.55	25.77	63392
Amazon[19]	1.27	11.89	410236
CitePatents[19]	1.20	8.77	3764117
RoadNet[19]	0.35	2.82	1965206

That is, $\sum_{i=1}^{N} 1/d_i = 0.50N$. Therefore V_2/V_1 is approximately two as shown in panels (B) and (C).

When $\beta = 100$, $\sum_{i=1}^{N} 1/d_i = 0.29N$, which is smaller than the previous case. Therefore, the gap between V_2 and V_1 is larger.

- **V_1 vs. Observed Variance:** V_1 is obtained from Taylor expansion of $1/V$ by ignoring the third term in Eq. 12. While the approximation varies from data to data, it shows that the gap is really negligible in these two simulated data, and in 18 real network as we will show in Section 4.

4. Experiments On Real Networks

4.1. Datasets

We conducted experiments on 18 real networks, most of them are from the Stanford SNAP graph collection [19]. Due to space limitation, for some network categories only one graph is reported if they have similar behaviour. For instance, citation graphs have similar degree distribution, similar coefficient of variation, and similar error ratios between RN, RE, and RW samplings. For these categories, we choose only one graph for each category. In the category of the Web graph datasets, RW sampling deviates greatly from RE sampling. To investigate the cause for such deviation, we investigated several Web graphs on the domains of Notre Dame, Stanford, and Berkley-Stanford. The Facebook data is one of the few exceptions where RE sampling is inferior to RN sampling. Therefore, we include two Facebook graphs. Complete data description and programs can be found at `http://cs.uwindsor.ca/~jlu/degreevar`. Their statistics are summarized in Table 2, sorted according to γ, the coefficient of variation of the degrees.

We make several observations on the datasets. First, most of them are scale-free networks as shown in Fig. 5. The frequency-degree slope is around 2, their corresponding degree-rank slope shall be around 1, the same slope we selected in our simulation studies. Some datasets, such as Facebook and Citation networks, have a curve that is reflected by the Zipf-Mandelbrot law we used. One exception is the RoadNet network that is closer to normal or log-normal distribution. There are irregular data distributions, such as Flickr and Amazon that have broken lines.

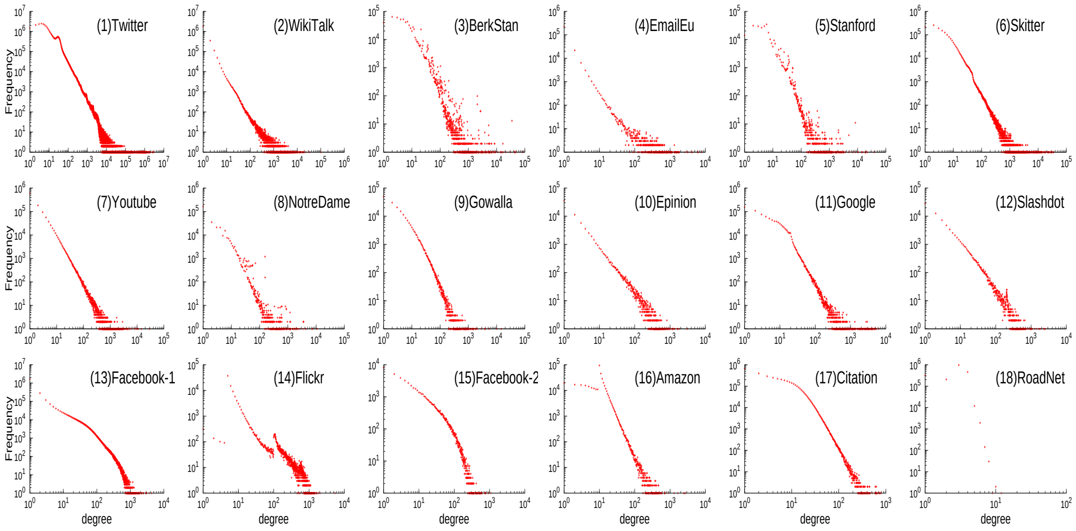


Figure 5: **Degree distributions of 18 graphs. Plots are sorted in decreasing order of coefficient of variation γ.**

Also, Web graphs (sub-figures 3, 5, 8) do not form a straight line in the upper part of the log-log plots, indicating irregularity in the graph structure. Albeit the varieties of the datasets, we will show that our result withstands without exception.

Second, it is interesting to note that two representative social networks Twitter and Facebook are in the two extremes of the spectrum of γ values, due to the way the networks are formed. Twitter dataset is much larger, and allows unlimited number of followers, while Facebook has an upper limit for the number of friends. Therefore Twitter is a scale-free network with large degree variation, while Facebook has a sharp dropping curve causing a low γ value. Besides, Facebook datasets are much smaller. Because of their structural difference, for Twitter data RE is a hundred times better than RN sampling, for Facebook data RE and RN samplings are similar.

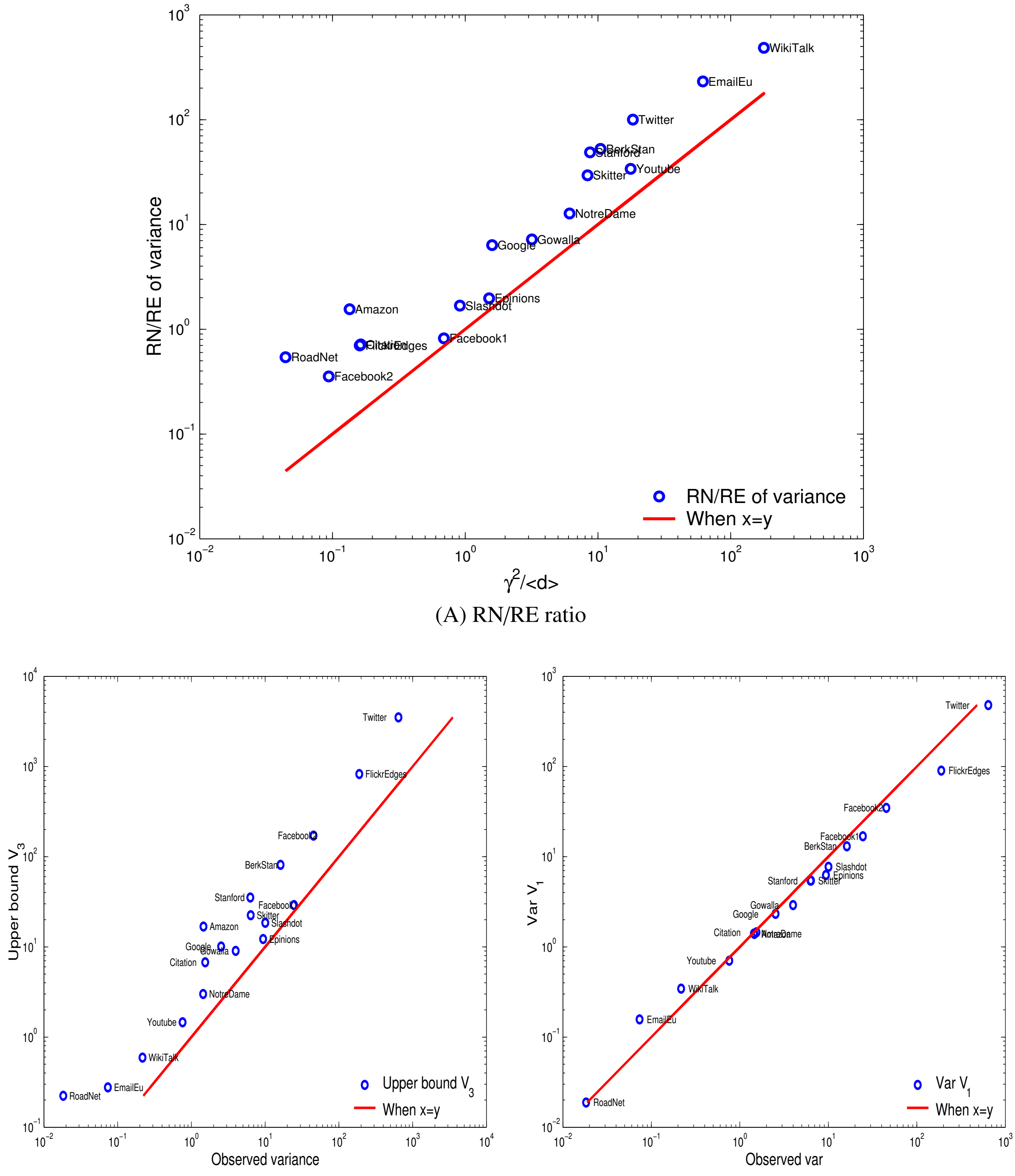
One technical detail needs to be mentioned is the sampling of Twitter data. It is the complete user network collected in 2009 [16], which has billions of edges that can not fit into computer memory. We use index engine Lucene to store the data in hard drive and use search engine to mimic the random sampling methods. We treat all the neighbours of a node as a 'document', and build a index for those 'documents', or neighbours. Then, graph sampling can be accomplished by searching the index.

4.2. RE vs. RN sampling

Despite their variety in degree distribution and topology, all the 18 datasets support our analytical result, i.e., RE improves the variance at least by a factor of $\gamma^2/\langle d\rangle$. We demonstrate this using a fixed sample size first in Fig. 6, then the trend over sample sizes in Fig. 7.

Fig. 6 panel (A) demonstrates our main result, i.e., the variance ratios between RN and RE samplings have an almost linear relation with $\gamma^2/\langle d\rangle$, whose Pearson's correlation coefficient is as high as 0.9867. In addition, the ratios are consistently higher than $\gamma^2/\langle d\rangle$, indicating that $\langle d\rangle^3$ is indeed an upper bound of the RE variance. The plot is in log-log scale so that the points are spread out along the axes. For this experiment, sample size n=100, and the variances are obtained with 20,000 repetitions. There are only five datasets whose RN/RE ratio is slightly below one, i.e., RE sampling is slightly worse than RN sampling. A closer inspection of these datasets shows that they all have small degree variations as shown in Table 2 and Fig. 5. Both of the citation and road networks are at the lower end of the γ values. The RoadNetwork has maximal 12 degrees, and its degrees follow a log-normal distribution. The Facebook network has an upper limit on the number friends, thus the maximal degree is abnormally small compared with its size. The Flickr network has an irregular degree distribution that has a large bump around degree 100.

Panels (B) and (C) in Fig. 6 are plotted to corroborate our conclusion. Panel (B) shows that V_3 is indeed the upper bound for the RE variance. For all the 18 datasets, V_3 is consistently larger than the observed variance. For some



(A) RN/RE ratio

(B) V_3

(C) V_1

Figure 6: **Variance of RE sampling for 18 datasets. (A) RN/RE ratio is always higher than $\gamma^2/\langle d\rangle$; (B) V_3, the upper bound, is greater than the observed variance; (C) V_1, the approximation obtained by Taylor expansion, is close to the observed variance. Sample size n=100. Variances are obtained from 20,000 repetitions.**

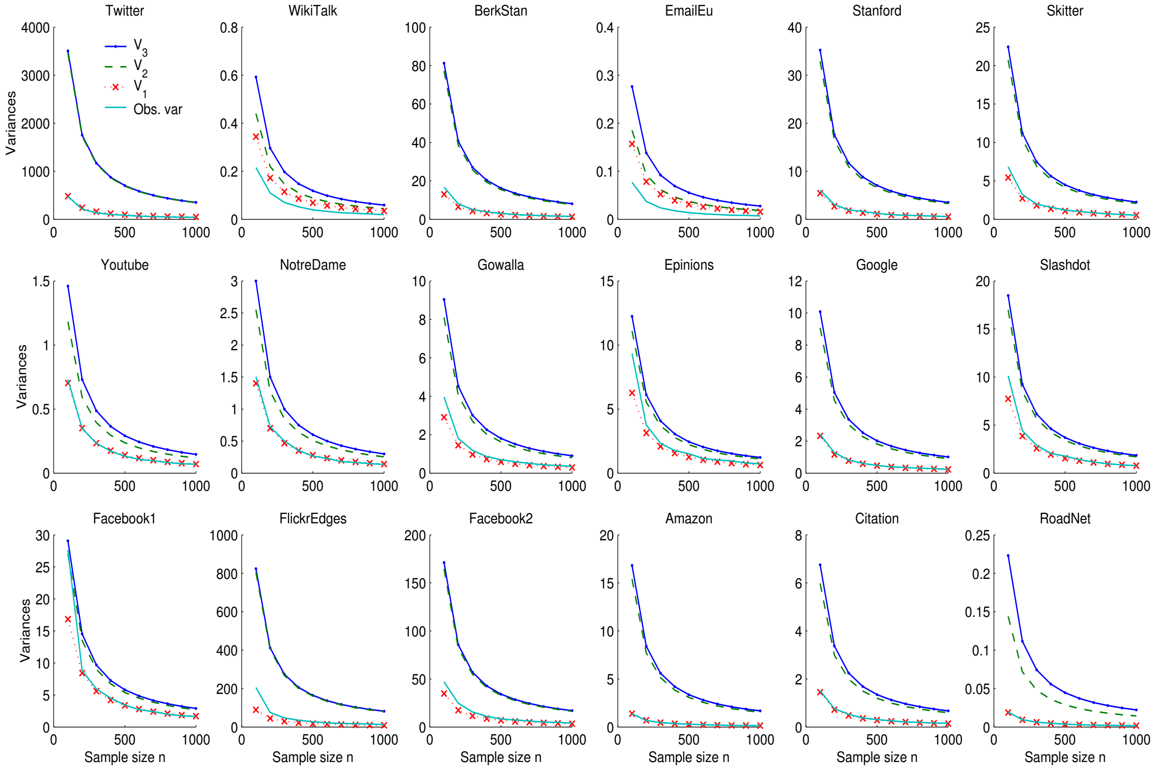


Figure 7: **Variances vs. sample size for 18 datasets. Sample size n ranges between 100 and 1000. Variances are obtained over 1000 repetitions. The upper bound V_3 is higher than the observed variance.**

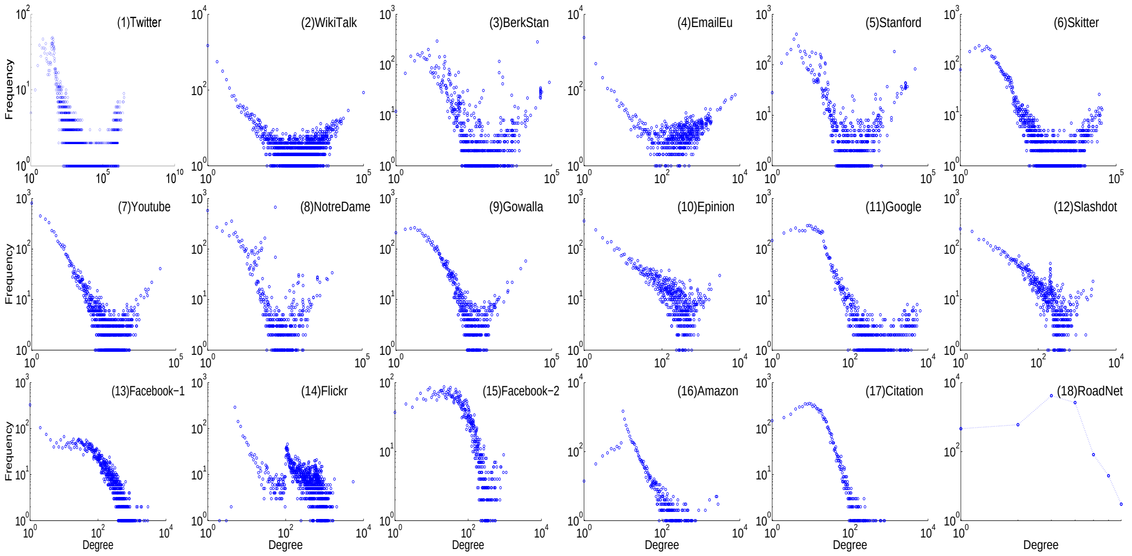


Figure 8: **Degree distributions of the samples obtained from RE sampling. n=8,000.**

datasets such as Epinion and Slashdots, the upper bound is rather close to the true variance. Panel (C) verifies that V_1, the approximation obtained by the Taylor expansion, is very close to the real variance.

Fig. 7 shows the trend of the variances over various sample sizes, and reconfirms the relationship between variances V_1, V_2, V_3, and the real one. It demonstrates that $V_3 > V_2 > V_1 \approx$ *observed variance* as expected. Overall, V_1 is very close to the observed variance, as the Taylor expansion can approximate the original function well. V_2 and V_3 are also close in general, since their difference is dictated by $\langle d\rangle/(\langle d\rangle - 1)$. When average degree $\langle d\rangle$ is small, as in RoadNet, WikiTalk and EmailEU, the difference between V_2 and V_3 is noticeable. The largest gap is between V_1 and V_2, which is determined by $N/\sum_{i=1}^{N} 1/d_i$. As we have shown in simulation studies in section 3.3, $N/\sum_{i=1}^{N} 1/d_i$ is around two when the exponent $\alpha = 1$ and average degree is 10. That is, V_2 is about twice as large as V_1. In real datasets, $N/\sum_{i=1}^{N} 1/d_i$ varies from data to data, ranging between 1.19 (WikiTalk) to 7.55 (Flickr).

Another perspective to understand the reduced variance of RE sampling is the sample distributions that are depicted in Fig. 8, where the sample size is 8000. It shows that most of the sample distributions have a "V" shape, indicating that the small nodes still follow a power law as in the original data, while the large nodes can be sampled many times. In RE sampling, the sampling probability $P(k)$ of degree k is determined by k and its frequency $f(k)$. Bearing in mind that $f(k) \propto k^{-\alpha}$, where α is normally around two, we have:

$$P(k) \propto k \times f(k) \propto k \times k^{-\alpha} = k^{-(\alpha-1)}. \tag{21}$$

Therefore, the sampling probability still follows a power law with the exponent $\alpha - 1$. Same as the power-law for degree distribution, the formula is accurate only when k is small. When k is large, $f(k) \ll 1$ in the formula. In real data, $f(k)$ can not be a fraction number. Instead, $f(k)$ must be zeros in most cases so that in average they can follow the power-law. When $f(k)$ is non-zero number such as one, $P(k)$ amplifies the value one by a factor of k, thereby generating the second ascending branch in the sampling frequency plot.

In other words, both small and large nodes are sampled multiple times but for different reasons. Small nodes are sampled because there are many of them. Although each individual small node has a very small probability of being sampled, collectively the large number of small nodes will guarantee that some will be sampled. On the other hand, large nodes are sampled because they have higher probability of being hit by random edges, even though there are only a few of them. Therefore both small and large nodes are well represented in the sample, resulting in small variance of the estimation. In RN sampling, large nodes are included by chance, inducing a large variance in estimation.

The datasets that do not have the "V" shape in RE sampling happen to be the ones not in favour of RE sampling. They do not have nodes that are large enough to be sampled many times. Their RE sample distributions are just similar to the original data, or to RN sample distribution. Therefore RE sampling does not have an advantage in this kind of data.

Two of the representative online social networks are Twitter and Facebook. It is interesting to see that they favour different sampling methods, one RE sampling and the other RN sampling. Moreover, their RN/RE ratios happen to be on the two extremes of the spectrum. Twitter has the third highest RN/RE ratio because it is scale-free and the largest network in our experiment. Facebook-2 has the lowest RN/RE ratio because it has a cap on the number of friends.

4.3. RW Sampling

RW sampling can be regarded as an approximation to RE sampling in that *asymptotically* the node sampling probability is proportional to its degree. The sampling probability is not exactly PPS, yet the PPS estimator is used. Therefore, the bias of the estimator can no longer be omitted in our evaluation. Both bias and variance should be considered, and they can be measured by RRMSE (Relative Root MSE) as defined below:

$$RRMSE(\widehat{\langle d\rangle}) = \frac{1}{\langle d\rangle}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\langle d\rangle}_i - \langle d\rangle\right)^2} \tag{22}$$

where $\widehat{\langle d\rangle}$ is an estimator, $\langle d\rangle$ is the true average degree, $\widehat{\langle d\rangle}_i$ is the estimation obtained in the i-th run. In our experiments, all the RRMSE data are obtained by 5000 independent runs, except for Twitter data that has 2000 runs due to its large size and the long computation time of the sampling. Fig. 9 shows the comparison of three estimators.

Our first observation is that RW is worse than RE consistently as expected, since it is an approximation of RE sampling. To understand exactly how much worse RW is, we plot RW/RE ratio in Fig. 10 along graph conductance Φ

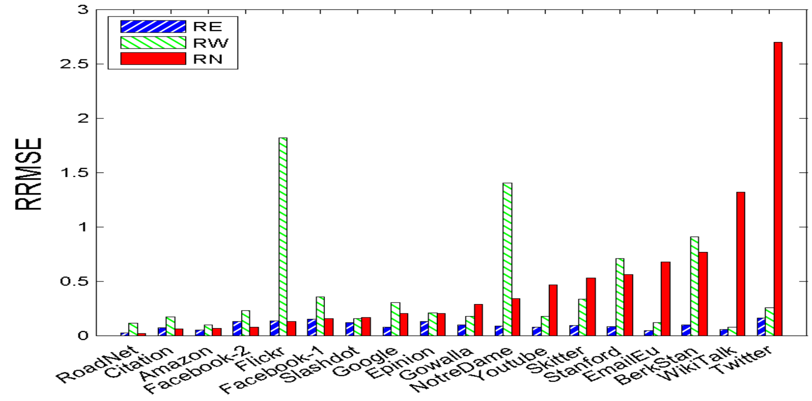


Figure 9: **RE, RW, and RN samplings on 18 graphs in terms of RRMSE. Sample size n=400, and RRMSEs are obtained over 5000 runs except for Twitter (2000 runs).**

[34], which can reflect mixing time of random walk. Very small graph conductance indicates the existence of loosely connected components. For datasets with small conductance such as NotreDame, Stanford, BerkStan and Flickr, RW is worse than RE sampling by a factor of ten in terms of RRMSE. In terms of variance, it is worse by a factor of a hundred. To develop the intuition for such poor performance, we plot their random walk traces when the estimations have large bias in Fig. 11. Each RW trace contains 10^4 steps. All four datasets, especially Flickr and NotreDame, have an extremely dense component that dangles loosely from the main component. This shows that RW sampling depends not only on the variation of the degrees, but also the topological structure of the graph. In the Flickr graph, there are two almost disconnected components. Random walks will happen mostly either in one of the component, and the corresponding estimations are the average degrees for one of the components, not the entire graph. In the NotreDame data, there is a very large star on the right that resembles the graph in Fig. 2, indicating many nodes are only connected to the centre of the star. When a random walk is trapped inside this star, the estimated average degree will be around two as shown in the example in section 3.1, no matter what the true value is.

Given this relationship between RW and RE sampling, our second observation is that RW outperforms RN only 1) when RE outperforms RN (or the degree variance is large); 2) when there is no loosely connected components (or the conductance is not very small). When the RE is worse than RN, RW will be also worse than RN. Therefore there is no need to test RW method or improve RW with its various extensions. When there are loosely connected components, we can modify the simple random walk sampling so that it can approximate RE sampling, e.g., by uniform random restart.

5. Conclusions

The size of the data, compounded by the power-law distribution, is changing the landscape of sampling practice. Uniform random sampling is no longer the method of choice. This does not happen until the data size reaches a threshold for a give degree distribution, as illustrated in Fig. 3 Panel D. The gap between RE and RN samplings grows almost linearly as the data size. It can be infinitely large in theory, and is orders of magnitude in observed data. Such a large difference is particularly important for web-based networks, such as online social networks and the deep web, where the sampling process is costly because of network traffic and daily quota.

It is remarkable to notice that it is uniform random node (RN) sampling that is on the downside of the comparison. In the past, great efforts are devoted to obtain uniform random samples using methods such as Metropolis-Hasting Random Walk and rejection sampling [2]. During the sampling process many nodes are visited, examined, and rejected. In the end these precious uniform random samples can be much worse than the samples obtained using low cost RE or RW methods.

While it is easy to understand that uniform random sampling has large estimation error for data with large variance, it is not straightforward to see whether RE sampling can reduce the variance for data of various distributions. We show that the variance of RE estimator has an upper bound $\langle d\rangle^3/n$. This upper bound is derived independent of the data

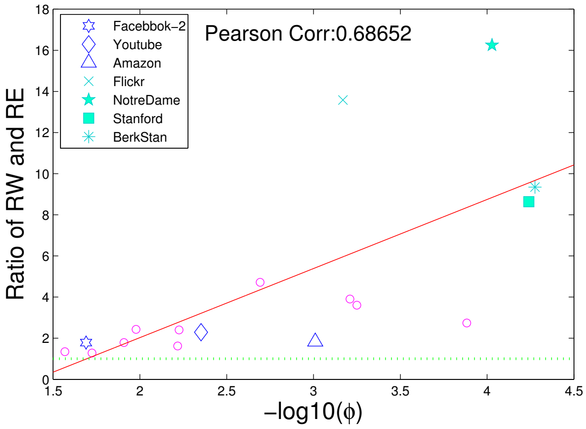


Figure 10: **Standard error ratio between RW and RE vs. graph conductance Φ for 18 datasets. Sample size is 400.**

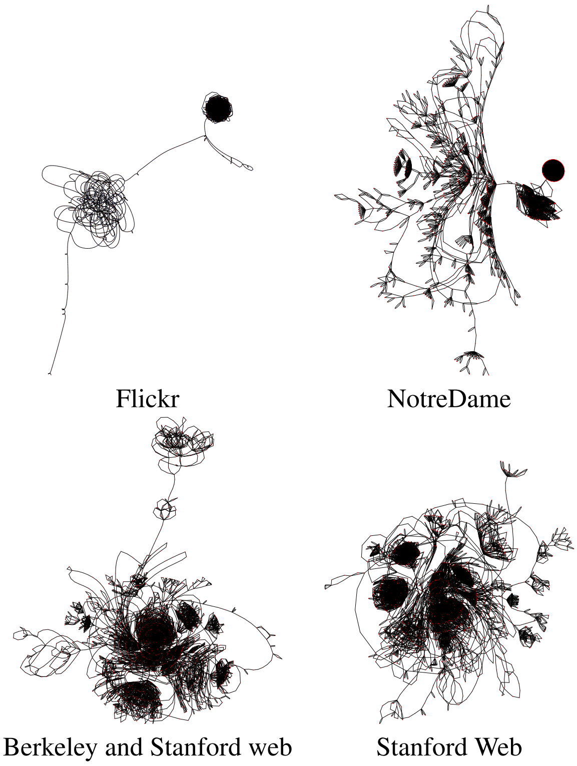


Figure 11: Random walks on four graphs, each has loosely connected components. Each random walk contains 10^4 steps.

distribution, and is close to the real variance when the data follows a power law distribution. We generate several synthetic scale-free networks to verify and explain our derivations, and use 18 real networks to support our result.

The upper bound of the variance implies that RE reduces the variance of RN sampling when the graph is large. First, the variance ratio between RN and RE samplings is at least $\gamma^2/\langle d\rangle$. Although the derivation involves several approximations, it is remarkable that the observed RN/RE ratio has a high linear correlation with $\gamma^2/\langle d\rangle$. The Pearson's correlation coefficient is 0.9867 among the 18 real networks we studied.

Second, the comparison between RN and RE depends on the values of γ^2 and $\langle d\rangle$. For a typical scale-free networks (degree vs. frequency power law slope is -2), we show that γ^2 grows almost linearly with data size for the same data distribution. When data size is small ($N < 6 \times 10^4$), γ^2 can be smaller than $\langle d\rangle$. However, when the data size grows, γ^2 is much larger than $\langle d\rangle$. In other words, RE is much better than RN for large graphs. Empirically we demonstrate the improvement ratio is greater than a hundred for Twitter, EmailEU, and WikiTalk. In theory, we project larger improvement ratio can be found. The dependency on data size may also explain why such variance reduction was not observed in the literature. Variance reduction happens only for very large data that are available only recently. If the data is not very large, RE may not be as good as RN even if the data is scale-free.

When RE sampling is not possible, we can use RW to approximate it in that both methods sample nodes with probability proportional to its size. The difference is that RW is a PPS sampling only asymptotically. Thus the performance of RW sampling differs from data to data. Our experiments show that in general RW sampling performs a little bit below RE sampling as expected, but sometimes it can be much worse, even worse than RN sampling when there are loosely connected components in the graph characterized by graph conductance.

In retrospect, RE sampling is not widely studied, probably because that in most real situations, nodes are the primary objects – they are represented explicitly, and can be searched, queried, and crawled. In other words, nodes can be sampled in various ways. Edges, on the other hand, come as secondary objects that reside in nodes, and can be accessed from the nodes only. In the Web, web pages (the nodes) can be sampled using various methods, while the edges are only revealed as a by-product when we crawl the Web from one page to another. In social networks, we sample people (the nodes), while the relations between people can be accessed from people, not the other way around. In software component networks, classes and objects are represented explicitly, while the relations between the classes can be obtained from those objects or classes.

New developments in the digitalized world are making RE sampling more common. When using random queries to sample documents, long documents are sampled more often. It is a PPS sampling, or RE sampling when we view the queries as edges connecting the documents. When using random messages/emails to sample users, It is a RE sampling for user network connected by messages/emails. In the semantic web, edges are explicitly represented as RDF triples.

6. Acknowledgements

The authors thank the anonymous reviewers for their thoughtful and detailed suggestions and comments. We would also like to thank Chase Chance for his comments.

7. References

[1] K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving random walk estimation accuracy with uniform restarts. In *Algorithms and Models for the Web-Graph*, pages 98–109. Springer, 2010.

[2] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. *Journal of the ACM*, 55(5):1–74, 2008.

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

[4] A. Broder and et al. Estimating corpus size via queries. In *CIKM*, pages 594–603. ACM, 2006.

[5] 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.

[6] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Trans. Inf. Syst.*, 19(2):97–130, 2001.

[7] A. Dasgupta, R. Kumar, and D. Sivakumar. Social sampling. In *SIGKDD*, pages 235–243. ACM, 2012.

[8] U. Feige. On sums of independent random variables with unbounded variance, and estimating the average degree in a graph. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 594–603. ACM, 2004.

[9] 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.

[10] O. Goldreich and D. Ron. On estimating the average degree of a graph. *Electronic Colloquim on Computational Complexity (ECCC)*, 2004.

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

[12] M. Jackson. *Social and economic networks*. Princeton University Press, 2008.

[13] L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *WWW*, pages 597–606. ACM, 2011.

[14] E. D. Kolaczyk. *Statistical analysis of network data*. Springer, 2009.

[15] M. Kurant, C. Butts, and A. Markopoulou. Graph size estimation. *arXiv preprint arXiv:1210.0460*, 2012.

[16] 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.

[17] S. Lawrence and C. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998.

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

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

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

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

[22] J. Lu and D. Li. Sampling online social networks by random walk. In *ACM SIGKDD Workshop on Hot Topics in Online Social Networks*, pages 33–40. ACM, 2012.

[23] J. Lu and D. Li. Bias correction in small sample from big data. *TKDE, IEEE Transactions on Knowledge and Data Engineering*, 25(11):2658–2663, 2013.

[24] 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.

[25] 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.

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

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

[28] M. E. J. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46:323, 2005.

[29] M. Papagelis, G. Das, and N. Koudas. Sampling online social networks. *Knowledge and Data Engineering, IEEE Transactions on*, 25(3):662–6761, 2013.

[30] A. Rasti and et al. Respondent-driven sampling for characterizing unstructured overlays. In *INFOCOM*, pages 2701–2705. IEEE, 2009.

[31] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *Annual conference on Internet measurement*, pages 390–403. ACM, 2010.

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

[33] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 298–305, Toronto, Canada, 2003. ACM.

[34] A. Sinclair and M. Jerrum. Conductance and the rapid mixing property for markov chains: the appr oximation of the permanent resolved. In *Proc. 20th ACM STOC*, pages 235–244, 1988.

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

[36] M. Stumpf, C. Wiuf, and R. May. Subnets of scale-free networks are not scale-free: sampling properties of networks. *PANAS*, 102(12):4221, 2005.

[37] 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.

[38] 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 *Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on*, pages 123–128. IEEE, 2011.

[39] Y. Wang, J. Liang, and J. Lu. Discover hidden web properties by random walk on bipartite graph. *Information Retrieval. Springer. 27 pages. in press.*, 2013.

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

[41] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.