Uniform Random Sampling Not Recommended For Large Graph Size Estimation

Jianguo Lu, Hao Wang

School of Computer Science, University of Windsor, Canada

Abstract

The norm of data size estimation is to use uniform random samples whenever possible. There have been tremendous efforts in obtaining uniform random samples using methods such as Metropolis-Hasting random walk or importance sampling [2]. This paper shows that, on the contrary to the common practice, uniform random sampling should be avoided when PPS (probability proportional to size) sampling is available for large data.

To develop intuition of the sampling process, we discuss the sampling and estimation problem in the context of graph. The size is the number of nodes in the graph; uniform random sampling corresponds to uniform random node (RN) sampling; and PPS sampling is approximated by random edge (RE) sampling. In this setting, we show that for large graphs RE sampling outperforms RN sampling with a ratio proportional to the normalized graph degree variance. This result is particularly important in the era of big data, when data are typically large and scale-free [3], resulting in large degree variance.

We derive the result by giving the variances of RN and RE estimators. Each step of the derivation is supported and demonstrated by simulation studies assuming power law distributions. Then we use 18 real-world networks to verify the result. Furthermore, we show that the performance of random walk (RW) sampling is data dependent and can be significantly worse than RN and RE. More specifically, RW can estimate online social networks but not Web graphs due to the difference of the graph conductance.

1. Introduction

Size estimation is a classic problem that has many applications, ranging from the war time problem of finding out the number of German tanks [14], to the more recent challenge of gauging the size of the Web and search engines [20, 2, 6, 38] and online social networks [18, 15]. The direct calculation of data size is often not possible or desirable for several reasons. Quite often, data are hidden behind some searchable interfaces and programmable web APIs, such as online social networks and deep web data sources. The access is limited, and the data in its entirety are not available [37, 18]. The data can be distributed, and there is no central data repository such as in the case of peer-to-peer networks [30] or the Web [20]. Even when the data are available in one place, there are requirements for fast just-in-time analysis of the data [17]. Regardless of a large variety of application scenarios, a common approach to solving these problems is to use samples to have a fast estimation of the data size, instead of slow and direct counting of the data.

Many datasets can be viewed as graphs, especially the ones extracted from the Web and online social networks such as Twitter and Facebook. These graphs are large, often distributed and hidden behind searchable interfaces. The sampling process requires sending queries that occupy network traffic. In addition, most data sources impose daily quotas. In such cases, the sample size has to be far less than the data size, and it is paramount to choose an efficient sampling and estimation method.

For ease of discussion, sampling is modelled in the context of a graph, where uniform sampling corresponds to uniform random node (RN) sampling, PPS (probability proportional to size) sampling corresponds to random edge (RE) sampling. In this setting, we define the size as the number of nodes in the graph. Random walk (RW) sampling approximates PPS sampling in that the sampling probability is proportional to its degree asymptotically.

State of the art The norm of size estimation is to use uniform random samples whenever possible. Real data sources seldom provide uniform random samples directly. Therefore, there have been tremendous efforts to obtain uniform random samples from the Web [16], search engine indexes [2], and online social networks [12], to name a few. These uniform random samples are costly, in that each valid sample may be accompanied by many invalid ones that are thrown away. Recently, it was empirically observed that, instead of obtaining those costly uniform random samples, RW sampling is actually better than RN sampling for size [18] and average degree estimation [26][10] on *some* datasets.

Our contribution This paper shows that the sampling methods for very large graphs should be different from the ones traditionally preferred. Instead of RW, we show that it is RE that is better than RN when the graph is very large. We demonstrate our conclusion not only empirically on 18 datasets and simulated data, but also analytically by showing that its variance is smaller in our setting. In addition, we delineated the details as for

- When is RE better than RN? RE is better than RN only when the graph is very large, and consequently, the sample size n has to be much smaller than the data size N. This is the scenario we assume, with application background such as estimating online social networks with a limited number of web-based queries.

- How much better is RE over RN? We demonstrate that there is an upper bound for the performance improvement, which is quantified by $\gamma^2 + 1$. Here γ is the coefficient of variation of node degrees. The upper bound is derived analytically, and confirmed empirically on 18 large data sets. The derivation uses the assumption that the data is very large.

- What can approximate RE sampling? When RE sampling is not available in practice, we need to resort to other methods to approximate RE (or PPS) sampling. RW is an option, but the performance varies widely from data to data. We find that RW can approximate the performance of RE for online social networks, but not for Web graphs.

This result is particularly important in the age of big data when large and scale-free networks are ubiquitous [3] [33]. These networks can have very large degree variance. In theory, γ^2 can be infinitely large when the slope of the scale-free network falls under certain range. In practice, we observe γ^2 as large as 1300 for the Twitter network in 2009 [27], meaning that potentially RE sampling can be better in three orders of magnitude in terms of variance. Such huge difference between the sampling methods will not only change the landscape of sampling practice, but also shift the research focus. In the past, people strive for uniform random samples [2]. Nowadays for very large data, we should take PPS samples, or develop sampling methods that can approximate PPS sampling.

2. The Background

Without loss of generality, this paper discusses the sampling methods and estimators in the setting of (un-)directed graph. We focus on three sampling methods on the graph, i.e., random node (RN), random edge (RE), and random walk (RW). In practice, these sampling methods can be implemented in a variety of ways, depending on the access interfaces provided by the data sources.

Take online social network as an example. Suppose that the nodes are the user accounts, an edge is a message linking two accounts. When a data source provides random access to messages, say searchable interface for messages, we can access the accounts that are connected by the message. Thereby random edge sampling is implemented. For random node sampling, some data sources may provide direct access to random accounts, or we can design a sampling scheme to get uniform random nodes. For example, in Facebook, Twitter and Weibo, user IDs are fixed-length numbers. Thus we can generate a random number within the ID space to probe valid random IDs. Appendix in [12] proves that such ID sampling will result in uniform random samples. Or, we can use rejection sampling or Metropolis-Hasting random walk to get random nodes [2, 12]. RW seems to be supported by most programmable web APIs, but the reality is more complicated. Web APIs, such as Twitter, typically do not return all the neighbouring nodes in one remote call. Instead, one call can obtain a small number of neighbours, and the number of calls have daily quota that is bound to IP addresses. With such restriction, it is costly to select a random neighbour when the out-degree is large, and some 'approximate' RW needs to be employed in the sampling. RE sampling, on the other hand, are also provided in various forms. A tweet or email is an edge that connects two users; a research paper links two authors in co-authorship network. In particular, some Web APIs provide direct random edge access. For instance, CiteSeerX, an academic paper search engine, assigns a citation number to all the citation edges in the network, ranging from 1 to 30 million, enabling the random edge sampling directly.

Table 1: Summary of notations

Notation	Meaning	Properties
N	number of nodes	
n	sample size	
f_j	number of nodes sampled exactly j times	$n = \sum jf_j$
C	number of collisions in n samples	$C = \sum \binom{j}{2} f_j$
d_i	degree of node i	
p_i	probability of node i being visited	$p_i = d_i / \sum d_i$ in RE sampling. $\sum_{i=1}^{N} p_i = 1$
$\langle d \rangle$	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	square of coefficient of variation	$\gamma^2 = \sigma^2 / \langle d \rangle^2 = \langle d^2 \rangle / \langle d \rangle^2 - 1 = \Gamma - 1$
d_{x_j}	degree of the j th sampled node	$x_j \in \{1, 2, \ldots, N\}$
$\langle d_x \rangle$	asymptotic mean degree of RE samples	$\langle d_x \rangle = \langle d^2 \rangle / \langle d \rangle$

In summary, it is difficult to obtain RN, RE, or RW samples directly in practice. Thereby, we need to know in advance what are the performance of those sampling methods, or how large the sample size has to be so that we can reach certain accuracy. For instance, if we know that the RE outperforms RN by a large margin, say by a factor of 100 according to the structure of the graph, we should go after PPS sampling no matter how 'approximate' it is.

2.1. Sampling Methods and Their Estimators

Given an undirected graph $G(V, E)$, where V is the set of nodes, and E the set of edges. Let $N = |V|$, the parameter we want to estimate. 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 graph is defined as $\sigma^2 = \langle d^2 \rangle - \langle d \rangle^2$, 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. 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:

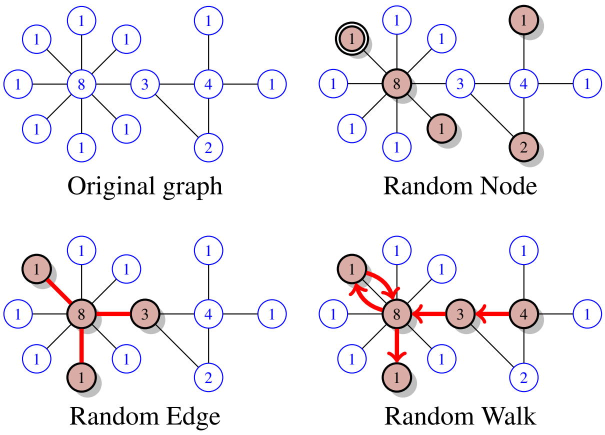
$$\gamma^2 = \frac{\sigma^2}{\langle d \rangle^2} = \frac{\langle d^2 \rangle}{\langle d \rangle^2} - 1. \tag{1}$$

Let $\Gamma = \gamma^2 + 1$. Table 1 summarizes the notations used in this paper.

Suppose that a sample of n nodes $(d_{x_1}, \ldots, d_{x_n})$ is taken from the graph, where $x_i \in \{1, 2, \ldots, N\}$ for $i = 1, 2, \ldots, n$. Among them, there are f_j nodes that are sampled exactly j times. Then, sample size $n = \sum jf_j$. Let C denote the number of collisions in the sample, i.e., $C = \sum \binom{j}{2} f_j$. Note that C is larger than the number of duplicates that is often used in capture-recapture methods [8]. Our task is to estimate N using the sample. Table 1 summarizes the notations used in this paper.

This paper focuses on three basic sampling methods, i.e., RN (random node), RE (random edge), and RW (random walk). In RN sampling, each node is sampled uniformly at random with replacement. In RE sampling, edges are selected with equal probability and two nodes incident to a random edge are collected. Thus, 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. Fig. 1 is an illustration of the sampling methods. The following subsections explain the corresponding estimators.

We shall emphasize that our method and conclusions are not limited to estimating the size of graphs. The result applies to any size estimation as long as PPS sampling is used. For instance, one concrete application is the estimation of the number of unique words in a corpus. If we sample word occurrences uniformly at random, popular words will be sampled more often. In fact, each word is sampled with probability proportional to its 'size' (word frequency). Since



Original graph

Random Node

Random Edge

Random Walk

Figure 1: A graph and three sampling methods to select six sample nodes. Nodes can be sampled multiple times.

word frequency follows a power-law with a very large variance, we should use PPS sampling to estimate vocabulary size according to our result. In addition, we can give the variance of this estimator. Thus, during the estimating process, depending on sample size, we can give the confidence interval of the estimation.

2.1.1. *RN Sampling*

Different sampling methods require different estimators. When nodes are sampled uniformly at random, each node is sampled with equal probability, i.e.,

$$p_i = \frac{1}{N}, \text{ for } i = 1, 2, \ldots, N. \tag{2}$$

When two nodes are chosen, the probability that a collision (the same node being selected twice) happens is

$$p = \sum_{i=1}^{N} p_i^2 = \frac{1}{N^2} \sum_{i=1}^{N} 1 = \frac{1}{N}. \tag{3}$$

Since there are $\binom{n}{2}$ pairs, the expected number of collisions is

$$\mathbb{E}(C) = \binom{n}{2} \sum_{i=1}^{N} p_i^2 = \binom{n}{2} \frac{1}{N}. \tag{4}$$

Thus, the RN estimator for N is

$$\widehat{N}_N = \binom{n}{2} \frac{1}{C}. \tag{5}$$

2.1.2. *RE Sampling*

When nodes are chosen with probability proportional to their sizes, the probability of choosing node i is $p_i = d_i/\tau$, where $\sum p_i = 1$. When two nodes are chosen independently at random with probability proportional to size d_i, the probability that a collision happens is

$$p = \sum_{i=1}^{N} p_i^2 = \frac{1}{\tau^2} \sum_{i=1}^{N} d_i^2 = \frac{\Gamma}{N}. \tag{6}$$

The expected number of collisions C is

$$\mathbb{E}(C) = \binom{n}{2} \sum_{i=1}^{N} p_i^2 = \binom{n}{2} \frac{\Gamma}{N}. \tag{7}$$

Thus, the RE estimator for N is

$$\widehat{N}_E = \binom{n}{2} \frac{\Gamma}{C}. \tag{8}$$

Thereby, we derived the RE estimator using Γ. The introduction of Γ in the estimator is important–it reveals the difference between the RE and RN estimators, consequently we can compare them. The same estimator in very different forms are used in [8, 18]. Our derivation is different, so that we can compare these two estimators for uniform and PPS samples. Comparing the estimators in equations 5 and 8, the only difference is that RE sampling produces Γ times more collisions using the same sample size. Consequently, the estimate is adjusted by a factor of Γ. When more collisions are observed, the accuracy of the estimation is also improved. Intuitively, RE method can outperform RN sampling by a factor of Γ. In reality, the performance improvement is upper-bounded by Γ as we will show in this paper.

The second issue is whether Γ is large enough to result in significant performance improvement for RE sampling. Our first observation is that when the graph being studied is regular, $\Gamma = 1$ and the RE estimator is reduced to the RN estimator. However, many networks are large and scale-free, inducing very large Γ. For instance, $\Gamma \approx 1300$ for the Twitter user network in the year of 2009 [27]. This large Γ makes the RE sampling the obvious choice.

The third issue is that Γ itself needs to be estimated. Γ is the ratio of the average degree of the sampled nodes and the average degree of the original graph, and can be estimated using the following formula [27]:

$$\widehat{\Gamma} = \frac{\widehat{\langle d_x \rangle}}{\widehat{\langle d \rangle}} = \frac{\sum_{i=1}^{n} d_{x_i}}{n} \frac{1}{\widehat{\langle d \rangle}}. \tag{9}$$

In turn, the average degree can be estimated by the harmonic mean with high accuracy [28]:

$$\widehat{\langle d \rangle} = \frac{n}{\sum_{i=1}^{n} 1/d_{x_i}}. \tag{10}$$

The details of the accuracy of average degree (and Γ) estimation is discussed in our previous paper [28], hence not included in this one. What we want to emphasize is that its RSE is far less than the RSE of $1/C$, which is the focus of this paper. [28] proved that the RSE of average degree estimation is upper bounded by $\langle d \rangle / \sqrt{n}$. Given that the average degree is normally small, and the sample size n is large for large data, the impact of the variance of Γ can be neglected. This can be also demonstrated by our 18 experiments reported in this paper. The purpose of having experiments in addition to the proof is to verify this assumption, and a simplification used in the derivation. In particular we choose large number of different graphs from a variety of areas to bolster our assumptions.

2.1.3. RW Sampling

In the literature, for instance in [18], an estimator equivalent to Eq. 8 was developed for samples obtained by random walk (RW). It is based on the assumption that RW sampling can approximate RE sampling in that, asymptotically, the sampling probability of a node is proportional to its degree. While RW can approximate RE sampling for well enmeshed fast mixing networks, it can differ greatly when the graph conductance is small. [18] suggested that RW sampling outperforms RN sampling on datasets IMDB, DBLP and Facebook. We prove that it is RE sampling, not RW sampling, that outperforms RN sampling. Empirically we repeated the experiments on these three datasets as well as 15 other networks. While it is true that for these three networks RW does outperform RN sampling, for some other datasets, especially the Web graphs formed by web pages and hyperlinks, we observe that RW is much worse than RE sampling.

2.2. Illustrating examples

The sampling and estimation methods can be illustrated using Fig.1, where $N = 13, \langle d \rangle = 2, \Gamma = 1.96$. Suppose that the sample degrees taken by RN, RE, and RW sampling methods are (1, 1, 1, 1, 2, 8), (1, 8, 1, 8, 3, 8), and (4, 3, 8, 1, 8, 1), respectively.

For RN sampling, there are four nodes being sampled once, one node sampled twice. Therefore, $f_1 = 4, f_2 = 1, C = 1$, and

$$\widehat{N_N} = \binom{n}{2}\frac{1}{C} = \frac{6 \times 5}{2} \times \frac{1}{1} = 15.$$

For RE sampling, three nodes are sampled once, and there is one node that is sampled three times. Therefore, $f_1 = 3, f_2 = 0, f_3 = 1, C = 3$, and

$$\begin{aligned}
\widehat{\langle d_x \rangle} &= \frac{1}{6}(1 + 8 + 1 + 8 + 3 + 8) = 4.83,\\
\widehat{\langle d \rangle} &= \frac{6}{\frac{1}{1} + \frac{1}{8} + \frac{1}{1} + \frac{1}{8} + \frac{1}{3} + \frac{1}{8}} = 2.21,\\
\widehat{\Gamma} &= 4.83/2.21 = 2.18,\\
\widehat{N_E} &= \binom{n}{2}\frac{\Gamma}{C} = 15 \times \frac{2.18}{3} = 10.90.
\end{aligned}$$

For RW sampling,

$$\begin{aligned}
f_1 &= 4, f_2 = 1, C = 1,\\
\widehat{\langle d_x \rangle} &= \frac{1}{6}(4 + 3 + 8 + 1 + 8 + 1) = 4.16,\\
\widehat{\langle d \rangle} &= \frac{6}{\frac{1}{4} + \frac{1}{3} + \frac{1}{8} + \frac{1}{1} + \frac{1}{8} + \frac{1}{1}} \approx 2.11,\\
\widehat{\Gamma} &= 4.16/2.11 = 1.97,\\
\widehat{N_W} &= \binom{n}{2}\frac{\Gamma}{C} = 15 \times \frac{1.97}{1} = 29.62
\end{aligned}$$

2.3. Sampling Other Graph Properties

Graph sampling is an active research area that starts with the basic sampling methods such as random node (RN), random edge (RE), random walk (RW), and their numerous variations and combinations [22], such as RW with uniform RN restart [1] [46]. Most work focuses on simple graphs, while some develops specific sampling methods targeting directed or even weighted graphs [36]. Quite often, this group of work studies the sampling schemes in an abstract level, independent of concrete application scenarios, so that the sampling problem can be understood better without hinderance from implementation details resulted from real applications. Some application details are discussed in subsection 2.4.

The key issues is graph sampling are what graph properties are preserved in the subgraph for each sampling method, what properties can be inferred if they are not preserved, and which sampling method is better for a particular objective. The properties that have been studied include aggregate functions (including data size [18] and average [28]), degree distributions [41], community [29, 47, 42], PageRank values [43], social network centralities [5]. For instance, [41] showed that power law slope is not preserved by RN sampling.

To answer the question regarding which sampling is better, numerous comparative studies have been conducted [35, 21, 22, 13]. For instance, [35] observed that random walk sampling can outperform MHRW (Metropolis-Hastings Random Walk) in the context of peer-to-peer networks, [21] showed that RN sampling performs better than RE sampling in approximating the clustering coefficient of the graph. In theory RW or MHRW can obtain samples from desired distributions [25]. In practice, we observed large networks that have very long mixing time, making it almost impossible to reach stationary distribution. Random jump can ameliorate the problem in general [1]. However, controlling the random restarting ratio have mixed results for different graphs.

The evaluation of these methods, especially for advanced properties, are mostly limited to empirical experiments on some datasets. One emerging direction is the impacts of graph size and topology on the sampling methods. Recently, we discovered that RE can outperform RN sampling in orders of magnitude on average degree [28] estimations for large and scale-free graph. A more interesting question is whether this phenomenon can be expanded to other properties. This paper is an advance on this direction: we show that data size estimation can be also improved greatly using random edge sampling.

This paper differs from our previous work described in [28] in that one estimates the average degree, while the other estimates size of the graph, i.e., the number of nodes in a graph. First, size estimation is a more complex problem than average degree estimation. Size estimation is based on collisions as well as the degree variation. Degree variation in turn depends on average degree. Thus, size estimation depends on the average degree estimation. Secondly, the estimators and their variances are different. We derived the variances for the RN and RE size estimators, and compared their differences.

2.4. Sampling Web Interfaces

Another set of related work is to apply sampling methods on real applications, in particular on hidden web estimation through restrictive web interfaces. Since the seminal work on estimating the size of the Web and search engines [20], query-based profiling of hidden data sources has been widely studied by the information retrieval community [7]. The key challenge is that the samples obtained from web interfaces have various types of biases [4]. For example, one of the biases is caused by non-uniform random sampling. To overcome this bias, vast amount of research has been done to obtain uniform random samples. One approach relies on Markov Chain Monte Carlo methods such as rejection sampling, importance sampling, and Metropolis-Hasting random walk (MHRW) [2]. The other approach is to exploit the web interface specifics, such as relatively small ID space in Facebook [13], prefix encoding in Youtube [48], and negation of queries [9]. Since uniform random samples are costly to obtain, there are works that adapt the estimators to account for the biased sample. For instance, [38] uses linear regression to adjust the estimator. Two problems remain open due to the high cost of the remote queries and the restrictive web interfaces: (a) how to utilize all the return results of remote queries. For instance, simple random walk sampling requires to select only one random neighbour, while in practice one remote call returns many documents or neighbours. We will develop sampling methods that do not throw away many returned data; (b) how to mimic traditional sampling methods using restrictive web interfaces: web interfaces impose many restrictions that can disable traditional graph sampling methods. For instance, a query can return only the top k matches. Many sampling methods, including simple random walk, requires that all the returns are available so that a random neighbour is selected. We will seek web interface sampling methods that can mimic those basic graph sampling methods.

3. Variances Of The Estimators

Estimators are normally evaluated in terms of bias, variance ($var(\widehat{N})$), and the combination of them, i.e., mean squared error (MSE). In [27], we discussed the bias problem, which is rather small in general. This paper focuses on the variances of the two estimators. We do not use Chebyshev's inequality for evaluation as some other papers do, because Chebyshev's inequality gives an upper bound that is valid for any data distribution. Consequently, experimental results can not be explained well using Chebyshev's inequality. We observed that the estimates are of normal distribution [44], thus there is a much tighter bounds. For instance, when relative standard error RSE$\left(= \sqrt{var(\widehat{N})}/N\right)$ is 0.1, the 95% confidence interval is roughly $\widehat{N} \pm 0.2\widehat{N}$. This is the why in our experiments the RSE values are around 0.1.

3.1. Variance of RN Sampling

We derive the variances using the classic Delta method. The key difference is the approximations we make due to the big data assumption. Otherwise, the Taylor expansion has a sequence of long terms, and loses the intuitive understanding. Let C, the number of collisions, be the random variable. The Taylor expansion of $1/C$ around $\mathbb{E}(C)$ is:

$$\frac{1}{C} = \frac{1}{\mathbb{E}(C)} - \frac{C - \mathbb{E}(C)}{\mathbb{E}(C)^2} + \frac{2}{\mathbb{E}(C)^3}\frac{(C - \mathbb{E}(C))^2}{2!}\dots \tag{11}$$

By applying *var* on Eq. 5, and taking the first two terms in the Taylor expansion, we have

$$var(\widehat{N_N}) \approx \frac{n^4}{4}var\left(\frac{1}{C}\right) = \frac{n^4}{4\mathbb{E}(C)^4}var(C). \tag{12}$$

When selecting two nodes randomly from a set of N nodes, the probability of having a collision is $p = 1/N$. When n number of sample nodes are selected, there are $\binom{n}{2}$ pairs. The number of collisions follows the binomial distribution $B(n(n-1)/2, 1/N)$ whose variance is

$$var(C) = \binom{n}{2}p(1-p) = \mathbb{E}(C)(1 - 1/N) \tag{13}$$

When N is large, $var(C) \approx \mathbb{E}(C)$. Substitute this into Eq. 12, and note that $n^2/(2\mathbb{E}(C)) = N$, we derive the following:

Lemma 1 (Variance of $\widehat{N_N}$)**.** *The estimated variance of RN estimator $\widehat{N_N}$ is*

$$\widehat{var}(\widehat{N_N}) \approx \frac{N^2}{\mathbb{E}(C)} \approx \frac{2N^3}{n^2}. \tag{14}$$

Reformulating the above result using RSE, we see that the accuracy of the estimation depends solely on the expected number of collisions:

$$RSE(\widehat{N_N}) = \frac{\sqrt{\widehat{var}(\widehat{N_N})}}{N} \approx \frac{1}{\sqrt{\mathbb{E}(C)}}. \tag{15}$$

Since the derivation employs several approximations, we conduct a simulation study to verify our result and understand its limitation. The simulation study is depicted in Fig. 2. The data size $N = 10^6$. Sample sizes range between 4472 and 14142, so that the expected collisions range between 10 and 100. For each sample size, estimations are repeated 1000 times to obtain the observed collisions and RSEs.

First, the simulation study shows that random variable C does follow the binomial distribution $B(n(n-1)/2, p)$ as depicted in panels (A) and (B) of Fig. 2. Both plots are histograms of the collisions, along with the corresponding binomial distributions. Panel (A) plots the histogram when $\mathbb{E}(C) = 10$, panel (B) is when $\mathbb{E}(C) = 100$.

Second, the observed variance of C fits the estimated variance very well over various sample sizes, as illustrated by panel (C). I.e., $\widehat{var}(C) \approx \mathbb{E}(C)$. Third, the observed RSE (or equivalently variance) fits the estimated RSE when sample size is not very small. From panel (D) we can see that RSE of $\widehat{N_N}$ is about $1/\sqrt{\mathbb{E}(C)}$ when $\mathbb{E}(C) > 20$. When $\mathbb{E}(C) = 10$, there is a gap between the estimated and observed RSEs, introduced by the Taylor expansion approximation. When $\mathbb{E}(C)$ is as small as 10, the third term in Eq.11 can be no longer omitted.

3.2. Variance of RE Sampling

The variance of RE estimator involves three variables, the collisions C, the estimated average degree $\widehat{\langle d \rangle}$ of the original graph, and the average degree of the sampled nodes $\widehat{\langle d_x \rangle}$. The variance of $\widehat{N_E}$ is too complicated to compare with that of $\widehat{N_N}$ without some assumptions. We assume that N is very large, and $C \approx 100$. Consequently $n = \sqrt{2NC/\Gamma}$. We can see that $C \ll n \ll N$. We restrict the collisions around 100 so that the corresponding $\widehat{N_N}$ estimator has RSE 0.1, or, the 95% confidence interval is $N \pm 0.2N$. Under such assumption, we can approximate the variance of $\widehat{N_E}$ as follows:

Lemma 2. *The variance of $\widehat{N_E}$ is*

$$var(\widehat{N_E}) \approx \frac{N^2}{\mathbb{E}(C)}\left(1 + \frac{2n\langle d^3 \rangle}{N\Gamma\langle d \rangle^3}\right) \tag{16}$$

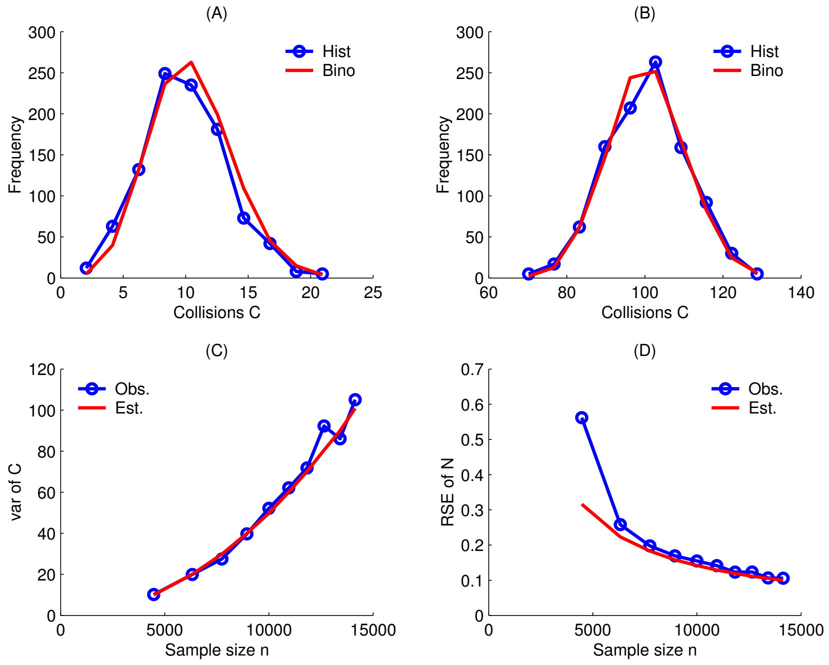


Figure 2: Variance of RN estimator.

Proof. See Appendix.

Comparing the variances for RN and RE samplings in Lemma 1 and Lemma 2, we have the following:

Theorem 1. *Given the same sample size n. The variance ratio between RN and RE sampling is:*

$$\frac{var(\widehat{N}_N)}{var(\widehat{N}_E)} \approx \Gamma\left(1 + \frac{2n\langle d^3\rangle}{N\Gamma\langle d\rangle^3}\right)^{-1} \tag{17}$$

We highlight two points regarding this result. First, when sample size $n \ll N$, the second term in Eq. 17 is small enough to be negligible. In this case, RE sampling outperforms RN sampling up to Γ folds in terms of variance, and $\sqrt{\Gamma}$ in terms of sample size.

Second, the second term grows with sample size n, indicating eventually RN will become better. The tipping point is

$$n = N\Gamma^2\langle d\rangle^3/(2\langle d^3\rangle). \tag{18}$$

When sampling large graphs, in general RE is better than RN, or $n < N\Gamma^2\langle d\rangle^3/(2\langle d^3\rangle)$, as we will show in our simulation studies and in 18 real networks. This is due to two reasons: 1) n is in the order of $\sqrt{2N/\Gamma}$ to generate enough collisions, or gain sufficient estimation precision. The ratio n/N is in the order of $O(1/\sqrt{N\Gamma})$. 2) Although in theory we can let n approach or even surpass N, the essence of sampling is to use a very small portion of the data to predicate the properties.

3.3. Simulation Study

Suppose that degree d_i follows an extended version of power law,

$$d_i = \frac{A}{(\beta + i)^\alpha}, \tag{19}$$

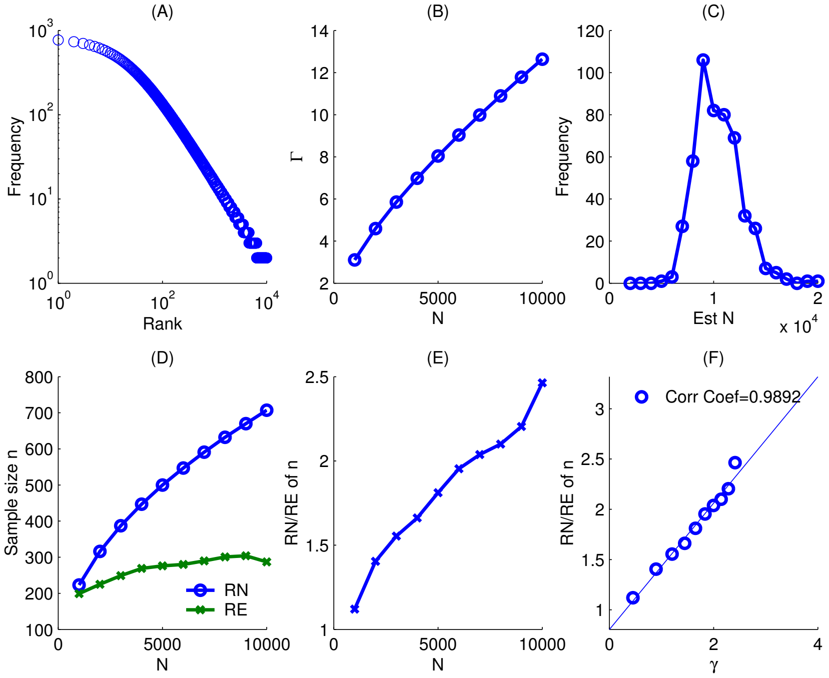


Figure 3: Variances when data size N changes. All the datasets have the same distribution. Panel (A) Degree distribution when $N = 10^4$. (B) Γ grows almost linearly with N. (C-F) Sample sizes needed to produce 0.2 RSE. (C) Histogram of 500 estimations using RE method. (D) Sample size for RE does not increase as fast as that of RN. Improvement ratio grows with N (in Panel E) and γ (in Panel F).

where A is a normalizing constant that satisfies

$$\sum_{i=1}^{N} d_i = A \sum_{i=1}^{N} \frac{1}{(\beta + i)^{\alpha}} = N\langle d \rangle, \tag{20}$$

and $\beta \ll N$ is a constant so that the top-portion of the log-log plot has a curve instead of a straight line. It is called Zipf-Mandelbrot law [32] that can model real-world data better. When $\beta = 0$, it is reduced to the standard power law. Note that the exponent α is for the degree-rank formulation. The corresponding frequency-degree version of power law has slope $-(\alpha + 1)$ [34]. Since the vast majority of networks have degree-frequency slope around -2 [33], in the following we derive the variance when the slope is exactly -2, i.e., $\alpha = 1$ in the degree-rank equation.

3.3.1. When Data Size Changes

We first demonstrate that the advantage of RE method grows with the data size, assuming the degree distribution remains the same. In this experiment, all the datasets follow Zipf-Mandelbrot law with $\beta = 50$ and $\alpha = 1$. Data sizes range between 10^3 and 10^4. Here we use relatively small data so that an overlap between RN and RE can occur, while the trend is still clear.

First, in Fig. 3 panel (A) we show the degree distribution for the graph when $N = 10^4$. We can see that the shape agrees with most real networks, particularly the flatter segment in the left upper corner of the plot. Panel (B) shows that Γ grows almost linearly with N for the same distribution. Panel (C-F) compares RE and RN samplings when RSE=0.2. All the data are obtained with 500 repetitions. Panel (C) gives an intuitive understanding for the 500 estimations when RSE=0.2 for the data $N = 10^4$. As expected, these estimations are unbiased, and follow a normal distribution. To achieve such RSE, RN method requires larger sample size when N is large, as panel (D) indicates. Sample size for RE grows very slow compared with RN. In particular, RN and RE methods are almost the same when the data is small (N=1000). To compare the growing speed, we plot the sample size ratios between RN and

RE methods. We can see that the ratio grows with data size N in Panel (E) , and with γ in panel (F). The Pearson correlation coefficient is 0.9892, indicating a linear relation between the ratio and γ, as implied by Theorem 1.

3.3.2. When Sample Size Changes

Here we show that the greatest advantage happens when sample size is small relative to the data size. This time we have a fixed data size $N = 10^4$, sample size ranges between 500 and 5000. We assume a sightly different distribution where $\beta = 0$ instead of 50 as in the previous example, so that the simplification gives us a better understanding for several parameters in Eq. 17.

Now that $\beta = 0$, the normalizing constant $A = N\langle d\rangle / \sum_{i=1}^{N} 1/i^{\alpha}$ can be characterized by the Riemann-zeta function $\zeta(\alpha) \approx \sum_{i=1}^{N} 1/i^{\alpha}$. In such distribution $N \approx A$, since the smallest degree $d_N = A/N = 1$. Utilizing the fact that $\zeta(1) \approx \ln N$, $\zeta(2) \approx 1.6$, and $\zeta(3) \approx 1.2$, we derive several approximations as below:

$$\langle d\rangle = \frac{1}{N}\sum_{i=1}^{N} d_i = \sum_{i=1}^{N}\frac{1}{i} = \zeta(1) \approx \ln N, \tag{21}$$

$$\langle d^2\rangle = \frac{1}{N}\sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N}\frac{N}{i^2} = N\zeta(2) \approx 1.6N, \tag{22}$$

$$\langle d^3\rangle = \frac{1}{N}\sum_{i=1}^{N} d_i^3 = \sum_{i=1}^{N}\frac{N^2}{i^3} = N^2\zeta(3) \approx 1.2N^2, \tag{23}$$

$$\Gamma = \langle d^2\rangle/\langle d\rangle^2 \approx 1.6N/\ln^2 N. \tag{24}$$

Substitute these equations into Eq. 17 we have

Corollary 1. *When the degree distribution follows a power law $d_i = A/i$, the variance ratio between RN and RE methods is*

$$\frac{var(\widehat{N}_N)}{var(\widehat{N}_E)} = \Gamma\left(1 + \frac{n}{0.8\langle d\rangle}\right)^{-1} \tag{25}$$

In other words, *RE* outperforms *RN* when

$$n \leq 0.8\langle d\rangle(\Gamma - 1) \tag{26}$$

$$\approx 0.8\langle d\rangle N/\ln^2 N \tag{27}$$

For large data, $n \ll N$. Therefore RE is better than RN method.

To verify our analysis, we generate synthetic data that follow the same distribution. In our synthetic data, $N = 10^4, \langle d\rangle = 10, \langle d^2\rangle = 1.71 \times 10^4, \langle d^3\rangle = 1.28 \times 10^8$, and $\Gamma = 153.82$. Compared with the formulas in Eq.'s 21 to 24, the synthetic data are rather close to those approximations. Its degree distribution is shown in Fig. 4 panel (A). On this synthetic data, we run PPS sampling for sample size ranging between 500 and 5000. For each sample size, we repeat the experiment a hundred times, record its RSE, and plot it in Fig. 4 as observed RSE. Also, we plot the RSE against sample size for RN sampling.

This experiment demonstrates that 1) Observed RSE fits well with the estimated RSE that is calculated by Eq. 30. 2) RE is better than RN sampling when sample size is small. When n grows closer to N, such performance improvement diminishes. The tipping point is $0.8\langle d\rangle(\Gamma - 1) = 0.8 \times 10 \times (\Gamma - 1) \approx 1222$ by Eq. 26, which is close to the crossing point in panel (B). Note in this simulation study we use rather small data (N=10000) to show the occurrence of the crossing point. When data size is larger, the advantage of RE method becomes more obvious.

4. Experiments on Real Networks

4.1. Datasets

We demonstrate our results on 18 datasets listed in Table 2. Most of them are from the Stanford SNAP graph collection [22]. Due to space limitation, for some network categories only one graph is reported if they have similar

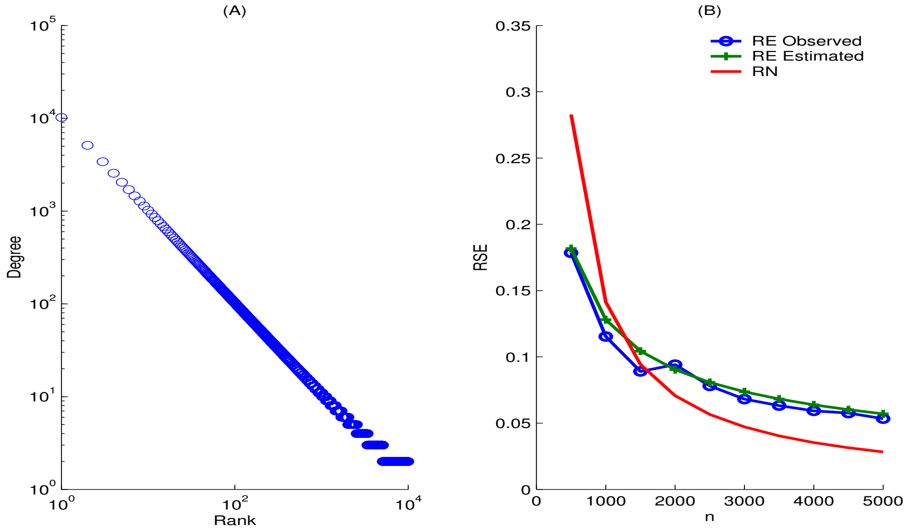


Figure 4: Variances when sample size *n* changes.

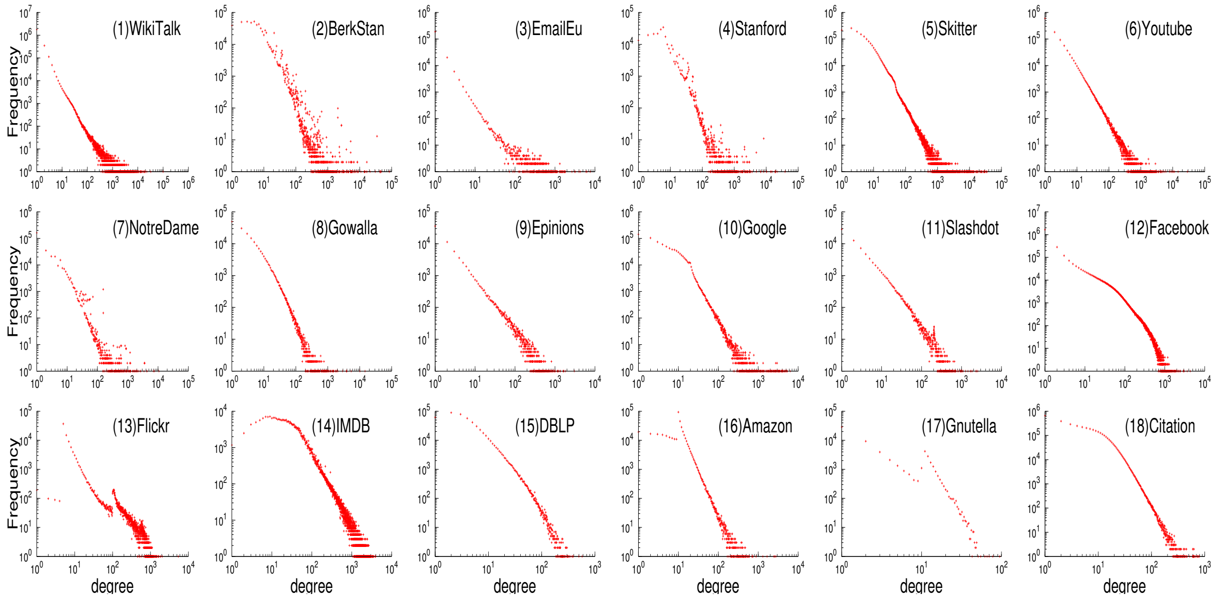


Figure 5: **Degree distributions of 18 real networks.**

Table 2: Statistics of the 18 real-world graphs, sorted in descending order of the coefficient of degree variation γ. Each graph has a citation indicating where the data is from. Φ is the conductance.

Graph	N($\times 10^3$)	γ or $\sqrt{\Gamma - 1}$	$\Phi(\times 10^{-5})$
WikiTalk [22]	2,388	26.32	2,700
BerkStan [22]	654	14.51	5.3
EmailEu [22]	224	13.66	13
Stanford [22]	255	11.51	5.8
Skitter [22]	1,694	10.46	56
Youtube [31]	1,134	9.64	440
NotreDame [22]	325	6.40	9.4
Gowalla [22]	196	5.54	1,200
Epinion [22]	75	4.02	610
Google [22]	855	4.00	62
Slashdot [22]	82	3.35	1,900
Facebook [45]	2,937	3.14	590
Flickr [22]	105	2.64	68
IMDB [3]	374	2.05	130
DBLP [11]	511	1.61	560
Amazon [22]	410	1.27	98
Gnutella [22]	62	1.21	9,100
CitePatents [22]	3,764	1.20	1,100

behaviour. For instance, citation graphs have similar degree distribution, similar coefficient of variation, and similar error ratios between RN, RE, and RW sampling. For these networks, we choose only one representative network for each category. In the category of the Web graph datasets, RW sampling deviates greatly from RE sampling. So we include several Web graphs, including the Web graph on the domains of Notre Dame, Stanford, and Berkley-Stanford, to investigate the cause for such deviation. Complete data description and programs can be found at `http://cs.uwindsor.ca/~jlu/size`, 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 degree distributions are similar to the ones we studied in simulation. 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. There are irregular data distributions, such as Flickr and Amazon that have broken trends in the plots.

Second, not all the scale-free networks are the same. They are very different in terms of γ, ranging between 1.20 to 26.32 (Γ ranges from 2.44 to 693.74). Third, Web graphs (sub-figures 4, 7, and 10) 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.

4.2. *RE vs. RN Sampling*

First, we compare the sample sizes needed to obtain the same RSE for all the datasets. We show that there is a strong correlation between $\sqrt{\Gamma}$ and RN/RE ratio. Fig. 6 plots the sample size ratio against $\sqrt{\Gamma}$ for the 18 datasets when $RSE = 0.2$ (panel A) and $RSE = 0.1$ (panel B).

The plot shows that 1) RE is better than RN consistently for all the datasets, as all the RN/RE ratio values are greater than one; 2) The ratio has a strong linear correlation with $\sqrt{\Gamma}$ as can be seen visually from the plot, and from the Pearson's correlation coefficient (0.98 when RSE=0.2 and 0.95 when RSE =0.1); 3) The improvement ratio is bounded from above by $\sqrt{\Gamma}$, as all the ratio values are below the line; 4) The markers are closer to the straight line in panel (A) than that of panel (B), validating our analysis in section 3.2 that the second term in Eq. 29 becomes more important with the growth of sample size.

Next, we use Fig. 7 to demonstrate the trend of the variances with growth of sample size, and compare the estimated variance vs. the observed ones for the 18 datasets. Sample sizes are chosen so that the largest observed RSE is approximately 0.2. So n varies from data to data. Larger RSEs are not considered because they do not produce meaningful estimations. Besides, the Taylor expansion approximation is not accurate when RSE is large as we have shown in Section 3.1. Fig. 7 shows that 1) the estimated variance agrees with the observed variance in general, especially for the datasets with small γ.

To summarize, albeit the great varieties of the datasets, RE sampling always outperforms RN sampling, and the ratio has a strong positive relation to $\sqrt{\Gamma}$ with very high correlation coefficient.

4.3. RW Sampling

RW sampling is more prevalent and supported by most real networks such as Twitter and Facebook [12]. It can be regarded as an approximation to RE sampling in that *asymptotically* the node sampling probability is proportional to its degree. Based on this assumption, the same RE estimator $\widehat{N_E}$ is used in this paper and others' such as [18, 37, 19, 26]. It was reported that RW is better than RN sampling for Twitter [26], DBLP, IMDB, and Facebook [18]. Now we run 18 datasets with 3000 repetitions. The sample size is $\sqrt{2NC}$ where C=100. i.e., the expected number of collisions is 100 for random node sampling. The comparison of three sampling methods is depicted in Figure 8. As Lemma 1 indicates, RSE of RN sampling is approximately $1/\sqrt{100} = 0.1$. For RE sampling, the same sample size will create more collisions, thereby less RSE according to Lemma 2.

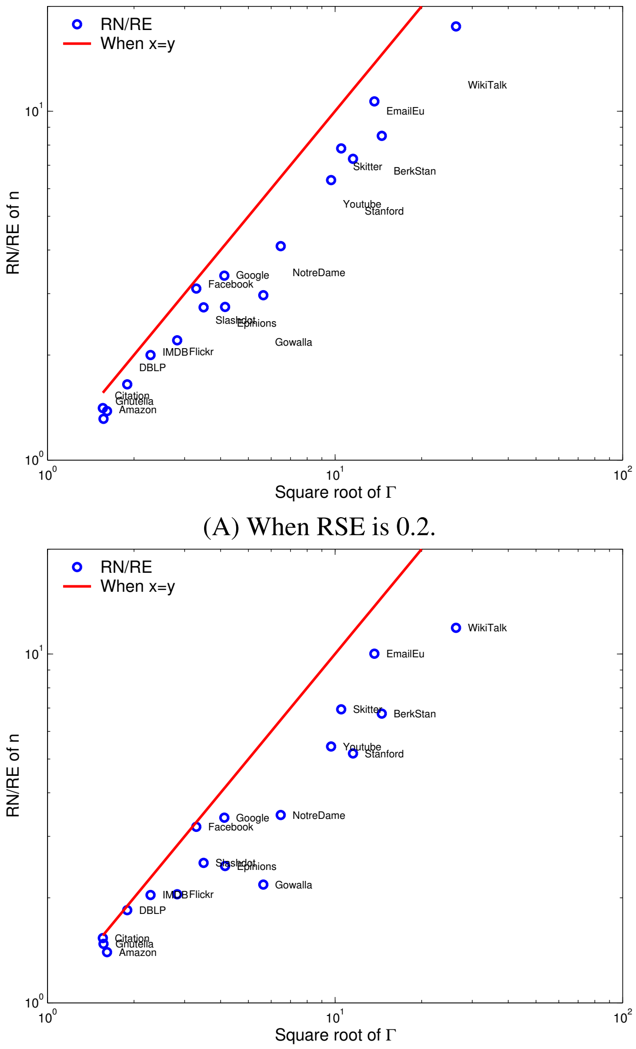
RW sampling does approximate RE sampling for many datasets, including the ones reported in the literature. However, there are several datasets (Stanford, NotreDame, BerkStan, Google, EmailEu, and Flickr) whose RW is very wrong. Most of them are Web graphs. Datasets NotreDame, Stanford, and BerkStan are the Web graphs in the domains of the universities of NorteDame, Stanford, and the combination of Berkeley-Stanford. Dataset Google is a sample Web graph collected by Google. EmailEu is a graph created from email senders and receivers. Flickr is a network created by picture sharing.

Our question is why these graphs defy RW sampling. Random walk sampling is based on the assumption that the nodes are sampled with probability proportional to its degree. This assumption can be hardly met in many real networks, mainly due to two reasons: 1) mixing time: sampling probability is proportional to its degree only after the mixing time. The mixing time can be very large when there are loosely connected components; 2) thinning rate: the estimator assumes that the nodes are sampled independently. In a random walk, a node selection is actually dependent on the previous nodes. To reduce such dependency, thinning is often applied, i.e., taking the samples every a few steps, while discarding the samples in between. More precisely, given a sequence of sampled nodes $(x_1, x_2, \ldots, x_n)$, there are correlations between the samples when they are obtained by random walk. To reduce such autocorrelation, we thin the chain by discarding all but every s-th sample. s is called the thinning rate. [24] reported that the medium of thinning rate is 40 among 21 papers that applied thinning. So we choose 40 as the thinning rate in this experiment. We also tried other thinning rates, with limited impact on the RW result.

The other more important factor is random walk mixing time, which is inversely proportional to the square of the conductance of the graph [40]. So we calculate the conductances of all the 18 graphs using SNAP graph API [23] , and plot their correlation with the RSE ratios between RW and RE sampling in Figure 9. It shows that there is a strong positive correlation between the performance of RW sampling and the log of the inverse of conductance, where the Pearson correlation is 0.8. Among the top four small conductance graphs (BerkStan, Stanford, NotreDame, and EmailEu), the conductances are in the order of 10^{-5}, and they are about ten times worse than RE sampling. On the other hand, most datasets have the ratio values close to 1, indicating that RW approximates RE sampling. Thereby it is also better than RN sampling.

For low conductance graphs, we may wonder whether longer burn-in period or random restart [1] will improve RW sampling. The answer is yes, but the performance of RW can be still far away from RE sampling. Imagine that there is a subgraph that is a bolas graph [25]–there is a long single path, connecting with a densely connected component. Suppose the size of this subgraph is k, the mixing time can be in the order of k^3 [25] in the worst case. That is, one such small component with size 100 will cost 10^6 steps to escape from the RW trap. Such large mixing time is impossible to implement, not to mention that k can be well above 100.

We demonstrate that such bolas subgraphs do exist in real networks in Figure 10. It shows the subgraphs obtained from random walks from three datasets (Flickr, EmailEu, and Stanford) whose conductances are low and one normal



(A) When RSE is 0.2.

(B) When RSE is 0.1.

Figure 6: RN/RE ratio of sample sizes is bounded from above by $\sqrt{\Gamma}$ for 18 networks. Panel (A) displays the ratio of sample sizes needed to achieve 0.2 RSE; panel (B) the ratios to achieve 0.1 RSE. RSE is obtained over 500 repetitions.

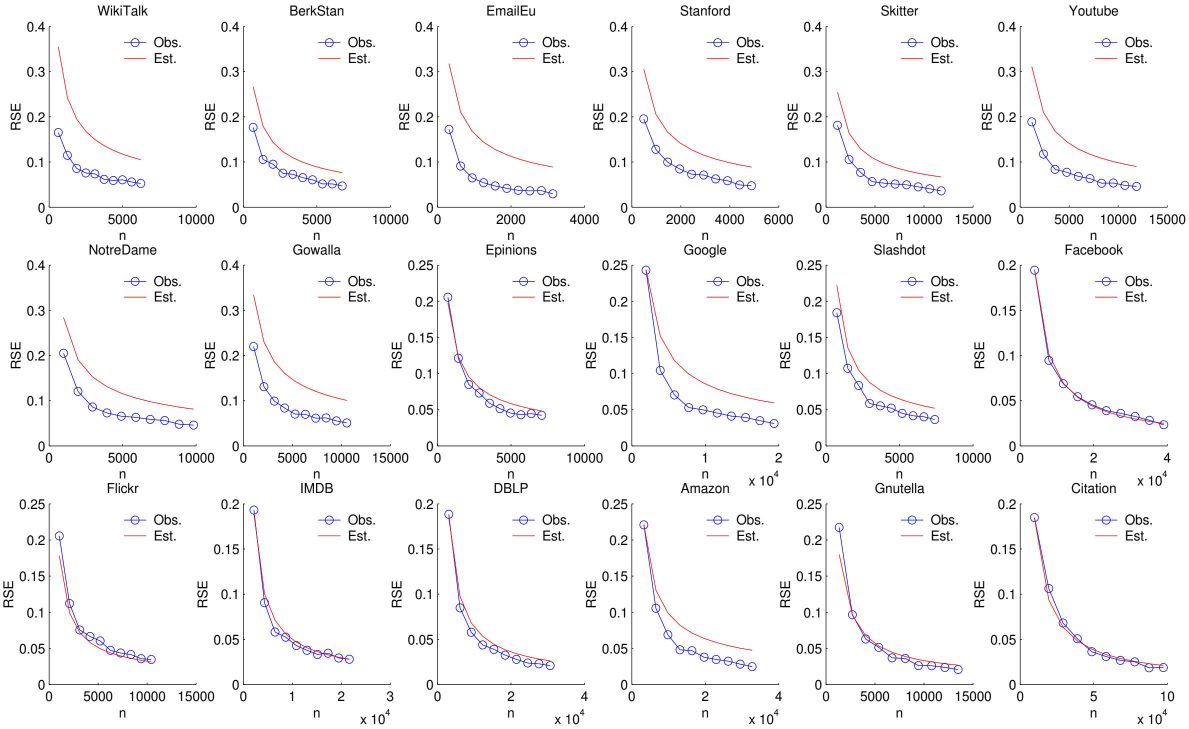


Figure 7: Observed vs. Estimated RSEs for the 18 datasets over various sample sizes.

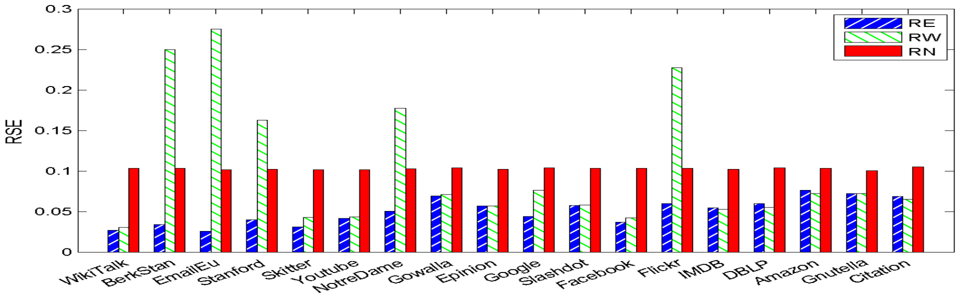


Figure 8: Comparison of three sampling methods. The sample size $n = \sqrt{2NC}$ where $\sqrt{C} = 10$. It shows that for RN sampling (red solid bars), the relative standard error is equal to $1/\sqrt{C} = 0.1$ across all the datasets. RE sampling is consistently smaller than RN sampling.RW sampling can approximate RE sampling for some datasets. For NotreDame etc. that have low conductance, RW is grossly wrong.

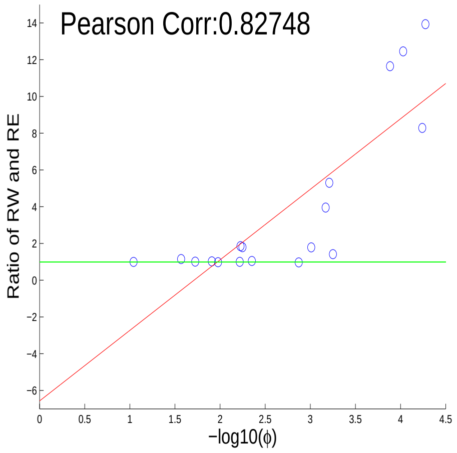


Figure 9: The ratio of RSEs between RW and RE samplings over the conductance Φ. For the four graphs with the lowest conductance, RW is around 10 times worse than RE sampling. Sample size $n = \sqrt{2N\mathbb{E}(C)}$ where $\mathbb{E}(C) = 100$. RSE is obtained over 3000 runs.

graph (Youtube) as a comparison. The node colour indicates the node degree in the original graph. It is clear that Flickr has two loosely connected components with a long narrow tube, indicated by the blue/green colour of the tube. What is more, the two components obviously have different average degrees, since one component is dominated by orange/red colour and the other by green/blue colour. It shows that RW will take long steps to escape from one component to the other. Depending on where it visited, RW will produce very different estimation.

EmailEu has a different topology even though its conductance is equally small. The subgraphs are mostly stars, maybe caused by group emails. RW will be trapped in those large stars. Web graphs, for instance Stanford web, have many bolas as subgraphs. A densely connected subgraph can be easily created using a few computer commands, such as automated generation of documents in JavaDoc or HTML version of PPT slides. Many bolas subgraphs will make the RW on the Web almost impossible.

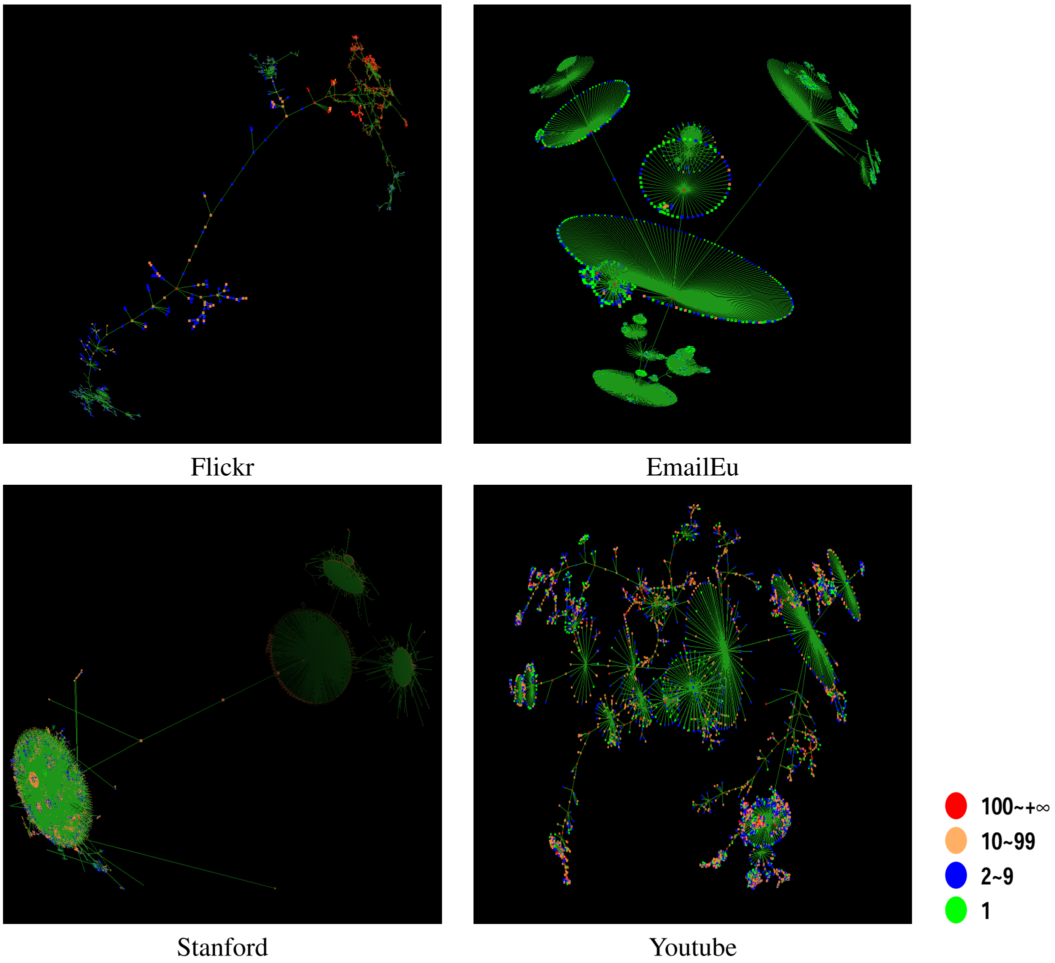
5. Discussions and Conclusions

The state of art in size estimation is to use uniform random samples whenever possible. We show that on the contrary to this common practice, PPS sampling outperforms uniform random sampling by a factor up to $\sqrt{\Gamma}$ for large data in terms of sample size.

In retrospect, this phenomenon was not observed in the past probably due to several reasons: 1) In traditional size estimation studies, Γ is typically small (between one and two), thus the difference is hardly discernible. Our result shows that the improvement ratio is up-bounded by Γ. Thus, when Γ is small, RE could be worse than RN. Even in scale-free networks, Γ in real networks may not be large due to the cut-off for the maximal values. For instance, Facebook has an up-limit of the number of followers, resulting small Γ value around two. Only recently we see large scale-free networks whose Γ value can be as high as 1000, such as Twitter and WikiTalk; 2) RE sampling is hardly studied in the past. Random walk sampling is often used, but it is only an approximation to PPS sampling. The comparison between RW and RN samplings often has a mixed results, failing to reveal a definite answer. In particular, RW on the Web graph is always worse than RN; 3) The result is true only for big data. In the synthetic data that assumes a power law distribution, we show that the improvement ratio grows almost linearly with the data size. When the data size is very small, RN can be better than RE even if the network is scale-free.

This paper gives the variances of random node and random edge sampling for graph size estimation. The result is surprisingly simple for RN sampling: the relative standard error is the reciprocal of the square root of the collisions. As a rule of thumb, if we want the 95% error bound to lie within the range $\pm 0.2N$, the expected number of collisions should be 100.

In RE sampling the large nodes tend to be sampled more often, resulting in higher collisions given the same sample size. It is easy to understand that RE sampling requires a smaller sample size to produce the same number


Flickr

EmailEu

Stanford

Youtube

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

of collisions, or the same standard error. What is more interesting is that we can quantify the difference using the coefficient of variation of node degrees. So the second rule of thumb is that the ratio of RSEs between RE and RN samplings has an upper bound $\sqrt{\Gamma}$.

To emphasize the importance of the result, we would like to point out that this result is not restricted to estimation problems in graph. It can be applied to any size estimation problems where there is no graph at all. In that case, RN sampling corresponds to uniform random sampling, RE sampling corresponds to PPS sampling. This paper is written in the setting of graph because: 1) there are many large datasets that are in the form of graphs; 2) graphs give a tangible explanation for PPS sampling; 3) RW exists in graph only, and it is a perfect example for a sampling method that approximates PPS sampling (or RE sampling).

Traditionally RW sampling is studied more often, but its relationship with RN sampling is hard to construct. With clear understanding of the relationship between RE and RN samplings, we can infer whether RW sampling is better than RN sampling. The third rule of thumb is that if the graph does not have loosely connected components, most probably RW will be better. This is because the random walk mixing time is small, and RW can approximate RE sampling. This explains why RW is better for the datasets (DBLP, IMDB, and Facebook whose conductances are high) in [18], and why various methods such as random restart need to be proposed to improve the simple random walk for datasets such as Flickr [37].

From another perspective, this paper explains the results obtained in the past experiments such as [18], and predicts future empirical results if there will be. There are many data sources that have different graph topologies and different sampling interfaces. These sampling interfaces enable various sampling methods and their approximations. We can envision that there will be numerous empirical results in the pipeline from the combinations of datasets, interfaces, sampling methods and the estimators. Our results can help people find the correct combination to produce excellent empirical conclusions.

As a corollary, this paper implies that RW sampling is not good for the estimation of the properties of the Web. For all the Web graphs we studied, including the ones listed in this paper (NotreDame, Stanford, BerkStan, Google), they all have loosely connected components, resulting in very large estimation error. This may explain why the Web is usually not sampled by RW.

This observation also reveals a fundamental distinction between the Web and online social networks such as Facebook and citation networks. The Web is created with the help of computer programs. A single computer instruction can spawn a large subgraph that is loosely connected to other parts. On the other hand, online social networks evolve more naturally with full participation of people. It is unlikely large loosely connected component can be engineered in movie actor networks, Facebook, Twitter, or citation networks. We conjecture that random walk works for the networks created by humans, but not for the networks created by computers.

We want to emphasize that our result is not restricted to graph size estimation. It applies to any data size estimation as long as the data are sampled uniform randomly or sampled proportional to their sizes. We use graph sampling to explain our result because 1) Graph model provides an intuitive way to explain different sampling methods; 2) Random Walk sampling is unique in graph and can be compared and evaluated against other sampling methods; 3) It fits naturally to many applications such as online social networks, query-based sampling of deep web and search engines where the documents and queries form a bipartite graph.

6. Acknowledgements

The research is supported by NSERC discovery grant (RGPIN-2014-04463).

7. Appendix: Proof of Lemma 2

Proof. The major component of the variance depends on the random variable C. Applying the same Taylor expansion on $1/C$ as in Lemma 1, we derive:

$$var(\widehat{N_E}) \approx N^2 \frac{var(C)}{\mathbb{E}(C)^2} \tag{28}$$

$var(C)$ is given by [39]:

$$var(C) = \binom{n}{2}\sum_{i=1}^{N} p_i^2 + n(n-1)(n-2)\sum_{i=1}^{N} p_i^3$$
$$- n(n-1)(n-3/2)(\sum_{i=1}^{N} p_i^2)^2.$$

It can be approximated by the following, bearing in mind the assumption that $\mathbb{E}(C) \approx 100 \ll n \ll N$, $p_i = d_i(N\langle d\rangle)^{-1}$, and $\binom{n}{2}\sum_{i=1}^{N} p_i^2 = \mathbb{E}(C)$:

$$var(C) \approx \mathbb{E}(C) + n^3\sum_{i=1}^{N} p_i^3 = \mathbb{E}(C)\left(1 + \frac{2n\langle d^3\rangle}{N\Gamma\langle d\rangle^3}\right). \tag{29}$$

Plugging Eq. 29 into Eq. 28, we obtain the result in Lemma 2:

$$var(\widehat{N_E}) \approx \frac{N^2}{\mathbb{E}(C)}\left(1 + \frac{2n\langle d^3\rangle}{N\Gamma\langle d\rangle^3}\right) \tag{30}$$

We can omit the variances of $\widehat{\langle d\rangle}$ and $\widehat{\langle d_x\rangle}$ because their RSEs are much smaller than that of C, which is

$$RSE(C) \approx \left[\frac{1}{\mathbb{E}(C)} + \frac{n^3}{\mathbb{E}^2(C)}\sum_{i=1}^{N} p_i^3\right]^{1/2}. \tag{31}$$

First, $RSE(\widehat{\langle d\rangle}) < \sqrt{\frac{\langle d\rangle}{n}}$ according to [28]. It is negligible compared with the first term of RSE(C), which is $1/\sqrt{\mathbb{E}(C)}$. For $\widehat{\langle d_x\rangle}$,

$$var(d_{x_i}) < \mathbb{E}(d_{x_i}^2) = \sum_{i=1}^{N} p_i d_i^2 = \langle d\rangle^2 N^2 \sum_{i=1}^{N} p_i^3. \tag{32}$$

Applying the fact that $\langle d_x\rangle = \langle d\rangle\Gamma$, we derive

$$RSE(\widehat{\langle d_x\rangle}) < \sqrt{\frac{N^2}{n\Gamma^2}\sum_{i=1}^{N} p_i^3} = \sqrt{\frac{n^3}{4\mathbb{E}^2(C)}\sum_{i=1}^{N} p_i^3}. \tag{33}$$

Hence, it is smaller than the second term of RSE(C), which is often negligible compared with the first term. □

8. References

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

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

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

[4] Krishna Bharat and Andrei Broder. 1998. A technique for measuring the relative size and overlap of public Web search engines. *Comput. Netw. ISDN Syst.* 30, 1-7 (1998), 379–388.

[5] Stephen P Borgatti, Kathleen M Carley, and David Krackhardt. 2006. On the robustness of centrality measures under conditions of imperfect data. *Social networks* 28, 2 (2006), 124–136.

[6] Andrei Broder and et al. 2006. Estimating corpus size via queries. In *CIKM*. ACM, 594–603. DOI:http://dx.doi.org/10.1145/1183614.1183699

[7] Jamie Callan and Margaret Connell. 2001. Query-based sampling of text databases. *ACM Trans. Inf. Syst.* 19, 2 (2001), 97–130. DOI:http://dx.doi.org/10.1145/382979.383040

[8] A. Chao, SM Lee, and SL Jeng. 1992. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics* 48, 1 (1992), 201–216.

[9] A. Dasgupta, G. Das, and H. Mannila. 2007. A random walk approach to sampling hidden databases. In *SIGMOD*. ACM, 629–640.

[10] Anirban Dasgupta, Ravi Kumar, and Tamas Sarlos. 2014. On estimating the average degree. In *Proceedings of the 23rd international conference on World wide web*. International World Wide Web Conferences Steering Committee, 795–806.

[11] dblp. 2013. from http://www.sommer.jp/graphs/. (2013).

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

[13] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. 2010. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*. Ieee, 1–9.

[14] Leo A Goodman. 1954. Some Practical Techniques in Serial Number Analysis. *J. Amer. Statist. Assoc.* 49, 265 (1954), 97–112.

[15] Stephen J Hardiman and Liran Katzir. 2013. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 539–550.

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

[17] Howie Huang, Nan Zhang, Wei Wang, Gautam Das, and A Szalay. 2012. Just-in-time analytics on large file systems. *Computer* 61, 11 (2012), 1651–1664.

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

[19] M. Kurant, C.T. Butts, and A. Markopoulou. 2012. Graph Size Estimation. *arXiv preprint arXiv:1210.0460* (2012).

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

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

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

[23] J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. 2009. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* 6, 1 (2009), 29–123.

[24] W.A. Link and M.J. Eaton. 2011. On thinning of chains in MCMC. *Methods in Ecology and Evolution* 3, 1 (2011), 112–115.

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

[26] J. Lu and D. Li. 2012. Sampling Online Social Networks by Random Walk. In *ACM SIGKDD Workshop on Hot Topics in Online Social Networks*. ACM, 33–40.

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

[28] Jianguo Lu and Hao Wang. 2014. Variance Reduction in Large Graph Sampling. *Information Processing and Management* 50, 3 (2014), 476–491.

[29] A.S. Maiya and T.Y. Berger-Wolf. 2010. Sampling community structure. In *WWW*. ACM, 701–710.

[30] Sandeep Mane, Sandeep Mopuru, Kriti Mehra, and Jaideep Srivastava. 2005. Network size estimation in a peer-to-peer network. *University of Minnesota, MN, Tech. Rep* (2005), 05–030.

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

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

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

[34] M. E. J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46 (2005), 323. DOI:http://dx.doi.org/10.1080/00107510500052444

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

[36] Alireza Rezvanian and Mohammad Reza Meybodi. 2016. Sampling algorithms for weighted networks. *Social Network Analysis and Mining* 6, 1 (2016), 60.

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

[38] Milad Shokouhi, Justin Zobel, Falk Scholer, and S. M. M. Tahaghoghi. 2006. Capturing collection size for distributed non-cooperative retrieval. In *SIGIR*. ACM, 316–323. DOI:http://dx.doi.org/10.1145/1148170.1148227

[39] Edward H Simpson. 1949. Measurement of diversity. *Nature* (1949).

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

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

[42] Chao Tong, Yu Lian, Jianwei Niu, Zhongyu Xie, and Yang Zhang. 2016. A novel green algorithm for sampling complex networks. *Journal of Network and Computer Applications* 59 (2016), 55–62.

[43] A. Vattani, D. Chakrabarti, and M. Gurevich. 2011. Preserving personalized pagerank in subgraphs. In *Proceedings of ICML*.

[44] Yan Wang, Jie Liang, and Jianguo Lu. 2014. Discover hidden web properties by random walk on bipartite graph. *Information Retrieval* 17, 3 (2014), 203–228.

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

[46] Xiao-Ke Xu and Jonathan JH Zhu. 2016. Flexible sampling large-scale social networks by self-adjustable random walk. *Physica A: Statistical Mechanics and its Applications* 463 (2016), 356–365.

[47] Seok-Ho Yoon, Ki-Nam Kim, Jiwon Hong, Sang-Wook Kim, and Sunju Park. 2015. A community-based sampling method using DPL for online social networks. *Information Sciences* 306 (2015), 53–69.

[48] J. Zhou, Y. Li, V.K. Adhikari, and Z.L. Zhang. 2011. Counting YouTube videos via random prefix sampling. In *SIGCOMM*. ACM, 371–380.