2013

Comparative Study of Sampling Methods for Online Social Networks

Hao Wang
University of Windsor

Comparative Study of Sampling Methods for Online Social Networks

by

Hao Wang

A Thesis
Submitted to the Faculty of Graduate Studies
through Computer Science
in partial fulfilment of the requirements for
the Degree of Master of Science at the
University of Windsor

Windsor, Ontario, Canada
2013

Comparative Study of Sampling Methods for Online Social Networks

by

Hao Wang

APPROVED BY:

Dr. Zhiguo Hu, External Reader
Department of Mathematics and Statistics

Dr. Luis Rueda, Internal Reader
School of Computer Science

Dr. Jianguo Lu, Advisor
School of Computer Science

Dr. Arunita Jaekel, Chair of Defense
School of Computer Science

17th May, 2013

Declaration of Co-Authorship/Previous Publication

1. Declaration of Co-Authorship

   I hereby declare that this thesis incorporates the outcome of joint research undertaken under the supervision of Dr. Jianguo Lu. The collaboration is covered in Chapters 2, 3, 4 and 5 of the thesis. In all cases, I conducted all the experiments including data collection and analysis.

   I certify that I have properly acknowledged the contribution of other researchers to my thesis, and have obtained written permission from the co-author to include the above material(s) in my thesis.

   I certify that, with the above qualification, this thesis, and the research to which it refers, is the product of my own work.

2. Declaration of Previous Publication

   This thesis includes four original papers that have been previously submitted to the journal and conference, as follows:

   I certify that I have obtained written permission from the copyright owner(s) to include the above published material(s) in my thesis. I certify that the above material describes work completed during my registration as graduate student at the University of Windsor.

   I declare that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis are fully acknowledged. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright

Thesis Chapter	Publication Title	Publication Status
2	Hao Wang, Jianguo Lu, *Uniform Random Sampling on Graph: To Be or Not To Be?*	Submitted to Information Processing and Management(IPM), 2013.
3	Hao Wang, Jianguo Lu, *Uniform Random Sampling not Recommended for Size Estimation.*	Submitted to Physica A: Statistical Mechanics and its Applications, 2013.
4	Hao Wang, Jianguo Lu, *Detect Inflated Follower Numbers in OSN Using Star Sampling.*	Submitted to The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013.
5	Hao Wang, Jianguo Lu, *What Do Large Networks Look Like?*	Submitted to The 19th ACM SIGKDD Conference on Knowledge, Discovery, and Data Mining, Workshop, 2013.

Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other Institution.

Abstract

The properties of online social networks are of great interests to the general public as well as IT professionals. Often the raw data are not available and the summaries released by the service providers are sketchy. Thus sampling is needed to reveal the hidden properties and structure of the underlying network. This thesis conducts comparative studies on various sampling methods, including Random Node(RN), Random Walk(RW) and Random Edge(RE) samplings. The properties to be discovered include the average degree and population size of the network. Additionally, this thesis proposes a new sampling method called STAR sampling and applies this method to an online social network Weibo. Furthermore, visualization of network structure is studied to explain the impact of network structure on the performance of sampling methods. We show that RE sampling is better than RN sampling in general. This result is supported by over 20 real-world networks.

Dedication

To my grandmother Suzhen Zhang(1927 - 2013)
for teaching me to be a honest and hard-work person.

Acknowledgements

I would like to acknowledge the important role of my thesis committee members and thank them for their enlightening and encouraging comments and reviews.

I wish to express my great gratitude to my supervisor Dr.Jianguo Lu for his valuable assistance and support during my thesis work, and for his persistent guidance throughout my study during master program.

My parents' unconditional love guide me through my life and I am grateful to my parents for providing me with flesh and soul. A heartfelt thanks goes out to my girlfriend Shuangshuang Dong for all your love, support and patience while we are seperated distantly.

To all who help me get through the graduate time.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The properties and structure of online social networks are of interest to a variety of stakeholders, including the general public as well as IT professionals. With the knowledge of the topology of the network, users can post their status on the network, so that their information can diffuse more effectively. Often the raw data are not available and the summaries released by the service providers are sketchy. Thus sampling is needed to reveal the hidden structure of the underlying data.

Online social networks are so large that exhaustive exploration of the network is infeasible. In fact, we can only obtain a small sample of the network and estimate the properties of the network using the sample.

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

$$d = \frac{1}{n} \sum_{i=1}^{n} d_i \tag{1.1}$$

The sample mean is an unbiased estimator of the population if the users can be selected randomly with uniform distribution. Unfortunately this is not the case in practice. When microbloggers are selected, they are often not picked randomly due to the limited access methods.

Average degree is just one of the many properties that are of interest. Other properties include the order (the number of nodes), the size (the number of edges) of the graph, the distribution of degrees, the diameter of the graph, the centralities commonly used in social network measurement such as betweenness, the closeness, the eigenvector centrality, the clustering coefficients. All those properties can be calculated with the presence of the complete data, even though some of the properties can not be computed efficiently.

Different sampling methods shall be employed depending on the properties we want to reveal. On the one hand, we need to learn the macroscopic properties such as size of the network and the average number of followers. These properties can be better discovered using uniform sampling, i.e., every account is sampled using equal probability. On the other hand, it is interesting to find out the top bloggers, their topological structure, or even clusters. Those top bloggers are easier to have higher probabilities of being sampled.

This thesis contains four papers addressing different aspects of OSNs analysis including degree estimation, size estimation, follower estimation and visualization. The following sections in this chapter introduce the background in general terms of four papers.

1.2 Average Degree Estimation

The norm of practice in estimating graph properties is to obtain uniform random (node) samples whenever possible. Often uniform random node (RN) samples are hard to obtain, henceforth the less costly simple random walk (RW) sampling is applied instead.

Chapter 2 contains our paper that tries to answer the question as for which method is better in estimating average degree, disregarding the extra cost to obtain uniform random samples by methods such as rejection sampling. Two basic sampling schemas are UR(Uniform Random) sampling and PPS(Probability Proportional to Size) sampling. In UR sampling, each node is sampled with equal probability, thus the sample $(d_{x1}, d_{x2}, \ldots, d_{xn})$ is uniform at random. The arithmetic mean is applied in estimating average degree:

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

In PPS sampling a node is sampled with probability proportional to its size. The harmonic mean is used:

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

Corresponding to these two basic sampling schemes, a graph can be sampled by RN(Random Node) and RE/RW(Random Edge/Random Walk) sampling methods. We conduct experiments on 18 real-worlk large networks and evaluate the accuracy of three sampling methods in terms of RRMSE(Relative Rooted Mean Square Error).

After comparing the results of the sampling methods, we find that when the network is large and scale-free, RE sampling is much better than RN sampling and even if the network is not large or not following the right distribution, we still can distinguish the good from the bad by using the coefficient of variation of the degree.

Since RE is not practical in some real applications, we use RW sampling instead. From the comparison, we also find that RW can be better than the costly RN in orders of magnitude for some datasets, yet it can be worse for some other datasets, depending on the degree variance and conductance of the graph. Furthermore, we show the ratio between RW and RE sampling depends on the structure of the graph which we use conductance to describe.

1.3 Size Estimation

Chapter 3 contains our paper that describes the performance of different sampling methods in estimating the network population. Population is the very fundamental property of the network and size estimation has been widely studied. As same as average degree estimation, we still compare RN, RW and RE sampling methods. The estimator for RN sampling is:

$$\widehat{N}_N = \binom{n}{2} \frac{1}{C} \approx \frac{n^2}{2C}.$$

While the estimator for RW and RE sampling is:

$$\widehat{N}_E = \Gamma \binom{n}{2} \frac{1}{C} \approx \widehat{\Gamma} \frac{n^2}{2C}$$

where C is the collision(s) in the sample and $\Gamma = \gamma^2 + 1$ is the coefficient of variation of the degree.

Based on these two estimators, we derive the variance for RN sampling and RE sampling. We verify the estimated variance with the observed variance from the experiment and they agree with each other very well. This gives us the confidence to further quantify the difference between RE and RN sampling methods. The only difference between RE and RN sampling is Γ and this suggests that with the same sample size n, RE sampling creates Γ times more collisions. Therefore the variance of $\widehat{N}_E$ is smaller by a factor of Γ. The experiment result confirms that $\sqrt{\Gamma}$ is the upper bound for the ratio of RSEs(Relative Standard Error) between RE and RN sampling methods.

We also use RW sampling to approximate RE sampling since RE is rarely possible in real applications. We find that random walk mixing time plays an important role in the performance of size estimation. If there is no loosely connected components existing, RW is better than RN, otherwise, RW will be increadibly bad.

1.4 Follower Estimation

Chapter 4 contains our paper which describes follower estimation by a high-efficient sampling method called STAR sampling and this sampling method derived from simple RW sampling. Different from simple RW sampling that takes only one sample at each step, STAR sampling takes all the neighbours as valid samples. As a result, It is more efficient than random walk sampling by a factor of the average degree. Moreover, STAR sampling is a kind of PPS sampling indicating large nodes will be sampled with higher frequency which are proportional to their degrees. We benifit from this character that we can easily focus on the sub graphs of the top users.

We apply STAR sampling in estimating the number of followers of the top nodes who have the most links in the graph. We first conduct experiment on six local datasets whose ground-truth are known. The results on six networks show that our method works very well as the empirical results according to the true values.

Then we apply our method on Weibo, the Chinese version of Twitter, whose properties are rarely studied albeit its enormous size and inuence. Before applying star sampling, we use ID sampling(known as Rejection-Acceptance Sampling) to estimate the size of Weibo. The sample obtained by ID sampling is a uniform random sample, as the by-product of the uniform random sample, the degree distribution has been plotted to give a direct impression of Weibo. Surprisingly, we find that the average in-degree is larger than the average out-degree.

We explain this by investigating the number of followers for the users in the network. After pouring the uniform random IDs, we get a STAR sample to estimate the numbers

of followers of the top 1000 users. In general the estimated follower number is consistent with the claimed number, but there are cases that the follower numbers are inflated by a factor up to 132.

1.5 Visualization

Chapter 4 contains our paper that uses random spanning tree to visualize large online social networks. Since visulization could bring too much benifit to realize the topology of large networks, many research works have been done to draw the structure of large graphs. However, most approaches can only handle small graphs with a couple hundred nodes and edges. To show the overall structure of OSNs with huge size, we reduce the number of nodes and edges by producing a representative subgraph. This subgraph is produced by simple random walk, due to its PPS character, the nodes are sampled with probability proportional to their degrees, so that large nodes with more connections have a higher probability of being sampled. The edges are reduced further using uniform random spanning tree. We use NCP(Network Community Profile) plots to explain that the subgraph produced by our method preserves the structure of original graph, thus it is representative. Lastly, we visualize six real-world networks and explain their conductance values by inspecting the visualizations.

Chapter 2

Uniform Random Sampling on Graph: To Be or Not To Be?

This paper was submitted as:

Hao Wang, Jianguo Lu, *Uniform Random Sampling on Graph: To Be or Not To Be?*. Information Processing and Management(IPM), 2013, submitted.

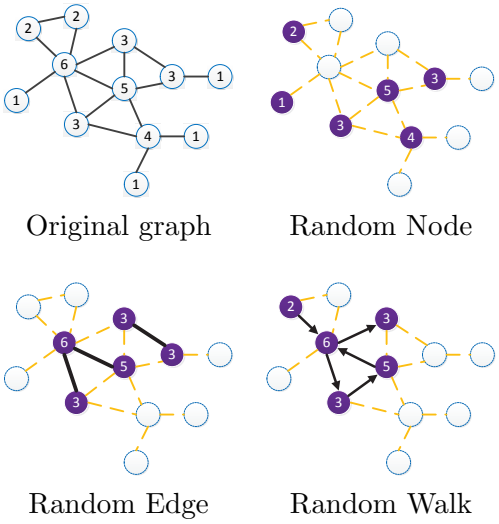


FIGURE 2.1: A graph and three sampling methods to select six sample nodes. The three sampling methods are random node (RN), random edge (RE), and random walk (RW). Nodes can be sampled multiple times as shown in sub-figures for RE and RW samplings.

2.1 Introduction

Many datasets can be viewed as graphs, especially the Web and online social networks such as Twitter and Facebook. These graphs are large, and sometimes are hidden behind searchable interfaces. Thus, the direct process of the graphs is not efficient or feasible, and sampling is the other option to reveal the hidden properties or structure of the underlying data. In the past, extensive research was carried out to explore the profile of search engines [2] and other data collections [3–5]. Many of them focused on obtaining uniform random samples [6, 7]. Recently the problem becomes more urgent due to the wide spread adoption of big data, resulting in a spate of research on this topic, such as [8–13] to name a few of them.

Two of the basic sampling methods are uniform random sampling and PPS (Probability-Proportional-to-Size) sampling. In uniform random sampling, each item is sampled with equal probability. In PPS sampling an item is sampled with probability proportional to its size. Corresponding to these two basic sampling schemes a graph can be sampled by random nodes (RN) and random edges (RE) as illustrated in Figure 2.1. In random node sampling, each node is sampled uniform randomly. In random edge sampling, two nodes incident to a random edge are collected. Random walk (RW) sampling approximates the random edge sampling by selecting the next random node in the current neighbourhood.

A fundamental question in graph sampling is which sampling method is better. There is no single answer to this question because it depends on the graph properties to be revealed, and the graph structure being investigated [10]. This paper tries to answer a narrowed-down question: which sampling method is better for estimating the average degree of a scale-free graph? By focusing on one property, we can give a more definite answer, and then expand the answer to other properties. Average degree itself is an important metric for any network. Furthermore, it can lead to the discovery of other properties such as the population size of a network [14, 15], the variation of the degrees, and even the threshold value for the occurrence of large components in message diffusion [16]. Section 2.3.4 will discuss the utilization of the average degree in the estimation of these properties in more details.

While it is easy to understand that RN sampling does not work well for scale-free networks due to its large variation of degrees, it is far from clear whether other sampling methods can reduce the variance. Take Twitter and Facebook, two popular online social networks, for example. They are both scale-free networks, yet they favour different sampling methods. Twitter prefers RE sampling, but Facebook is the other way around. What is more surprising is that for Twitter network, RE sampling is hundreds of times better than RN sampling in terms of sample size. Such huge difference has a great impact on the theory and practice of big data analysis. In practice, the selection of the correct sampling method can save the cost in orders of magnitude. In theory, new sampling methods need to be developed to exploit this difference. Naturally, we ponder when RE is better than RN sampling, and how much better it can be in reality.

The main contribution of this paper is that the normalized variance of the degrees dictates the sampling method we should use. More precisely, the ratio of the estimation errors between RE and RN samplings can be quantified by the coefficient of variation (γ) of the degrees of the data. Their Pearson's correlation coefficient is as high as 0.9354 among 18 networks we studied. In theory, the variance and γ may not exist when the slope of the power law distribution is between 1 and 2. In practice, all the real networks have a maximal degree, resulting in a bounded variance for each network. Among 18 networks we studied, most of them have γ ranging between 1 and 36. In other words, RE can be roughly 36 times better than RN sampling in terms of estimation error, or hundreds of times better in terms of sample size. On the other hand, RN sampling can outperform RE sampling when γ is small, as shown by four datasets in our experiments described in Section 2.4.

This empirical result can be justified by our derivations based on the assumption of the scale-free degree distribution. We prove that if the data satisfy Zipf's law with exponent one, the variance of RN sampling grows linearly with the data size, while the variance

of RE sampling grows logarithmically. In RE sampling, the variance of the estimator is dependent on the variance of the reciprocals of the degrees instead of the degrees themselves as shown in Section 2.3. The variance of the reciprocals has not been studied in literature. Our experiments on 18 real-world networks show that it is indeed smaller or similar to the variance of the degrees.

Based on the analysis on RE and RN samplings, we extend the comparison to random walk (RW) sampling, since RW sampling is often preferred in real applications [17, 18]. RW sampling is an approximation to RE sampling in that RW also samples nodes with probability proportional to their degrees, but only asymptotically. Because of this difference, we show that RW sampling is always worse than RE sampling for all the datasets. For some networks, RW sampling is very close to RE sampling, while others are much worse. The question is when RW can approximate RE sampling.

This paper shows that the ratio of standard errors between RW and RE sampling is dependent on the conductance of the graph. When the conductance is small, there are loosely connected components, causing RW being trapped in a component. However, it can be still better than RN sampling when there are no loosely connected components, or the conductance of the graph is not very small. When there are only two options to select from, namely RW or RN, the choice will be rather difficult because there are two factors we need to take into consideration: γ and conductance Φ.

Our results on these basic sampling methods also shed light on the directions to devise new sampling methods. Since RE is the best sampling method but may not be supported in real applications, what we need to do is to approximate RE sampling as much as possible based on the provided sampling interfaces. Simple random walk is one of the approximations [19] [17], but it may suffer from random walk traps due to loosely connected components. To overcome this problem, simple random walk can be improved by uniform random restart as verified in [20]. To make it closer to RE sampling, we can select the random restart node with probability proportional to its degree, and make the random restart more frequent. In the extreme case when random walk restarts after every one step, the RW sampling method morphs into RE sampling.

2.2 Related work

Graph sampling has been widely studied [10, 13], and finds its applications in online social networks [9, 12, 17, 18], real social networks [21, 22], web graphs [6], and search engine index and deep web [7, 23, 24]. The typical underlying techniques include Metropolis Hasting Random Walk (MHRW) [25] for uniform sampling and Random Walk (RW)

[19] for unequal probability sampling. The norm of the practice is to use uniform random node (RN) samples whenever possible. Only recently there are a few work on the comparison between RN and RW sampling.

Some research compared RW sampling with MHRW sampling [17, 26] instead of uniform random samples. Although MHRW does produce uniform random samples, it incurs additional unknown cost that is usually rather high. Therefore it is easier to observe that RW can be better than MHRW sampling. We compare RW with uniform random samples ignoring the cost of obtaining these random samples, thereby removed the noise introduced by MHRW. Rasti et al. observed that random walk sampling can outperform MHRW in the context of peer-to-peer networks [26], Gjoka et al. showed that RW (called re-weighted random walk in their paper) and MHRW are comparable [17]. We make a stronger claim that RW can outperform RN sampling even when the cost of uniform random sampling is ignored.

One of the few direct comparisons between RW and RN sampling is done by Katzir et al. [14] for the estimation of network size, not for average degree. They showed that RW sampling could outperform RN sampling in synthesized data and several real-world networks. We show that the result is data dependent– RW outperforms RN sampling only when the graph does not have loosely connected components. Instead of comparing RW and RN directly, we break it down into two subproblems, i.e., the comparison between RN and RE, and the comparison between RW and RE.

Our earlier work on the comparison between RW and RN samplings on Twitter data [16] motivated the studies conducted in this paper. [16] found that on Twitter data RW sampling is much better than RN sampling. Our further study on dozens of other datasets finds that it is not always true. We identify 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 insures that RW sampling can approximate RE sampling.

The harmonic mean estimator was first derived and studied in depth by Salganik et al. [21] to estimate the properties of hidden population 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 [27] for two sampling schemes, i.e., random sampling and the degree dependent sampling of the nodes. One result of the paper is that in random node sampling the degree distribution still follows power law if the original network is scale-free, with a steeper slope. For average degree estimation, both [28] and [29] used uniform random sampling of the nodes. [28] discussed the lower bound of the estimation. Based on this result, [29] proposed a sampling scheme that put more weight on the nodes that have less probability of being sampled.

TABLE 2.1: Summary of notations

Notation	Meaning	Properties
N	population size	
n	sample size	
d_i	degree of node i	
τ	volume of all the nodes	$\tau = \sum_1^N d_i = N\langle d \rangle$
d_{x_j}	degree of the j th sampled node	$x_j \in \{1, 2, \ldots, N\}$
p_i	probability of node i being visited	$p_i = d_i/\tau$, $\sum_1^N p_i = 1$
$\langle d \rangle$	mean degree	$\langle d \rangle = \tau/N$
$\langle d^2 \rangle$	mean of the squared degrees	$\langle d^2 \rangle = \sum_1^N d_i^2/N$
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	$\langle d_E \rangle = \langle d^2 \rangle/\langle d \rangle$

The impact of sampling methods (sampling by node, edge, and random walk) on the discovery of graph properties has also been studied in [10, 27, 30, 31]. 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, [10] investigated several network characteristics like the distribution of connected components. [31] showed that random node sampling performs better than random edge sampling in approximating the clustering coefficient of the graph. There are also works to find representative subgraphs that preserve community [32] or page rank values [33]. A related area is the data stream algorithms [34] that use a snapshot of the data to predict overall structural properties.

In contrast to the traditional sampling in ecology and social studies, the diversity of the access interfaces to web data collections opens up opportunities for designing sampling schemes that take advantages of interface specifics. For instance, [17] samples valid Facebook IDs from an ID space of 9 digits, utilizing the Facebook implementation details that make the number of invalid IDs not much bigger than the valid ones; [35] leverages the prefix encoding of Youtube links; [36] depends on the negation of queries to break down the search results; [37] deals with the return limit of the search engines.

2.3 Sampling methods and their estimators

2.3.1 The problem

Suppose that in a graph there are N number of nodes labeled from 1 to N. Node i has a degree $d_i, i \in \{1, 2, \ldots, N\}$. Let the total number of degree is $\tau = \sum_{i=1}^N d_i$, and

the mean of degrees is $\langle d \rangle = \tau/N$. The variance σ^2 of the degrees in the population is defined as [38]

$$\sigma^2 = \langle d^2 \rangle - \langle d \rangle^2 \tag{2.1}$$

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

$$\gamma^2 = \frac{\sigma^2}{\langle d \rangle^2} = \frac{\langle d^2 \rangle}{\langle d \rangle^2} - 1. \tag{2.2}$$

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. There are different ways to take the samples, notably by RN, RE, and RW samplings. Different sampling method may require its own estimator as described in the following subsections. Table 2.1 summarizes the notations used in this paper.

2.3.2 RN sampling

In random node (RN) sampling, the sample $(d_{x1}, \ldots, d_{xn})$ is uniform random. The arithmetic mean is an unbiased estimator as defined below:

$$\widehat{\langle d \rangle}_{RN} = \frac{1}{n} \sum_{i=1}^{n} d_{xi} \tag{2.3}$$

Although it is an unbiased estimator, the problem is that its variance can be very large for scale-free networks. The degrees of most real life networks are close to Zipf's distribution, inducing a large variation of the degrees. What we need is to quantify the variance so that we know how good the estimator is. Unfortunately it is hard to predict the variance because 1) real data does not fit exactly the Zipf's law; 2) the exponent and cut-off value vary from data to data.

Nonetheless, we can assume a distribution to gain some understanding of the variance. By inspecting the degree frequency distributions of the 18 graphs, we find that most of them can be described using Zipf-Mandelbrot law $d_i = A/(\alpha + i^\beta)$ [39], where A, α and β are constants. β is the constant for the slope and we assume it is one to simplify the problem. Note that this is the rank-degree exponent, and the corresponding degree-frequency exponent is $\beta + 1 = 2$. This is a variation of Zipf's law, adding α to account for the drooping curve that exists in most real data. α is a small value relative to N that corresponds to the turning point in the rank-degree plot. When $\alpha = 0$, the distribution

is reduced to the simplified Zipf's law, and the log-log plot turns to be straight line. With such assumptions, we can have

Theorem 1. Suppose the degrees follow Zipf's law with exponent one, i.e., $d_i = \frac{A}{\alpha+i}$. The variance of the random node estimator is

$$var(\widehat{\langle d \rangle}_{RN}) \approx \frac{\langle d \rangle^2}{n} \left(N \left[(\alpha+1) \ln^2 \frac{N+\alpha}{1+\alpha} \right]^{-1} - 1 \right). \tag{2.4}$$

Proof. See appendix. □

The intuitive understanding of the theorem is that the variance grows almost linearly with the data size N, in the order of $O(N/ln^2 N)$. In other words, 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 graph using uniform random sampling.

As an illustrative example, consider the 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 graph in a much larger scale is also found in real NotreDame web graph as shown in Figure 2.7. The average degree is $(N-1+N-1)/N \approx 2$, assuming $1/N \approx 0$. Most of the 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 nodes 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.

2.3.3 RE and RW Sampling

In random edge (RE) sampling, each edge has an equal probability of being sampled, and the two incident nodes of the selected edge are taken. In random walk (RW) sampling a node is selected randomly from its current neighbourhood. In both RE and RW sampling, nodes are sampled with probability proportional to their degrees. For this kind of samples, arithmetic mean estimator tends to overestimate the average degree $\langle d \rangle$ by $(\gamma^2 + 1)$ times. Borrowing the techniques from PPS (Probability Proportional

to Size) sampling that is based on Hansen-Hurwitz estimators [40], the harmonic mean should be used for these samples:

$$\widehat{\langle d \rangle}_{RE} = \widehat{\langle d \rangle}_{RW} = n \left[\sum_{i=1}^{n} \frac{1}{d_{x_i}} \right]^{-1} \tag{2.5}$$

For the detailed derivation of this estimator, we refer to [21]. In the idealized case when the degrees follow exactly Zipf's law, we have the following theorem that can highlight the reduced variance of the estimator:

Theorem 2. When the degrees follow Zipf's law whose exponent is one, the variance of the estimator is

$$var(\widehat{\langle d \rangle}_{RE}) \approx \frac{\langle d \rangle^2}{n} \left(\frac{1}{2} \ln \frac{N+\alpha}{1+\alpha} - 1 \right). \tag{2.6}$$

Proof. See appendix. □

Comparing the estimators $\widehat{\langle d \rangle}_{RN}$ and $\widehat{\langle d \rangle}_{RE}$, we can see that the variance of $\widehat{\langle d \rangle}_{RE}$ grows logarithmically with graph size N, while $\widehat{\langle d \rangle}_{RN}$ increases in the order of $O(N/ln^2 N)$, almost linearly with N when N is large. In other words, in order to make the variance commensurate to the real value $\langle d \rangle^2$, the sample size n should be in the order of N for $\widehat{\langle d \rangle}_{RN}$, but merely $\ln N$ for $\widehat{\langle d \rangle}_{RE}$.

Example 1. The sampling and estimation methods can be illustrated using Figure 2.1. The average degree of the graph is 2.7. The sample degrees taken by RN, RE, and RW sampling methods are (1,2,3,5,3,4), (3,3,6,5,6,3), and (2,6,3,5,6,3), respectively. The estimations for RN, RE, and RW samples are:

$$\widehat{\langle d \rangle}_{RN} = \frac{1+2+3+5+3+4}{6} = 3$$
$$\widehat{\langle d \rangle}_{RE} = \frac{6}{\frac{1}{3}+\frac{1}{3}+\frac{1}{6}+\frac{1}{5}+\frac{1}{6}+\frac{1}{3}} \approx 3.9$$
$$\widehat{\langle d \rangle}_{RW} = \frac{6}{\frac{1}{2}+\frac{1}{6}+\frac{1}{3}+\frac{1}{5}+\frac{1}{6}+\frac{1}{3}} \approx 3.5$$

2.3.4 Other estimators

Average degree $\langle d \rangle$ can be used to derive other properties. For instance, the coefficient of variation of the degrees can be decided by:

$$\gamma^2 = \frac{\langle d_E \rangle}{\langle d \rangle} - 1, \tag{2.7}$$

TABLE 2.2: 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	# Nodes	γ	$\langle d \rangle$	Max degree
Twitter [41]	41,652,230	35.95	70.51	2,997,652
WikiTalk[10]	2,394,385	26.34	3.89	100,029
BerkStan[10]	685,230	14.69	19.41	84,230
EmailEu[10]	265,009	13.93	2.75	7,636
Stanford[10]	281,903	11.79	14.14	38,625
Skitter[10]	1,696,415	10.46	13.08	35,455
Youtube[42]	1,138,499	9.65	5.25	28,754
NotreDame[10]	325,729	6.40	5.25	10,721
Gowalla[10]	196,591	5.54	9.67	14,730
Epinion[10]	75,879	4.02	10.69	3,044
Google[10]	875,713	4.02	9.87	6,332
Slashdot[10]	82,168	3.35	12.27	2,552
Facebook-1[43]	2,937,612	3.14	14.27	4,356
Flickr [10]	105,936	2.65	43.43	*5,425*
Facebook-2 [44]	63,731	1.56	25.64	*1,098*
Amazon[10]	410,236	1.27	11.89	2,760
CitePatents[10]	3,774,768	1.20	8.75	*793*
RoadNet[10]	1,965,206	0.35	2.82	*12*

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 the following estimator [14–16]:

$$\widehat{N} = (\gamma^2 + 1)\frac{n^2}{2C}, \tag{2.8}$$

where n is the sample size, C is the number of collisions in the samples. When $\gamma = 0$, every node has an equal probability of being sampled and the above estimator is reduced to $\widehat{N} = n^2/(2C)$, a well-known equation in birthday paradox.

2.4 Experiments

2.4.1 Datasets

We conducted experiments on dozens of large networks we can find. Most of them are from Stanford SNAP graph collection [10]. 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 sampling. 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. 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. Facebook data is one of the few exceptions that RE sampling is not obviously better than RN sampling. Therefore we include two Facebook graphs that can be found. More complete data description and programs can be found at http://cs.uwindsor.ca/∼jlu/graph.

Altogether 18 graphs are reported and their statistics are summarized in Table 2.2. They are sorted according to γ, the coefficient of variation of the degrees. γ is proportional to the RRMSE of RN sampling, and decides whether RE is better than RN sampling. In the last a few datasets, the maximal degrees are smaller relative to their data sizes, causing small γ value. We highlight four datasets with italic font, whose RE sampling is not as good as RN sampling.

The degree distributions give an overview of the data and are shown in Figure 2.2. Among these graphs, most of them have a long-tail distribution, resulting in large coefficient of variation. Graphs in the last row have small γ values because the maximal degrees are small compared to their data sizes. In particular, RoadNet graph has maximal degree 12. They are scale-free networks since their log-log plots obviously deviate from straight-lines. For these graphs RE sampling does not have clear advantage. 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. For these graphs, simple RW sampling should be avoided.

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 allows unlimited number of followers, while Facebook has an up-limit for the maximal number of friends. Therefore Twitter is a scale-free network with large degree variation, while Facebook has a sharp dropping curve causing low γ value. Because of their structural difference, for Twitter data RE is hundreds of times better than RN sampling in terms of sample size, for Facebook data RE and RN samplings are similar.

The estimators are evaluated by RRMSE (Relative Rooted MSE), which is defined as below:

$$RRMSE(\widehat{\langle d\rangle}) = \frac{1}{\langle d\rangle}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\langle d\rangle}_i - \langle d\rangle\right)^2} \tag{2.9}$$

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. 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. Twitter data is the complete user network collected in 2009 [41]. It has

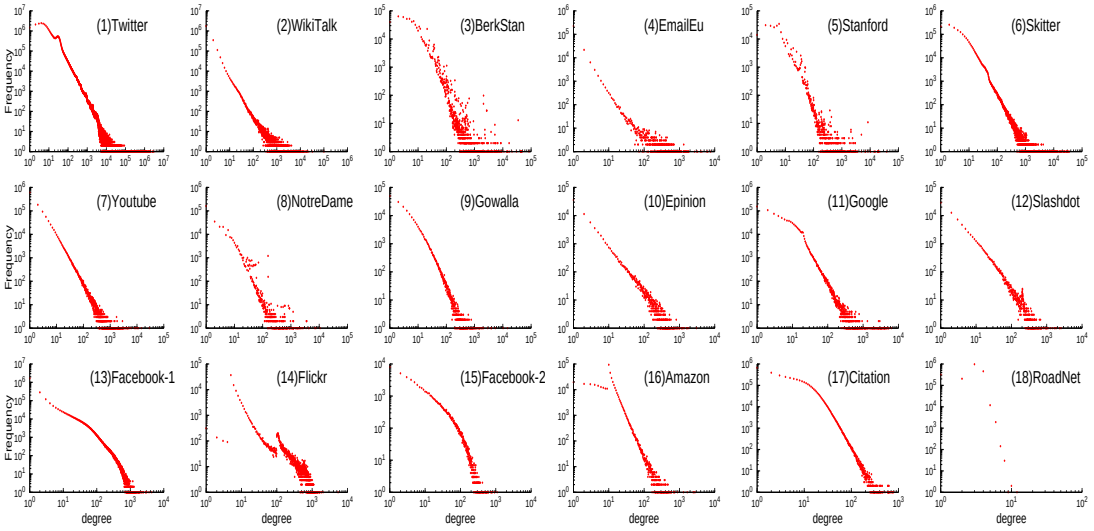


FIGURE 2.2: Degree distributions of 18 graphs. Plots are sorted in decreasing order of coefficient of variation γ. Graphs in the last row have small γ values because the maximal degree is small compared to the data size. For these graphs RE sampling does not have obvious advantage. Web graphs (sub-figures 3, 5, 8) do not form a straight line in the upper part of the log-log plots, indicating irregularity in the graph structure. For these graphs simple RW sampling should be avoided.

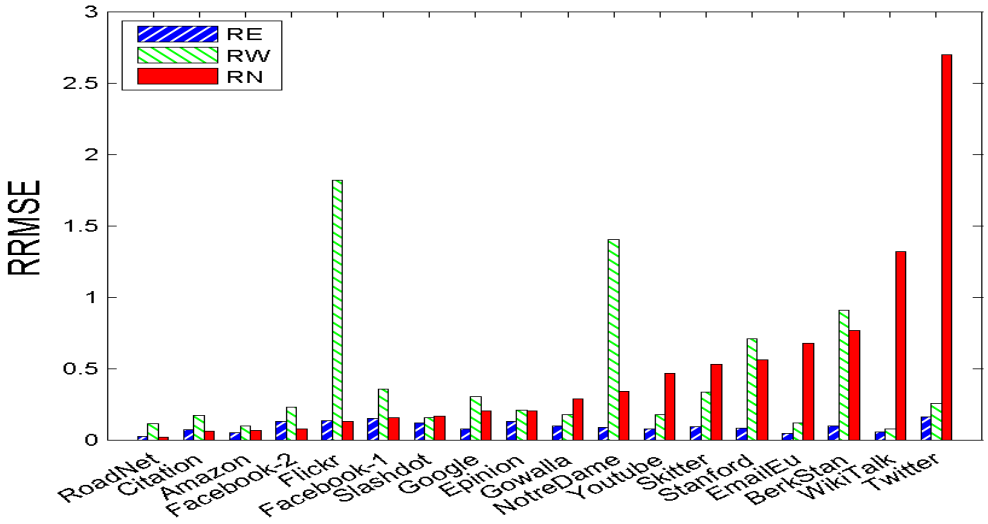
billions of edges that can not fit into computer memory. We use index engine Lucene to store the data in hard drive and use search engine to mimic the random sampling methods.

2.4.2 RE vs. RN sampling

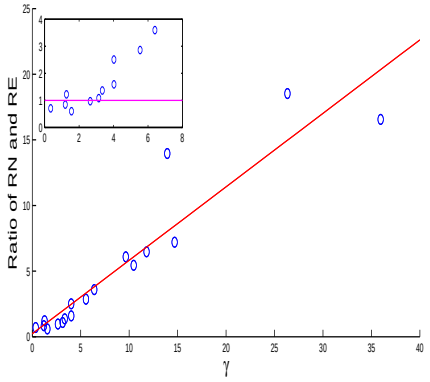
Panel A in Figure 2.3 shows that RE outperforms RN sampling on most of the data. The estimation error of RN sampling is proportional to γ as expected. Consequently, the figure shows that RRMSE of RN sampling grows from RoadNet up to Twitter, since the datasets are sorted according to γ.

In contrast to the monotonic increase of RN sampling, RRMSE of RE sampling remains mostly a constant as Theorem 2 indicates. Because of this, the RRMSE ratio between RN and RE sampling grows almost linearly with the coefficient of variation γ as shown in panel (B) of Figure 2.3. For Twitter and WikiTalk networks, RN sampling is around 15 times worse than RE sampling in terms of RRMSE. When translated into sample size, that means 225 ($= 15^2$) times more samples are needed to produce the same confidence interval as RE sampling.

This huge difference between RN and RE sampling will change the practice of sampling, especially in big data. Here we are not talking about a few percentage of improvement. It is a saving in several orders of magnitude. We project the advantage of RE sampling



(A) RRMSE of three sampling methods for 18 graphs.



(B) RRMSE ratio of RE and RN as a function of γ.

FIGURE 2.3: Summary of the three sampling methods on 18 graphs. (A) Comparison in terms of RRMSEs. (B) The advantage of RE grows as a linear function of γ. It shows that RRMSE of RN/RE grows almost linearly with γ. Pearson's coefficient of correlation is 0.9354. Inset: there are four cases where RN/RE is smaller than one. These four graphs from left to right are RoadNet, Citation, Facebook-2, and Flickr. In both (A) and (B) the sample size is 400, and RRMSEs are obtained over 5000 runs except for Twitter data that has 2000 runs.

will become even more prominent with the growth of data size. Although most data exhibit power law distributions with similar exponent, their coefficient of variations grow with the data size, therefore the savings of RE sampling.

There are only four datasets whose RE sampling is slightly worse than RN sampling. Panel B in Figure 2.3 shows that four datasets are below the horizontal line 1. A closer inspection on these datasets shows that they all have small degree variations as shown in Figure 2.2. RoadNetwork has maximal 12 degrees, and its degrees show a log-normal distribution. Facebook has a limit on the number friends, thus the maximal degree is

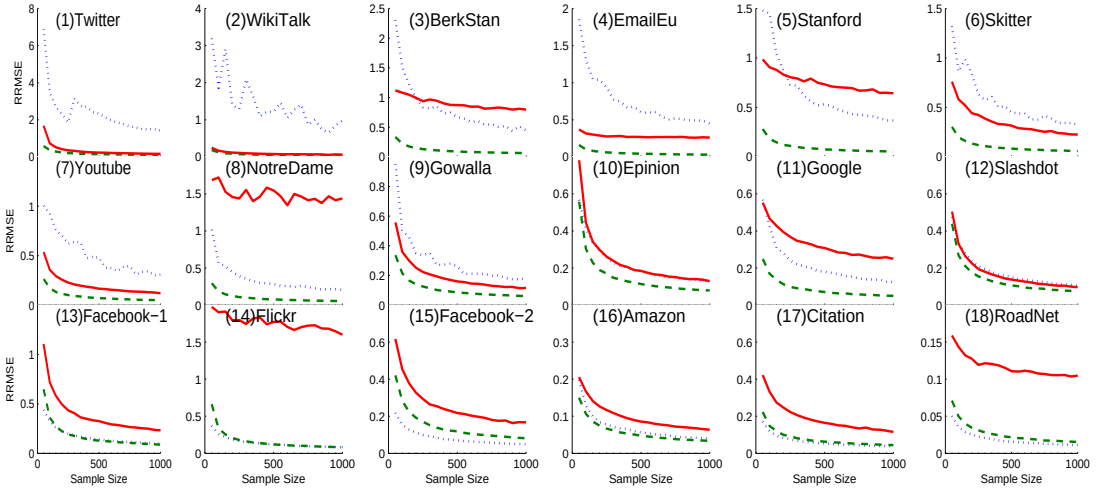


FIGURE 2.4: RRMSEs of RN, RE, and RW samplings as a function of sample size for 18 graphs. The dotted, dashed, and solid lines are for RN(...), RE(−−), and RW(–) samplings respectively. It shows that in most cases the sample size does not change the relative positions of the sampling methods. The exceptions are the web graphs 3 and 5 where RW sampling does not improve with the increase of sample size because of the random walk traps.

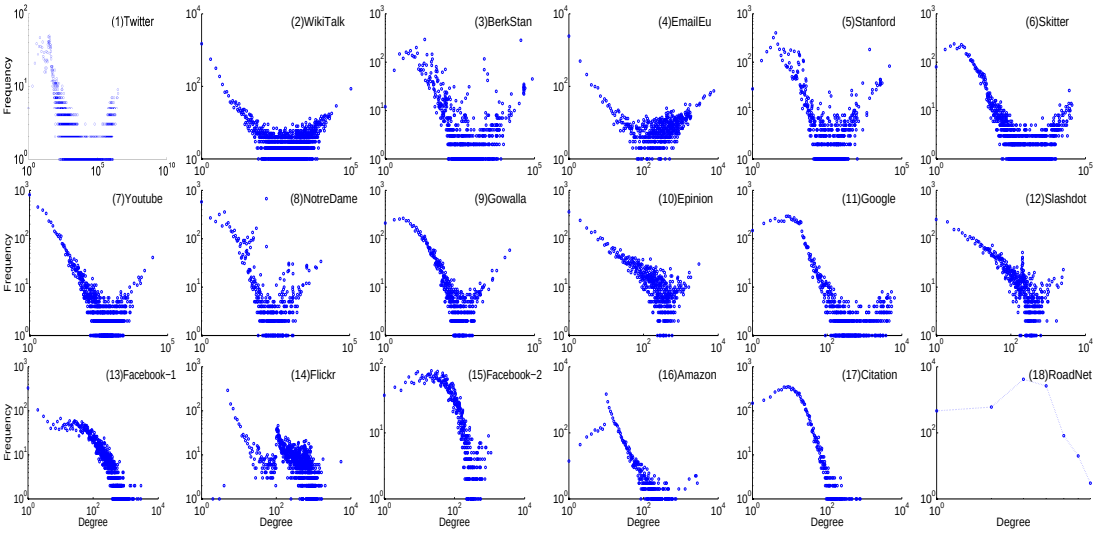


FIGURE 2.5: The degree distributions of the samples obtained from RE (Random Edge) samplings. n=8,000. The log-log plots in the first two rows exhibit a "V" shape, where the sampled small nodes resemble the distribution of the original graph, while the sampled large nodes have a tail pointing upwards. These plots in the first two rows indicate that both small and large nodes are well represented in the sample. The plots in the last row indicate that the sample distribution is similar to the original distribution, therefore the RRMSE of RE sampling is similar to that of RN sampling.

abnormally small compared with its size. Flickr has an irregular degree distribution that has a large bump around degree 100.

Another perspective to understand the reduced variance of RE sampling is its sample distributions in Figure 2.5, where the sample size 8000. It shows that most of the sample distributions have a "V" shape, indicating that the small nodes still follow power law roughly as in the original data, while the large nodes can be sampled many times. 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 the long tail, do not have very large nodes, and do not have large nodes that are 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 second 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.

2.4.3 RW vs. RN sampling

In many practical situations, RE sampling is not easy to implement, while RW sampling is supported by most real networks such as Twitter. RW sampling can be regarded as an approximation to RE sampling in that *asymptotically* the node sampling probability is proportional to its degree. The difference between RW and RE is dependent on the mixing time, the steps to reach closely enough the stationary distribution. The mixing time, in turn, is inversely proportional to the square of the conductance of the graph [45]. Let V be the set of nodes of a graph. The conductance of a subset of nodes S of

V is

$$\Phi(S) = \frac{\sum_{i \in S, j \in V \setminus S} A_{ij}}{min(A(S), A(V \setminus S))} \tag{2.10}$$

where A is the adjacency matrix of the graph, and $A(S) = \sum_{i \in S, j \in V} A_{ij}$. The conductance of the graph is

$$\Phi = min_S \Phi(S). \tag{2.11}$$

Our experiments as depicted in Figure 2.3 (A) show that RW is worse than RE consistently as expected. To have a detailed comparison between RW and RE samplings, Figure 2.6 plots the ratio of RRMSEs (RW/RE) against graph conductance for 18 datasets. The sample size for both RN and RW samplings is 400, and each RRMSE is obtained from 2000 runs. It shows that all the ratios are above the dashed green line for the value of one, indicating RE is always better than RW. When the conductance is not very small (the left section of the plot), overall RW can approximate RE sampling well, thereby outperforms RN sampling. When the conductance is small, indicating the existence of loosely connected components, RW can be dramatically worse than RE. The ratio RW/RE can be as large as that of RN/RE, thereby offsets the advantage gained by PPS (probability proportional to size) sampling, making RW and RN incomparable. Taking NotreDame and Flickr are example, their RW samplings are around 15 times worse than RE sampling in terms of RRMSE. When measured in sample size, they can be $15^2 = 225$ times worse. This reveals the reason why there are mixed results for the comparison between RW and RN samplings. Both RN and RW can be 10 times worse than RE, but for different reasons. RN is worse because of the large degree variance, while RW is because of the existence of loosely connected component indicated by small conducatnce Φ. It is remarkable that both degree variance and loosely connect component can be the dominant factor.

To find out the reason for the poor performance of RW sampling for some datasets, we plot the random walk traces in Figure 2.7 for six networks. Three of them (Flickr, NotreDame, and Stanford) have low graph conductance, while three others (Facebook, Amazon and Youtube) have high conductance as comparison. These six networks are also highlighted in Figure 2.6 using different markers. Figure 2.8 shows the conductances over the size of the subcomponents. For the networks in the first row, their lowest conductance are smaller than 10^{-3}. For Flickr data, the conductance value dips down only when the component size is large ($\approx 10^4$). This is reflected in its random walk trace where there are two components clearly separate by a long single edge link. Both NotreDame and Stanford have many loosely connected components, as shown by

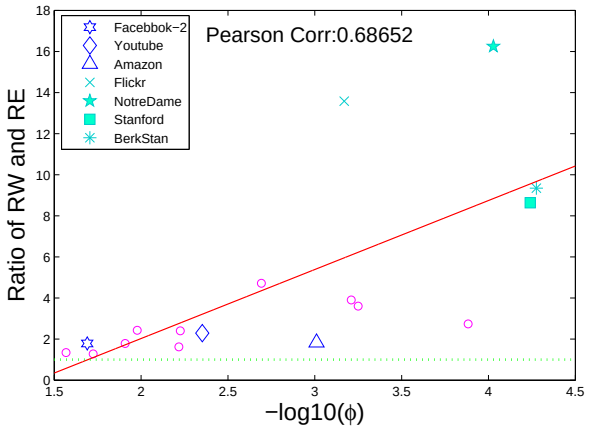


FIGURE 2.6: Standard error ratio between RW and RE vs. graph conductance Φ for 18 datasets. Sample size is 400.

many low conductance values over a variety of component sizes. Those three networks are in contrast to the well enmeshed networks Facebook-2 and Youtube, whose the conductances are high. Amazon also has a low conductance ($\approx 10^{-3}$), but reaches the lowest point only when the component size is around 100. This small component has less impact on the overall network structure as shown in the RW trace, and little impact on the performance of RW as shown in Figure 2.6.

2.5 Conclusions

This paper shows the importance of selecting the appropriate sampling method –the difference between the sampling methods can be infinitely large in theory and orders of magnitude in observed data. Such a large difference will have great impact on the sampling practice, especially for web-based networks such as online social networks 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 [7]. 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 simple random walk that are supported by many online data sources.

RN sampling is not always inferior to RW or RE sampling. When the data has a normal distribution RN sampling should be the method of choice. When the network is large and scale-free with Zipf law exponent one, we show that RE sampling is much better.

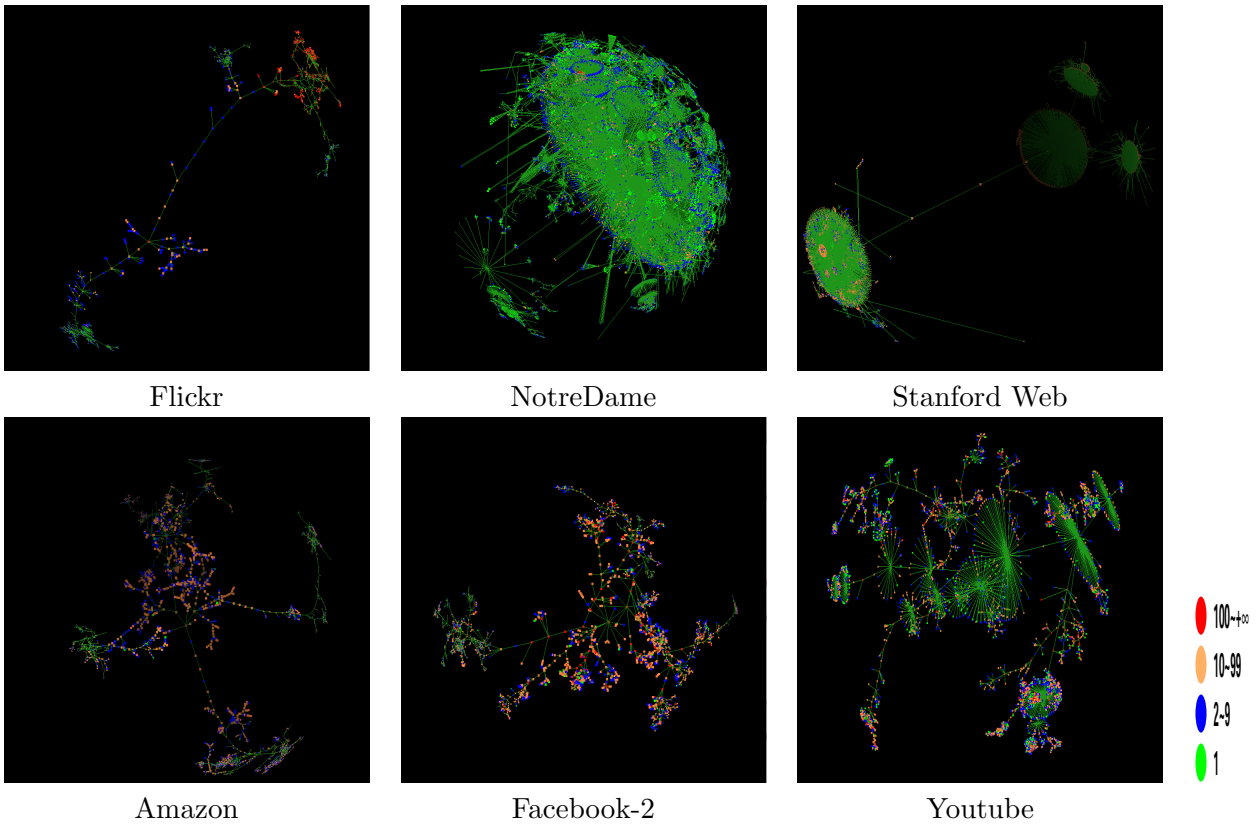


FIGURE 2.7: (Best viewed in colour) Random walks on six networks. Flickr, NotreDame and Stanford have loosely connected components while Amazon, Facebook and Youtube are well enmeshed. Each random walk contains 6×10^4 nodes except NotreDame which has 15×10^4 nodes. Node colour indicates the degree of the node. Green=1; Blue=2∼9; Yellow=10∼99; Red=100+.

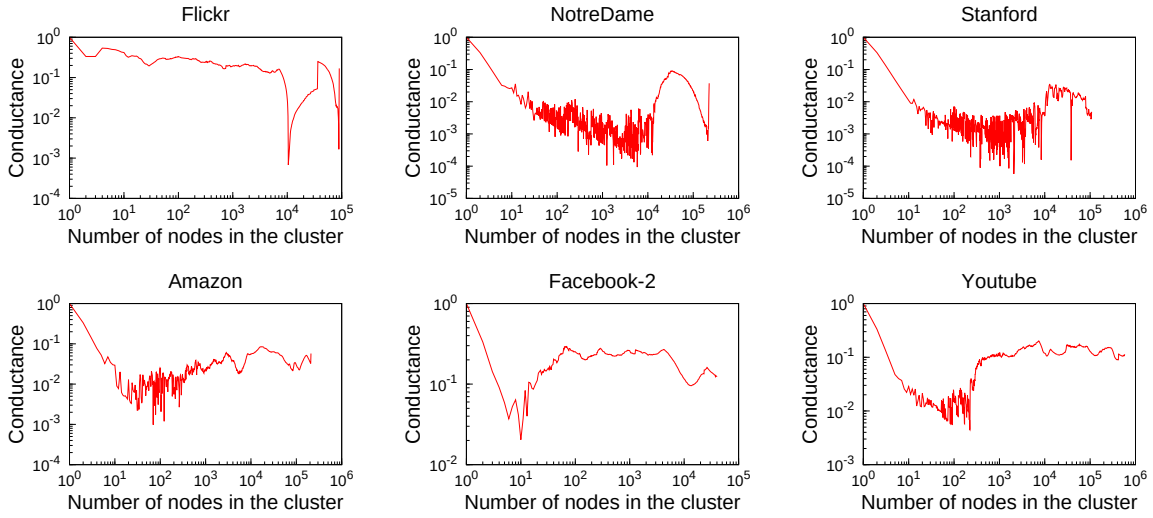


FIGURE 2.8: Conductance $\Phi(S)$ over $|S|$, the size of the the components, for six networks. Plots are drawn using SNAP API described in [1].

For data not following some distribution exactly, we suggest to use γ, the coefficient of variation of the degrees, to select the sampling method. 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 estimation error of RE sampling varies slightly across all the data we examined.

Although we have a clear answer for the selection of sampling methods between RN and RE samplings, RE sampling may not be supported in some real applications. RW sampling is a more practical method that can approximate RE sampling in that both methods sample nodes with probability proportional to its size. The difference is that in RW sampling this is true 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.

This paper focuses on average degree estimation so that the detailed analysis and comparison could be done. It is the first building block on top of which other properties could be derived. For instance, population size estimation is built on top of average degree estimation as shown in Section 2.3.4. In addition to the degrees sampled, size estimation depends on the collisions observed. This additional dimension of variation makes the evaluation of the sampling methods not so straightforward.

Overall, we study the most basic sampling methods for the simplest property of a graph so that we can draw conclusive results. Based on this result, we can develop more advanced sampling methods for more complex properties. For sampling methods, with the understanding that RE sampling is the best for scale-free networks, we can design a sampling method that can approximate RE sampling as much as possible, such as random walk with restart. For graph properties to be estimated, we can also discover friends of friends and Gini coefficient in addition to population size [16], and we expect that RE sampling would be better for some other structural graph properties such as clustering coefficient.

2.6 Acknowledgements

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

2.7 Appendix

Both Theorem 1 and Theorem 2 assume that the degrees follow the Zipf's-Mandelbrot law [39] which states that if the degrees d_i are sorted in descending order, then

$$d_i = \frac{A}{\alpha + i}, \tag{2.12}$$

where α and A are constants. $\alpha \ll N$. All the degrees sum up to τ, i.e.,

$$\sum_{1}^{N} d_i \approx \int_{1}^{N} \frac{A}{\alpha + x} dx \approx A \ln(\frac{\alpha + N}{\alpha + 1}) = A \ln B = \tau, \tag{2.13}$$

where we use $B = (\alpha+N)/(\alpha+1)$ to make our derivations more concise. The normalizing constant $A = \tau / \ln B$. Besides, $\sum_1^N d_i^2$ can be approximated by the following since N is a very large number:

$$\sum_{i=1}^{N} d_i^2 \approx \int_{1}^{N} \frac{A^2}{(\alpha + x)^2} dx \approx \frac{A^2}{\alpha + 1}. \tag{2.14}$$

2.7.1 Proof of Theorem 1

Proof. Based on Equations 2.13 and 2.14, the variance of all the degrees is

$$\begin{aligned} \sigma^2 = \langle d^2 \rangle - \langle d \rangle^2 &= \langle d \rangle^2 \left[N \frac{\sum_1^N d_i^2}{(\sum_1^N d_i)^2} - 1 \right] \\ &\approx \langle d \rangle^2 \left[\frac{N}{(\alpha + 1) \ln^2 B} - 1 \right]. \end{aligned} \tag{2.15}$$

Using central limit theorem,

$$var(\widehat{\langle d \rangle}_{RN}) = \frac{\sigma^2}{n} = \frac{\langle d \rangle^2}{n} \left[\frac{N}{(\alpha + 1) \ln^2 B} - 1 \right]. \tag{2.16}$$

□

2.7.2 Proof of Theorem 2

Proof. When nodes are sampled with simple random walk, the asymptotic probability of the node i being visited is $p_i = d_i / \tau$. When n nodes $(x_1, x_2, \ldots, x_n)$ are sampled, where each $x_i \in \{1, \ldots, N\}$, the Hansen-Hurwitz size estimator of the population size

N is [38]:

$$\widehat{N}_H = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_{x_i}} = \frac{\tau}{n}\sum_{1}^{n}\frac{1}{d_{x_i}}, \tag{2.17}$$

and the variance of $\widehat{N_H}$ is [38]:

$$var(\widehat{N_H}) = \frac{1}{n}\sum_{i=1}^{N} p_i\left(\frac{1}{p_i} - N\right)^2. \tag{2.18}$$

Replacing p_i with d_i/τ and expanding d_i with $A/(\alpha + i)$, we have

$$var(\widehat{N_H}) = \frac{1}{n}\left(\frac{\tau}{A}\sum_{1}^{N} i - N^2\right) \approx \frac{N^2}{n}\left(\frac{\ln B}{2} - 1\right). \tag{2.19}$$

The Taylor expansion of $\widehat{\langle d\rangle}_{RE}$ around N is

$$\widehat{\langle d\rangle}_{RE} = \frac{\tau}{\widehat{N_H}} = \tau\left(\frac{1}{N} - \frac{\widehat{N_H} - N}{N^2} + \dots\right). \tag{2.20}$$

By the Delta method, the variance of $\widehat{\langle d\rangle}_{RE}$ is

$$var(\widehat{\langle d\rangle}_{RE}) = \tau^2\frac{var(\widehat{N_H})}{N^4} = \frac{\langle d\rangle^2}{n}\left(\frac{\ln B}{2} - 1\right). \tag{2.21}$$

□

Chapter 3

Uniform Random Sampling not Recommended for Size Estimation

This paper was submitted as:

Hao Wang, Jianguo Lu, *Uniform Random Sampling not Recommended for Size Estimation.* Physica A: Statistical Mechanics and its Applications, 2013, submitted.

3.1 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 [46], to the more recent problem of gauging the size of the Web and search engines [2, 7, 47, 48] and online social networks [14]. In the era of big data, the first thing we want to know is how big the data is. The size is of interests to general public and decisions makers, and determines the way for IT practitioners to perform data analysis and data mining.

The direct calculation of data size is often not for possible or desirable for several reasons. The data can be distributed and there is no central data deposit, such as in the case of peer-to-peer networks or the Web [2]. Even when the data are available in one place, there are requirements for fast just-in-time analysis of the data. Quite often, data are hidden behind some searchable interfaces and programmable web APIs, such as online social networks where the access is limited and the data in its entirety is not available [14, 18]. Regardless of a large variety of application scenarios, a common approach to solving these problems is to use a sample 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 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 the sending of queries that occupies network traffic. In addition, most data sources impose daily quotas. In such case it is paramount to choose an efficient sampling and estimation method.

The norm of sampling practice in general, and size estimation in particular, is to use uniform random samples whenever possible. This paper shows that, on the contrary, uniform random sampling should be avoided when PPS (probability proportional to size) sampling is available. For ease of discussion, sampling is modelled in the context of graph, where uniform sampling corresponds to uniform random node (RN) sampling, PPS sampling corresponds to random edge (RE) sampling. In this setting, we prove that RN sampling is always inferior to RE sampling, and show that the performance ratio between RN and RE samplings can be quantified by the relative degree variance of the nodes. Since large and scale-free networks can have very large degree variance, RE sampling can excel RN sampling in orders of magnitude. Furthermore, we show that random walk (RW) sampling can approximate RE sampling when the graph conductance is not very small (equivalently when there are no loosely connected components).

Very recently, we gave the size estimator for RE sampling in [15]:

$$\widehat{N}_E \approx \hat{\Gamma}\frac{n^2}{2C}. \tag{3.1}$$

Here n is the total number of sampled nodes, C is the number of collisions, and $\Gamma = \gamma^2+1$ where γ is the coefficient of variation of node degrees. [15] analyzed and corrected its bias. This paper derives its variance, and consequently proved that it is better than RN sampling.

More specifically, *our main contributions* are: 1) we derive the variance of the size estimators for RN sampling and RE sampling, and analytically show that RE sampling is better than RN sampling up to a factor $\sqrt{\Gamma}$ in terms of sample size. This result is also supported empirically by 18 large real-world networks; 2) Empirically we show that RW can approximate RE sampling in many datasets, and fail for networks with loosely connected component. We show that the ratio between RW and RE is dependent on the conductance of the graph.

3.2 Background and Related Work

3.2.1 Random Node Sampling

The traditional and widely applied size estimator is the Petersen estimator [49] that assumes the nodes are sampled uniformly at random. It is also used for the estimation of the size of WWW [2, 50] and other online data sources [5]. Suppose that we sample n_1 number of distinct nodes first, then sample another n_2 number of distinct nodes. In both sampling occasions, assume that each node has the equal probability of being sampled. Among two captures, there are D number of duplicates. The Petersen estimator is

$$\widehat{N}_P = \frac{n_1 n_2}{D} \tag{3.2}$$

This estimator is also the starting point for the capture-recapture method that is well studied in ecology [49]. When there are multiple capture occasions, several estimators are developed. Among them is the approximate Maximum Likelihood Estimator (MLE) $\widehat{N}_D$ that is given by Darroch [51]:

$$n - D = N\left(1 - e^{-\frac{n}{N}}\right), \tag{3.3}$$

Where n is the total number of sampled nodes including duplicates, D is the duplicates. In the context and random graph theory, this equation has also been used to predict the

isolated nodes in random graph when nodes are connected randomly [52]. Unfortunately it does not have a simple closed form solution [52], i.e., it can not be solved algebraically for N. In online social network studies, [53] used numeric method to find the solution to this estimator. In deep web size estimation, [23] gives an approximate solution for N that reveals a power law governing the data not being sampled and the overlapping rate.

Alternatively, there is an estimator based on the number of collisions [54] instead of number of duplicates in Equation 3.3. Let f_i denote the number of nodes that are sampled exactly i times. Collisions $C = \sum_{i=1}^{+\infty} \binom{i}{2} f_i$, while duplicates $D = \sum_{i=1}^{+\infty} (i-1) f_i$. The random node estimator $\widehat{N}_N$ is

$$\widehat{N}_N = \binom{n}{2} \frac{1}{C} \approx \frac{n^2}{2C}. \tag{3.4}$$

All these estimators assume that nodes are sampled with equal probability, i.e., they are in the category of RN sampling. They have similar performance, and analytically Equation 3.3 can approximate the other two estimators by applying Taylor expansion on the right side of the equation. In literature, most approaches use one of these estimators, while the major research challenge is to obtain the uniform random samples using algorithms such as Metropolis-Hasting random walk [7].

3.2.2 Sampling Nodes With Unequal Probability

When the node sampling probability is not uniform, the estimation becomes notoriously difficult. If we continue to use the estimators in the previous section, there will be a negative bias because the same population will induce more collisions (or duplicates) when sampling probabilities are not equal. Some researches adjust the bias by devising new estimators [23, 47, 48]. Broder et al.[47] assigned less weight to large documents being sampled; Shokouhi et al. [48] run regression on past data to establish the relationship between the homogeneous and heterogeneous data; Lu et al. [23] went a step further by using γ, the degree of heterogeneity, to adjust the discrepancy. All those estimators are mostly empirical and data dependent.

On the other hand, unequal sampling probability gives us a great advantage for size estimation when we known that the sampling probability is proportional to its degree (or PPS sampling). We gave the following estimator in [15], which can be also derived

from [14, 54]:

$$\widehat{N}_E = \Gamma \binom{n}{2} \frac{1}{C} \approx \widehat{\Gamma} \frac{n^2}{2C}, \tag{3.5}$$

where $\Gamma = \gamma^2 + 1$. Comparing Equations 3.9 and 3.5, we can see that the only difference is Γ. It means that we can estimate the size as if the nodes were sampled uniformly at random, then increase the estimation by a factor of Γ. When Γ is small, as in the case of most studies in Ecology where Γ is typically in the range of 1..2, the difference between $\widehat{N}_N$ and $\widehat{N}_E$ is small. In large scale-free networks, Γ can be as large as 1000, as in the case of Twitter user network. This results in striking difference between $\widehat{N}_N$ and $\widehat{N}_E$. Given the same sample size n and graph size N, large Γ also induces more collisions, consequently higher accuracy as we will explain. In other words, the advantage of PPS sampling becomes more prominent when Γ is large.

We want to emphasize that Equation 3.5 can be applied only in PPS sampling, or random edge sampling. In literature, for instance in [14], an estimator equivalent to the above was developed for samples obtained by random walk (RW), 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 from RE sampling when the graph conductance is low. [14] 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.

The estimator $\widehat{N}_E$ is not well studied in literature, partially because $\widehat{\Gamma}$ is not easy to estimate in traditional applications such as wildlife population estimation [49]. In applications such as online social networks, PPS sampling or its approximations are possible, making the estimation of Γ feasible. Among the n number of samples obtained by RE sampling, suppose that their degrees are $d_{x1}, d_{x2}, \ldots, d_{xn}$. Let $\langle d \rangle$ and $\langle d_E \rangle$ denote the average degree of the graph and average degree obtained by RE sampling, respectively. We showed that Γ can be estimated by [15]

$$\widehat{\Gamma} = \frac{\widehat{\langle d_E \rangle}}{\widehat{\langle d \rangle}}, \tag{3.6}$$

where

$$\widehat{\langle d_E \rangle} = \frac{1}{n} \sum_{i=1}^{n} d_{xi}, \tag{3.7}$$

$$\widehat{\langle d \rangle} = \frac{n}{\sum_{i=1}^{n} 1/d_{xi}}. \tag{3.8}$$

3.2.3 Evaluation of Estimation Methods

Estimators are normally evaluated in terms of bias ($bias(\widehat{N})$), variance ($var(\widehat{N})$), and/or the combination of them, i.e., mean squared error ($MSE(\widehat{N}) = bias(\widehat{N})^2 + var(\widehat{N})$). In most of size estimation research, the evaluations are empirical, for instances in [14, 16, 26, 55]. Empirical evaluation is data dependent, and the result may not be extended to other datasets. This problem is more acute for network size estimation where network topology differs greatly. For instance, both [14] and [16] observed that RW outperforms RN sampling in terms of MSE when the same number of samples are taken, one on DBLP and Facebook networks, the other on Twitter data. On other networks such as the Web graphs, we find that RW is actually much worse than RN sampling.

To draw more conclusive results, we derive the variances for RN and RE estimators, thereby their performance can be compared analytically. We show that RE sampling is guaranteed to outperform RN sampling. What is more, the improvement ratio can be quantified by the relative degree variance. On the other hand, the comparison between RN and RW samplings depends on both degree variation and graph conductance.

3.2.4 Graph Sampling

It has been an open problem to decide which is a better sampling method for graph [10, 27]. There are three basic sampling methods for graph, namely, random node (RN), random edge (RE), and random walk (RW) [10, 30], and various combinations and improvements [20]. Most of the research focuses on the comparison of RW sampling with RN sampling [14, 16, 26], Metropolis-Hasting Random Walk [17, 26], and variations of RW sampling [18]. [31] is one of the few studies on RE sampling. For data size estimation, recently it was observed empirically that RW sampling could be better than RN sampling for some datasets such as Twitter [16], DBLP, Facebook, and IMDB [14]. No definitive analytical answer was obtained because the performance of RW sampling method depends on many factors including the degree distribution, graph topology, and the property to be discovered.

Instead of comparing RW and RN directly, this paper uses RE sampling as a bridge to connect them. Unlike RW sampling where the node sampling probability is dependent

on many factors, in particular the mixing time of random walk, RE sampling selects a node with probability proportional to its degree. Thanks to this property, we prove analytically that RE is better than RN sampling, and quantify the improvement rate by the coefficient of variation (γ) of node degrees. Furthermore, we show that RW can approximate RE sampling when graph conductance (Φ) is not very small, thereby it is better than RN sampling. When conductance is small, indicating that RW mixing time is long, RW sampling can fail grossly for some datasets. Our result is supported by 18 representative real-world networks collected from various sources.

3.2.5 Other Size Estimation Approaches

In contrast to the traditional sampling in ecology and social studies, the diversity of the access interfaces to web data collections opens up opportunities for designing sampling schemes that take advantages of interface specifics. For instance, [17] samples valid Facebook IDs from an ID space of 9 digits, utilizing the Facebook implementation details that make the number of invalid IDs not much bigger than the valid ones; [35] leverages the prefix encoding of Youtube links; [36] depends on the negation of queries to break down the search results; [37] deals with the return limit of the search engines. Dasgupta et al. use random walk in query space to probe database properties [56] [36]. This paper differs those more practical applications in that they focus on the methods to find the (mostly uniform) samples, while we discuss the performance of different estimators regardless of the sampling details.

3.3 Random Node (RN) Sampling

When n number of nodes are selected with equal probability with replacement from a graph, the total number of nodes N can be estimated by random node estimator $\widehat{N}_N$, which is derived from the classic birthday problem and used in several papers [14, 15]:

$$\widehat{N}_N = \binom{n}{2}\frac{1}{C} \approx \frac{n^2}{2C}, \tag{3.9}$$

where C is the total number of collisions. The approximation holds because we only consider big data in this paper where the size N and sample size n are both large. Therefore $n^2 \approx n(n-1)$. What we are interested in is the variance of the estimator. Based on the assumption that N is large, we derive

Graph	N($\times 10^3$)	γ or $\sqrt{\Gamma - 1}$	$\Phi(\times 10^{-5})$
WikiTalk [10]	2,388	26.32	2,700
BerkStan [10]	654	14.51	5.3
EmailEu [10]	224	13.66	13
Stanford [10]	255	11.51	5.8
Skitter [10]	1,694	10.46	56
Youtube [42]	1,134	9.64	440
NotreDame [10]	325	6.40	9.4
Gowalla [10]	196	5.54	1,200
Epinion [10]	75	4.02	610
Google [10]	855	4.00	62
Slashdot [10]	82	3.35	1,900
Facebook [43]	2,937	3.14	590
Flickr [10]	105	2.64	68
IMDB [57]	374	2.05	130
DBLP [58]	511	1.61	560
Amazon [10]	410	1.27	98
Gnutella [10]	62	1.21	9,100
CitePatents [10]	3,764	1.20	1,100

TABLE 3.1: Statistics of the 18 real-world graphs, sorted in descending order of the coefficient of degree variation γ. Φ is the conductance.

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

$$\widehat{var}(\widehat{N}_N) \approx \frac{N^2}{E(C)} \approx \frac{2N^3}{n^2} \tag{3.10}$$

Proof. See Appendix. □

Intuitively, the expected number of collisions decides the accuracy of the estimation. For instance, when the expected collisions $E(C) = 100$, the estimated variance is approximately $N^2/100$ and the standard error (SE) is $0.1N$. Therefore the 95% confidence interval is $N \pm 1.96 \times SE = N \pm 0.196N$. In other words, the range is within [0.8N, 1.2N] when there are 100 collisions. When the number of collisions is 400, the estimated relative standard error (RSE) is $1/20 = 0.05$, so the confidence interval is about $N \pm 0.1N$.

We conducted experiments on 18 datasets listed in Table 1 to validate Lemma 1. Since all the datasets have very close results, in Figure 3.1 we only show the result for one of the graphs, a Facebook graph from [43]. The red (upper) lines depict the estimated variance given by Equation 3.10 and the observed variance with 3000 runs for RN sampling. It demonstrates that Equation 3.10 can predict the variance very well. The blue lines are for RE sampling that will be discussed in the next section. The sample sizes ranges

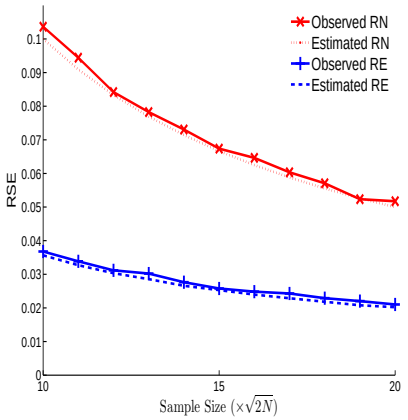


FIGURE 3.1: Relative standard errors of RE and RN samplings on Facebook data when sample size ranges between $10 \times \sqrt{2N}$ and $20 \times \sqrt{2N}$, where $\sqrt{2N} = 2423$. The red lines are for RN sampling and blue lines are for RE sampling.

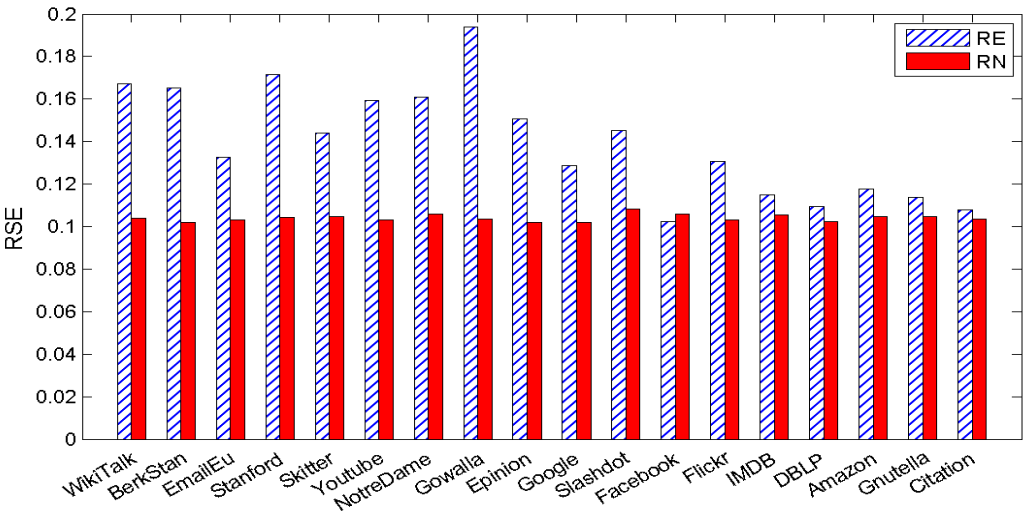


FIGURE 3.2: Comparison of RE and RN in terms of standard error when the sample size of RE is $\sqrt{\Gamma}$ times smaller than that of RN, for 18 datasets. The expected C for both RE and RN sampling is 100.

between $10 \times \sqrt{2N}$ and $20 \times \sqrt{2N}$, so that the expected collisions are between 100 and 400. In all the other datasets, Lemma 1 can also predict the variance accurately as corroborated in Figure 3.2 and 3.4.

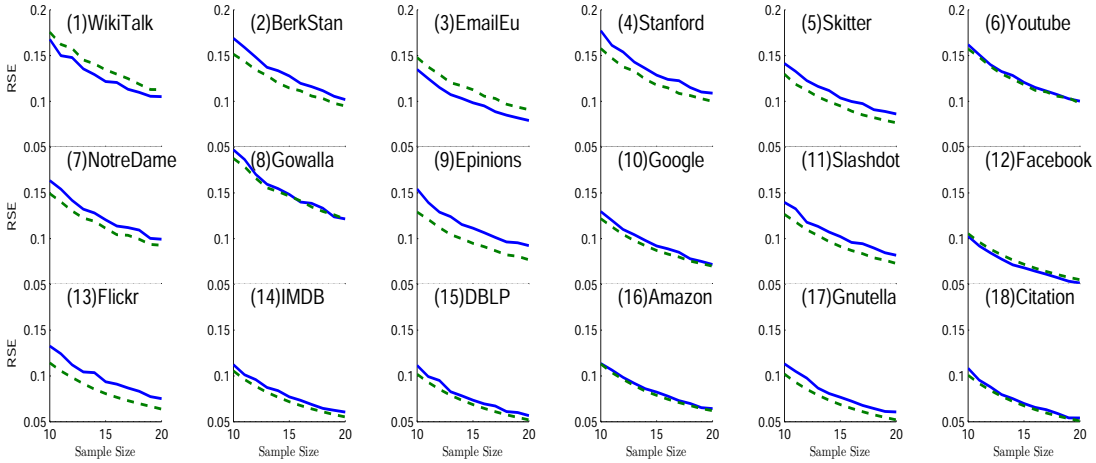


FIGURE 3.3: Estimated and observed RSE of RE sampling with the growth of sample size over 18 datasets. The sample size ranges between $10 \times \sqrt{2N/\Gamma}$ and $20 \times \sqrt{2N/\Gamma}$, i.e., the expected collisions are between 100 and 400.

3.4 Random Edge (RE) Sampling

In RE sampling, each edge is sampled with equal probability. When an edge is selected, two nodes incident to the edge are chosen as samples. Although traditionally RE sampling is hard to implement, RE sampling becomes possible in modern web applications, such as sampling two bloggers by picking a random message connecting them. In this sampling scheme, each node is selected with probability proportional to its degree. Large nodes have higher probability being selected, resulting in higher number of collisions for the same sample size in RN sampling. In this sampling method, the number of nodes N can be estimated by the random edge estimator $\widehat{N}_E$ [14, 15]:

$$\widehat{N}_E \approx \hat{\Gamma}\frac{n^2}{2C}, \tag{3.11}$$

where $\Gamma = \gamma^2 + 1$, and γ is the coefficient of variation of the degrees in the graph.

This estimator has been evaluated only empirically in the context on RW sampling [14]. In order to understand its performance, we need to derive the variance as below:

Lemma 2 (Variance of $\widehat{N}_E$). The estimated variance of RE estimator $\widehat{N}_E$ is

$$\begin{aligned}\widehat{var}(\widehat{N}_E) &= \frac{N^2}{E(C)}\left(1 + \frac{n\Gamma CV^2(\Gamma)}{2N}\right) \\ &= \frac{2N^3}{n^2\Gamma}\left(1 + \frac{n\Gamma CV^2(\Gamma)}{2N}\right),\end{aligned} \tag{3.12}$$

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

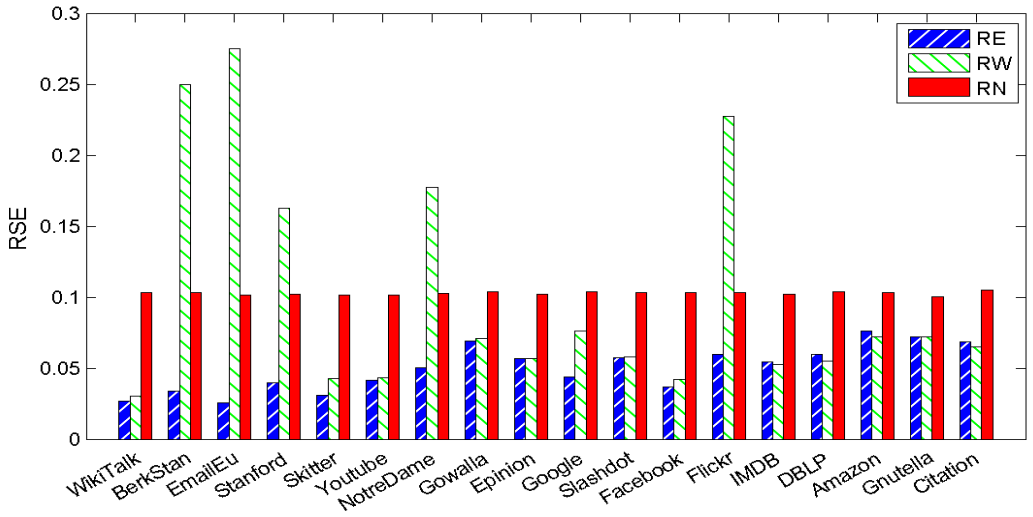


FIGURE 3.4: Comparison of three sampling methods. The sample size $n = \sqrt{2NC}$ where $\sqrt{C} = 10$. It shows that for RN sampling (red solid bars), the relative standard error is equal to $1/\sqrt{C} = 0.1$ across all the datasets. RE sampling is consistently smaller than RN sampling, and decreases with the growth of γ. RW sampling can approximate RE sampling for some datasets. For NotreDame etc. that have low conductance, RW is grossly wrong.

Proof. See Appendix. □

When N is very large, the ratio n/N can be very small to produce enough collisions. So the second term in Equation 3.12 can be omitted when Γ is small, rendering Equation 3.12 as

$$\widehat{var}(\widehat{N}_E) = \frac{N^2}{E(C)} = \frac{2N^3}{n^2\Gamma}. \tag{3.13}$$

In this case, the variance is $N^2/E(C)$, or the relative standard error (RSE) is $1/\sqrt{E(C)}$. When $E(C) = 100$, RSE should be around 0.1. Figure 3.3 shows the estimated and observed RSE over 18 datasets. The sample size ranges between $10 \times \sqrt{2N/\Gamma}$ and $20 \times \sqrt{2N/\Gamma}$, i.e., the expected collisions are between 100 and 400. As expected by Lemma 2, both the estimated and observed RSEs are around 0.1 when $\sqrt{E(C)} = 10$ and Γ is small for the datasets in the last row of the figure. For datasets with larger Γ, the second term in Equation 3.12 takes charge and becomes more dominant. That explains why the estimated (and observed) RSEs of the datasets in the first row of the figure are greater than 0.1. Similarly, when sample size $n = 20 \times \sqrt{2N/\Gamma}$, i.e., when $E(C) = 400$, the RSE is around 0.05 as expected for datasets with small Γ. Again for larger Γs, the RSE is higher than $1/\sqrt{E(C)}$.

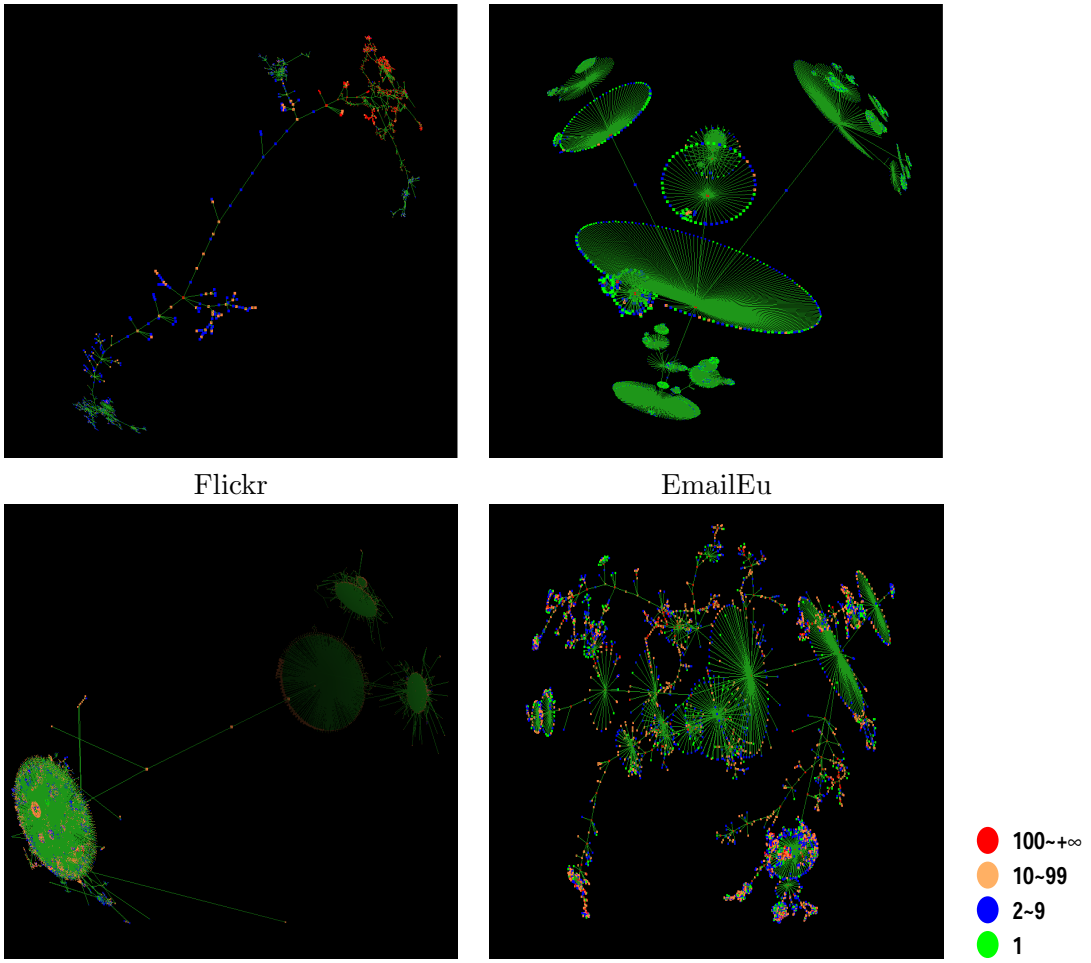


FIGURE 3.5: (Better viewed in colour) Subgraphs obtained by RW sampling from Flickr, EmailEu, Stanford and Youtube. Each subgraph contains 60,000 nodes. Node colour represents its degree in the original graph. Green=1; Blue=2 $\sim$ 9; Orange= 10$\sim$99; Red=100$\sim$ ∞.

Comparing Equations 3.10 and 3.13, the only difference is Γ. Given the same sample size n, RE sampling creates Γ times more collisions. Therefore the variance of $\widehat{N}_E$ is smaller by a factor of Γ. i.e.,

Theorem 3 (RN vs. RE). To achieve the same variance of $\widehat{N}_E$, $\widehat{N}_N$ needs to use at most $\sqrt{\Gamma}$ times more samples.

These results can be explained from two perspectives using Figures 3.1 and 3.2. In these figures and other plots throughout this paper, all the observed variances are calculated from 3000 repetitions. The variance is depicted using relative standard error (RSE) so that it is normalized by data size and different datasets are comparable in y-axis. First, we focus on *one dataset* using Figure 3.1 that compares the variances of RE and RN sampling side by side for the Facebook data. It shows that 1) the estimated variance

agrees with the observed variance very well; 2) when $n = 10\sqrt{2N}$, i.e., the expected number of collisions is 100 for RN sampling, the RSE is about 3 times higher than that of RE sampling, which is a good approximation to our Theorem 1; 3) It also shows that the RSE of RN sampling drops with faster speed, causing the diminishing advantage of RE sample as the sample size grows.

Secondly, we compare them on all the 18 datasets for a fixed value of expected collisions in Figure 3.2. In this figure $E(C) = 100$ for both RE and RN sampling, i.e., the sample size of RE sampling is $\sqrt{\Gamma}$ times smaller than that of RN sampling. As indicated by Theorem 1, RE sampling and RN sampling has similar RSEs. But RE can have higher variance when Γ is large since the second term in Equation 3.12 can no longer be omitted.

3.5 Random Walk (RW) Sampling

Although RE sampling is possible nowadays, RW sampling is more prevalent and supported by most real networks such as Twitter and Facebook [17]. RW sampling 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 [14, 16, 18, 55]. It was reported that RW is better than RE sampling for Twitter [16], DBLP, IMDB, and Facebook [14]. 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 3.4. 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 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 the 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

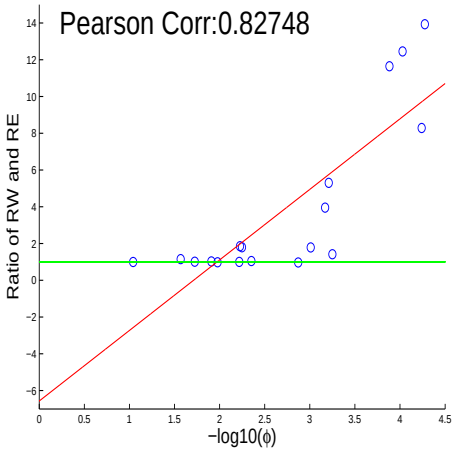


FIGURE 3.6: The ratio of RSEs between RW and RE samplings over the conductance Φ. For the four graphs with the lowest conductance, RW is around 10 times worse than RE sampling. Sample size $n = \sqrt{2NE(C)}$ where $E(C) = 100$. RSE is obtained over 3000 runs.

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 random walk a node selection is actually dependent on the previous node. 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 disregarding all but every s-th sample. s is called the thinning rate. [59] 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 [45]. So we calculate the conductances of all the 18 graphs using SNAP graph API [1] , and plot their correlation with the RSE ratios between RW and RE sampling in Figure 3.6. 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 [20] 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 which is a bolas graph [19]–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 [19] 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 3.5. It shows the subgraphs obtained from random walk from three datasets (Flickr, EmailEu, and Stanford) whose conductances are low and one normal graph (Youtube) as a comparison. The node colour indicates the degree of the node. 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 such as Stanford has many bolases 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.

3.6 Discussions and Conclusions

This paper gives the variances of random node and random edge sampling for graph size estimation. The result is surprisingly simple: 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. This rule applies for both RN sampling and RE sampling. However, 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 of collisions, or the same standard error. What is more interesting is that we can quantify how much less samples are needed 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 $\sqrt{\gamma^2 + 1}$.

Traditionally RE sampling is rarely possible. Therefore, it was hardly studied and compared with. But it is a bridge to connect RN and RW samplings. 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 [14], and why various methods need to be proposed to improve the simple random walk for datasets such as Flickr [18].

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.

3.7 Appendix

3.7.1 Proof of Lemma 1

Proof. Expanding the definition for $\widehat{N}_N$ we have

$$var(\widehat{N}_N) = var\left(\frac{n^2}{2C}\right) = \frac{n^4}{4} var\left(\frac{1}{C}\right) \tag{3.14}$$

The variance of $1/C$ can be derived using Taylor expansion of $1/C$ around $\mu = E(C)$ as below:

$$\frac{1}{C} = \frac{1}{\mu} - \frac{C-\mu}{\mu^2} + \frac{2}{\mu^3}\frac{(C-\mu)^2}{2!} \cdots$$

When selecting two nodes, the probability that the same node i is visited twice is $1/N^2$. Among all the nodes, the probability of having a collision is $p = \sum_{i=1}^{N} 1/N^2 = 1/N$. Since there are $\binom{n}{2}$ pairs in a sample of size n, the number of collisions follows the binomial distribution $B(n(n-1)/2, 1/N)$ whose variance is

$$var(C) = \binom{n}{2} p(1 - 1/N) = \mu(1 - 1/N) \tag{3.15}$$

When N is large, $var(C) \approx \mu$. Since C follows binomial distribution, $E(C - \mu)/\mu \ll 1$ when μ is not a very small number, causing the third term above is negligible compared with the second term. Therefore

$$\frac{1}{C} \approx \frac{1}{\mu} - \frac{C - \mu}{\mu^2}, \tag{3.16}$$

and

$$var\left(\frac{1}{C}\right) \approx \frac{var(C)}{\mu^4} \approx \frac{1}{\mu^3}.$$

Substitute the above into Equation 3.14, we have:

$$var(\widehat{N}_N) \approx \frac{n^4}{4\mu^3} = \frac{N^2}{\mu}. \tag{3.17}$$

□

3.7.2 Proof of Lemma 2

Proof. Let random variables $X = \hat{\gamma_n}^2 + 1$ and $Y = n^2/(2C)$, where $\widehat{\langle\gamma^2\rangle}_n$ is the estimated γ when the sample size is n. Applying the formula for variance of a product of two random variables, we have

$$\begin{aligned} var(\widehat{N}_E) &= \bar{Y}^2 var(X) + \bar{X}^2 var(Y) + var(X)var(Y) \\ &= \frac{N^2}{\bar{X}^2} var(X) + \bar{X}^2 \frac{N^2}{\bar{X}^2 \mu} + var(X)\frac{N^2}{\bar{X}^2 \mu} \\ &= \frac{N^2}{\mu}\left(\mu\frac{var(X)}{\bar{X}^2} + 1 + \frac{var(X)}{\bar{X}^2}\right) \\ &= \frac{N^2}{\mu}\left(1 + (\mu + 1)\frac{var(X)}{\bar{X}^2}\right) \end{aligned}$$

where $\bar{X}$ and $\bar{Y}$ are the expectations of X and Y. Let CV(X) denote the coefficient of variation of X, which is a constant for a given dataset. $var(X) = CV^2(X)\bar{X}^2/n$.

Substitute this into the above equation, we have

$$\begin{aligned} var(\widehat{N}_E) &= \frac{N^2}{\mu}\left[1 + (\mu+1)\frac{CV^2(X)}{n}\right] \\ &\approx \frac{N^2}{\mu}\left[1 + \frac{nXCV^2(X)}{2N}\right] \end{aligned}$$

When data size is large, $n \ll N$, the second term in the above equation can be omitted, resulting in

$$var(\widehat{N}_E) \approx \frac{N^2}{\mu}$$

□

Chapter 4

Detect Inflated Follower Numbers in OSN Using Star Sampling

This paper was submitted as:

Hao Wang, Jianguo Lu, *Detect Inflated Follower Numbers in OSN Using Star Sampling.* The 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, submitted.

4.1 introduction

The properties of online social networks (OSNs) are of interests to a variety of stakeholders, including general public as well as IT professionals [60]. Often the raw data are not available and the summary released by the service providers are sketchy. OSNs are so large that exhaustive exploration of the network is infeasible. Instead, we can only obtain a small portion of the network and estimate the properties of the network using the sample [17] [61] [13].

There are many studies on sampling methods for OSNs. Two of the basic sampling methods are uniform random sampling and PPS (probability proportional to size) sampling. Uniform random sampling is the norm of the practice, the method opted for whenever possible, also the method not easy to implement in many applications. In OSN studies, it is often realized by uniform ID sampling [61] [17], or Metropolis-Hasting random walk [17].

Some properties, such as the top bloggers, are innately not suitable for uniform random sampling, especially for scale-free networks where most of the bloggers have small number of followers [57]. It was widely accepted [62], as well as demonstrated in this paper, that most OSNs are scale-free networks. Uniform random sampling gives each blogger an equal probability of being sampled, meaning that top bloggers have no more chance of being sampled than other people. Consequently, the sample is mostly comprised of small accounts. Most top bloggers are not even sampled at once, let alone to study their properties.

Therefore, there is a need to use PPS sampling to sample the large microblog accounts more often. PPS sampling is hard to implement directly using existing OSN access methods. Most OSNs support random walk sampling, which can approximate PPS sampling in the sense that the sampling probability of a node(account) is proportional to its degree asymptotically [19]. It is not efficient in that in every random walk step, only one random sample is obtained from all the neighbours of the current node. This problem becomes more acute in OSN sampling where each step involves remote access to the API through internet, and sometimes there are daily quotas for the total number of accesses allowed. When one API call retrieves all the neighbouring nodes (followees), it is too costly to select only one of them and discarded all the others.

Thus, we propose star sampling that is an efficient approximation to PPS sampling. It selects random nodes first using ID sampling that is enabled by several OSNs, including Weibo OSN being studied. Then, for each random node we select all its neighbours connected by outgoing links, as if expanding the node to a star. In this way, the sampling process is faster by a factor of the average degree of the nodes. Yet, it is a kind of PPS

sampling as we will demonstrate in this paper. This method also avoids other difficulties in random walk sampling, such as dead-ends, infinite loops, and isolated components [19].

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

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

4.2 Background and Related Work

4.2.1 OSN Access Methods

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

- By probing account IDs: In some microblog sites such as Weibo and Twitter, microblogger's account can be accessed using http request such as www.weibo.com/1234567890, where the number is the account ID. Because every account can be accessed using an ID, and the ID space is not very large (a 10 digit number for Weibo), uniform random accounts can be found by generating a random number within the ID space. This method is used to obtain uniform random nodes (accounts) from Facebook [17], Youtube [35] and Weibo [61] .
- By crawling using web API: Most OSNs provide programmable web APIs, typically supporting programmers to navigate in the network, such as getting the out-going and in-coming links. New blogger data can be obtained by following the links provided in the current account.
- By crawling HTML pages and screen scraping: Instead of using more organized web APIs, OSN data can be also directly extracted from its HTML pages. By following

the hyperlinks imbedded inside the web pages, we can find the neighbours of the current blogger.

- By sending queries: Most OSNs provide searchable interfaces, either by providing an API or an HTML form. In either way, we can send queries and retrieve matched pages.

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

4.2.2 Graph sampling

An OSN can be modelled as a graph, where an account(or a blogger) is a node, and nodes are connected by following relationship. In general, a graph can be sampled by random node, random edge, and random walk. Their comparative studies are conducted in [31] [27] [10].

4.2.2.1 Uniform Random Node Sampling

In this sampling method, each node is sampled with equal probability. It can be realized by selecting the nodes directly, as in random ID sampling, or by following the links using certain strategies such as Metropolis-Hasting random walk [25] [63] [17].

4.2.2.2 Random edge sampling

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

4.2.2.3 Random walk sampling

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

4.2.3 Weibo and other OSN sampling

There are very few papers depicting the landscape of Weibo OSN despite its enormous size and influence. Very recently [61] uses uniform samples to estimate the properties of Weibo OSN. They report a wide range of estimated properties such as number of accounts, active accounts according to messaging information, and geographic distributions. Due to the limitation of uniform random sampling, they are not able to find out the degree distribution, the number of followers of top bloggers. We apply both uniform random sampling and PPS sampling, thereby obtain more interesting results.

Similar OSNs are extensively studied, including Facebook [17] [64] [65] [66] and Twitter [67] [62] [41] [68] [15]. Compared to this groups of work, we are not aware of the star sampling method proposed in this paper, neither any study on properties of top bloggers.

4.3 Uniform ID Sampling and Size Estimation

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

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

$$\hat{N} = \frac{v}{n} U, \tag{4.1}$$

whose approximate relative standard deviation (RSD) is

$$RSD(\widehat{N}) = \sqrt{1/v}. \tag{4.2}$$

We refer to Appendix for the proof. Intuitively, the 95% confidence interval depends on v, the number of valid IDs being sampled. For instance, when $v = 10^4$, 95% confidence interval is

$$\widehat{N} \pm 1.96 \times RSD \times \widehat{N} \approx \widehat{N} \pm 0.02\widehat{N}.$$

In the case of Weibo each user account ID is a 10-digit number, i.e., $U = 10^{10}$. Although some accounts have account names, they are still have a 10-digit ID that is accessible by

Algorithm 1: Uniform ID sampling

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

our method. As Equation 4.2 shows, the success of ID sampling hinges on the value of v, which in terms is decided by the ration of N/U. If the universe U were very large (say, by allowing for arbitrary length of letters), most of the randomly generated IDs would be invalid ones. Such low ratio will render the ID sampling infeasible. Fortunately, in the case of Weibo, the ratio is rather large and the probability of success is $21104/848969 \approx 0.025$. To expedite the speed, the DNS resolution is done once and cached for later use.

We run ID sampling in December 2011. Figure 4.1 shows four independent sampling processes, along with the projected error bound derived from Equation 4.2. Each process has sample size around 500,000. Overall, the estimation of the total number of accounts is 243 million (95% Confidence interval is between 238 million and 247 million).

Our results coincide with the size estimation reported in [61], where 269 millions of account are projected in January 2012. Within one month, we observe an increase about 9 % of user accounts. Another observation we have is that relatively small number of samples are needed to reach an accurate estimation of the user account number.

4.3.1 Degree and message distributions

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

It is reported that a uniform random sample can reflect the distributions of the original data [27]. Figure 4.2 shows the distributions of the in-degree, out-degree, and messages. All are in log-log plot since they have long tails. Each data is plotted in two ways: The degree-rank plots in the first row focus on the top nodes, while the frequency-degree plots on the second row focus on the nodes with small degrees. In a degree-rank plot, all the nodes are sorted according to its degree in increasing order, then a rank is assigned

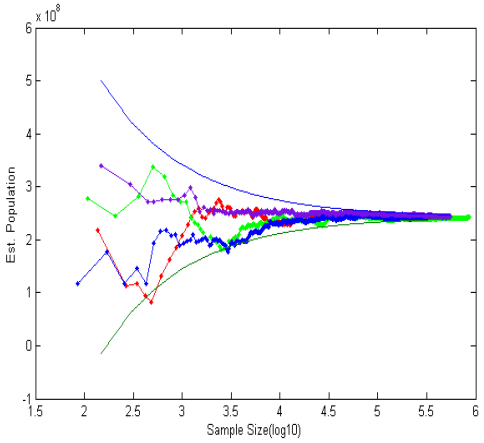


FIGURE 4.1: Estimated number of accounts against sample size. The estimation stabilizes when only 20,000 random IDs are tested.

to each node. In frequency-degree plot, the occurrence frequency of a degree is plotted against the degree.

Figure 4.2 (A) shows that the out-degree has a limit around two thousand. Its corresponding frequency-degree plot in subplot (D) shows that there are more than 10^5 nodes that have only one out-going edge among the one million sampled nodes.

In-degrees are closer to power-law distribution, similar to most other networks such as Twitter [41], Facebook [17], and the Web graph [69]. Subplot (B) shows an almost straight line with exponent one, similar to that of Twitter data [41]. This plot also demonstrates that uniform random sampling can only reveals the shape of the follower distribution, not the details of top bloggers. For instance, there are only two of the sampled accounts that have follower number greater than one million. Using star sampling discussed in the next section, we found that there are 691 millionaires who have more than one million followers. Subplot (E) is the corresponding frequency-degree plot that shows most of the bloggers have small number of followers. For instance, in the sampled nodes, there are more than 10^5 number of nodes/bloggers who have only one follower.

Subplots (C) and (F) describe the message distribution among the samples. Subplot (C) is the rank-message plot, describing that the number of messages descreases quickly. The top sampled blogger sends close to 10^5 number of messages. Overall, the curve fit better with Mendelbrot law [39].

The standard deviations are 103.2708 and 2916.8887 for in-degrees and out-degrees, respectively. From the uniform random samples we can estimate the average in-degree

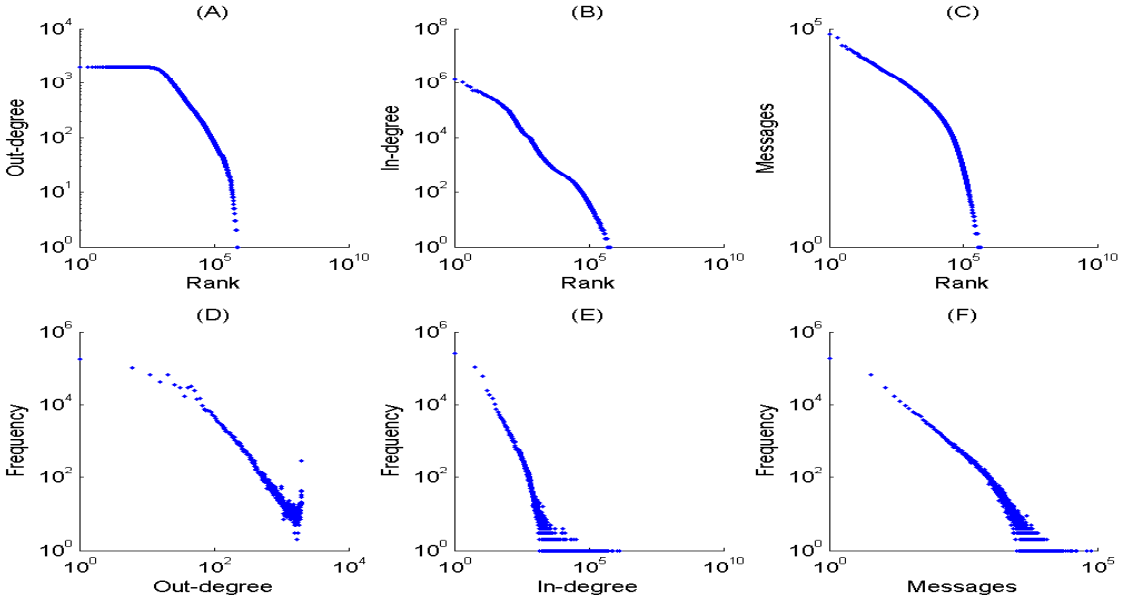


FIGURE 4.2: Estimated out-degree, in-degree, and message distributions of Weibo.

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

$$\widehat{\langle d \rangle}^{out} = \frac{1}{n}\sum_{i=1}^{n} d_i^{out} = 32.1083$$

$$\widehat{\langle d \rangle}^{in} = \frac{1}{n}\sum_{i=1}^{n} d_i^{in} = 54.3973$$

Surprisingly, the average in-degree is markably larger than the average out-degree. Such inconsistency can be caused by several sources. One may be the inflated follower number as suggested in the next section.

4.4 Star sampling and Follower Number Estimation

Fake OSN followers has become a multimillion dollar business. In Twitter zombi followers are sold in large quantities ranging from thousands to millions. There are robots to generate zombies to follow designated bloggers, and there are also tools to detect the percentage of zombi followers[1]. This paper addresses another type of fake follower number, where the zombis are not added, instead the follower number is artificially inflated.

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

Algorithm 2: Star sampling

Input: Valid uniform IDs V, sample size n;
Output: Sample sequence S.
i=0;
S is empty;
while $i < n$ **do**
    x is a valid ID randomly selected from V;
    Expand node x into a star that contains edges $\{x \to x_1, x \to x_2, \ldots, x \to x_k\}$;
    Add nodes $i_1, i_2, \ldots, i_k$ to sample sequence S;
    i=i+k;
end

4.4.1 Star sampling

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

Since we already have the uniform random IDs, and every remote API call gets back all the neighbours, we utilize all the neighbouring nodes by employing the following star sampling as described in Algorithm 2: select a set of uniform random nodes, and expand each node as a star that contains all the neighbours. Put all the nodes in the neighbours into the sample.

Intuitively, star sampling approximates random edge sampling in that a set of edges are selected randomly. The centre of the star is discarded because its selection probability is uniform. The other nodes being pointed to are selected because their selection probability is roughly proportional to their in-degrees.

Suppose that in the directed graph G, there are N number of nodes labeled as $1, 2, \ldots, N$. Define the volume of the graph $vol(G)$ as the sum of all the in-degrees, i.e., $vol(G) = \sum_1^N d_i$, where d_i denote the in-degree of node i. Given a sample containing n nodes $\{x_1, x_2, \ldots x_n\}$, where $x_i \in \{1, 2, \ldots, N\}$. Suppose that node i occurs f_i number of times in the sample. Our task is to estimate the number of followers (in-degrees) F_i of node i.

The probability node i being sampled is

$$p_i = \frac{d_i}{vol(G)} \tag{4.3}$$

The number of times node i is selected after n sample nodes are taken can be approximated by the binomial distribution as below:

$$P(k) = \binom{n}{k} p_i^k (1 - p_i)^{n-k} \tag{4.4}$$

It is well known that the expected number of captures of node i is $E(f_i) = np_i$. Thus, the occurrence probability of node i in the population can be estimated using the following, which can be also derived from maximal likelihood method:

$$\hat{p}_i = \frac{f_i}{n}. \tag{4.5}$$

The number of the followers of node i is estimated by

$$\widehat{\langle d \rangle}_i = \hat{p}_i \; vol(G) = \frac{f_i \; vol(G)}{n} = \frac{f_i}{n} \widehat{N} \widehat{\langle d \rangle}^{out}. \tag{4.6}$$

Because of the binomial distribution, the variance of f_i is

$$var(f_i) = np_i(1 - p_i) \approx np_i. \tag{4.7}$$

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

$$var(\widehat{\langle d \rangle}_i) = var(f_i) \; vol(G)^2/n^2 = f_i vol(G)^2/n^2. \tag{4.8}$$

Hence the relative standard deviation is

$$RSD(\widehat{\langle d \rangle}_i) = \sqrt{1/f_i}. \tag{4.9}$$

Equation 4.9 gives the guideline to select the sample size so that satisfactory estimation can be obtained. For instance, if we want the 95% confidence interval to be within $\widehat{\langle d \rangle}_i \pm 0.2\widehat{\langle d \rangle}_i \approx \widehat{\langle d \rangle}_i \pm 1.96 \times \sqrt{1/f_i}\hat{d}_i$, f_i needs to be greater than 100. We use this guideline to design our experiments.

4.4.2 Pilot study on local datasets

The accuracy of the star sampling depends on the assumption that nodes are sampled with probability proportional to their degrees. In one extreme case where every star has only exactly one out-going edge, star sampling resembles random edge sampling, and it is a PPS sampling. In the other extreme when a star is very large and covers all the remaining nodes in the network, all those nodes are equally sampled. Therefore, it becomes uniform random node sampling. In most cases stars are of moderate size, since the vast majority of the nodes are small ones because of the scale-free nature of the network.

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

To validate our assumption, we carried out a pilot experiment on local datasets. Since the ground truth values are known, the estimator and the variance can be evaluated. Our local datasets are six networks whose statistics are summarized in Table 4.1. Star sampling are applied to each network. We evaluate the estimation performance on the top 15 nodes for each network, by comparing their empirical variance with the theoretical variance, and empirical average with the true value as demonstrated in Figure 4.3.

The 95% error bounds are calculated from Equation 4.9, the box plots are obtained from 100 repetition of the experiments. The average of 100 estimations fit well with the true value across all the networks and all the top 15 nodes. This indicates that star sampling is indeed unbiased. In addition, most of the estimations fall within the estimated error bound, demonstrating that star sampling can approximate PPS sampling. The sample size is controlled so that each of the 15-th node can be sampled at least 50 times. Depending on the degree of the 15-th node and the overall degree distribution, varying sample size is needed for each network (50K for WikiTalk, 200K for Skitter, 80K for Youtube, 80K for NotreDame, 200K for Stanford and 40K for EmailEU).

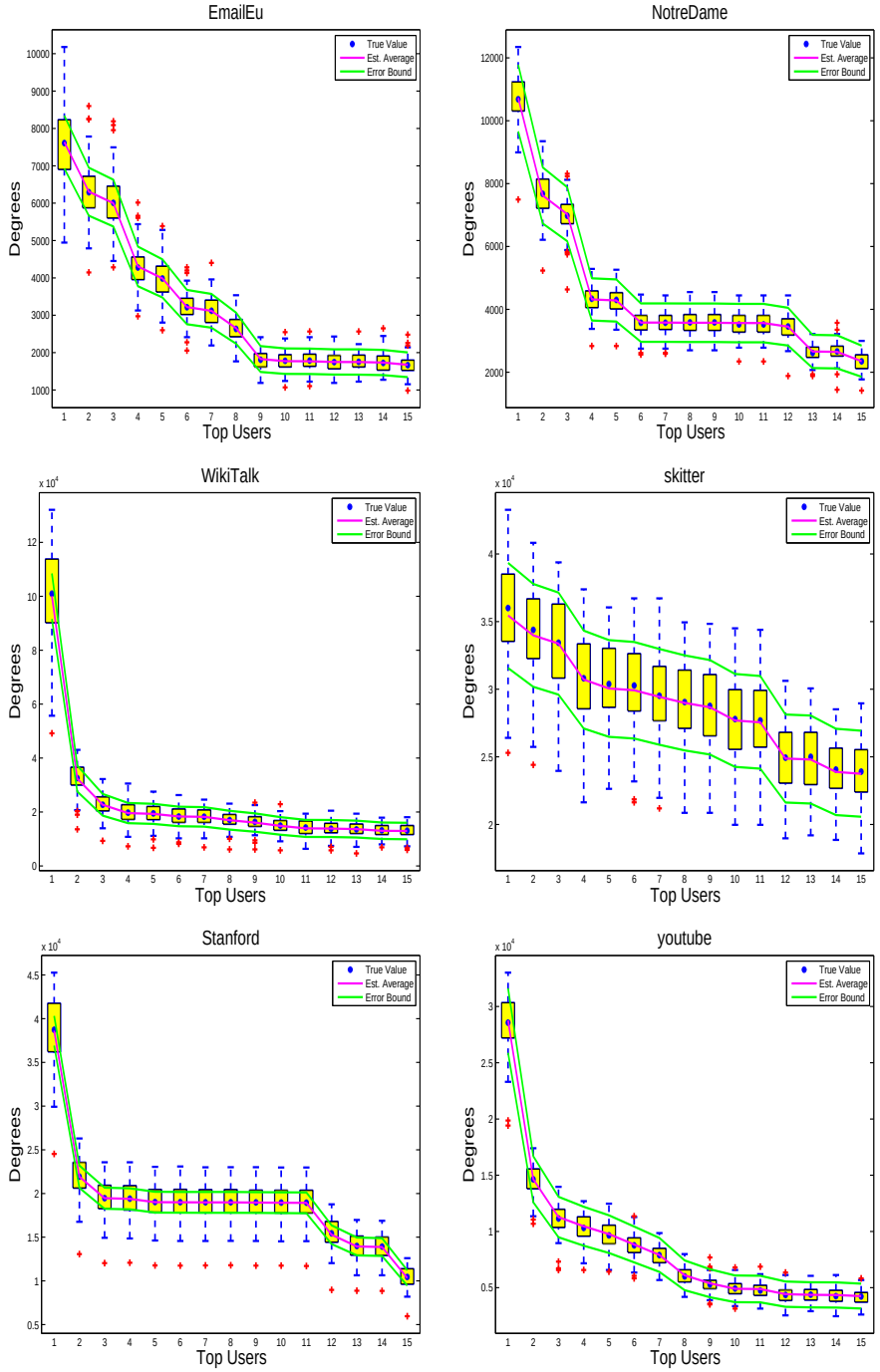


FIGURE 4.3: Degree estimation of six networks using star sampling. Boxplots are obtained from 100 repeated experiments.

Network	# Nodes	γ	$\langle d \rangle$	Max degree
WikiTalk[10]	2,394,385	26.34	3.89	100,029
EmailEu[10]	265,009	13.93	2.75	7,636
Stanford[10]	281,903	11.79	14.14	38,625
Skitter[10]	1,696,415	10.46	13.08	35,455
Youtube[42]	1,138,499	9.65	5.25	28,754
NotreDame[10]	325,729	6.40	5.25	10,721

TABLE 4.1: Statistics of the 6 real-world graphs, sorted in descending order of the coefficient of degree variation $\gamma = variance/\langle d \rangle^2$.

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

TABLE 4.2: Estimation for the top 10 Weibo accounts. f_i: capture frequency of the account i; d_i claimed in-degree or number of followers; $\widehat{\langle d \rangle}_i$: estimated number of followers.

4.4.3 Results for Weibo data

During October 2011 and January 2012, we selected 1,184,964 number of uniform random nodes, on average each random node has 32.08 number of out-going links. We expand each uniform random node as a star, and collect all the nodes pointed by the out-going edges as the sample. Overall 38,019,277 number of sample nodes are collected, including duplicates. The largest account has claimed 23,335,290 number of followers, and is captured 85,016 times. We reckon according to Equation 4.9 that around 100 of captures are required to produce meaningful estimation. Thus, we take only the top 10,000 accounts, the lowest has 16,038 followers and is captured 65 times.

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

$$ratio = (d_i - \widehat{\langle d \rangle}_i)/\widehat{\langle d \rangle}_i, \tag{4.10}$$

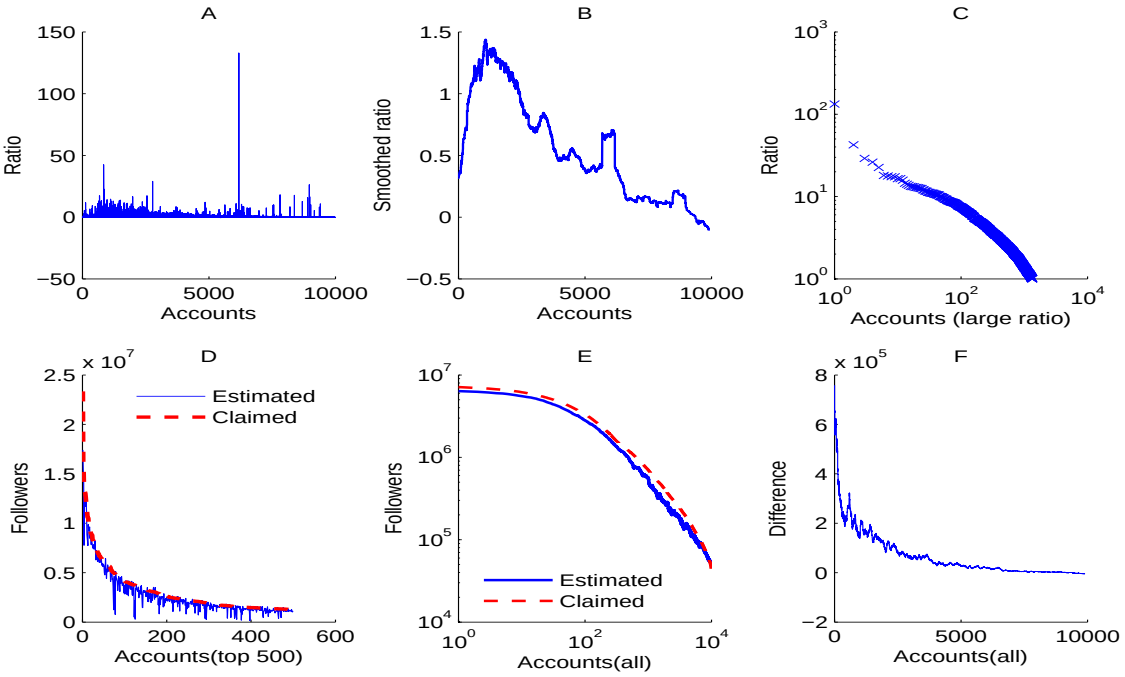


FIGURE 4.4: Weibo followers estimation. Panel A: inflation ration over 10^4 top accounts. Panel B: the smoothed version of A. Panel C: All the accounts whose inflation ratio is higher than one. Panel D: top 500 accounts. Panel E: comparison of top 10^4 accounts, smoothed. Panel F: difference between the claimed and estimated followers. Smoothed.

where d_i is the claimed number of followers (in-degree), and $\widehat{\langle d \rangle}_i$ is the estimated number of followers. Table 4.2 listed the estimations for the top 10 accounts in Weibo.

For the claimed top 10,000 accounts, there are in total 6,069 accounts whose ratio is between -0.2 and 0.2, 52 is smaller than -0.2, and 3,930 is larger than 0.2. The minimal ratio value is -0.482, while the highest is 132.8413. Overall the total number of claimed followers is 23% more than the estimated followers. The claimed follower numbers are taken at the end of the experiment, while the whole sampling process spans a few months. Given the dynamic nature of the network, especially the fast increasing number of new accounts and followers, it is understandable that overall claimed number is higher than the estimated number.

However, there are many accounts with very high inflation rate. The inflation rate for all 10,000 accounts are plotted in Figure 4.4 (A), where the accounts are sorted by their claimed follower numbers in decreasing order. Figure 4.4 (B) is the corresponding smoothed ration plot using moving average with 500 window size. Those two plots show that there are higher inflation rate for accounts ranked between 1,000 to 2,000. In general, the inflation rate drops for smaller accounts. Altogether, there are 194 accounts whose ratios are greater than five (1,342 accounts greater than one), a very

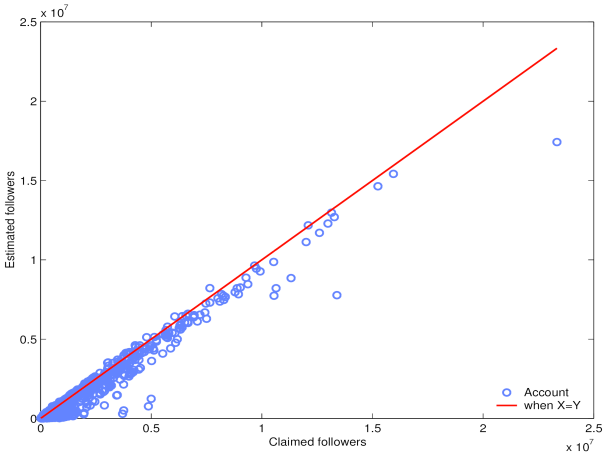


FIGURE 4.5: Estimated followers vs. claimed followers. The Pearson correlation coefficient is 0.9797.

large discrepancy that is hard to explain. We plot those 1,342 accounts in log-log scale in Figure 4.4 (C). Interestingly enough, the inflation rate also follows power law.

Figures 4.4 (D) depicts the comparison between the estimated and claimed follower numbers for the top 500 accounts. There are spikes pointing downwards, indicting the accounts having high inflation rate. Figure 4.4(E) gives an overall picture of all the accounts, both lines are smoothed using window size 100 using log-log plot. Figure 4.4(F) is the smoothed difference (claimed -estimated) for all the accounts. The smoothing window size is also 100.

Lastly, we draw the correlations between the claimed and estimated followers in Figure 4.5. The inflated number does not change greatly the overall landscape of the rankings of the accounts. The estimation is closely related to the claimed number as evidenced by the high Pearson correlation coefficient 0.9797. However, it is also clear from the plot that some accounts deviate a lot from the estimations.

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

4.5 Discussions and Conclusions

This paper proposes the star sampling to estimate properties of online social networks. Some network properties prefer PPS sampling, which is not easy to carry out for most online social networks. Random walk can approximate PPS sampling, but it collects only one random node in its neighbours. Star sampling improves the performance of random

walk sampling by a factor of $\langle d \rangle$, the average degree of the network. We demonstrate on six local datasets that star sampling approximates PPS sampling very well.

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

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

4.6 Acknowledgement

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

4.7 Appendix

The occurrence probability P of the valid ID can be estimated by the maximum likelihood estimator:

$$\hat{P} = \frac{v}{n}. \tag{4.11}$$

Therefore, the number of accounts can be estimated by

$$\hat{N} = \hat{P}U = \frac{v}{n}U. \tag{4.12}$$

Although it is an unbiased estimator, what matters is the variance, and whether it is feasible in our application. Note that the number of times v of hitting a valid ID follows a binomial distribution whose variance is

$$var(v) = nP(1-P). \tag{4.13}$$

Consequently, the variance of $\widehat{N}$ is

$$var(\widehat{N}) = \frac{P(1-P)}{n}U^2. \tag{4.14}$$

Given that P is a rather small value, the relative standard deviation is

$$RSD(\widehat{N}) = \sqrt{\frac{1-P}{nP}} \approx \sqrt{1/v}. \tag{4.15}$$

Chapter 5

What Do Large Networks Look Like?

This paper was submitted as:

Hao Wang, Jianguo Lu, *What Do Large Networks Look Like?*. The 19th ACM SIGKDD Conference on Knowledge, Discovery, and Data Mining, Workshop, 2013, submitted.

5.1 Introduction

The topology of large network is hard to visualize, yet it is crucial for data mining applications. If we plot a network with millions of nodes, not to mention hundreds of millions of them, it will be hard to discern the community structure no matter what graph layout is used, and how powerful the computer is. To reveal the visual cues to the structure of the network, we need to reduce the number of nodes and edges by producing a representative subgraph.

Network visualization has been widely studied [70]. Most approaches can only handle graphs of size up to hundreds of nodes and thousands of edges [71]. Beyond this limit, it will be hard to discern the nodes from edges, preventing the discovery of patterns in the graph. Since the tree layout algorithm has the simplest complexity to implement, it is a common practice to reduce the number of edges by turning the graph into a tree, especially a spanning tree representation [72]. The crucial issue is which spanning tree is more representative of the original graph. A spanning tree obtained by breadth-first search will distort the structure of the original network. Numerous efforts have been devoted in adding weights and finding the minimal spanning tree. When the network is very large, computationally it may not be feasible to compute the minimal spanning tree.

Instead of artificially tweaking the parameters for a better spanning tree, we argue that a uniform random spanning tree should be a more natural choice. A typical algorithm to find the uniform spanning tree borrows the idea from random walk [73], therefore, the complexity of the algorithm is the same as the random walk cover time. Although for uniform random graphs the cover time is in the order of $O(NlogN)$, where N is the number of nodes in the network, real-world networks are often scale-free and clustered. Thus we need to prepare for the worst case complexity which is $O(N^3)$ [19]. Obviously, the cost is too high for large networks if we use that algorithm directly.

We observe that it is not necessary to keep all the nodes to reveal the topological structure of large networks. The number of nodes also need to be cut down for very large networks even when tree representation is adopted. We can imagine that a large network has many layers of meshes lying in stack. When all the layers are plotted, the nodes and the structure are obscured by the meshes. If we plot only one random mesh, the crucial nodes and the structure are revealed.

Such random mesh can be obtained by casting the edges uniform randomly–each edge has the same probability of being selected. When an edge is casted, two nodes incident to the edge are collected. In this way, a node will be selected with probability proportional to its degree size. Since random edge selection is not supported in many online social

networks, we use simple random walk to approximate the process, considering that the node selection probability is the same asymptotically as random edge sampling [19]. Based on this random walk we simultaneously generate the corresponding random spanning tree. Thereby we reduce the number of nodes and edges at the same time efficiently.

The evaluation of the visualization also imposes a challenge. Since the entire network can not be effectively plotted, the visual comparison between the sub-graph and the original graph is impossible. In particular we would like to see whether the community structure can be visualized. For this purpose we use NCP (network community profile) [1] to evaluate the visualization. As a result, we find that our visualization corresponds to NCP very well.

Contributions 1) We propose an efficient algorithm to visualize large networks. It can scale to very large networks when they are scale-free and crucial nodes and subsequent structure can be surfaced quickly using random walk; 2) We demonstrate that the visualization can preserve the community structure by comparison to the NCP; 3) The random spanning tree algorithm is adapted into our random walk node sampling process, reducing the potential high complexity ($O(N^3)$) to a linear algorithm.

5.2 Our method

There are at least two ways to select the representative nodes in a graph: by selecting the nodes uniform randomly, or selecting the nodes with probability proportional to their sizes (PPS). When uniform random node selection is applied, most of the nodes will be small nodes with low degrees due to the scale-free nature of the network. The large node with many connections most probably will not be sampled and omitted in the subgraph. Thus we use PPS sampling to obtain the representative nodes, where large nodes have higher probability of being selected. Simple random walk is an efficient sampling method that is supported by many real online social networks, and node sampling probability is proportional to its size asymptotically. Since our random walk is rather long (6×10^4 distinct nodes in our experiments) and well exceeds the mixing time of the graph, the sampling probability can approximate PPS sampling.

Even when the number of nodes are reduced, the network structure is still being obscured by excessive number of edges. Various methods have been proposed to reduce edge size, such as turning the graph into a spanning tree [70, 72]. We propose to use random spanning tree, which can be generated using random walk as illustrated in Algorithm 3. It was originally given by [73], and can be explained as follows: we perform the simple

Algorithm 3: Random Spanning Tree

Data: Graph G;
Result: Random spanning tree T of size n.
Let n_0 be a uniform random node from G;
mark n_0;
while i*j*n **do**
    neighbours(n_{i-1})= all the neighbours of n_{i-1};
    n_i is a random node of neighbours(n_{i-1}) ;
    if n_i *is not marked* **then**
        i++;
        mark n_i;
        add edge (n_{i-1}, n_i) to T ;
    end
end

random walk as usual, but add an edge to the tree only when the edge does not form a loop. According to [73], we have the following surprising result:

Theorem 4. Among all the spanning trees of graph G, T is one of the uniform random sample.

It may take very long random walk to cover all the nodes of a graph, especially when the graph is scale-free and clustered. The worst case complexity is in the order of $O(N^3)$. Since node selection also uses random walk, we combine the two random walks together to trim nodes and edges simultaneously, avoiding the need to cover all the nodes.

When the random spanning tree is plotted using two-dimensional layouts such as the well-known spring model, the structure is still cluttered for trees containing tens of thousands nodes. We use 3D hyperbolic layout [71] to ameliorate the problem.

5.3 Community structure

We demonstrate our method on the discovery of community structure. The community structure is measured using NCP (network community profile) plot proposed in [1]. We refer to Figure 5 in [1] for a good explanation of NCP, where complete small network visualizations are compared side-by-side with NCP. That figure explains that NCP corresponds well to network visualization in small size ($\sim$ 100 nodes), while we show that the profile is also reflected in our visualization for large networks consisting of millions of nodes.

In network studies, one important measurement for network structure is its conductance, which can be used to characterize the spectral gap and random walk mixing time [45].

TABLE 5.1: Statistics of the six networks, each has a citation indicating where the data is from. $\langle d \rangle$ is the average degree, CV stands for coefficient of variation.

Network	# Nodes	CV	$\langle d \rangle$	Max degree
Flickr [10]	105,936	2.65	43.43	5,425
NotreDame[10]	325,729	6.40	5.25	10,721
Stanford[10]	281,903	11.79	14.14	38,625
Amazon[10]	410,236	1.27	11.89	2,760
Facebook [44]	63,731	1.56	25.64	1,098
Youtube[42]	1,138,499	9.65	5.25	28,754

The conductance is defined as follows: Let V be the set of nodes of a graph. The conductance of a subset of nodes S of V is

$$\Phi(S) = \frac{\sum_{i \in S, j \in V \setminus S} A_{ij}}{min(A(S), A(V \setminus S))} \tag{5.1}$$

where A is the adjacency matrix of the graph, and $A(S) = \sum_{i \in S, j \in V} A_{ij}$. The conductance of the graph is $\Phi = min_S \Phi(S)$. NCP not only looks at the minimal graph conductance, but also the component conductance over the component size.

We conducted experiments on dozens of large networks we can find. Most of them are from Stanford SNAP graph collection [10]. Due to space limitation, we only report the comparison with NCP on six networks[1]. Their statistics are summarized in Table 5.1. To demonstrate the scalability of our method, we plot a subgraph obtained from the complete Twitter user network that contains 4.1×10^7 nodes and 1.4×10^9 edges [41] in Figure 5.1. The overall structure clearly differs from other networks plotted in Figure 5.3. In contrast to the well enmeshed Facebook network, Twitter has a string of super large nodes(bloggers) stacking on each other. Each super node has its own circle of fans with little interaction between them. The veracity of such topology is not easy to verify using NCP, because NCP can not be calculated due to the huge size. However, we can gain some confidence from other relatively smaller networks where NCP can be computed as shown in Figures 5.2 and 5.3.

Figure 5.2 shows the NCP plots, the conductances over the size of the subcomponents for the original six networks. They are plotted using SNAP API [1]. The insets (in red colour) are the NCP plots obtained from the corresponding subgraphs. We can see that the NCP from subgraph resembles the shape of NCP from the original graph.

Our visualizations of these networks are plotted in Figure 5.3. The colour of the nodes represents the node degree in the original network. Among the six networks, three of

[1]Complete data description and programs can be found at http://cs.uwindsor.ca/∼jlu/visualization.

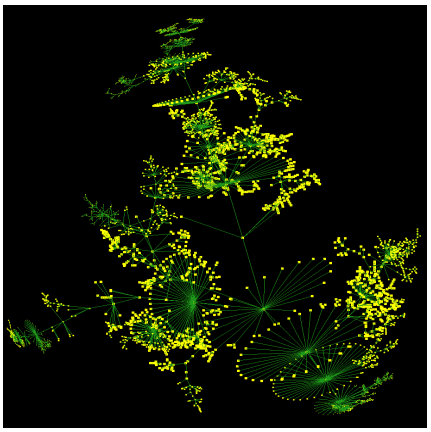


FIGURE 5.1: Visualization of Twitter user network.

them (Flickr, NotreDame, and Stanford) have low graph conductance, while three others (Facebook, Amazon and Youtube) have high conductance as comparison.

Overall, each visualization corresponds well to its NCP plot of the original network. Several networks are remarkably different from others. Take the first network, Flickr, for example. The NCP plot of the original network in Figure 5.2 shows a sharp dip ($\sim 10^{-3.5}$) around the component size 10^4, indicating that there is a large component separated from the remaining part. Our visualization in Figure 5.3 reflects this dumbbell structure clearly. There is a long link connecting these two components, the nodes along the link are mostly of blue and green colour, indicating that the passage between those two components is narrow in the original network. These two components are well enmeshed, coinciding with the NCP plot showing that for most component sizes the conductance is rather large (above 10^{-2}).

NotreDame and Stanford web graphs exhibit a different pattern in both visualization and NCP plots. In their NCP plots, there is a low conductance when the component size is commensurate to the total size. Correspondingly, in the visualization there are clusters of similar sizes. In NCP plots, there are many low conductances when the component size is small. Correspondingly, in the visualization there are many small clusters that is obviously different from the Flickr network.

Amazon, Facebook, and Youtube networks have high conductances as shown in their NCP plots. Correspondingly their visualizations show well enmeshed networks. Note that although the minimal conductance of Amazon network is rather small, the cut happens when the component size is around 100, well below the total size. Therefore, its visualization does not show large clusters.

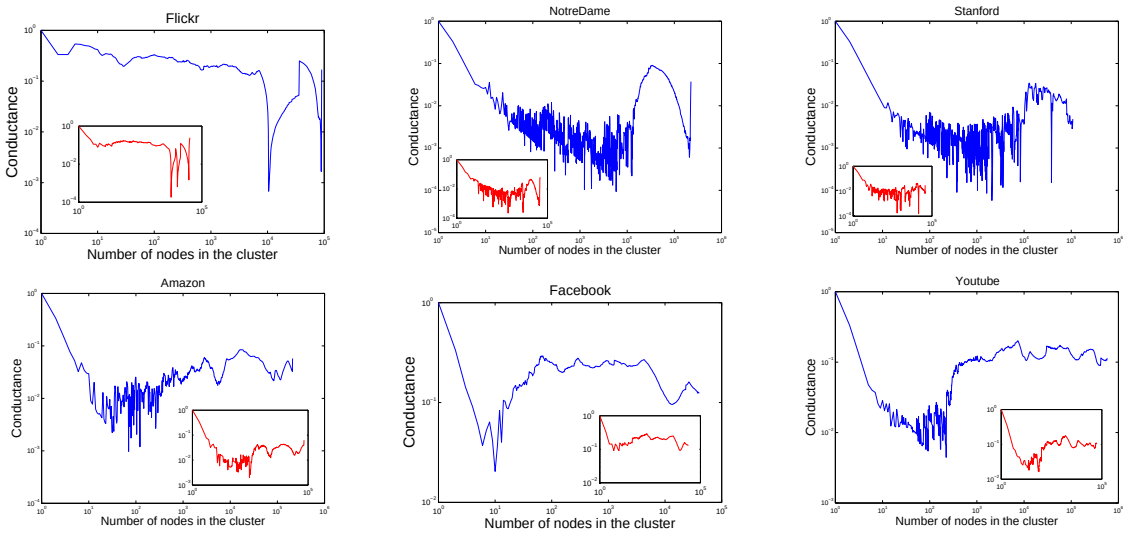


FIGURE 5.2: Conductance $\Phi(S)$ over $|S|$, the size of the the components, for six networks. Insets: The corresponding NCP plots obtained from the subgraphs.

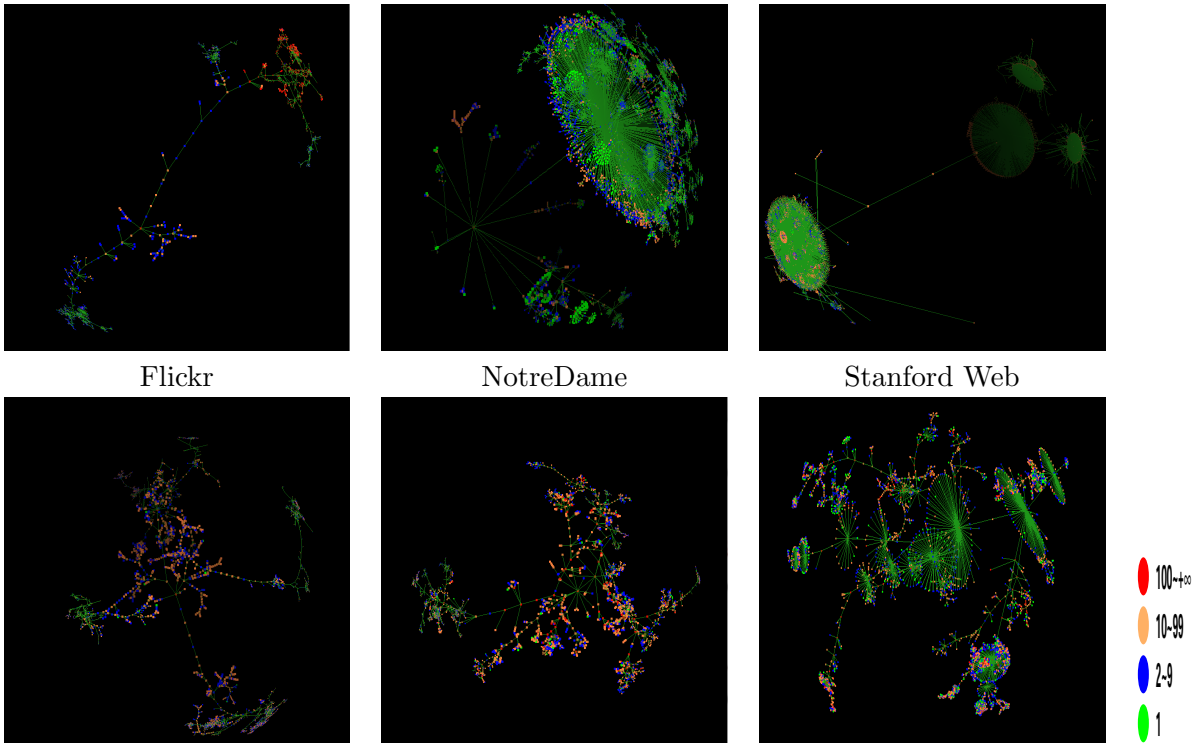


FIGURE 5.3: (Best viewed in colour) Visualization of six networks. The networks in the first row (Flickr, NotreDame, and Stanford) are clustered, while the networks in the second row (Amazon, Facebook and Youtube) are well enmeshed. Node colour indicates the node degree in the original network. More graphs can be found at http://cs.uwindsor.ca/∼jlu/visualization.

5.4 Conclusions

We demonstrate a practical method to visualize the structure of large networks. The method reduces both the number of nodes and edges of the network dramatically, yet it retains the global topology of the networks. More importantly, our method is very efficient, and works even when the data in its entirety is not available as long as simple random walk is supported.

This is the first attempt to use random spanning tree to reduce the size of the graph for visualization purpose. Direct application of the random spanning tree algorithm does not scale. By combining the random spanning tree algorithm with PPS node sampling, we propose a very efficient algorithm to reduce both the number of nodes and the number of edges leveraging the scale-free nature of the networks. 3D layout is also essential to capture the overall structure.

The calculation of NCP requires the access of the entire data, and may not be feasible for very large networks. As a companion to NCP (network community profile), our fast visualization method sheds a light for the prediction of NCP using only a small sample of the data.

Bibliography

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

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

[3] Andrei Broder, Marcus Fontura, Vanja Josifovski, Ravi Kumar, Rajeev Motwani, Shubha Nabar, Rina Panigrahy, Andrew Tomkins, and Ying Xu. Estimating corpus size via queries. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 594–603. ACM, 2006.

[4] Jamie Callan and Margaret Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems (TOIS)*, 19(2):97–130, 2001.

[5] Luo Si and Jamie 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. ACM, 2003.

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

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

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

[9] Manos Papagelis, Gautam Das, and Nick Koudas. Sampling online social networks. 2011.

[10] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636. ACM, 2006.

[11] Bruno Ribeiro and Don Towsley. On the estimation accuracy of degree distributions from graph sampling. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 5240–5247. IEEE, 2012.

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

[13] 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.

[14] L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *Proceedings of the 20th International Conference on World Wide Web*, pages 597–606. ACM, 2011.

[15] J. Lu and D. Li. Bias correction in small sample from big data. 2012.

[16] Jianguo Lu and Dingding Li. Sampling online social networks by random walk. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, pages 33–40. ACM, 2012.

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

[18] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proceedings of the 10th Annual Conference on Internet Measurement*, pages 390–403. ACM, 2010.

[19] László Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993.

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

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

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

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

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

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

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

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

[28] 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.

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

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

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

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

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

[34] L.S. Buriol, G. Frahling, S. Leonardi, A. Marchetti-Spaccamela, and C. Sohler. Counting triangles in data streams. In *PODS*, pages 253–262. ACM, 2006.

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

[36] A. Dasgupta, X. Jin, B. Jewell, N. Zhang, and G. Das. Unbiased estimation of size and other aggregates over hidden web databases. In *SIGMOD*, pages 855–866. ACM, 2010.

[37] M. Zhang, N. Zhang, and G. Das. Mining a search engine's corpus: efficient yet unbiased sampling and aggregate estimation. In *SIGMOD*, pages 793–804. ACM, 2011.

[38] S.K. Thompson. *Sampling*. Wiley, 2012.

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

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

[41] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM, 2010.

[42] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[43] C. Wilson, B. Boe, A. Sala, K.P.N. Puttaswamy, and B.Y. 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.

[44] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna 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.

[45] 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.

[46] William I Goodman. The future of staff planning. *Journal of the American Planning Association*, 22(1):24–29, 1956.

[47] Andrei Broder and et al. Estimating corpus size via queries. In *CIKM*, pages 594–603. ACM, 2006. ISBN 1-59593-433-2. doi: 10.1145/1183614.1183699.

[48] Milad Shokouhi, Justin Zobel, Falk Scholer, and S. M. M. Tahaghoghi. Capturing collection size for distributed non-cooperative retrieval. In *SIGIR*, pages 316–323. ACM, 2006. ISBN 1-59593-369-7. doi: 10.1145/1148170.1148227.

[49] S.C. Amstrup, T.L. McDonald, and B.F.J. Manly. *Handbook of capture-recapture analysis*. Princeton Univ Press, 2005.

[50] K. Bharat and A. Broder. Estimating the relative size and overlap of public web search engines. In *WWW*, 1998.

[51] JN Darroch. The multiple-recapture census: I. estimation of a closed population. *Biometrika*, 45(3/4):343–359, 1958.

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

[53] S. Ye and S.F. Wu. Estimating the size of online social networks. *International Journal of Social Computing and Cyber-Physical Systems*, 1(2):160–179, 2011.

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

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

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

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

[58] dblp. from http://www.sommer.jp/graphs/. 2013.

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

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

[61] King-wa Fu and Michael Chau. Reality check for the chinese microblog space: a random sampling approach. *PLOS ONE*, 8(3):e58356, 2013.

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

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

[64] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna 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.

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

[66] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.

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

[68] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, pages 56–65. ACM, 2007.

[69] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[70] Ivan Herman, Guy Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualization: A survey. *Visualization and Computer Graphics, IEEE Transactions on*, 6(1):24–43, 2000.

[71] Tamara Munzner. Drawing large graphs with h3viewer and site manager. In *Graph Drawing*, pages 384–393. Springer, 1998.

[72] Chaomei Chen and Steven Morris. Visualizing evolving networks: Minimum spanning trees versus pathfinder networks. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 67–74. IEEE, 2003.

[73] David J Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM Journal on Discrete Mathematics*, 3(4):450–465, 1990.

Vita Auctoris

NAME:	Hao Wang
PLACE OF BIRTH:	Wuhan, China
YEAR OF BIRTH:	1988
EDUCATION:	University of Windsor Windsor, Ontario, Canada 2011-2013 M.Sc.
	China University of Geosciences Wuhan, China. 2007-2011 B.Sc.