Uniform Random Sampling Not Recommended

Jianguo Lu
School of Computer Science
University of Windsor
jlu@uwindsor.ca

Hao Wang
School of Computer Science
University of Windsor
wang115o@uwindsor.ca

Dingding Li
Department of Economics
University of Windsor
dli@uwindsor.ca

ABSTRACT

We show that uniform random sampling is not as effective as PPS (probability proportional to size) sampling in many estimation tasks. In the setting of (graph) size estimation, this paper demonstrates that random edge sampling outperforms random node sampling, with a performance ratio proportional to the normalized graph degree variance. This result is particularly important in the era of big data, when data are typically large and scale-free, resulting in large degree variance. We derive the result by first giving the variances of random node and random edge estimators. A simpler and more intuitive result is obtained by assuming that the data is large and degree distribution follows a power law.

ACM Reference Format:
Jianguo Lu, Hao Wang, and Dingding Li. 2018. Uniform Random Sampling Not Recommended. In *Proceedings of The 2018 Web Conference Companion (WWW'18 Companion).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3186240

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 [7], to the more recent challenge of gauging the size of the Web and search engines [1, 3, 12, 20] and online social networks [8, 11]. The direct calculation of data size is often not possible or desirable for several reasons. Quite often, data are hidden behind some searchable interfaces and programmable web APIs, such as online social networks and deep web data sources. The access is limited, and the data in its entirety are not available [11, 19]. The data can be distributed, and there is no central data repository such as in the case of peer-to-peer networks [17] or the Web [12]. Even when the data are available in one place, there are requirements for fast just-in-time analysis of the data [10]. Regardless of a large variety of application scenarios, a common approach to solving these problems is to use samples to have a fast estimation of the data size, instead of slow and direct counting of the data.

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

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

The norm of size estimation is to use uniform random samples whenever possible. Real data sources seldom provide uniform random samples directly. Therefore, there have been tremendous efforts to obtain uniform random samples from the Web [9], search engine indexes [1], and online social networks [6], to name a few. These uniform random samples are costly, in that each valid sample may be accompanied by many invalid ones that are thrown away. Recently, it was empirically observed that, instead of obtaining those costly uniform random samples, RW sampling is actually better than RN sampling for size [11] and average degree estimation [14][5] on *some* datasets.

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

- When is RE better than RN? RE is better than RN only when the graph is very large, and consequently, the sample size n has to be much smaller than the data size N. This is the scenario we assume, with application background such as estimating online social networks with a limited number of web-based queries.
- How much better is RE over RN? We demonstrate that there is an upper bound for the performance improvement, which is quantified by $\gamma^2 + 1$. Here γ is the coefficient of variation of node degrees. The upper bound is derived analytically, and confirmed empirically on 18 large data sets. The derivation uses the assumption that the data is very large.
- What can approximate RE sampling? When RE sampling is not available in practice, we need to resort to other methods to approximate RE (or PPS) sampling. RW is an option, but the performance varies widely from data to data. We find that RW can approximate the performance of RE for online social networks, but not for Web graphs.

This result is particularly important in the age of big data when large and scale-free networks are ubiquitous [2] [18]. These networks can have very large degree variance. In theory, γ^2 can be

WWW'18 Companion, April 23–27, 2018, Lyons, France

ACM ISBN 978-1-4503-5640-4/18/04..
https://doi.org/10.1145/3184558.3186240

infinitely large when the slope of the scale-free network falls under certain range. In practice, we observe γ^2 as large as 1300 for the Twitter network in 2009 [15], meaning that potentially RE sampling can be better in three orders of magnitude in terms of variance. Such huge difference between the sampling methods will not only change the landscape of sampling practice, but also shift the research focus. In the past, people strive for uniform random samples [1]. Nowadays for very large data, we should take PPS samples, or develop sampling methods that can approximate PPS sampling.

2 SAMPLING METHODS AND THEIR ESTIMATORS

Given an undirected graph $G(V, E)$, where V is the set of nodes, and E the set of edges. Let $N = |V|$, the parameter we want to estimate. Nodes are labeled as $1, 2, \ldots, N$, and their corresponding degrees are $d_1, d_2, \ldots, d_N$. The volume of the graph is $\tau = \sum_{i=1}^{N} d_i$, the average degree is $\langle d \rangle = \frac{1}{N} \sum_{i=1}^{N} d_i = \tau/N$. The variance σ^2 of the degrees in the graph is defined as $\sigma^2 = \langle d^2 \rangle - \langle d \rangle^2$, where $\langle d^2 \rangle = \sum_{i=1}^{N} d_i^2/N$ is the second moment, i.e., the arithmetic mean of the square of the degrees. The coefficient of variation (denoted as γ) is defined as the standard deviation, or the square root of the variance, normalized by the mean of the degrees:

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

Let $\Gamma = \gamma^2 + 1$.

Suppose that a sample of n nodes $(d_{x_1}, \ldots, d_{x_n})$ is taken from the graph, where $x_i \in \{1, 2, \ldots, N\}$ for $i = 1, 2, \ldots, n$. Among them, there are f_j nodes that are sampled exactly j times. Then, sample size $n = \sum j f_j$. Let C denote the number of collisions in the sample, i.e., $C = \sum \binom{j}{2} f_j$. Note that C is larger than the number of duplicates that is often used in capture-recapture methods [4]. Our task is to estimate N using the sample. Table **??** summarizes the notations used in this paper.

This paper focuses on three basic sampling methods, i.e., RN (random node), RE (random edge), and RW (random walk). In RN sampling, each node is sampled uniformly at random with replacement. In RE sampling, edges are selected with equal probability and two nodes incident to a random edge are collected. Thus, RE sampling is a kind of PPS (probability proportional to size) sampling in that each node is sampled with probability proportional to its degree. RW sampling selects the next node in the current neighbourhood uniformly at random. Its node selection probability is proportional to the degree asymptotically.

2.1 RN Sampling

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

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

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

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

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

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

Thus, the RN estimator for N is

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

2.2 RE Sampling

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

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

The expected number of collisions C is

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

Thus, the RE estimator for N is

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

Thereby, we derived the RE estimator using Γ. The introduction of Γ in the estimator is important–it reveals the difference between the RE and RN estimators, consequently we can compare them. The same estimator in very different forms are used in [4, 11]. Our derivation is different, so that we can compare these two estimators for uniform and PPS samples. Comparing the estimators in equations 5 and 8, the only difference is that RE sampling produces Γ times more collisions using the same sample size. Consequently, the estimate is adjusted by a factor of Γ. When more collisions are observed, the accuracy of the estimation is also improved. Intuitively, RE method can outperform RN sampling by a factor of Γ. In reality, the performance improvement is upper-bounded by Γ as we will show in this paper.

The second issue is whether Γ is large enough to result in significant performance improvement for RE sampling. Our first observation is that when the graph being studied is regular, $\Gamma = 1$ and the RE estimator is reduced to the RN estimator. However, many networks are large and scale-free, inducing very large Γ. For instance, $\Gamma \approx 1300$ for the Twitter user network in the year of 2009 [15]. This large Γ makes the RE sampling the obvious choice.

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

formula [15]:

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

In turn, the average degree can be estimated by the harmonic mean with high accuracy [16]:

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

3 VARIANCES OF THE ESTIMATORS

Estimators are normally evaluated in terms of bias, variance ($var(\widehat{N})$), and the combination of them, i.e., mean squared error (MSE). In [15], we discussed the bias problem, which is rather small in general. This paper focuses on the variances of the two estimators. We do not use Chebyshev's inequality for evaluation as some other papers do, because Chebyshev's inequality gives an upper bound that is valid for any data distribution. Consequently, experimental results can not be explained well using Chebyshev's inequality. We observed that the estimates are of normal distribution [21], thus there is a much tighter bounds. For instance, when relative standard error RSE$\left(= \sqrt{var(\widehat{N})}/N\right)$ is 0.1, the 95% confidence interval is roughly $\widehat{N} \pm 0.2\widehat{N}$. This is the why in our experiments the RSE values are around 0.1.

3.1 Variance of RN Sampling

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.2 Variance of RE Sampling

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

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

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

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

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

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

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

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

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

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

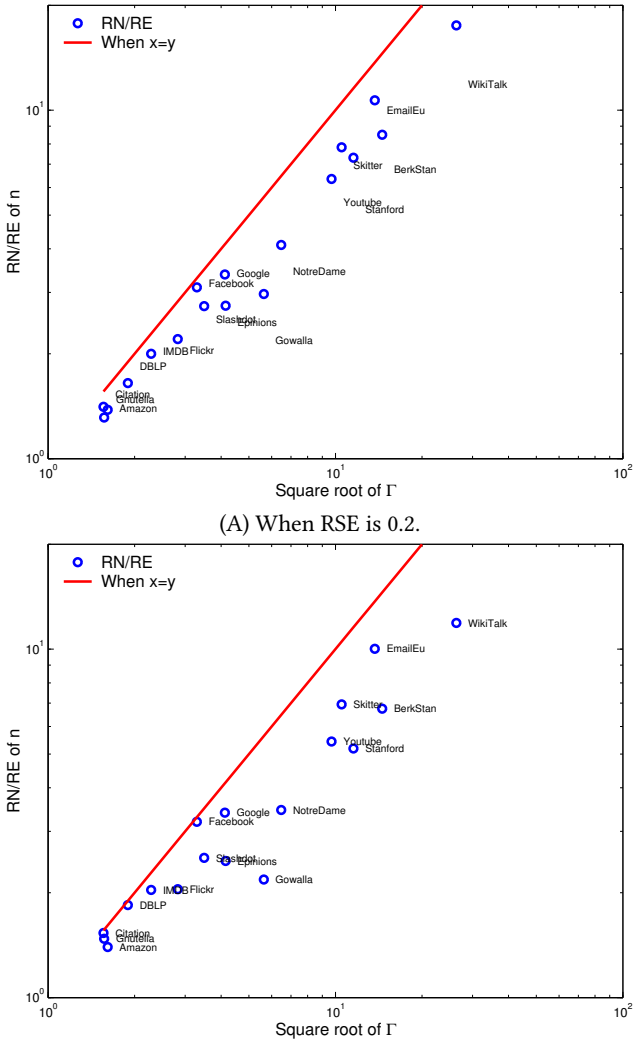
4 EXPERIMENTS ON REAL NETWORKS

We demonstrate our results on 18 datasets.Most of them are from the Stanford SNAP graph collection [13]. 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 networks, we choose only one representative network for each category. In the category of the Web graph datasets, RW sampling deviates greatly from RE sampling. So we include several Web graphs, including the Web graph on the domains of Notre Dame, Stanford, and Berkley-Stanford, to investigate the cause for such deviation. Complete data description and programs can be found at http://cs.uwindsor.ca/~jlu/size,

We compare the sample sizes needed to obtain the same RSE for all the datasets. We show that there is a strong correlation between $\sqrt{\Gamma}$ and RN/RE ratio. Fig. 1 plots the sample size ratio against $\sqrt{\Gamma}$ for the 18 datasets when $RSE = 0.2$ (panel A) and $RSE = 0.1$ (panel B).

The plot shows that 1) RE is better than RN consistently for all the datasets, as all the RN/RE ratio values are greater than one; 2) The ratio has a strong linear correlation with $\sqrt{\Gamma}$ as can be seen visually from the plot, and from the Pearson's correlation coefficient (0.98 when RSE=0.2 and 0.95 when RSE =0.1); 3) The improvement ratio is bounded from above by $\sqrt{\Gamma}$, as all the ratio values are below the line.

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



(A) When RSE is 0.2.

(B) When RSE is 0.1.

Figure 1: RN/RE ratio of sample sizes is bounded from above by $\sqrt{\Gamma}$ for 18 networks. Panel (A) displays the ratio of sample sizes needed to achieve 0.2 RSE; panel (B) the ratios to achieve 0.1 RSE. RSE is obtained over 500 repetitions.

5 DISCUSSIONS AND CONCLUSIONS

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

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

6 ACKNOWLEDGEMENTS

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

REFERENCES

[1] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. *Journal of the ACM*, 55(5):1–74, 2008.
[2] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
[3] A. Broder and et al. Estimating corpus size via queries. In *CIKM*, pages 594–603. ACM, 2006.
[4] A. Chao, S. Lee, and S. Jeng. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. *Biometrics*, 48(1):201–216, 1992.
[5] A. Dasgupta, R. Kumar, and T. Sarlos. On estimating the average degree. In *Proceedings of the 23rd international conference on World wide web*, pages 795–806. International World Wide Web Conferences Steering Committee, 2014.
[6] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. A walk in facebook: Uniform sampling of users in online social networks. *Arxiv preprint arXiv:0906.0060*, 2009.
[7] L. A. Goodman. Some practical techniques in serial number analysis. *Journal of the American Statistical Association*, 49(265):97–112, 1954.
[8] S. J. Hardiman and L. Katzir. Estimating clustering coefficients and size of social networks via random walk. In *Proceedings of the 22nd international conference on World Wide Web*, pages 539–550. ACM, 2013.
[9] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. *Computer Networks*, 33(1-6):295–308, 2000.
[10] H. Huang, N. Zhang, W. Wang, G. Das, and A. Szalay. Just-in-time analytics on large file systems. *Computer*, 61(11):1651–1664, 2012.
[11] L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *WWW*, pages 597–606. ACM, 2011.
[12] S. Lawrence and C. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998.
[13] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *SIGKDD*, pages 631–636. ACM, 2006.
[14] J. Lu and D. Li. Sampling online social networks by random walk. In *ACM SIGKDD Workshop on Hot Topics in Online Social Networks*, pages 33–40. ACM, 2012.
[15] J. Lu and D. Li. Bias correction in small sample from big data. *TKDE, IEEE Transactions on Knowledge and Data Engineering*, 25(11):2658–2663, 2013.
[16] J. Lu and H. Wang. Variance reduction in large graph sampling. *Information Processing and Management*, 50(3):476–491, 2014.
[17] S. Mane, S. Mopuru, K. Mehra, and J. Srivastava. Network size estimation in a peer-to-peer network. *University of Minnesota, MN, Tech. Rep*, pages 05–030, 2005.
[18] M. Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010.
[19] B. Ribeiro and D. Towsley. Estimating and sampling graphs with multidimensional random walks. In *Annual conference on Internet measurement*, pages 390–403. ACM, 2010.
[20] M. Shokouhi, J. Zobel, F. Scholer, and S. M. M. Tahaghoghi. Capturing collection size for distributed non-cooperative retrieval. In *SIGIR*, pages 316–323. ACM, 2006.
[21] Y. Wang, J. Liang, and J. Lu. Discover hidden web properties by random walk on bipartite graph. *Information Retrieval*, 17(3):203–228, 2014.