# Concise Papers

## Bias Correction in a Small Sample from Big Data

Jianguo Lu and Dingding Li

**Abstract**—This paper discusses the bias problem when estimating the population size of big data such as online social networks (OSN) using uniform random sampling and simple random walk. Unlike the traditional estimation problem where the sample size is not very small relative to the data size, in big data, a small sample relative to the data size is already very large and costly to obtain. We point out that when small samples are used, there is a bias that is no longer negligible. This paper shows analytically that the relative bias can be approximated by the reciprocal of the number of collisions; thereby, a bias correction estimator is introduced. The result is further supported by both simulation studies and the real Twitter network that contains 41.7 million nodes.

**Index Terms**—Big data, online social networks, small sample, bias, size estimation

## 1 Introduction

In the era of big data, the size of data is often in the magnitude of billions. Examples of such big data include online social networks (OSN) such as Facebook, pages on the web, the deep web, and the semantic web. Most of the time, the direct access to the entire data is neither possible nor computationally feasible, forcing people to probe the properties of the data by looking at a sample [16]. Because of the huge size of the data, quite often even a sufficient sample is too costly to obtain considering the network traffic involved and daily quota imposed. For practical consideration, we are often limited to the smallest possible sample.

This paper studies the size estimation using simple random walk when the sample size is limited due to the high cost of sampling. We choose simple random walk sampling because it is supported by most OSN interfaces [13], [11], [25], and it is more efficient compared with uniform random samples achieved by rejection samplings or Metropolis-Hasting sampling [21].

The basic idea of population size estimation is based on the collisions during a random walk or repeated samplings. It is rooted in a classical birthday-paradox problem, in a capture-recapture method developed in ecology [1], and in the Erdos random graph [9]. In terms of random walk sampling on a network, a node can be visited multiple times during a random walk. When each node has an equal probability of being visited, a collision occurs when the sample size is in the order of $O(\sqrt{2N})$ (see (14)), where $N$ is the total population size. Even when a sample is large in its number, the collisions can be rather small when the data are big. If the number of collisions is barely above 1, we call the sample a small one relative to the data.

- *J. Lu is with the School of Computer Science, University of Windsor, 401 Sunset Avenue Windsor, Ontario N9B 3P4, Canada. E-mail: jlu@uwindsor.ca.*
- *D. Li is with the Economics Department, University of Windsor, Chrysler Hall North, 401 Sunset Avenue Windsor, Ontario N9B 3P4, Canada. E-mail: dli@uwindsor.ca.*

*Manuscript received 10 July 2012; revised 18 Oct. 2012 accepted 19 Oct. 2012; published online 7 Nov. 2012.*
*Recommended for acceptance by D. Agrawal.*
*For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2012-07-0485.*
*Digital Object Identifier no. 10.1109/TKDE.2012.220.*

For instance, given a network comprised of one million nodes, we need to visit around 4500 nodes before, on average, 10 collisions can occur. The number of the collisions lies mostly between 3 and 17 according to its 95 percent confidence interval. Considering each node visit requires multiple remote calls to the server over the network, the cost of obtaining this sample is rather high. Yet, the collisions can be close to zero. Relative to the size of the total population, this is a small sample.

When only a small sample is affordable, we need to utilize what we have to give the best estimation. One thing often overlooked is that there is a bias in the estimators used in the literature, and the bias is rather large when the sample size is small. Continuing our previous example, the small sample can induce a bias as large as 10 percent.

This paper is based on the following estimator $\widehat{N}$ that is given in [20], and also can be derived from [3], [13]

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

where $n$ is the sample size, $\gamma$ is the coefficient of variation of the degrees of the network, and $C$ is the number of collisions. We show that $\widehat{N}$ is biased upward and its relative bias, the bias normalized by the population size, can be approximated by the reciprocal of the expectation of $C$. Based on this, we derived the bias correction estimator $\widehat{N}^*$ as

$$\widehat{N}^* = (\gamma^2 + 1)\binom{n}{2}\frac{1}{C+1}. \tag{2}$$

This result is demonstrated by simulation studies and supported by real Twitter data.

## 2 Related Work

Population size estimation has been widely studied in ecology [3] and social studies [23], and more recently in computer science for estimating the size of the web [15], databases [12], web data sources [19], [18], [8], [7], [27], [2], and online social networks [20], [13], [11], [25].

The starting point of population estimation is the well-known Lincoln-Petersen estimator [1] that can be applied when there are two sampling occasions and every node has equal probability of being sampled:

$$\hat{N}_{LP} = \frac{n_1 n_2}{d}, \tag{3}$$

where $n_1$ is the number of nodes sampled in the first capture occasion, $n_2$ is the number of nodes sampled in the second occasion, and $d$ is the duplicate among those two samples. For the Lincoln-Petersen estimator, the bias correction has been addressed by Chapman [4], [24] by suggesting the following Chapman estimator:

$$\hat{N}_{Chap} = \frac{(n_1+1)(n_2+1)}{d+1} - 1. \tag{4}$$

The derivation is based on the hypergeometric distribution of the repeated elements since the Lincoln-Peterson estimator assumes the sampling without replacement, which is different from the sampling with replacement assumed by the $\widehat{N}$ estimator.

The assumptions of the Lincoln-Petersen estimator can be hardly met in reality. It is extended in two dimensions: one is allowing multiple sampling occasions, and the other is supporting heterogeneity in capture probability.

When there are more than two sampling occasions and each time only one sample is taken, Darroch [6] derived that the approximate maximum likelihood estimator (MLE), $\widehat{N}_D$, is the solution of the following equation:

$$u = N\left(1 - e^{-\frac{n}{N}}\right), \tag{5}$$

where $n$ is the total sample size and $u = n - d$ is the number of unique data items that have been sampled.

This equation has also been used to predict the isolated nodes in a random graph when edges are randomly added [22]. Since it does not have a simple closed form solution [22], [6], its bias correction is not discussed in the literature. In OSN studies, Ye and Wu [25] used the numeric method to find the solution to this estimator. Lu and Li [19] gave an approximate solution to (5) as follows:

$$P = OR^{-2.1}, \tag{6}$$

where $P = 1 - u/N$ is the percentage of the data not being sampled yet, $OR = n/u$ is the overlapping rate between the total sample size and unique data items being sampled. Intuitively, there is a power law governing the fraction of the unsampled data, and it is solely dependent on the overlapping rate.

When the data are heterogeneous and the capture occasions are more than 2, the estimation is notoriously difficult, mainly due to the lack of knowledge of $\gamma$. Therefore, (1) as an estimator for $N$ was not seen in ecology, let alone the correction of bias. Instead, the same equation was used by Chao et al. [3] in reverse way to estimate $\gamma$ as follows:

$$\widehat{\gamma^2} = N_0 C \binom{n}{2}^{-1} - 1, \tag{7}$$

where $N_0$ is a bootstrapped estimation for $N$ by another estimator.

Note that $\widehat{N}$ can be approximated by (5) when $\gamma = 0$ and the sample size is small. It follows by applying Taylor expansion on the right-hand side of (5), and approximating duplicates $d$ by collisions $C$.

The bias correction in this paper reminds us of the legendary Good-Turing smoothing [10] in word frequency estimation and Enigma code breaking. In particular, among a string of adjusted estimators there is an add-one smoothing [26] that looks related to our method. But these two methods are different in that we are adjusting the bias, while their methods try to save the probability space to account for unseen word types.

## 3 Preliminaries

Given a graph of $N$ nodes labeled as $(1, 2, \ldots, N)$. A sample of the nodes $(x_1, x_2, \ldots, x_n)$, $x_i \in \{1, \ldots, N\}$, is taken by a simple random walk on the graph, where node $x_{i+1}$ is selected randomly from the neighbors of the proceeding node $x_i$. In addition to the node ids, we assume that their corresponding degrees $(d_{x_1}, d_{x_2}, \ldots, d_{x_n})$ are also obtained. Our task is to estimate $N$ based on the sample.

Depending on the sampling scheme, the probability of a node being included in a sample may not be equal. In simple random walk sampling, a node with a larger degree will have higher probability of being sampled. The sampling probability $p_i$ of node $i$ is asymptotically proportional to its degree $d_i$ [17], i.e.,

$$p_i = \frac{d_i}{\tau}, \tag{8}$$

where $\tau = \sum_{i=1}^{N} d_i = N\langle d \rangle$.

The heterogeneity of the sampling probability or the node degrees can be measured by the coefficient of variation (CV, denoted as $\gamma$ hereafter), which is defined as the normalized standard deviation of the degrees:

$$\gamma^2 = \frac{\mathrm{var(d)}}{\langle d \rangle^2} = \frac{\langle d^2 \rangle}{\langle d \rangle^2} - 1. \tag{9}$$

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

$$E(C) = \binom{n}{2} p, \tag{10}$$

and its variance is

$$var(C) = \binom{n}{2} p(1-p) = E(C)(1-p). \tag{11}$$

The collision probability $p$ can be translated into the heterogeneity of the data measured by $\gamma$ using (8) and (9):

$$p = \sum_{i=1}^{N} p_i^2 = \frac{1}{\tau^2} \sum_{i=1}^{N} d_i^2 = \frac{\langle d^2 \rangle}{N \langle d \rangle^2} = \frac{\gamma^2 + 1}{N}. \tag{12}$$

Combining (12) and (10), we obtain the expected mean of collisions as follows:

$$E(C) = \binom{n}{2} \frac{\gamma^2 + 1}{N}. \tag{13}$$

When every node in the network has the same probability of being visited, $\gamma = 0$ and $p = p_i = 1/N$, the above formulation is reduced to the well-known birthday-paradox problem where

$$E(C) = \binom{n}{2} \frac{1}{N} \approx \frac{n^2}{2N}. \tag{14}$$

In other words, on average $\sqrt{2N}$ number of samples are needed to produce a collision.

In the case of big data, the variance can be simplified further. Given a network with a fixed $\gamma$, $p$ tends to zero when $N$ tends to infinity according to (12). It follows from (11) that

$$\lim_{N \to \infty} var(C) = E(C). \tag{15}$$

## 4 The Estimators

### 4.1 The Biased Estimator

From (13), the population size can be described by

$$N = (\gamma^2 + 1) \binom{n}{2} \frac{1}{E(C)}. \tag{16}$$

Since $E(C)$ is unknown, it can be estimated by the observed collisions $C$. This gives us the estimator

$$\widehat{N} = (\gamma^2 + 1) \binom{n}{2} \frac{1}{C}, \tag{17}$$

where $C$ is calculated as follows: Let $f_i$ denote the number of individuals that are visited exactly $i$ times, $C = \sum_{i=1}^{+\infty} \binom{i}{2} f_i$. Note that $C$ can be larger than the number of duplicate visits $d = \sum_{i=1}^{+\infty} (i-1) f_i$, especially when the sample size is large.

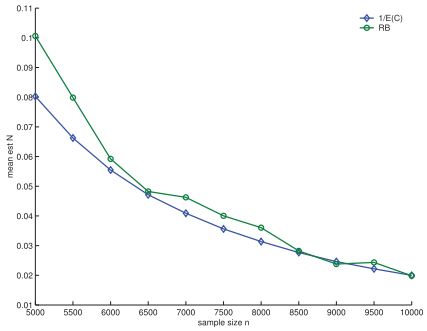


Fig. 1. RB and 1/E(C) against sample sizes in simulation study. It shows that $\widehat{N}$ is biased upward, and the relative bias can be approximated by the reciprocal of E(C).

Estimator $\widehat{N}$ is biased. The expected value of the estimator is

$$E(\widehat{N}) = E\left[(\gamma^2+1)\binom{n}{2}\frac{1}{C}\right] = (\gamma^2+1)\binom{n}{2}E\left(\frac{1}{C}\right). \tag{18}$$

Comparing (16) and (18), the only difference is between $1/E(C)$ and $E(1/C)$. It is well known [5] that the expectation of the reciprocal of a random variable is greater than the reciprocal of its expectation, if the random variable is nondegenerate and positive, i.e.,

$$E\left(\frac{1}{C}\right) > \frac{1}{E(C)}. \tag{19}$$

In other words, $\widehat{N}$ has a positive bias. What we need to know is exactly how large is the bias, or what is the relative bias ($RB$) of $\widehat{N}$ that is defined as follows:

$$RB = \frac{E(\widehat{N})-N}{N} = \frac{E(\frac{1}{C})-\frac{1}{\mu}}{\frac{1}{\mu}}, \tag{20}$$

where we use $\mu$ to denote $E(C)$ so that the deduction in the following is more succinct.

### 4.2 Bias Correction

The expected value of $1/C$ can be derived using the Taylor expansion of $1/C$ around $\mu$ as follows:

$$\frac{1}{C} = \frac{1}{\mu} - \frac{C-\mu}{\mu^2} + \frac{2}{\mu^3}\frac{(C-\mu)^2}{2!} - \frac{6}{\mu^4}\frac{(C-\mu)^3}{3!}\cdots.$$

Applying linearity of expectation, the expected value of $1/C$ is

$$E\left(\frac{1}{C}\right) = \frac{1}{\mu} - \frac{E(C)-\mu}{\mu^2} + \frac{E(C-\mu)^2}{\mu^3} - \frac{E(C-\mu)^3}{\mu^4}\cdots.$$

Note that the second-central moment is the variance, and the third-central moment $E(C-\mu)^3$ is

$$\binom{n}{2}p(1-p)(1-2p) \approx \binom{n}{2}p \approx var(C). \tag{21}$$

Thus by (15),

$$E\left(\frac{1}{C}\right) \approx \frac{1}{\mu} + \frac{var(C)}{\mu^3} - \frac{var(C)}{\mu^4} + \cdots = \frac{1}{\mu}\left(1 + \frac{1}{\mu} - \frac{1}{\mu^2}\right) \tag{22}$$

TABLE 1
Bias in Simulation Studies

| n ($\times 10^3$) | E(C) | $1/E(C)$ (%) | RB (%) $\widehat{N}$ | RB (%) $\widehat{N}^*$ |
|---|---|---|---|---|
| 5.0 | 12.4599 | 8.0257 | 10.0625 | 0.2753 |
| 5.5 | 15.0968 | 6.6239 | 7.9858 | 0.2244 |
| 6.0 | 18.0315 | 5.5459 | 5.9218 | -0.3300 |
| 6.5 | 21.2193 | 4.7127 | 4.8240 | -0.4025 |
| 7.0 | 24.4469 | 4.0905 | 4.6238 | 0.1457 |
| 7.5 | 28.0729 | 3.5622 | 4.0035 | 0.1524 |
| 8.0 | 31.8902 | 3.1358 | 3.6039 | 0.2486 |
| 8.5 | 36.1460 | 2.7666 | 2.8205 | -0.1075 |
| 9.0 | 40.6068 | 2.4626 | 2.3789 | -0.2132 |
| 9.5 | 45.0772 | 2.2184 | 2.4341 | 0.1072 |
| 10.0 | 50.0428 | 1.9983 | 1.9841 | -0.0968 |

$N = 10^6$. *Sample size* $n$ *is between* 5000 *and* $10^4$. *Repeated* $10^4$ *times.*

Substituting (22) for E(1/C) in (20), we derive the following theorem:

**Theorem 1.** *The relative bias of* $\widehat{N}$ *can be approximated by the reciprocal of* $E(C)$, *i.e.,*

$$RB = \frac{1}{E(C)} + \mathcal{O}\left(\frac{1}{E(C)^2}\right) \approx \frac{1}{E(C)}. \tag{23}$$

Fig. 1 depicts the relative bias against the sample size, when $N = 10^6, \gamma = 0$, and $n$ takes the ranges between 5000 and $10^4$. For each sample size, the experiment is repeated $10^4$ times. $RB$ and $E(C)$ are approximated from the $10^4$ experiments. It shows that $\widehat{N}$ has a positive bias, which tapers off as the sample size grows. Its relative bias agrees with the reciprocal of $E(C)$, especially when $E(C)$ is large. When $E(C)$ is small, we can see that $RB$ is greater than $1/E(C)$ as indicated in (23).

From the relative bias, we can derive the adjusted estimator if we replace $\mu$ by $C$:

$$\widehat{N}^* = \frac{\widehat{N}}{1+RB} \quad \text{(by (20))}$$

$$= (\gamma^2+1)\binom{n}{2}\frac{1}{C}\frac{1}{1+\frac{1}{\mu}} \quad \text{(by (23))}$$

$$= (\gamma^2+1)\binom{n}{2}\frac{1}{C+1}. \tag{24}$$

### 4.3 Illustrative Example

We use a fictitious example to gain intuitive understanding of the bias of $\widehat{N}$ and the adjusted estimator $\widehat{N}^*$ . Suppose that the expected value for collisions is $E(C) = 10$. Let $A = (\gamma^2+1)\binom{n}{2}$, and the true size of population is $N = A/E(C) = 0.1A$. The expected standard deviation of C is $\sqrt{10} \approx 3.3$. Suppose that we carried out three experiments, and observed three values for collisions which are 6, 10, and 14. Notice that their mean is exactly 10, indicating that the sampling is unbiased. The mean of $\widehat{N}$ is

$$\langle\widehat{N}\rangle = \frac{A}{3}\sum_{i=1}^{3}\frac{1}{C_i} = \frac{A}{3}\left(\frac{1}{6}+\frac{1}{10}+\frac{1}{14}\right) = 0.1127A.$$

Notice that there is a positive bias even though the observed collisions are unbiased. On the other hand, the mean of the adjusted estimates $\widehat{N}^*$ is

$$\langle\widehat{N}^*\rangle = \frac{A}{3}\sum_{i=1}^{3}\frac{1}{C_i+1} = \frac{A}{3}\left(\frac{1}{7}+\frac{1}{11}+\frac{1}{15}\right) \approx 0.1001A,$$

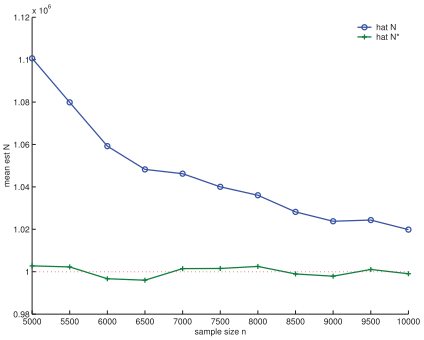


Fig. 2. $\widehat{N}$ and $\widehat{N}^*$ over $10^4$ runs for various sample sizes in simulation study. The red dotted line is the true value.

which is much closer to the real value. The relative biases of these two estimators are 11.27 percent for $\widehat{N}$ and 0.14 percent for $\widehat{N}^*$.

### 4.4 Simulation Studies

Before evaluating the estimators $\widehat{N}$ and $\widehat{N}^*$ in real random walk, we first conduct simulation studies where elements are selected randomly with a uniform distribution, i.e., every element has the same probability of being selected. Thus, $\gamma = 0$ in (1) and (24).

In our experiment, the total population is $N = 10^6$. Sample sizes tested are between 5,000 and $10^4$. The minimal sample size is set as 5,000 to guarantee the existence of at least one collision for every test. For each sample size, $10^4$ tests are run, and relative biases (RB) for two estimators are calculated from these $10^4$ tests.

Table 1 gives an overview of the experiments. It shows that indeed $\widehat{N}$ is biased upward, especially when the sample size is small. When $n = 5{,}000$, the collision mean is around 12, resulting in a high bias ($\mathrm{RB} = 0.10$).

Fig. 2 depicts the trends of the $\widehat{N}$ and $\widehat{N}^*$ with the growth of the sample size. It shows that $\widehat{N}^*$ fluctuates around the true value, while $\widehat{N}$ has a large bias when sample size is small. When the sample size is 5,000, on average among $10^4$ runs there are about 12 collisions, and the relative bias is around 10 percent.

Fig. 3 shows the distributions of the estimations when the sample sizes are 5,000, 5,500, 6,000, and 6,500 in sub-figures A, B, C, and D, respectively. In all the four sub-figures, we can see that $\widehat{N}^*$ has more concentration around the true value. In particular, it has a smaller number of very large estimations. For instance, in Fig. 3A there are more than 200 estimations of $\widehat{N}$ that are higher than 2 million, while $\widehat{N}^*$ has much smaller number of large estimations. With the growth of the sample size, the difference between $\widehat{N}$ and $\widehat{N}^*$ diminishes.

## 5 Random Walk on Twitter Data

We tested estimators $\widehat{N}$ and $\widehat{N}^*$ on the Twitter network data that are provided by Kwak et al. [14], characterizing the complete Twitter network as of July 2009. The data contain about 1.47 billion edges and 41.7 million nodes or users, occupying around 20 gigabytes hard drive space. Since they are too large to fit into the memory of commodity computers, we index them using Lucene, a popular index engine. Then the random walk sampling is performed on the index that is stored in the hard drive. Since random walk works better in an undirected graph, we remove the direction in Twitter data. Note that the Twitter graph is almost surely connected because its average degree is 70, far greater than the threshold value $ln(N) = ln(41,700,000) \approx 17$ for a graph to be connected [22]. The Matlab program and data are available at http://cs.uwindsor.ca/~jlu/bias.

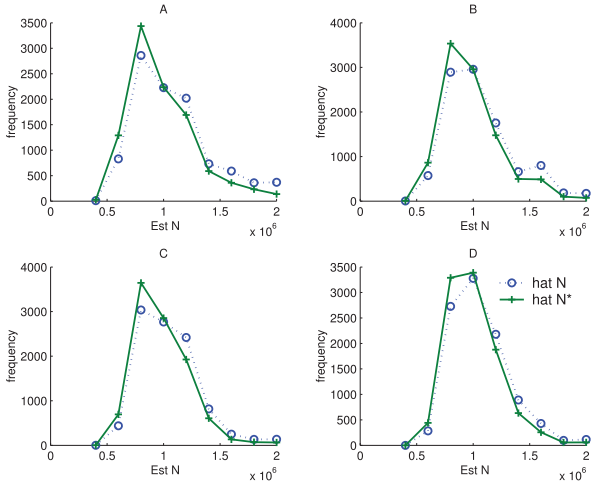


Fig. 3. Distribution of the estimations by $\widehat{N}$ and $\widehat{N}^*$ in simulation study, when $\mathrm{n} =$ 5,000, 5,500, 6,000, and 6,500 in sub-figures A, B, C, and D, respectively.

### 5.1 Estimate $\gamma$

Unlike the simulation studies presented in the last section, where $\gamma = 0$, in real network the node degree varies and we need to estimate $\gamma$. In the area of capture-recapture research [3], [19], it has been a perplexing problem for the population estimation of heterogeneous data whose capture probabilities are unequal, mainly due to the difficulty of estimating the heterogeneity.

Let $d_{x_i}$ be the degree of the node $x_i$ being sampled, where $i = 1, 2 \ldots, n$. The asymptotic mean of the degrees obtained by a random walk is

$$\langle d_x \rangle = \sum_{i=1}^{N} p_i d_i = \frac{\langle d^2 \rangle}{\langle d \rangle}, \tag{25}$$

which can be estimated by its sample mean:

$$\widehat{\langle d_x \rangle} = \frac{1}{n} \sum_{i=1}^{n} d_{x_i}. \tag{26}$$

The population mean of the degrees can be estimated by the harmonic mean of the sample degrees [23], [20]

$$\widehat{\langle d \rangle} = \frac{n}{\sum_{i=1}^{n} 1/d_{x_i}}. \tag{27}$$

According to (9), we have

$$\gamma^2 + 1 = \frac{\langle d^2 \rangle}{\langle d \rangle^2} = \frac{\langle d_x \rangle}{\langle d \rangle}. \tag{28}$$

Hence, the estimator for $\gamma^2$ is

$$\widehat{\gamma^2} + 1 = \frac{1}{n^2} \sum_{i=1}^{n} d_{x_i} \sum_{i=1}^{n} 1/d_{x_i}. \tag{29}$$

### 5.2 Results

In our experiments, the sample size ranges between 400 and 3,600. The smallest sample size is set as 400 so that it can induce at least one multiple visit to a node. Although the true population is rather large ($4.17 \times 10^7$), we do not need 5,000 samples as in the case of random simulation because of the heterogeneity of the degrees.

For each sample size, we run 500 random walks. Since both estimators $\widehat{N}$ and $\widehat{N}^*$ rely on collisions very much, extra caution should be taken to avoid spurious collisions caused by random walk. For instance, if a node A is only connected to node B, a visit

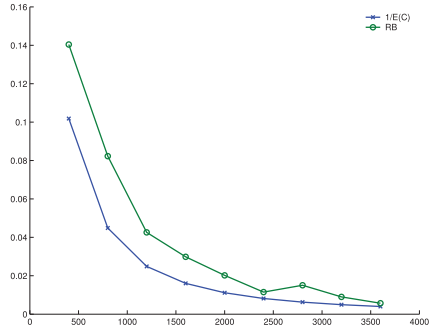


Fig. 4. Relative bias of $\widehat{N}$ in Twitter data for various sample sizes, and its comparison with 1/E(C).

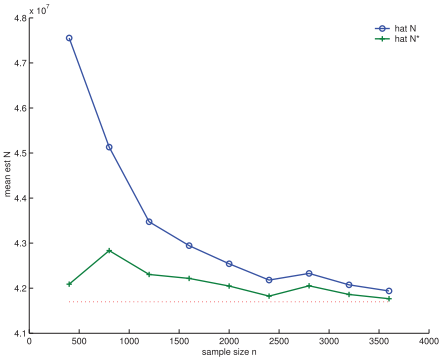


Fig. 5. $\widehat{N}$ and $\widehat{N}^*$ in Twitter for various sample sizes. The red dotted line is the true value.

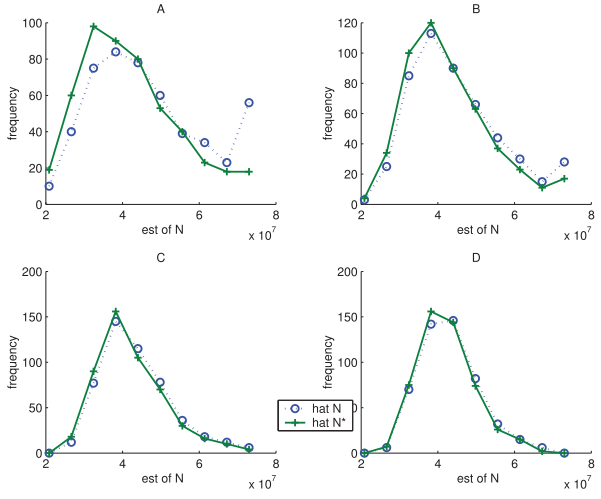


Fig. 6. Distributions of 500 estimations for Twitter data when sample sizes are 400, 800, 1,200, and 1,600 in sub-figures A, B, C, and D, respectively.

to A will cause node B visited twice. To avoid such loops, we take samples spaced every few steps apart.

Overall, the results conform well to our simulation studies. Fig. 4 shows that the relative bias of $\widehat{N}$ is close to the reciprocal of E(C) for various sample sizes. Consequently, $\widehat{N}^*$ corrects the bias quite well as shown in Fig. 5. It is clear that the bias diminishes as the sample size grows. Fig. 6 depicts the distribution of the estimations for the four smallest sample sizes. Table 2 summarizes the details of the results.

## 6 Conclusions

Estimators are usually evaluated by both bias and variance. The purpose of this paper is not to evaluate the overall performance of the estimator. Instead, it shows that there is a bias, and the bias of $\widehat{N}$ can be too large to neglect when sample size is small relative to the big data being studied. We derive the bias of the estimator $\widehat{N}$, and empirically demonstrate the result using simulations and real Twitter data. The derivation is based on the unique formulation of $\widehat{N}$ presented in this paper. Although $\widehat{N}$ in other forms were already given in [3], [13], we are the first to explicitly describe the estimator in terms of collisions $C$ and the coefficient of variance $\gamma$. It is this formulation that leads to the derivation of the bias.

Traditionally, $\widehat{N}$ is not widely used because it needs the estimation of $\gamma$, which is also a treacherous problem. However, in the unique setting of online data, the degrees of the sampled nodes are often available, whereas in social studies the friends of a drug addict are hardly collectable. Taking this advantage in OSN sampling, we can estimate correctly the average degree, and thereby the coefficient of variation $\gamma$.

Our bias correction formula works for both uniform random sampling and random walk sampling. The bias is dependent on the expected number of collisions. For random walk sampling where $\gamma$ is large, the sample size does not need to be very large to induce lots of collisions. Thus, the bias problem is not so prominent as illustrated in Twitter data. In uniform random sampling, the sample has to be much larger to cause the same number of collisions. Our experiments show that for a data of $10^6$ nodes, there is 10 percent of over estimation even when the sample size is as large as 5,000.

TABLE 2
Bias in Twitter Data

| n ($\times 100$) | E(C) | 1/E(C) (%) | RB (%) $\widehat{N}$ | RB (%) $\widehat{N}^*$ |
|---|---|---|---|---|
| 4 | 9.8186 | 10.1847 | 14.0388 | 0.9343 |
| 12 | 40.2164 | 2.4865 | 4.2570 | 1.4510 |
| 20 | 89.6493 | 1.1155 | 2.0186 | 0.8328 |
| 28 | 159.2846 | 0.6278 | 1.5061 | 0.8479 |
| 36 | 249.3307 | 0.4011 | 0.5709 | 0.1576 |

$N = 4.17 \times 10^7$.

## Acknowledgments

The authors thank the support from Natural Sciences and Engineering Research Council of Canada (NSERC) and Social Sciences and Humanities Research Council of Canada (SSHRC).

## References

[1] S. Amstrup, T. McDonald, and B. Manly, *Handbook of Capture-Recapture Analysis*. Princeton Univ Press, 2005.
[2] A. Broder et al., "Estimating Corpus Size via Queries," *Proc. 15th ACM Int'l Conf. Information and Knowledge Management (CIKM)*, pp. 594-603, 2006.
[3] A. Chao, S. Lee, and S. Jeng, "Estimating Population Size for Capture-Recapture Data When Capture Probabilities Vary by Time and Individual Animal," *Biometrics*, vol. 48, pp. 201-216, 1992.
[4] D. Chapman, "Some Properties of the Hypergeometric Distribution with Applications to Zoological Sample Censuses," *Univ. California Publications Statistics*, vol. 1, pp. 131-59, 1951.
[5] C.L. Chiang, "On the Expectation of the Reciprocal of a Random Variable," *The Am. Statistician*, vol. 20, no. 4, p. 28, 1966.
[6] J. Darroch, "The Multiple-Recapture Census: I. Estimation of a Closed Population," *Biometrika*, vol. 45, nos. 3/4, pp. 343-359, 1958.
[7] A. Dasgupta, G. Das, and H. Mannila, "A Random Walk Approach to Sampling Hidden Databases," *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD)*, pp. 629-640, 2007.
[8] A. Dasgupta, X. Jin, B. Jewell, N. Zhang, and G. Das, "Unbiased Estimation of Size and Other Aggregates over Hidden Web Databases," *Proc. ACM Int'l Conf. Management of Data (SIGMOD)*, pp. 855-866, 2010.
[9] P. Erdos and A. Rényi, "On the Evolution of Random Graphs," *Publication of Math. Inst. of Hungarian Academy of Sciences*, vol. 5, pp. 17-61, 1960.
[10] W. Gale and G. Sampson, "Good-Turing Frequency Estimation without Tears," *J. Quantitative Linguistics*, vol. 2, no. 3, pp. 217-237, 1995.
[11] 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.

[12] P.J. Haas, J.F. Naughton, S. Seshadri, and L. Stokes, "Sampling-Based Estimation of the Number of Distinct Values of an Attribute," *Proc. 21st Int'l Conf. Very Large Data Bases (VLDB)*, pp. 311-322, 1995.
[13] L. Katzir, E. Liberty, and O. Somekh, "Estimating Sizes of Social Networks via Biased Sampling," *Proc. 20th Int'l Conf. World Wide Web (WWW)*, pp. 597-606, 2011.
[14] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?," *Proc. 19th Int'l Conf. World Wide Web (WWW)*, pp. 591-600, 2010.
[15] S. Lawrence and C. Giles, "Searching the World Wide Web," *Science*, vol. 280, no. 5360, pp. 98-100, 1998.
[16] J. Leskovec and C. Faloutsos, "Sampling from Large Graphs," *Proc. 12th ACM Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD)*, pp. 631-636, 2006.
[17] L. Lovász, "Random Walks on Graphs: A Survey," *Combinatorics, Paul Erdos is Eighty*, vol. 2, no. 1, pp. 1-46, 1993.
[18] J. Lu, "Ranking Bias in Deep Web Size Estimation Using Capture Recapture Method," *Data and Knowledge Eng.*, vol. 69, no. 8, pp. 866-879, 2010.
[19] J. Lu and D. Li, "Estimating Deep Web Data Source Size by Capture-Recapture Method," *Information Retrieval*, vol. 13, no. 1, pp. 70-95, 2010.
[20] J. Lu and D. Li., "Sampling Online Social Networks by Random Walk," *Proc. First ACM Int'l Workshop Hot Topics on Interdisciplinary Social Networks Research (SIGKDD)*, pp. 33-40, 2012.
[21] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *The J. Chemical Physics*, vol. 21, no. 6, pp. 1087-1092, 1953.
[22] M. Newman, *Networks: An Introduction*. Oxford Univ. Press, Inc., 2010.
[23] M. Salganik and D. Heckathorn, "Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling," *Sociological Methodology*, vol. 34, no. 1, pp. 193-240, 2004.
[24] J. Wittes, "On the Bias and Estimated Variance of Chapman's Two-Sample Capture-Recapture Population Estimate," *Biometrics*, vol. 28, pp. 592-597, 1972.
[25] S. Ye and S. Wu, "Estimating the Size of Online Social Networks," *Int'l J. Social Computing and Cyber-Physical Systems*, vol. 1, no. 2, pp. 160-179, 2011.
[26] C. Zhai and J. Lafferty, "A Study of Smoothing Methods for Language Models Applied to Information Retrieval," *ACM Trans. Information Systems*, vol. 22, no. 2, pp. 179-214, 2004.
[27] J. Zhou, Y. Li, V. Adhikari, and Z. Zhang, "Counting Youtube Videos via Random Prefix Sampling," *Proc. ACM SIGCOMM*, pp. 371-380, 2011.