2013

Bias Correction in Small Sample from Big Data

Jianguo Lu
University of Windsor

Dingding Li

Bias Correction in Small Sample from Big Data

Jianguo Lu [1], Dingding Li [2]
[1] School of Computer Science, University of Windsor
[2] Department of Economics, University of Windsor
Email: {jlu,dli}@uwindsor.ca
401 Sunset Avenue, Windsor, Ontario N9B 3P4. Canada

***Abstract*—This paper discusses the bias problem when estimating the population size of big data such as online social networks (OSN) using 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. When small samples are used, there is a bias that is no longer negligible. This paper shows analitically 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**

I. 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 [15]. Because of the huge size of the data, quite often even a sufficient sample is too big to obtain. For practical consideration, we are often limited to the smallest possible sample.

This paper studies the size estimation using simple random walk when 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 [12] [10] [23], and it is more efficient compared with uniform random samples achieved by rejection samplings or Metropolis-Hasting sampling [19].

The basic idea of population size estimation is based on the collisions during a random walk or repeated samplings. It is rooted in classical birthday paradox problem, in capture-recapture method developed in ecology [1], and in Erdos random graph [8]. 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 equation 14), where N is the total population size. We call a sample is small if the number of collisions is barely above one.

For instance, giving 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% 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 a best estimation. One thing often overlooked is that there is a bias in the estimators used in 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%.

This paper is based on the following estimator $\widehat{N}$ that can be derived from [3] [12]:

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

where n is the sample size, γ 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 upwards 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.

II. Related work

Population size estimation has been widely studied in ecology [3] and social studies [21], and more recently in computer science for estimating the size of the web [14], databases [11], web data sources [7] [25] [2], and online social networks [12] [10] [23].

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, d is the duplicates among those two samples.

For Lincoln-Petersen estimator the bias correction has been addressed by Chapman [4] [22] 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 Lincoln-Peterson estimator assumes the sampling without replacement, which is different from the sampling with replacement assumed by $\widehat{N}$ estimator.

The assumptions of Lincoln-Petersen estimator can be hardly met in reality. It is extended in two dimensions: one is allowing multiple sampling occasions, 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:

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

where n is the total sample size, and d is the duplications. This equation has also been used to predict the isolated nodes in random graph when edges are randomly added [20]. Since it does not have a simple closed form solution [20] [6], its bias correction is not discussed in literature. In OSN studies, [23] used numeric method to find the solution to this estimator.

When the data is heterogeneous and the capture occasions are more than two, the estimation is notoriously difficult, mainly due to the lack of knowledge of γ. Therefore Equation 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 γ as below:

$$\widehat{\gamma^2} = N_0 C \binom{n}{2}^{-1} - 1 \tag{6}$$

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

In the estimation of digitalized networks such as OSN, the sampling probability for each node can be (partially) decided by the degrees. Unlike traditional sampling schemes where sampling probability of animals are different but the exact variance is impossible to quantify accurately, in OSN simple random walk sampling we know not only the exact degree of the node being visited, but also that the sampling probability is proportional to its degree. With this knowledge, we can obtain the value of γ, thereby estimator $\widehat{N}$ can be applied. Not surprisingly Katzir et al. [12] used a similar equation to estimate the size of OSNs:

$$\widehat{N_K} = \frac{1}{2C}\sum_{i=1}^{n} d_{x_i} \sum_{i=1}^{n} 1/d_{x_i} \tag{7}$$

which can be transformed to $\widehat{N}$ as we will show in Section V-A (eq 29). [12] showed that it is a consistent estimator, but does not mention the bias problem.

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

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. The resulting estimators can be unbiased. For instance, [10] 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 valid ones; [25] takes advantage of the prefix encoding of Youtube links; [7] depends on the negation of queries to break down the search results. Compared to this group of estimation methods, our approach is rather generic, in that it works on any access interfaces as long as the interface supports simple random walk.

The bias correction in this paper reminds us the legendary Good-Turing smoothing [9] in word frequency estimation and Enigma code breaking. In particular amongst a string of adjusted estimators there is an add-one smoothing [24] 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.

III. 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 neighbours 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 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 [16], 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 Coefficient of Variation (CV, denoted as γ hereafter), which is defined as the normalized standard deviation of the degrees:

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

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 γ using equations 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 equations 12 and 10 we obtain the expected mean of collisions as below:

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

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 another word, 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 γ, p tends to zero when N tends to infinity according to equation 12. It follows from equation 11 that:

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

IV. The estimators

A. The biased estimator

From Equation 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 sample size is large.

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

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

Comparing equations 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 non-degenerate 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 μ to denote $E(C)$ so that the deduction in the following is more succinct.

B. Bias correction

The expected value of $1/C$ can be derived using Taylor expansion of $1/C$ around μ as below:

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

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, 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 Equation 15

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

Substituting Equation 22 for E(1/C) in Equation 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.,

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

Figure 1 depicts the relative bias against sample size, when $N = 10^6, \gamma = 0$, 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 equation 23.

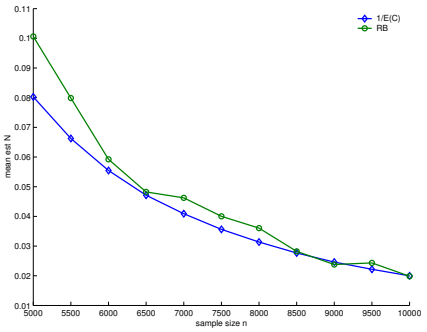

Fig. 1. RB and 1/E(C) against sample sizes in simulation study. It shows that $\widehat{N}$ is biased upwards, and the relative bias can be approximated by the reciprocal of E(C).

From the relative bias, we can derive the adjusted estimator if we replace μ by C:

$$\begin{aligned}\widehat{N}^* =& \frac{\widehat{N}}{1+RB} && \text{(by Eq 20)}\\ =&(\gamma^2+1)\binom{n}{2}\frac{1}{C}\frac{1}{1+\frac{1}{\mu}} && \text{(by Eq 23)}\\ =&(\gamma^2+1)\binom{n}{2}\frac{1}{C+1} && (24)\end{aligned}$$

C. 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, observed three values for collisions which are 6, 10, 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

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

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

D. 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 uniform distribution, i.e., every element has the same probability being selected. Thus $\gamma = 0$ in Equations 1 and 24.

TABLE I
BIAS IN SIMULATION STUDIES. $N = 10^6$. SAMPLE SIZE n IS BETWEEN 5000 AND 10^4. REPEATED 10^4 TIMES.

n ($\times 10^3$)	E(C)	$1/E(C)$ (%)	RB (%) $\hat{N}$	$\hat{N}^*$
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

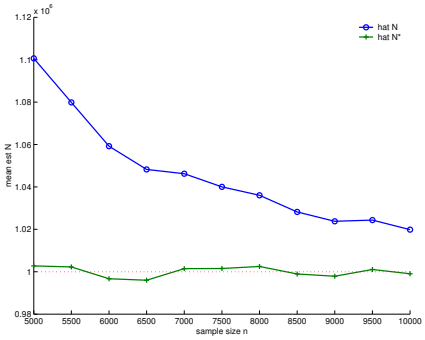

Fig. 2. $\widehat{N}$ and $\widehat{N}^*$ over 10^4 runs for various sample size in simulation study. Red dotted line is the true value.

In our experiment the total population is $N = 10^6$. Sample sizes tested are between 5000 and 10^4. The minimal sample size is set as 5000 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 I gives an overview of the experiments. It shows that indeed $\widehat{N}$ is biased upwards, especially when the sample size is small. When $n = 5000$, the collision mean is around 12, resulting in high bias (RB=0.10).

Figure 2 depicts the trends of the $\widehat{N}$ and $\widehat{N}^*$ with the growth of 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 5000, on average among 10^4 runs there are about 12 collisions, and the relative bias is around 10%.

Figure 3 shows the distributions of the estimations when the sample sizes are 5000, 5500, 6000, and 6500 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 smaller number of very large estimations. For instance in figure A there are more than 200 estimations of $\widehat{N}$ are higher than 2 millions, 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.

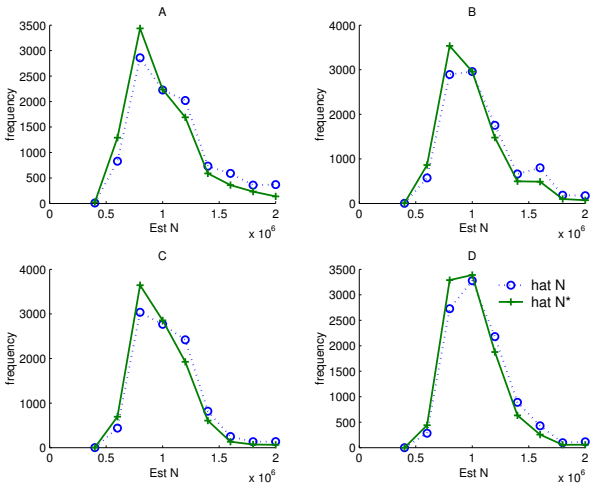

Fig. 3. Distribution of the estimations by $\widehat{N}$ and $\widehat{N}^*$ in simulation study, when n=5000 , 5500, 6000 and 6500 in sub-figures A, B, C, and D respectively.

V. 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. [13], 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 are stored in hard drive. Since our method is better to be used in undirected graph, we remove the direction in Twitter data. The matlab program and data are available at http://cs.uwindsor.ca/∼jlu/bias.

A. Estimate γ

Unlike the simulation studies presented in the last section where $\gamma = 0$, in real network the node degree varies and we need to estimate γ. In the area of capture-recapture research [3] [17], 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 [21][18]

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

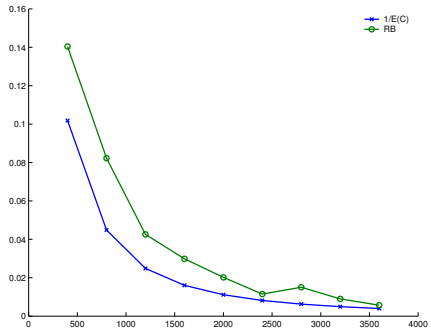

Fig. 4. Relative bias of $\widehat{N}$ in Twitter data for various sample sizes, and its comparison with 1/E(C).

TABLE II
Bias in Twitter data. $N = 4.17 \times 10^7$.

n ($\times 100$)	E(C)	1/E(C) (%)	RB (%) $\hat{N}$	RB (%) $\hat{N}^*$
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

According to equation 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 γ^2 is

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

B. Results

In our experiments the sample size ranges between 400 and 3600. The smallest sample size is set as 400 so that it can induce at least one multiple visits to a node. Although the true population is rather large (4.17×10^7), we do not need 5000 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 to A will cause node B visited twice. To avoid such loops, we take samples spaced every a few steps apart.

Overall the results conform well to our simulation studies. Figure 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 Figure 5. It is clear that the bias diminishes as the sample size grows. Figure 6 depicts the distribution of the estimations for the four smallest sample sizes. Table II summarizes the details of the results.

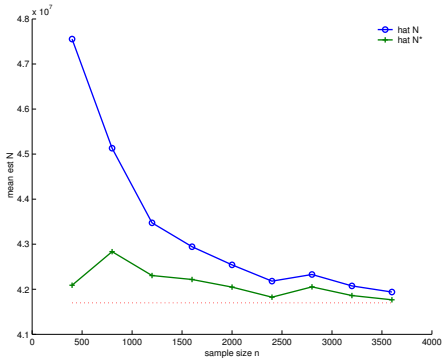

Fig. 5. $\widehat{N}$ and $\widehat{N}^*$ in Twitter for various sample sizes. The red dotted line is the true value.

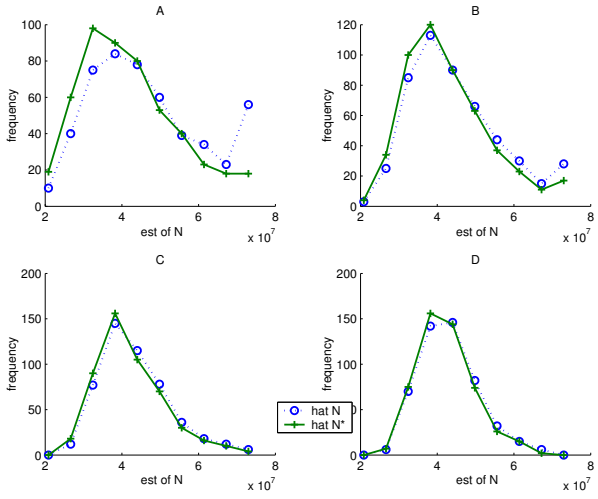

Fig. 6. Distributions of 500 estimations for Twitter data when sample sizes are 400, 800, 1200, and 1600 in sub-figures A, B, C, and D respectively.

VI. Conclusions

This paper shows that the the bias of $\widehat{N}$ can be too big 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] and [12], we are the first to explicitly describe the estimator in terms of collisions C and coefficient of variance γ. 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 γ, 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, thereby the coefficient of variation γ.

VII. Acknowledgements

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

References

[1] S. Amstrup, T. McDonald, and B. Manly. *Handbook of capture-recapture analysis*. Princeton Univ Press, 2005.

[2] A. Broder and et al. Estimating corpus size via queries. In *CIKM*, pages 594–603. ACM, 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*, pages 201–216, 1992.

[4] D. Chapman. Some properties of the hypergeometric distribution with applications to zoological sample censuses. *Univ. Calif. Publ. Stat*, 1:131–59, 1951.

[5] C. L. Chiang. On the expectation of the reciprocal of a random variable. *The American Statistician*, 20(4):p. 28, 1966.

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

[7] 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.

[8] P. Erdos and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[9] W. Gale and G. Sampson. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

[10] 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.

[11] P. J. Haas, J. F. Naughton, S. Seshadri, and L. Stokes. Sampling-Based estimation of the number of distinct values of an attribute. In *VLDB*, pages 311–322, 1995.

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

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600. ACM, 2010.

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

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

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

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

[18] 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.

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

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

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

[22] J. Wittes. On the bias and estimated variance of chapman's two-sample capture-recapture population estimate. *Biometrics*, pages 592–597, 1972.

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

[24] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):214, 2004.

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