Noname manuscript No.
(will be inserted by the editor)

Discover Hidden Web Properties by Random Walk on Bipartite Graph

Yan Wang · Jie Liang · Jianguo Lu

Abstract This paper proposes to use random walk to discover the properties of the deep web data sources that are hidden behind searchable interfaces. The properties, such as the average degree and population size of both documents and terms, are of interests to general public, and find their applications in business intelligence, data integration and deep web crawling. We show that simple random walk (RW) can outperform the uniform random (UR) samples disregarding the high cost of uniform random sampling. We prove that in the idealized case when the degrees follow Zipf's law, the sample size of UR sampling needs to grow in the order of $O(N/ln^2N)$ with the corpus size N, while the sample size of RW sampling grows logarithmically. Reuters corpus is used to demonstrate that the term degrees resemble power law distribution, thus RW is better than UR sampling. On the other hand, document degrees have lognormal distribution and exhibit a smaller variance, therefore UR sampling is slightly better.

Keywords Hidden data source, deep web, random walk, graph sampling, estimator, Zipf's law.

1 Introduction

Searchable forms are ubiquitous on the web. Many web sites, especially the large ones, have searchable interfaces such as HTML Forms or programmable web APIs. The data hidden behind a searchable interface constitute a hidden web data source [5] that can be accessed by queries only. The profiles of a hidden data source, including the average degrees and the total population size for both terms and documents in a data source, are of great interest to general public and business

Yan Wang
School of Information, Central University of Finance and Economics, Beijing, China. E-mail: dayanking@gmail.com

Jie Liang
BiblioCommons Inc., Toronto, Canada. E-mail: jie.liang@outlook.com

Jianguo Lu
School of Computer Science, University of Windsor, Windsor, Canada. E-mail: jlu@uwindsor.ca and State Key Laboratory for Novel Software Technology at Nanjing University.

competitors [24], to data crawlers [35,49,34,21,41], and to virtual data integrators such as vertical portals and meta search engines [47,45]. In business intelligence, people would like to know the number of users in Facebook, their average number of follower and their variations. In distributed information retrieval there is a need to profile the data sources before deciding where the queries should be sent to [8, 45]. In deep web crawling, it needs to know how many documents are there so that it can decide when to stop the crawling [35,51].

Discovering these properties has been a long lasting challenge [8], mainly due to the unequal probability of the data being sampled, or the heterogeneity of the data. Consequently it is difficulty or costly to obtain the uniform random samples [2,4]. This paper shows that instead of using uniform random (UR) samples, the biased sample obtained by simple random walk (RW) on the document-term graph can perform better.

For instance, we may want to learn the average document frequencies of the terms in a data source, or the average degree of the terms in its term-document graph. Average degree can be used to derive other properties such as degree variance and population size as we will show in Section 4. In turn average degree and population size reveal the total number of terms of the hidden corpus.

Given N number of terms labeled as $1, 2, \ldots, N$, and their degrees $d_1, d_2, \ldots, d_N$. The average degree is

$$\langle d \rangle = \frac{1}{N} \sum_{i=1}^{N} d_i. \tag{1}$$

One obvious but often impractical estimation method is via Uniform Random (UR) sampling, i.e., select a set of terms $\{x_1, x_2 \ldots, x_n\}$ where $x_i \in \{1, 2, \ldots, N\}$ randomly with equal probability, count their degrees $\{d_{x_1}, d_{x_2}, \ldots,\ d_{x_n}\}$ for each term, and calculate the sample mean as the estimate of the population mean:

$$\widehat{\langle d \rangle}_{SM} = \frac{1}{n} \sum_{i=1}^{n} d_{x_i}. \tag{2}$$

The sample mean estimator $\widehat{\langle d \rangle}_{SM}$ is unbiased if the terms or documents are *homogeneous*, i.e., they can be selected randomly with equal probability. Unfortunately this is not the case in most practice. Popular terms have a higher probability being sampled if terms are selected randomly from a document. Similarly, large documents tend to have a higher probability of being sampled if they are selected by random queries.

To analyze such *heterogeneous* data where elements have unequal probabilities of being sampled, various sampling methods have been studied for hidden data sources including search engine indexes [2], and in related areas such as the Web [20], graphs [26,3], online social networks [15,40], and real social networks [44, 50]. The typical underlying techniques include Metropolis Hasting Random Walk (MHRW) [36] for uniform sampling and Random Walk (RW) [28] for unequal probability sampling. MHRW is reported rather good at obtaining a random sample. However, in the sampling process many nodes are retrieved, examined, and rejected. The cost is rather high especially for hidden data sources. The samples are retrieved by queries that occupy network traffic, let alone the daily quotas impose by data providers. Thus a practical sampling method should include all the samples even if they induce bias.

Even when random samples are obtained, the sample mean estimator has a high variance because the degree distribution of the terms usually follows Zipf's law [55] [37]. Most terms have small degrees, while a few of them have huge degrees. The inclusion/exclusion of a huge term such as a stop word in a sample will make the estimation diverge.

We propose to use harmonic mean, instead of arithmetic mean, of the sample as the estimator of the average degree of documents and terms:

$$\widehat{\langle d \rangle}_H = n \left[\sum_{i=1}^{n} \frac{1}{d_{x_i}} \right]^{-1}. \tag{3}$$

Here the subscript H indicates that it is the harmonic mean, and that it can be derived from the traditional Hansen-Hurwitz estimator [19].

The sample for this estimator is obtained by low cost simple random walk where the node selection probability is asymptotically proportional to its degree. It is rather common to use Hansen-Hurwitz related estimators when selection probabilities are not equal for elements in the population. But usually people use PPS (Probability Proportional to Size) sampling because of the unavailability of random samples [44]. This paper shows that $\widehat{\langle d \rangle}_H$ can be better than the sample mean estimator even when uniform random samples are available–it has a very small bias, and the variance is smaller than the sample mean estimator for terms and is only slightly larger for the documents.

The crux of population size estimation is the heterogeneity of the data–documents and terms have unequal probabilities of being sampled. Yet the degree of the heterogeneity, called Coefficient of Variation (CV, denoted as γ hereafter), is difficult to predict in traditional sampling studies where the accurate degree is hard to quantify. In our setting the degrees of sampled documents and terms are easy to obtain, thereby the average degree is ascertained accurately thanks to the estimator $\widehat{\langle d \rangle}_H$. Thus, the coefficient of variation can be estimated by

$$\widehat{\gamma}^2 + 1 = \frac{1}{\langle d \rangle n} \sum_{1}^{n} d_{x_i}. \tag{4}$$

With the knowledge of γ, the population size can be obtained by

$$\widehat{N} = (\widehat{\gamma}^2 + 1) N_0 = (\widehat{\gamma}^2 + 1) \binom{n}{2} \frac{1}{C}. \tag{5}$$

Here N_0 is an estimator for homogeneous data, and $\binom{n}{2}\frac{1}{C}$ is one of the N_0 estimators. C is the collisions of the nodes happened during sampling.

Our main *contributions* in this paper can be summarized as follows:

- For average degree estimation we show that RW sampling can outperform UR sampling, even ignoring the high cost of obtaining the uniform random samples.
- We show that RW is not always better than UR sampling. We give the condition when RW could be better.
- We show that average degree is an important property that can lead to the discovery of the population size;

- We solved the open problem to correct the bias in capture-recapture method. It is well known in the area of capture-recapture method that there is a negative bias when the data is heterogeneous. The problem to quantify and consequently correct this bias has never been solved. We show that the population size can be estimated first as if the data were homogeneous, then multiply the estimation by $\gamma^2 + 1$. In particular, we show that it is a practical approach because of the success estimation of the mean degree that leads to the discovery of γ^2.

In the following we will first introduce the related work, especially the background of population size estimation. Then we model the query-based sampling as a random walk on a bipartite graph. Section 4 introduces two estimators, one for average degree and the other for population size. We prove that in the idealized case where term degrees follow Zipf's law with exponent one, our proposed estimator $\widehat{\langle d \rangle}_H$ is much better than $\widehat{\langle d \rangle}_{SM}$ when the corpus size is large. The experiments section dissects the Reuter corpus with details of the data distributions, sample distributions, and estimation results with various sample sizes. Then we give an intuitive explanation for why $\widehat{\langle d \rangle}_H$ can reduce the variance.

2 Related work

Query based profiling of hidden data sources has been studied ever since the occurrence of web query interfaces. One of the early influential works is the estimation of search engine size [24]. The problem can be further classified by the syntax of the queries allowed and the types of data bases sitting behind. Queries can be simple key words [9,46,7,31,53], boolean expressions [24], or even SQL queries [13, 14,18]. The data sources can be text databases such as a collection of documents [9,46,7,3], or structured data in the form of relational database tables [13,14,18]. Our paper assumes simple keyword interfacing with textual database.

The background of this research is the population size estimation and sampling that have been widely studied in other disciplines [48] especially in ecology [1] and social studies [44], and more recently in computer science for estimating the size of the web [24], databases [18], web data sources [53,14,20,54,7], and online social networks [22,15,52].

2.1 Average degree estimation

At the first sight, the average degree estimation problem seems neither important nor difficult. In reality it is an important problem in that 1) it leads to the discovery of the coefficient of variation (CV), in turn CV can be used in the population size estimation; 2) it reveals the overall data size. Average degree estimation is also a difficult problem because uniform random samples may not be directly available due to the restricted sampling interface as in the query-based sampling. Although there are studies to obtain the uniform random sample using rejection method or MHRW [3], the cost will be rather high. Therefore there are studies to use the biased sample directly, and use harmonic mean to adjust the bias. The detailed derivation of the harmonic mean estimator can be found in [44] where the purpose is to sample hidden population such as drug-addicts. In this setting it is impossible to evaluate the estimator because neither the true value nor the sampling probability can be verified. The biased samples are taken because uniform random sampling is impossible. The harmonic mean estimator is derived to correct the bias, not to improve the performance of the estimation.

In the area of peer-to-peer network [42] and online social network [23,16], the re-weighted random walk that resembles harmonic mean was used and empirically compared with MHRW, but not with uniform random samples.

Our work is the first to show that RW can outperform UR samples disregarding the cost of obtaining these samples. In addition, we give the conditions when RW could be better. The preliminary result was also published in our workshop paper [32] where the data is online social network. This paper reports our recent progresses on the following aspects: 1) we experiment on text bipartite graph instead of the non-bipartite graph representing social networks; 2) we show that RW is not always better than UR sampling, and analytically give the conditions when RW outperforms UR sampling; 3) the experiments verifie that when the degree distribution follows power law, RW sampling is much better than UR sampling. When the degree distribution is log-normal, UR is slightly better than RW sampling; 4) in addition to average degree estimation, population size estimation is discussed and experimented in detail; 5) we added the intuitive explanation as for why UR can be better than RW sampling.

2.2 Population size estimation

2.2.1 Capture-Recapture Method

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:

$$\hat{N}_{LP} = \frac{n_1 n_2}{d}, \tag{6}$$

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 two samples. 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, as will be discussed in the next two subsections.

Albeit its simplicity and severe restriction, most of the existing work used the capture-recapture sampling method and the corresponding Lincoln-Petersen estimator in one form or another. The classic work is the estimation of the Web and the search engines described in [25] and [6]. Both approaches use queries to capture documents, and count the duplicates between the two captures. In [25], Lawrence and Giles were aware that the estimation is not accurate, therefore they presented the estimation as *relative* size, not the absolute size. Bharat et al. [6] investigated the causes for the inaccuracy, in particular the unequal probability of documents being matched by queries (called query bias in their paper). They proposed to alleviate the bias by obtaining uniform random samples, say, from the search engine directly as privileged users instead of public searchable interfaces.

Gulli and Signorini [17] used the same method in a larger scale, by building query lexicon from dmoz.com directory that contains 4 million pages.

Continuing in this direction, other innovative methods are proposed for those two captures, not necessarily by two queries. Nonetheless, the underlying estimator is still Lincoln-Petersen estimator, and the bias problem remains un-tackled. Both Si et al. [47] and Kunder [1] used query frequencies to estimate search engine sizes. They take a sample set of documents with size (n_1), and another sample set of documents that contain a specific query. Say the second capture is of size n_2, which is actually the document frequency of the query, and it may be provided by search engines. The overlapping d is the the intersection of those two sets, i.e., the number of documents that match the query in the sample documents. The advantage of this method is that n_2 does not need to be calculated by the sampler. The disadvantage is also obvious: in both sampling occasions the documents are not sampled with equal probability. What is worse, the document frequency returned by search engines are often inflated, sometimes in orders of magnitude.

In [7] Broder et al. also use Lincoln-Petersen estimator, but each capture is $defined$ as the documents covered by many queries. Because the number of queries is very large, it is not possible to obtain n_1 and n_2 directly by actually submitting the queries. Instead n_1 and n_2 themselves are estimated.

Equation 6 is ubiquitous and applied in various forms. Quite often even the users may not be aware that they are actually applying the basic capture-recapture method. For instance, ID sampling is used to estimate Facebook population by leveraging the fact that each ID is a 9-digit number [15]. The estimation method is to select a number uniformly at random in the range 1 to 10^9, then probe the server to check whether it is a valid ID. Suppose that the total number of valid IDs is n_1, the probings being sent is n_2, and valid ones among the samples are the duplicates between the two sets, denoted as d. Then, according to 6, we have $10^9 = n_1 n_2/d$. When n_2 and d are available, we can use the equation to estimate n_1, the number of valid IDs.

2.2.2 Multiple Capture-Recapture Method

When there are more than two sampling occasions and each time only one sample is taken, Darroch [12] derived that the approximate Maximum Likelihood Estimator (MLE), $\widehat{N}_D$, is the solution of the following equation:

$$n - d = N\left(1 - e^{-\frac{n}{N}}\right), \tag{7}$$

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 [38]. Unfortunately it does not have a simple closed form solution [38] [12], i.e., it can not be solved algebraically for N. In online social network studies, [52] used numeric method to find the solution to this estimator. [31] gives an approximate solution for N that reveals a power law governing the data not sampled and the overlapping rate. In a simpler form, it states that the percentage P of the data *not* sampled decreases in the power of the overlapping rate $R =$

[1] `http://www.worldwidewebsize.com/`

$n/(n-d)$, i.e.,

$$P = R^{-2.1}. \tag{8}$$

2.2.3 Unequal Sampling Probability

When the data is heterogeneous, i.e., elements have unequal probabilities of being sampled, the estimation becomes notoriously difficult. One approach to solving this problem is to obtain the uniform random sample [3] using algorithms such as Metropolis-Hasting random walk, then traditional estimators are applied on the random sample. The other approach uses the biased sample to save the sampling cost, but adjust the bias by devising new estimators [7,46,31]. Broder et al.[7] assigned less weight to large documents being sampled; Shokouhi et al. [46] run regression on past data to establish the relationship between the homogeneous and heterogeneous data; Lu et al. [29][31] [33] went a step further by using γ, the degree of heterogeneity, to adjust the discrepancy.

The problem is that estimating γ itself is equally challenging. Therefore Equation 5 as an estimator for N was not seen in ecology. Instead, the same equation was used by Chao et al. [10] in a reverse way to estimate γ as below:

$$\widehat{\gamma^2} = N_0 C \binom{n}{2}^{-1} - 1, \tag{9}$$

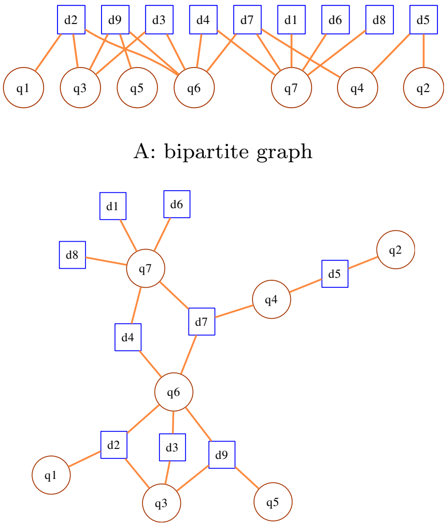
where N_0 is a rough estimation for N assuming the data is homogeneous. This method was demonstrated [10] on small data where γ^2 is typically around one. That is, the ratio between N and N_0 is around two. In our large and power law data γ^2 can go up to hundreds, making the traditional estimator biased downwards by hundreds of times smaller than the real value.

In the estimation of digitalized networks such as hidden web data sources, 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 the simple random walk on the term-document graph 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. [22] used a similar equation to estimate the size of online social networks:

$$\widehat{N_K} = \frac{1}{2C} \sum_{1}^{n} d_{x_i} \sum_{1}^{n} 1/d_{x_i}, \tag{10}$$

which can be transformed to estimator $\widehat{N}$. [22] showed that it is a consistent estimator.

Note that estimator $\widehat{N}$ can be approximated by equations either 7 or 8 when $\gamma = 0$, sample size is small, and collisions C can be approximated by duplicates d. The approximation can be established by applying Taylor expansion on the right hand side of equation 7.



A: bipartite graph

B: same graph as (A) in spring model layout

Fig. 1 Hidden data source as a bipartite graph

2.3 Other size estimation methods

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, [15] 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; [54] levarages the prefix encoding of Youtube links; [14] depends on the negation of queries to break down the search results; [30][53] deals with the return limit of the search engines. Dasgupta et al. use random walk in query space to probe database properties [13] [14], which is different from our random walk in that 1) they suppose the SQL like syntax of the queries that support boolean expressions. 2) The random walk is on the query space constructed by the boolean operators, not the document-term graphs in our paper.

3 Problem definition

Following [51] [39] [53], a hidden data source can be modelled as a document-term bipartite graph $G = (D, T, E)$, where nodes are divided into two separate sets, D the set of documents, and T the set of terms. Every edge in E links a term and a document. There is an edge between a term and a document if the term occurs in the document.

Let d_i^D denote the degree of the document node i , for $i \in \{1, 2, \ldots, |D|\}$, i.e., the number of distinct terms in document i. Let d_j^T denote the degree (or document frequency) of the term node j for $j \in \{1, 2, \ldots, |T|\}$. The volume τ of the documents

D (or the volume of terms T) is

$$\tau = \sum_{i=1}^{|D|} d_i^D = \sum_{j=1}^{|T|} d_j^T .$$

The mean degree of documents D is $\langle d^D \rangle = \tau/|D|$, and the mean degree of terms T is $\langle d^T \rangle = \tau/|T|$.

The goal of this paper is to estimate $\langle d^D \rangle, \langle d^T \rangle, |D|$ and $|T|$ using a sample. When it is clear from the context, we will omit the superscript D and T, using $\langle d \rangle$ to denote the average degree for documents or terms, and N to denote the size of the population $|D|$ or $|T|$. We use *terms* and *queries* interchangeably with slight different connotations: a lexicon in a document is a term, when a term is sent to a searchable interface it is called a query.

Example 1 (Bipartite graph of hidden web) *Figure 1 gives an example of a hidden data source that is represented as a bipartite graph, where* $D = \{d_1, d_2, \ldots, d_9\}$ *and* $T = \{q_1, q_2, \ldots, q_7\}$. $\langle d^T \rangle = 18/7$, *and* $\langle d^D \rangle = 18/9$.

A simple random walk on the document-term graph is described in Algorithm 1. First a seed query is selected randomly from a dictionary and the list of the matched document URLs are retrieved. From the list we select randomly one of the URLs and download the corresponding document. From the downloaded document a query is selected randomly and sent to the data source. The process is repeated until n number of sample documents and n number of sample terms are obtained. In the samples the documents or the terms can be visited multiple times. In other words it is a sampling with replacement.

During the random walk process, we do not need to explore all matched documents. Instead, we can first ask for the number of matches m, generate a random number r between 1 and m, then directly access the page containing the r-th document. Therefore for each sample document and term at most two queries are needed, one to get the degree of the query, the other to get the page containing the r-th document.

Algorithm 1: Random Walk Sampling

```
Input: $t_0$=seed term, sample size $n$
Output: Sample documents $D_s$ and their degrees; Sample terms $T_s$ and their degrees.
$D_S = T_S$=empty lists;
i =1;
while $i \leq n$ do
    add $t_i$ and its degree to $T_s$ ;
    $d_i$=one random document that matches $t_i$ ;
    add $d_i$ and its degree to $D_s$ ;
    $t_{i+1}$=one random term in document $d_i$;
    $i++$;
end
return $D_s$ and $T_s$;
```

Example 2 (Random walk) *If the sample size n is 5 and the seed term is q_1. A random walk result can be*

$$T_s = ((q_1, 1), (q_6, 5), (q_7, 5), (q_7, 5), (q_6, 5))$$
$$D_s = ((d_2, 3), (d_4, 2), (d_1, 1), (d_7, 3), (d_3, 2))$$

4 Estimators

This paper focuses on two properties, the average degree and the total population size for both terms and documents. When uniform random samples are available, the former property can be estimated by the sample mean, the latter by capture-recapture methods [1]. However, uniform random samples are not easy to obtain. It is well known that in random walk large documents and queries have higher probability of being visited. Asymptotically the sampling probability of a document or a term is proportional to its degree. Therefore we need to use estimators developed for such samples whose sampling probability is proportional to their sizes.

We first develop the estimation of average degree, including the average length (number of distinct terms) of the documents and the average size (or document frequency) of terms. Based on the average degree, the estimator of population size (total number of terms and documents) is derived.

Table 1 summarizes the notations used in this paper.

Table 1 Summary of notations

Notation	Meaning	Properties
N	population size	
n	sample size	
d_i	degree of node i	
τ	volume of all the document/term 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	square of coefficient of variation	$\gamma^2 = \sigma^2/\langle d\rangle^2 = \langle d^2\rangle/\langle d\rangle^2 - 1$
$\langle d^W\rangle$	asymptotic mean degree of random walk	$\langle d^W\rangle = \langle d^2\rangle/\langle d\rangle$

4.1 Average degree

Suppose that in the document-term graph there are N number of document nodes. Node i has a degree $d_i, i \in \{1, 2, \ldots, N\}$. Let the total number of document degree is $\tau = \sum_{i=1}^N d_i$, and the mean of document degrees is $\langle d\rangle = \tau/N$.

The variance σ^2 of the degrees in the population is defined as [48]

$$\sigma^2 = \langle d^2\rangle - \langle d\rangle^2, \tag{11}$$

where $\langle d^2 \rangle$ is the 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:

$$\gamma^2 = \frac{\sigma^2}{\langle d \rangle^2} = \frac{\langle d^2 \rangle}{\langle d \rangle^2} - 1. \tag{12}$$

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.

4.1.1 Sample mean estimator

If a uniform random sample $(d_{x_1}, \ldots, d_{x_n})$ is obtained, the sample mean is an unbiased estimator as defined below:

$$\widehat{\langle d \rangle}_{SM} = \frac{1}{n} \sum_{i=1}^{n} d_{x_i}. \tag{13}$$

The variance of the estimator $\widehat{\langle d \rangle}_{SM}$ is [48]

$$var(\widehat{\langle d \rangle}_{SM}) = \frac{\sigma^2}{n}. \tag{14}$$

The problem with this sample mean estimator is that the uniform random sample is not easy to obtain. Moreover, its variance is too large to be of practical application if the degrees have a large variance. It is well established that the degree of the terms follows Zipf's law, causing the population variance σ^2 of the term degrees very large.

More specifically, if the degrees follow the Zipf's law strictly, the variance of the sample mean estimator can be described by the following theorem

Theorem 1 *Suppose the degrees follow Zipf's law with exponent one, i.e., $d_i = \frac{A}{\alpha + i}$, where A and α are constants. The variance of the sample mean estimator is*

$$var(\widehat{\langle d \rangle}_{SM}) \approx \frac{\langle d \rangle^2}{n} \left(N \left[\alpha \ln^2 \frac{N + \alpha}{1 + \alpha} \right]^{-1} - 1 \right). \tag{15}$$

Proof See appendix.

4.1.2 Harmonic mean estimator

When sampling probability is not equal for each unit, a common approach is to use Hansen-Hurwitz estimators [48]. In the case where the sampling probability of a node is proportional to its degree, the estimator for degree mean $\widehat{\langle d \rangle}_H$ is the harmonic mean of the degrees:

$$\widehat{\langle d \rangle}_H = n \left[\sum_{i=1}^{n} \frac{1}{d_{x_i}} \right]^{-1}. \tag{16}$$

We refer to Salganik et al. [44] for detailed derivations of the estimator in the setting of respondent driven sampling. Also it can be derived as a special case of importance sampling [27].

Unlike the unbiased estimator $\widehat{\langle d \rangle}_{SM}$, $\widehat{\langle d \rangle}_{H}$ is biased. According to Cochran [11] the bias is on the order of $1/n$. Since the sample size n in our setting is far greater than one in general, the bias is negligible.

Its variance can be derived from the variance of Hansen-Hurwitz estimator using the Delta method, resulting in:

$$\widehat{var}(\widehat{\langle d \rangle}_{H}) = \frac{s_v^2}{\overline{v}^4 n}, \tag{17}$$

where $v_i = 1/d_{xi}$, $\overline{v}$ and s_v^2 are the sample mean and variance of v_i's. This estimated variance will be supported by our experiments in Section 5.

In the idealized case when the degrees follows exactly with 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*

$$var(\widehat{\langle d \rangle}_{H}) = \frac{\langle d \rangle^2}{n} \left(\frac{1}{2} \ln \frac{N+\alpha}{1+\alpha} - 1 \right). \tag{18}$$

Proof See appendix.

Comparing the variances of estimators $\widehat{\langle d \rangle}_{SM}$ and $\widehat{\langle d \rangle}_{H}$, we can see that the variance of $\widehat{\langle d \rangle}_{H}$ grows logarithmically with corpus size N, while $\widehat{\langle d \rangle}_{SM}$ 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 $\widehat{\langle d \rangle}_{SM}$, but merely $\ln N$ for $\widehat{\langle d \rangle}_{H}$.

Example 3 (Degree estimation) *For our example, the harmonic mean estimation for average degree of terms is*

$$\widehat{\langle d \rangle}_{H} = \frac{n}{\sum 1/d_i} = \frac{5}{\frac{1}{1} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5} + \frac{1}{5}} = \frac{25}{9} = 2.7778.$$

For documents the estimated average degree is

$$\widehat{\langle d \rangle}_{H} = \frac{n}{\sum 1/d_i} = \frac{5}{\frac{1}{3} + \frac{1}{2} + \frac{1}{1} + \frac{1}{3} + \frac{1}{2}} = \frac{30}{16} = 1.875.$$

4.2 Population size estimation

The population size can be estimated as follows,

$$\widehat{N} = (\gamma^2 + 1)\widehat{N_0} = (\gamma^2 + 1) \binom{n}{2} \frac{1}{C}, \tag{19}$$

where $\widehat{N_0}$ is the estimation of N if the samples are taken uniform randomly, n is the sample size, C is the number of collisions, γ is Coefficient of Variation (CV) of the degrees. Let f_i denote the number of individuals that are visited exactly i times.

$$C = \sum_{i=1}^{+\infty} \binom{i}{2} f_i.$$

The derivation of the estimator is given in Appendix 9.3. It can be also derived as a special case of Eq 3.20 in [10]. But that equation is used to estimate γ instead of N. Katzir et al. [22] used an equivalent formula but in a very different form.

Note that this is a biased estimator as we pointed out in [33]. The relative bias is approximately $1/C$, and can be corrected using the following estimator. When collisions C is large, such bias can be neglected. This paper uses the estimator in equation 19 to focus on the other bigger bias, i.e., the bias introduced by γ.

$$\widehat{N}^* = (\gamma^2 + 1) \binom{n}{2} \frac{1}{C+1}, \tag{20}$$

The elegance of the equation is that when the samples are uniform, $\gamma = 0$, and the estimator is reduced to the traditional birthday paradox or capture-recapture method. When the sample is obtained by random walk, the sampling probability is not equal among all the documents or terms, resulting in $\gamma > 0$. The population size can be estimated as if the samples were taken uniformly, then multiplied by $\gamma^2 + 1$. This was not seen in literature as far as we are aware.

The reason of this formulation being overlooked may due to the challenge of determining γ in traditional estimation problems. In ecology and social studies the degree of a node can not be quantified accurately. For instance, it is hard to determine the friends of a drug-addict. This makes it impossible to perform a simple random walk in the graph. In our setting of deep web data sources, we know exactly the number of documents a query matches, and the number of terms a document contains. Leveraging this information, the heterogeneity γ of the data can be obtained by simple random walk as below. Asymptotically the mean of the degrees obtained by a random walk is

$$\langle d^W \rangle = \sum_{i=1}^{N} p_i d_i = \frac{\langle d^2 \rangle}{\langle d \rangle}, \tag{21}$$

where $p_i = d_i/\tau$ is the selection probability of node i. Hence

$$\gamma^2 + 1 = \frac{\langle d^W \rangle}{\langle d \rangle}, \tag{22}$$

where $\langle d^W \rangle$ can be estimated by its sample mean

$$\widehat{\langle d^W \rangle} = \frac{1}{n} \sum_{i=1}^{n} d_{x_i}, \tag{23}$$

and $\langle d \rangle$ can be estimated by its harmonic mean $\widehat{\langle d \rangle}_H$. Combining the two equations we derive the estimator for γ as follows:

$$\widehat{\gamma}^2 + 1 = \frac{\widehat{\langle d^W \rangle}}{\widehat{\langle d \rangle}} = \frac{1}{n^2} \sum_{1}^{n} d_{x_i} \sum_{1}^{n} 1/d_{x_i}. \tag{24}$$

Table 2 Summary of the Datasets

Data	# Docs	# Terms	Document Degree		Term Degree	
			Average	γ	Average	γ
Reuters	806,791	380,465	109.97	0.72	232.97	16.12
NG20	11,059	84,644	190.45	0.82	24.88	9.08
Wiki	9,989	7,213	19.3	1.09	26.74	3.86
KDDCUP	17,118	980,039	393.36	6.84	6.77	10.91

The convenience of the method is that only one random walk is needed to obtain both $\langle d \rangle$ and $\langle d^W \rangle$.

Example 4 (Population size estimation) *Continuing on our example data source for terms.*

$$\widehat{\gamma^2} + 1 = \frac{\widehat{\langle d^W \rangle}}{\widehat{\langle d \rangle}} = \frac{21/5}{25/9} = 1.512.$$

n=5, $f_1 = 1(q_1)$*,* $f_2 = 2$ *(*q_6 *and* q_7*), therefore*

$$C = \sum_{i=1}^{\infty} \binom{i}{2} f_i = 2,$$

$$\widehat{N} = 1.512 \times \frac{4 \times 5}{2} \times \frac{1}{2} = 7.56.$$

5 Experiments

5.1 Datasets

Our method is tested against several datasets including Reuters newswires [43] (Reuters), newsgroups [2] (NG20), KDDCUP 2013 research papers (KDDCUP) [3], and a subset of Wikipedia (Wiki). The statistics of these datasets are summarized in Table 2. In this experiment a term is a sequence of letters and is case-insensitive. The term population N is the total number of distinct terms that are collected in all the documents in the corpus. The degree of a term is the document frequency of the term, i.e., the number of documents that contain the term. The degree of a document is the number of distinct terms in that document.

We list γ, the normalized standard deviation of the degrees, for each dataset. It is well known that term degrees (document frequencies) follow power law, while document degrees follow lognormal law. Reflected in our datasets, γ for term degree is larger than that of the document degrees. As a verification, we show distributions of Reuters, for both term degrees and document degrees in Figure 2. In order to show both ends of the distributions, we plot the degree against its rank in sub panel (A) and (C), as well as the frequency against its degree in sub panels (B) and (D). The former plot has a better view of the top degrees, while the latter

[2] Available at `http://qwone.com/~jason/20Newsgroups/`

[3] `http://www.kaggle.com/c/kdd-cup-2013-author-paper-identification-challenge`. Our data contains 17,118 publication venues and the keywords (980,039) occurred in the venues.

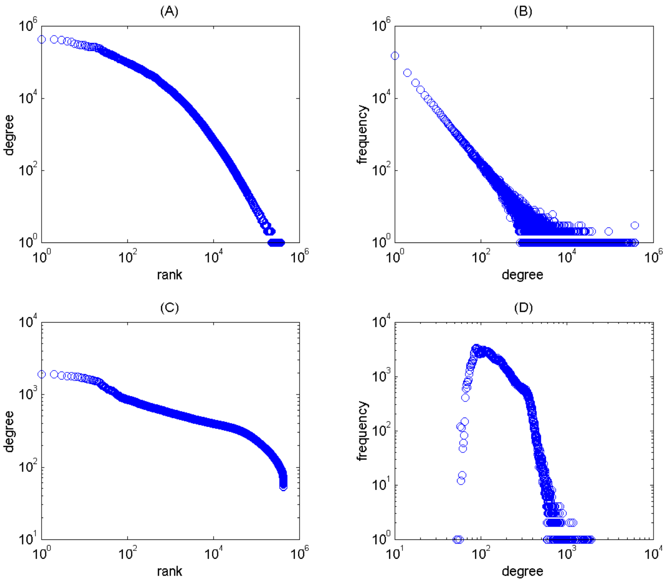


Fig. 2 Degree distribution of terms (panels A and B) and documents (panels C and D) in Reuters corpus. Panels (A) and (C) are degree vs .rank plots. (B) and (D) are frequency vs. degree plots. It shows that terms have a larger variance than documents.

depicts better the small terms or documents. Clearly the distributions of term degrees and document degrees are different, which is the cause of different values for γ, and different results on average degree estimation for terms and documents. The term degrees obviously follow Zipf's law, while document degrees are more like log-normal distribution. In addition, document degrees sit in a very narrow range (min is 6, max is 1659) compared with term degrees (1 to 434202). Therefore the heterogeneity of those two kinds of degrees are very different. CV of term degrees is 16, and CV of document degree is merely 0.7.

5.2 Summary of the Results

First, we evaluate the estimators in terms of relative standard error (RSE) that is defined as below:

$$RSE(\hat{d}) = \frac{1}{\mathbb{E}(\widehat{d})}\left[\mathbb{E}(\hat{d} - \mathbb{E}(\hat{d}))^2\right]^{1/2}, \tag{25}$$

where $\mathbb{E}(X)$ is the expectation of X, i.e., the mean of all the possible values, which can be approximated by the sample mean when the number of values is large. We omit bias or MSE because these estimators have negligible biases when the sample size is not small, and the degree estimator for uniform random sampling does not have bias. In the next subsection, we will give more detailed evaluation for Reuters data in terms of both bias and variance.

	Average Degree				Population			
	Documents		Terms		Documents		Terms	
	UR	RW	UR	RW	UR	RW	UR	RW
Reuters	**0.0215**	**0.0292**	0.2209	0.1741	0.4114	0.1474	0.1815	0.1736
NG20	**0.0580**	**0.0454**	0.6424	0.2979	0.2321	0.1757	0.3126	0.1112
Wiki	**0.0776**	**0.0746**	0.2730	0.2054	0.2683	0.1187	0.2661	0.1927
KDDCUP	0.7716	0.2779	0.4839	0.2010	0.4617	0.2877	0.2655	0.0507

Table 3 Relative standard errors of the estimations. For average degrees, the sample size is fixed at $n = 200$(5000 for Reuters), and the observed variance is obtained from 2000 repetitions (200 for Reuters). For population size, sample size varies with the true population size ($\propto \sqrt{2N}$).

Table 5.2 summarizes the relative standard errors for the four combinations of two estimators for both terms and documents. For average degree estimation, the sample size is 200 (5000 for Reuters). RSE is obtained from 2000 repetitions (200 for Reuters). For population size estimation, the sample size varies with real data size N, approximately in the order of $\sqrt{2N}$. That is the number needed to produce some collisions.

Overall, we observe that random walk (RW) outperforms uniform random (UR) samples, by obtaining smaller variance using the same sample size. There are also cases where RW is worse than or close to UR, for instance in document degree estimation in Reuters, NG20 and Wiki datasets. This is because γ for these datasets are rather small, ranging between 0.7 and 1.1. When γ is larger, RW demonstrate a larger advantage, such as in KDDCUP data.

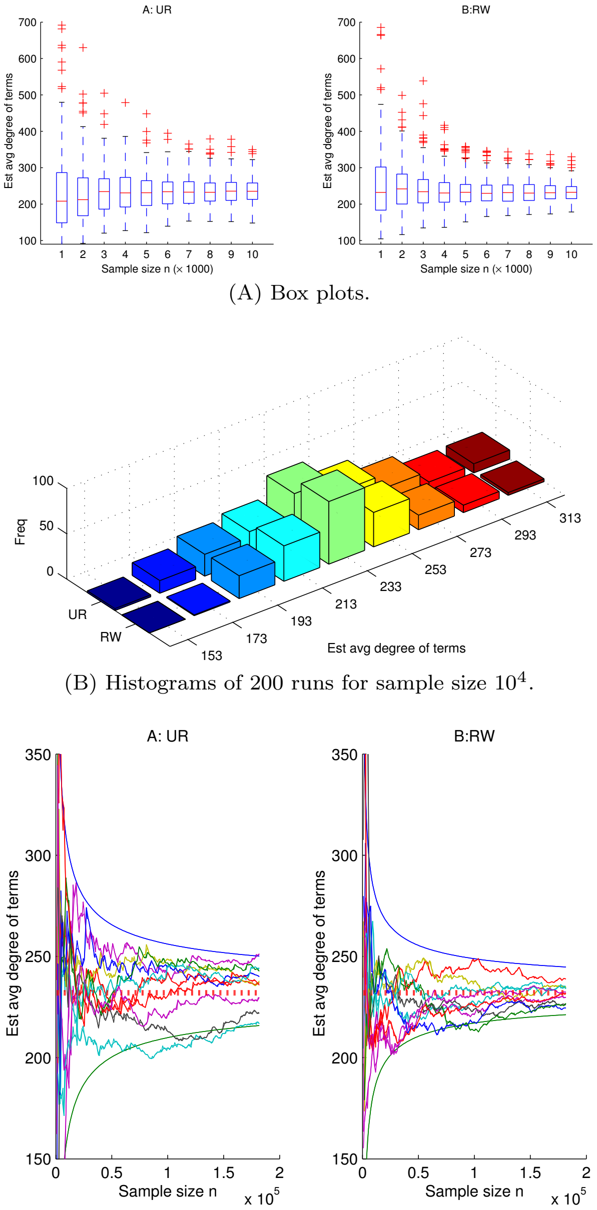
5.3 Average Degree of Reuters

In the following we focus on the detailed analysis using one dataset, the Reuters data. Estimators are normally evaluated in terms of bias, standard error (SE), and rooted mean squared error(RSME). In the case of average degrees they are defined as

$$\begin{aligned} Bias(\hat{d}) &= \mathbb{E}(\hat{d}) - d, \\ SE(\hat{d}) &= \left[\mathbb{E}(\hat{d} - \mathbb{E}(\hat{d}))^2\right]^{1/2}, \\ RMSE(\hat{d}) &= \left[\mathbb{E}(\hat{d} - d)^2\right]^{1/2}, \end{aligned}$$

Since $Bias^2 + SE^2 = RMSE^2$, and bias is negligible compared to SE according to Section 4 and our first experiment below, we have $SE \approx RMSE$. Thus except the first experiment where Bias, SE and RMSE are reported, in the remaining experiments we report SE only.

Average degrees are estimated on both UR and RW samples using sample mean estimator $\widehat{\langle d \rangle}_{SM}$ defined in Equation 2 and harmonic mean estimator $\widehat{\langle d \rangle}_{H}$ defined in Equation 3, respectively. Since the degrees of terms have a much larger CV than that of documents, RW estimator is better than the UR for terms, but slightly worse for documents.



(A) Box plots.

(B) Histograms of 200 runs for sample size 10^4.

(C) Error bound.

Fig. 3 Average term degree estimation by UR and RW samplings. Panel A: box plots for various sample sizes ranging between 10^3 and 10^4. Data consist of 200 runs for each sample size. It shows RW has a smaller variance for all different sample sizes. Panel B: histograms focusing on the last box in panel A when n=10^4 for UR and RW. It shows that the estimations by RW and UR follow normal distribution whose mean is the true value 233, and RW has a smaller variance. Panel C: 10 estimation processes along with the estimated 95% error bound calculated from Equations 14 and 17 respectively. It shows that mostly the estimations are within the error bound, and RW has a smaller variance.

Table 4 Estimations of average term degree $\langle d\rangle = 233$ over 200 runs for various sample size n.

n	Bias		Standard error		RMSE	
$\times 10^3$	UR	RW	UR	RW	UR	RW
1	4.3217	28.4480	132.6423	113.7653	**132.7130**	**117.2856**
2	-3.4330	17.7063	85.4701	74.0365	85.5394	76.1347
3	2.9543	9.8360	62.3450	59.5591	62.4153	60.3699
4	4.0269	4.7295	55.7902	47.2623	55.9361	47.4995
5	1.0870	1.8336	51.4717	40.5863	51.4833	40.6279
6	1.9411	1.0162	45.1416	**36.7005**	45.1835	36.7146
7	1.4922	0.1831	41.8117	33.2903	41.8385	33.2908
8	2.0617	0.3111	39.9589	30.7277	40.0123	30.7292
9	3.6235	1.5489	38.4063	29.4498	38.5777	29.4907
10	3.3618	1.3060	**37.0325**	27.3982	37.1855	27.4295

5.3.1 Average Degree of Terms

The two estimators are tested on the data for 10 different sample sizes ranging between 10^3 and 10^4. For each sample size we repeat the experiment for 200 times and the results are plotted in Figure 3. Panel A compares UR and RW using box plots. It shows that RW has a smaller variance consistently for all sample sizes. Panel B plots the distribution of the estimations in the last box of panel A when sample size is 10^4. It demonstrates that 1) both RW and UR estimations follow normal distribution with the same mean value (233); 2) RW has a smaller variance than UR.

Since the estimations follow normal distribution, the 95% error bound can be calculated as roughly twice of the standard error described in Equations 14 and 17. Panel C plots the error bounds along with 10 large samples, each with size up to 2×10^5. Although 10 sampling processes are hardly discernible from each other in the plot, what we want to show is that mostly they are within the error bounds as predicted by Equations 14 and 17. The plot validates the equations for estimated variances, and gives another perspective explaining why RW is better than UR. We will elaborate this further in Section 6. This plot also helps us determine how large the sample should be to achieve a satisfactory estimation.

The bias, standard error, and rooted mean squared error of the two estimators are tabulated in Table 4. It shows that indeed $\widehat{\langle d\rangle}_H$ has a very small bias as expected in Section 4.1.2 for most sample sizes except for the smallest ones. When the sample size is very small ($n \approx 1000$), RW has a positive bias. A closer inspection of the experiment data reveals that the small terms dictate the outcome of $\widehat{\langle d\rangle}_H$, but they can be hardly visited within a small number of random walk steps. Nonetheless, even for small samples where n=1000 the overall indicator RMSE of RW is 117, still smaller than that of UR (132). We also experimented with sample size n=500 where RMSE of UR is slightly better than RW.

Another perspective to interpret the results is looking at the sample sizes that reach similar SEs from RW and UR methods. For instance when the sample size is 6,000, the SE of RW is 36.7005. On the other hand, UR needs more than 10,000 samples to achieve the same SE.

Table 5 Estimated average document degree $\langle d \rangle = 109.97$. Data obtained from 200 runs for various sample size n ranging between 500 and 5,000.

$n(\times 100)$	Standard error	
	UR	RW
5	7.4210	10.3475
10	5.2573	7.2885
15	**4.2589**	5.9236
20	3.7398	5.0035
25	3.3266	4.5959
30	3.0740	**4.1151**
35	2.8880	3.9370
40	2.6676	3.6564
45	2.4589	3.3721
50	2.3763	3.2211

5.3.2 Average Degree of Documents

Table 5 lists the standard errors of the 200 runs on document degrees for sample sizes ranging between 500 and 5,000. First of all both estimators ($\widehat{\langle d \rangle}_{SM}$ and $\widehat{\langle d \rangle}_{H}$) perform better than term degree estimation because of the small variation of document degrees (CV=0.7). Since the standard error of UR is already rather small, there is little chance for RW to beat the UR samples.

However, uniform random sampling is only slightly better than random walk sampling for the same sample size. If we include the cost of obtaining the random samples using algorithm such as Metropolis-Hasting Random Walk, the total cost shall be much higher since samples can be rejected many times. In simple random walk we only need to double the sample size to achieve better result achieved by uniform random samples. For instance, the standard error of 3000 random walk samples is 4.1151, while 1500 uniform random samples can achieve similar standard error (4.2589).

5.4 Population Size

In this experiment again random walk (RW) samples are compared with uniform random (UR) samples. For RW samples, the population size N for both documents and terms are estimated using Equation 19, where γ is estimated using Equation 24. For UR samples the same estimator is used except that $\gamma = 0$.

Tables 6 and Table 7 show the standard errors for term population and document population, respectively. For term population size, the estimations may be infinite for small sample sizes 500, 1000 and 1500 due to zero conflict (C=0). For document population size, the smallest sample size is 2500, no longer 500 as in other experiments because small size may not induce collision in both sampling methods. Correspondingly the number of repetitions of the tests in this experiment is reduced to 20.

In both term and document population size estimations, RW works better than UR in terms of standard error, even for the document size estimation where CV is not large. This is because for the same sample size n RW has larger expected

Table 6 Population size of terms. Data obtained from 200 runs for various sample size n ranging between 500 and 5000.

$n(\times 100)$	Standard error($\times 10^5$)	
	UR	RW
5	NaN	2.1133
10	NaN	1.4647
15	NaN	1.1659
20	2.5481	0.9953
25	1.7894	0.8832
30	1.3004	0.8545
35	1.1187	0.7975
40	0.9463	0.7256
45	0.8208	0.6906
50	0.6969	0.6634

Table 7 Population size of documents. Data obtained from 20 runs for various sample size n ranging between 2500 and 25000.

$n(\times 100)$	Bias($\times 10^5$)		Standard error ($\times 10^5$)	
	UR	RW	UR	RW
25	3.7417	1.8063	8.8771	9.7239
50	1.2461	-0.2711	3.3212	1.1986
75	0.3536	-0.2045	1.6406	1.0799
100	0.0421	-0.1480	1.0858	0.7245
125	-0.1507	-0.0905	0.7640	0.5768
150	0.0022	-0.0722	0.7331	0.5832
175	0.0733	-0.0415	0.7017	0.5386
200	0.0573	0.0055	0.7009	0.4784
225	0.0358	0.0058	0.6003	0.4295
250	0.0164	0.0493	0.5012	0.3994

collision, therefore smaller relative variance. In addition, RW needs smaller sample size to produce non-infinite estimations. In UR sampling the sample size needs to be greater than $\sqrt{2N}$ [4] so that collisions can occur and the estimate is not infinite. For random walk sampling, large documents have higher probability of being visited, thus smaller sample size can also induce collisions.

6 Discussions

This paper shows that the biased sampling can be better than uniform sampling. In the past, people try to obtain uniform samples whenever possible, and resort to biased sampling such as PPS (Proportional To Size) sampling only when uniform sampling is impossible [44] or costly. The results of this paper suggest that in the context of hidden data sources, random walk sampling instead of uniform sampling should be used, even when uniform random samples are readily accessible.

We explain this using average term degree estimation as an example. The sample distributions of the degrees are depicted in Figure 4. Panel A is the degree

[4] by Equation 38 when $\gamma = 0$.

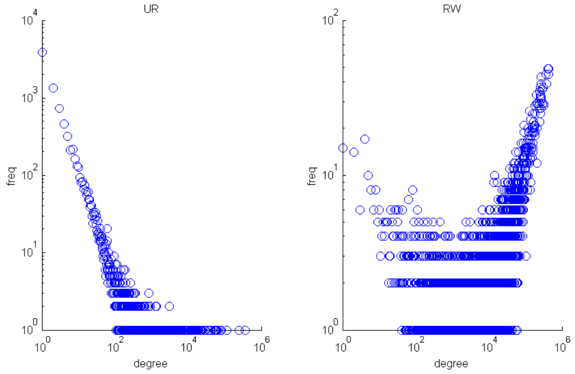


Fig. 4 The term degree distributions of the samples obtained from uniform random and random walk samplings. n=10,000.

distribution for uniform random sample, which resembles the distribution of the population as expected. Panel B describes the distribution obtained from random walk sampling. The ”Y” shape plot shows that the small terms still follow power law roughly, while the large terms, the popular words, can be sampled many times. In other words, both small terms and large terms are sampled multiple times but for different reasons. Rare terms are sampled because there are large number of them, even though each term has a very small probability of being sampled. Large terms are sampled because they have higher probability of being visited, even though there are only a few of them. Unlike uniform random samples where some types of terms, especially the very popular words, are included by chance, in random walk samples both rare words and popular words are well represented in the sample.

We elaborate this point further using a simplified fictitious example to gain an intuitive understanding of the method. Instead of the full spectrum of the degrees, we assume a polarized scenario that contains only two kinds of nodes–one million of small nodes with degree one and ten large nodes with degree one million. This mimics the scale-free graph that has many small nodes and a few very large nodes. Suppose that the sample size n is 10^4 (1% of the population). In both UR and RW sampling, the expected estimations are close to 11, the true value.

In UR sampling, the probability of a node being visited is $p \approx 1/10^6$ when one sample node is taken. When $n = 10^4$ samples are taken, the number of times a node is sampled follows binomial distribution B(n,p) whose expectation is $np = 10^{-2}$. Thus the expected number of small nodes being sampled is $10^{-2} \times 10^6 = 10^4$, the expected number of large nodes being visited is $10^{-2} \times 10 = 0.1$. However, we can not have 0.1 number of node. Instead, most probably a large node is sampled zero or one time. Either case the estimation deviates from the real mean greatly as shown in Table 8. In Case 1 the large node is not sampled, resulting in the estimation 10. In Case 2 the large node is sampled once, resulting in the estimation 100 which is way larger than the expectation 11.

On the other hand, in RW sampling the probability of a node being sampled is proportional to its degree. For a small node, the probability being sampled when one sample is taken is $p_s = 1/(11 \times 10^6)$. The probability of a large node being

Table 8 Illustrative example where the graph contains 10^6 small nodes whose degree is one, and 10 large nodes whose degree is 10^6 . Sample size $n = 10^4$.

		Node type		
		Small (d=1)	**Large** ($d = 10^6$)	**Tota**l
Data	N	10^6	10	$\sim 10^6$
	τ	10^6	10^7	11×10^6
	$\langle d\rangle$			~ 11
UR	E(n)	10^4	0.1	10^4
	$E(\sum d_i)$	10^4	10^5	11×10^4
	$E(\langle d\rangle)$			~ 11
Case 1	n	10^4	0	10^4
	$\sum d_i$	10^4	0	10^4
	$\widehat{\langle d\rangle}_{SM}$			$\sim$ **10**
Case 2	n	9999	1	10^4
	$\sum d_i$	9999	10^6	$\sim 10^6$
	$\widehat{\langle d\rangle}_{SM}$			$\sim$ **100**
RW	E(n)	909	9091	10000
	$E(\sum 1/d_i)$	909	0.01	909
	$E(\langle d\rangle)$			$\sim$ **11**

sampled is 10^6 times larger, i.e., $p_l = 10^6/(11 \times 10^6)$. Thus the expected number of times a small node being sampled is $np_s = 1/(11 \times 10^2)$, the expected total number of small nodes being sampled is $10^6/(11 \times 10^2) = 909$. The expected number of large nodes being sampled is $10 \times np_l = 9091$. Since the expected values are way larger than one, the estimates will not deviate a lot from the expected values.

7 Conclusions

This paper tackles the estimations of the average degree, the degree heterogeneity, and the population size in hidden web data sources. We show that the three proposed estimators are dependent on each other– population size is dependent on the heterogeneity, and in turn the heterogeneity relies on the average degree. Such decomposition of the estimation problem has not only the pedagogical significance, but more importantly, a large problem is divided into two smaller ones, and each sub-problem can be approached with different methods, not necessarily by random walk.

The highlight of the paper is not the random walk estimators. Rather, it is the comparison with the uniform random sampling. It shows that when the data follow Zipf's law, the variance of the UR method diverges with the corpus size, while the variance of RW sampling grows logarithmically. In real graphs with moderate high CV such as term degrees in Reuters, the RW method is already much better than the UR samples, let alone the high cost of obtaining those uniform samples. In [32] we show that with higher CV and larger graphs, RW is orders of magnitude better than uniform random samples.

This paper shows that the behaviour of the RW method depends on the heterogeneity of the data. For term degrees whose variance is large (CV=16), the RW method has a big lead over the UR method. For document degrees where the variance is small (CV=0.7), the RW method is slightly worse than the UR samples if the cost of random sampling is excluded. In ecology studies the data reported usually have small heterogeneity whose γ^2 is around one. In our big data γ^2 is in the scale of hundreds or thousands. This big difference entails new methods that are drastically different from the traditional estimators such as the ones developed in [10].

For the population size estimation, RW is better than UR for both terms and documents in two perspectives. One is that in UR sampling the sample size needs to be greater than $\sqrt{2N}$ so that collisions can occur and the estimate is not infinite. For RW sampling, since large documents have higher probability of being visited in the random walk, smaller sample size can also induce collisions, consequently produce non-infinite estimates. Secondly, the standard error of random walk is smaller than that of the uniform sampling because the expected collisions are larger in RW.

8 Acknowledgements

The authors thank Dingding Li for helpful discussions. The work is supported by NSERC (Natural Sciences and Engineering Research Council of Canada) and State Key Laboratory for Novel Software Technology at Nanjing University.

References

1. S. Amstrup, T. McDonald, and B. Manly. *Handbook of capture-recapture analysis.* Princeton Univ Press, 2005.
2. Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. In *Proceedings of the 15th international conference on World Wide Web*, pages 367–376, Edinburgh, Scotland, 2006. ACM.
3. Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. *Journal of the ACM*, 55(5):1–74, 2008.
4. Z. Bar-Yossef and M. Gurevich. Efficient search engine measurements. *ACM Transactions on the Web (TWEB)*, 5(4):1–48, 2011.
5. M. K. Bergman. White paper - the deep web: Surfacing hidden value. *Journal of Electronic Publishing*, 7(1), 2001.
6. K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. *Comput. Netw. ISDN Syst.*, 30(1-7):379–388, 1998.
7. A. Broder and et al. Estimating corpus size via queries. In *CIKM*, pages 594–603. ACM, 2006.
8. J. Callan and M. Connell. Query-based sampling of text databases. *ACM Trans. Inf. Syst.*, 19(2):97–130, 2001.
9. J. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. *SIGMOD Rec.*, 28(2):479–490, 1999.
10. 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.
11. W. Cochran. *Sampling techniques.* John Wiley, 1977.
12. J. Darroch. The multiple-recapture census: I. estimation of a closed population. *Biometrika*, 45(3/4):343–359, 1958.
13. A. Dasgupta, G. Das, and H. Mannila. A random walk approach to sampling hidden databases. In *SIGMOD*, pages 629–640. ACM, 2007.
14. 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.

15. 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.
16. M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1872–1892, 2011.
17. A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903. ACM, 2005.
18. 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.
19. M. Hansen and W. Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 1943.
20. M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. *Computer Networks*, 33(1-6):295–308, 2000.
21. P. G. Ipeirotis, L. Gravano, and M. Sahami. Probe, count, and classify: categorizing hidden web databases. In *SIGMOD*, pages 67–78. ACM, 2001.
22. L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *WWW*, pages 597–606. ACM, 2011.
23. M. Kurant, A. Markopoulou, and P. Thiran. Towards unbiased bfs sampling. *Selected Areas in Communications, IEEE Journal on*, 29(9):1799–1809, 2011.
24. S. Lawrence and C. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998.
25. S. Lawrence and C. L. Giles. Searching the world wide web. *Science*, 280(5360):98–100, Apr. 1998.
26. J. Leskovec and C. Faloutsos. Sampling from large graphs. In *SIGKDD*, pages 631–636. ACM, 2006.
27. J. Liu. *Monte Carlo strategies in scientific computing.* Springer Verlag, 2008.
28. L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.
29. J. Lu. Efficient estimation of the size of text deep web data source. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1485–1486. ACM, 2008.
30. J. Lu. Ranking bias in deep web size estimation using capture recapture method. *Data & Knowledge Engineering*, 69(8):866–879, 2010.
31. J. Lu and D. Li. Estimating deep web data source size by capture–recapture method. *Information Retrieval*, 13(1):70–95, 2010.
32. 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.
33. J. Lu and D. Li. Bias correction in small sample from big data. *TKDE, IEEE Transactions of Knowledge and Data Engineering, in Press*, 2013.
34. J. Lu, Y. Wang, J. Liang, J. Chen, and J. Liu. An approach to deep web crawling by sampling. In *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, volume 1, pages 718–724. IEEE, 2008.
35. J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. Google's deep web crawl. *Proc. VLDB Endow.*, 1(2):1241–1252, 2008.
36. 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.
37. M. Montemurro. Beyond the zipf–mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3):567–578, 2001.
38. M. Newman. *Networks: an introduction.* Oxford University Press, Inc., 2010.
39. C. Olston and M. Najork. Web Crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010.
40. M. Papagelis, G. Das, and N. Koudas. Sampling online social networks. *Knowledge and Data Engineering, IEEE Transactions on*, (99):1–1, 2011.
41. S. Raghavan and H. Garcia-Molina. Crawling the hidden web. In *VLDB*, pages 129–138. Morgan Kaufmann Publishers Inc., 2001.
42. A. 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.
43. T. Reuters. Reuters coprus. http://about.reuters.com/researchandstandards/corpus/, December 2008.

44. M. Salganik and D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1):193–240, 2004.
45. M. Shokouhi and L. Si. *Federated search.* Now Publishers, 2011.
46. 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.
47. L. Si, R. Jin, J. Callan, and P. Ogilvie. A language modeling framework for resource selection and results merging. In *Proceedings of the 11th CIKM*, pages 391–397. ACM, 2002.
48. S. Thompson. *Sampling.* Wiley, 2012.
49. Y. Wang, J. Lu, J. Liang, J. Chen, and J. Liu. Selecting queries from sample to crawl deep web data sources. *Web Intelligence and Agent Systems*, 10(1):75–88, 2012.
50. C. Wejnert and D. Heckathorn. Web-based network sampling. *Sociological Methods & Research*, 37(1):105–134, 2008.
51. P. Wu, J. Wen, H. Liu, and W. Ma. Query selection techniques for efficient crawling of structured web sources. In *ICDE*. IEEE, 2006.
52. 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.
53. 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.
54. J. Zhou, Y. Li, V. Adhikari, and Z. Zhang. Counting youtube videos via random prefix sampling. In *SIGCOMM*, pages 371–380. ACM, 2011.
55. G. Zipf. Human behavior and the principle of least effort. 1949.

9 Appendix

Both Theorem 1 and Theorem 2 assume that the degrees follow the Zipf's-Mandelbrot law [37] which states that if the term degrees d_i are sorted in descending order, then

$$d_i = \frac{A}{\alpha + i}, \tag{26}$$

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

$$\sum_1^N d_i \approx \int_1^N \frac{A}{\alpha + x} dx \approx A \ln(\frac{\alpha + N}{\alpha + 1}) = A \ln B = \tau, \tag{27}$$

where we use $B = (\alpha + N)/(\alpha + 1)$ to make our derivations more concise. Therefore 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:

$$\sum_1^N d_i^2 \approx \int_1^N \frac{A^2}{(\alpha + x)^2} dx \approx \frac{A^2}{\alpha + 1}. \tag{28}$$

9.1 Proof of Theorem 1

Proof Based on Equations 27 and 28, the variance of all the degrees is

$$\begin{aligned} \sigma^2 = \langle d^2 \rangle - \langle d \rangle^2 &= \langle d \rangle^2 \left[N \frac{\sum_1^N d_i^2}{(\sum_1^N d_i)^2} - 1 \right] \\ &\approx \langle d \rangle^2 \left[\frac{N}{(\alpha + 1) \ln^2 B} - 1 \right]. \end{aligned} \tag{29}$$

Using Equation 14 the variance of $\widehat{\langle d \rangle}_{SM}$ is

$$var(\widehat{\langle d \rangle}_{SM}) = \frac{\langle d \rangle^2}{n}\left[\frac{N}{(\alpha+1)\ln^2 B} - 1\right]. \tag{30}$$

9.2 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 [48]:

$$\widehat{N}_H = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_{x_i}} = \frac{\tau}{n}\sum_{1}^{n}\frac{1}{d_{x_i}}, \tag{31}$$

and the variance of $\widehat{N_H}$ is [48]:

$$var(\widehat{N_H}) = \frac{1}{n}\sum_{i=1}^{N}p_i\left(\frac{1}{p_i} - N\right)^2. \tag{32}$$

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

$$var(\widehat{N_H}) = \frac{1}{n}\left(\frac{\tau}{A}\sum_{1}^{N} i - N^2\right) \approx \frac{N^2}{n}\left(\frac{\ln B}{2} - 1\right). \tag{33}$$

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

$$\widehat{\langle d \rangle}_H = \frac{\tau}{\widehat{N_H}} = \tau\left(\frac{1}{N} - \frac{\widehat{N_H} - N}{N^2} + \ldots\right). \tag{34}$$

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

$$var(\widehat{\langle d \rangle}_H) = \tau^2\frac{var(\widehat{N_H})}{N^4} = \frac{\langle d \rangle^2}{n}\left(\frac{\ln B}{2} - 1\right). \tag{35}$$

9.3 Population size estimation

Nodes are selected during random walk. 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 binomial distribution $B(n(n-1)/2, p)$ whose mean is

$$E(C) = \binom{n}{2}p. \tag{36}$$

The collision probability p can be translated into the heterogeneity of the data measured by γ using the definition of γ in Equation 12 :

$$p = \sum_{i=1}^{N} p_i^2 = \frac{1}{\tau^2} \sum_{i=1}^{N} d_i^2 = \frac{1}{N} \frac{\langle d^2 \rangle}{\langle d \rangle^2} = \frac{1}{N}(\gamma^2 + 1). \tag{37}$$

Combining Equations 37 and 36 we obtain the expected number of collisions is:

$$E(C) = \binom{n}{2} \frac{\gamma^2 + 1}{N}. \tag{38}$$

Hence the population size can be described by

$$N = (\gamma^2 + 1) \binom{n}{2} \frac{1}{E(C)}. \tag{39}$$

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

$$\widehat{N} = (\gamma^2 + 1) \binom{n}{2} \frac{1}{C}. \tag{40}$$