2013

Estimating Deep Web Properties by Random Walk

Sajib Kumer Sinha
University of Windsor

Estimating Deep Web Properties by Random Walk

by

Sajib Kumer Sinha

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

Windsor, Ontario, Canada
2013

Estimating Deep Web Properties by Random Walk

by

Sajib Kumer Sinha

APPROVED BY:

Dr. Yunbi An, External Reader
Odette School of Business

Dr. Dan Wu, Internal Reader
School of Computer Science

Dr. Jianguo Lu, Advisor
School of Computer Science

Dr. Subir Bandyopadhyay, Chair of Defense
School of Computer Science

May 24, 2013

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.
I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis.
I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

The deep web is the part of World Wide Web that is hidden under form-like interfaces and can be accessed by queries only. Global properties of a deep web data source such as average degree, population size need to be estimated because the data in its entirety is not available. When a deep web data source is modelled as a document-term bipartite graph, the estimation can be performed by random walks on this graph. This thesis conducts comparative studies on various random walk sampling methods, including Simple Random Walk (SRW), Rejection Random Walk (RRW), Metropolis-Hastings Random Walk (MHRW) and uniform random sampling. Since random walks are conducted by queries in searchable interfaces, our study has focused on the overall sampling cost and the estimator performance in terms of bias, variance and RRMSE in this particular setting. From our experiments performed on Newsgroup data we find that MHRW results higher variance and RRMSE especially when the degree distribution follows the power law. On the other hand RRW performs worse in terms of query cost as it rejects too many samples. Compared to MHRW and RRW, SRW has low variance and RRMSE. Besides, SRW outperforms the real uniform random samples when the distribution follows the power law.

Dedication

To my mother Rina Sinha, father Ranada Prasad Sinha and my loving sister Susmita Sinha for their unconditional love, support and endless encouragements. Also to the workers of my motherland Bangladesh for their contribution to my country...

Acknowledgements

Firstly, I would like to thank my parents. Without their encouragement, support and love, it would not be possible for me to reach here.

I would like to express my deepest appreciation to my advisor, Dr. Jianguo Lu, for his motivation, support and invaluable suggestions in guiding me towards the successful completion of this research work. Without his generous funding and advice, it would be hard for me to finish this research.

I would like to express my great gratitude to Dr. Yunbi An, Odette School of Business, and Dr. Dan Wu, School of Computer Science for giving me corrections and constructive criticism to improve the quality of this research, for their patience in arranging the time of my proposal and defense, and for being in the committee, and Dr. Subir Bandyopadhyay for serving as the chair of the committee.

I gratefully acknowledge the assistance of Md. Rajibul Mian, Gunjan Soni and Numanul Subhani for their helpful discussion and support.

Finally, I want to extend my gratitude to my friends, the faculty members and staff of the School of Computer Science for their friendly suggestions and support during my study at University of Windsor.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Deep Web

The deep web [9] is the part of WWW which have no specific hyper-links to extract and are not indexable by the search engines. These are the pages which are generated dynamically from the back-end data sources and can be extracted only by its search interfaces. All dynamic pages behind the search engines, content without in-links, limited access content, scripted content, contextual web are part of the deep web.For example, The Leddy Library site of university of Windsor, Google's index, The New York times site, all are part of the deep web resources. Deep web properties estimation and evaluation such as size, degree distribution, corpus freshness evaluation, spam evaluation, security evaluation and many more are buzzing issues for many researchers and organizations [28]. Besides deep web properties are important parameters of many more algorithms in distributed Information Retrieval system [12, 41]. In real application web marketing is a major concern for all business organi-

zations and these kinds of deep web analysis can be more beneficial for those marketing people to determine the importance and influence of a particular web source in the real market. For example, size estimation can help to determine which online library is rich with books, which social network covers the maximum individuals, which search engine corpus is more updated and content rich, which blogger is more influential upon the society.

When deep web is represented as a document-term graph, degree distribution, average degree, etc. are of great interest to the researchers to estimate other properties such as population size. But calculating average degree is not that straightforward as the deep web data in its entirety is not accessible and much larger than the surface web. Besides, it is not efficient to crawl and determine its different properties mentioned above as the size of the deep web itself is an important parameter for the deep web crawler and extractor [23, 34, 16]. Moreover, we have the issues of network bandwidth, we have limited number of permitted queries over a search interface, limited number of access from a certain IP and many more. As a result estimation is needed to determine those properties. Hence, the sampling comes into consideration which is very popular regarding this matter.

1.2 Deep Web Sampling

Sampling is a statistical technique in which a small part of large population has been selected to estimate some properties of the whole population [42].

The selection of samples depends on the sampling design. There are lot of sampling techniques in use for different estimation process in different areas. As we have black box access to the deep web data [7] via its publicly available search interface, query based sampling [13] is required.

Query based sampling was first proposed by Callan et al [13] for acquiring resource description of databases. Resource description mainly consists of vocabulary and frequency information [12]. In query based sampling a query term needs to be submitted to the search interface and samples can be obtained randomly from the matched documents. Here matched document refers to the document which contains that submitted query term. Based on the probability of a node to be sampled, sampling techniques can be categorized into two different categories called Probability Proportion to Size (PPS) and Uniform Random(UR) sampling.

In PPS sampling probability of being sampled is proportional to its size, what means larger documents or more frequent terms are more likely to be sampled. In contrast, for UR sampling each document will have the equal probability to be sampled. In this research average degree of documents $\langle deg^d \rangle$ and terms $\langle deg^t \rangle$ will be estimated, which can be helpful to derive whole population size and degree variance.

Assume we have N number of documents with their corresponding degree deg_i^d where $i \in \{1,2,3... \ N\}$. In this case the average degree is

$$\langle deg^d \rangle = \frac{1}{N} \sum_{i=1}^{N} deg_i^d \tag{1.1}$$

One straightforward way of estimation is via Uniform Random (UR) sampling where each document or term has the equal probability to be sampled.

Hence the average of the total population can be estimated by the arithmetic mean of the obtained samples, which can be called as sample mean estimator $\widehat{\langle deg \rangle}_{SM}$. For n number of UR document samples $\{d_1, d_2, d_3, ...d_n\} \in D$ with corresponding degree deg_i^d, the sample mean estimator will be

$$\widehat{\langle deg^d \rangle}_{SM} = \frac{1}{n} \sum_{i=1}^{n} deg_i^d \tag{1.2}$$

The $< \widehat{deg^d} >_{SM}$ is unbiased if samples are truly uniform random. But using simple query based sampling UR samples can not be obtained for the heterogeneity of the degree which rather gives PPS samples. To overcome the issue of heterogeneity various Monte Carlo simulation methods such as rejection sampling, Metropolis Hasting algorithm, importance sampling, maximum degree method have been used in different areas including search engine index [11, 7, 8], surface web [22], graphs [30], online social network [17, 37], real social network [40, 44], etc. But these methods are not always efficient. Because, Rejection sampling results higher query cost as it rejects too many samples. Also Metropolis Hasting algorithm gets stuck in a smaller portion of a large graph as it remains in the same state because of rejection. Hence biased samples come into consideration.

To estimate average degree from the biased PPS samples harmonic mean estimator is being used by many researchers in deep web properties estimation [33], in social network analysis [18, 26, 32] and peer to peer network analysis [38], which can be derived from Hansen-Hurwitz estimator [20]. This harmonic estimator also has been used in sociology to estimate drug addicts [40]. For n number of PPS document samples $\{d_1, d_2, d_3, ...d_n\} \in D$ with corre-

sponding degree deg_i^d, the harmonic mean estimator will be

$$\widehat{\langle deg^d \rangle}_H = n \left[\sum_{i=1}^{n} \frac{1}{deg_i^d} \right]^{-1} \tag{1.3}$$

As the simple query based sampling such as simple random walk does not have any rejection procedure (rejection of sample or state) it can cover more of the graph with less query cost.

1.3 Thesis Problem and Contribution

Several research have been performed to estimate deep web properties such as average degree. But the question which techniques perform better for the estimation remains unanswered. Besides there was no explicit empirical studies on the cost of those sampling techniques. In this research our problem can be defined as

Given a deep web data source, how to estimate the average degree of the documents and terms using UR and PPS sampling, and which method can be considered as the better one?

To solve this problem, we have experimented and evaluated various sampling techniques, including UR sampling and biased PPS sampling. We have estimated the average degree of documents and terms using both sampling methods. Given the limited access capabilities provided by deep web data sources, UR samples are usually hard to obtain. For obtaining UR samples We have experimented with two UR sampling method called Rejection Random Walk(RRW) and Metropolis Hasting Random Walk (MHRW) on document-term bipartite graph. We observed that MHRW has higher vari-

ance in estimation compared to all other methods. Because, MHRW gets stuck and covers a small part of the graph in each iteration. Hence estimation becomes biased based on that covered area. Also RRW waste too many samples because of its acceptance rejection procedure. Since UR sampling is costly and inefficient, we have also experimented with one PPS sampling method called Simple Random Walk. We have found that biased SRW performs better than the RRW and MHRW for both documents and terms. For better comparison we obtain real UR samples directly from the index and estimate the average degree as well. Here we observe that UR performs better, when the distribution has less heterogeneity. For documents UR performs better as the document degree distribution follow the log normal form and for terms where the degree distribution follows power law SRW outperforms UR. We have also explained the cost of both RRW and MHRW in terms of rejection rate and found sample rejection rate is the average degree of the distribution and state rejection rate of MHRW also dependent on the average degree.

1.4 Thesis Organization

The rest of the thesis is organized as following. Chapter 2 discussed about some major related works that have been performed before. In Chapter 3 we have explained some useful terms which can be handy for our analysis and discussed our all approaches with example. Chapter 4 consists of our experiments and results. Lastly in Chapter 5 we have stated our concluding comments and future works.

Chapter 2

Related Work

Query based sampling for surfacing deep web properties has been studied since the advent of the query based search interfaces. Different research works on query based sampling have been performed with different types of data such as search engines index [4, 2, 11, 7, 8] or relational database tables [14, 15]or social networks [17, 24, 46]. All query based sampling approaches can be divided into two parts called Random Query and Random Walk sampling.

2.1 Random Query based approaches

Random query is a lexicon based approach where query needs to be selected randomly from the lexicon or collection of queries and submitted to the search interface. After that, one random document is being selected as a sample from the matched documents.

Random Query based PPS sampling

Bharat and Broder [10] first realized the necessity of obtaining random pages from a search engine's index for calculating the relative size and overlapping between two search engines. To solve this problem they first introduced the lexicon based approach where conjunctive and disjunctive queries are being generated randomly from the lexicon and run in the public interface of the Search engine.

In the same year another influential research had been performed by Lawrence and Giles [28] where they have estimated the size of search engines by random query selected from user query logs. But both of these methods were biased towards content rich highly ranked documents as both of these methods are PPS sampling.

Callan and Connell [12] have proposed a query based sampling algorithm for acquiring resource description of the relational databases based on the concept of Bharat and Broder [10]. Here initially one query term is selected at random and run on the database. Based on the returned top N number of documents resource description is updated. This process run many times based on different query terms until stop condition reached. This algorithm also have the option of choosing number of query terms, how many documents to examine per query and the stop condition of the sampling.

Bar-Yossef and Gurevich [6] first used importance sampling method with PPS samples which is more similar to ours approach. The authors define two new estimators called Accurate Estimator and Efficient Estimator to estimate the target distribution of the sample documents based on the Importance

sampling methods. The Accurate estimator uses approximate weights and it requires sending some queries to a search engine and the Efficient Estimator uses deterministic approximate weights and it does not need any query. They use the Rao-Blackwell theorem with the importance sampling method to reduce estimation variance.

Random Query based UR sampling

In 2006 Bar-Yossef and Gurevich [5] introduced the concept of the Query pool and used one of the Monte Carlo simulation methods called Rejection sampling with random query to obtain uniform random samples from the search engine's index. In their pool based approach they have applied rejection sampling twice. First they applied rejection sampling to select a query from the query pool which overcomes the ranking bias and again they applied rejection sampling to select document from the matched documents which overcomes the degree bias. Because of applying rejection procedure twice their query cost is very large.

After that Broder et al. [11] used the Pool based concept of Bar-Yossef and Gurevich [5] and introduced their new algorithm with new low variance estimator where they have used the concept of traditional Peterson estimator [1]. They have carefully crafted the importance sampling method with naive estimator to reduce the bias. The authors propose two approaches based on a basic low variance and unbiased estimator. Their first method requires a uniform random sample from the underlying corpus and after getting the uniform sample, the corpus size is computed by the basic estimator. For random

sampling they use the rejection sampling method. The second approach is based on two query pools where both pools are uncorrelated with respect to a set of query terms. Next, using these two query pools, the corpus size is estimated using the low variance estimator that taken into account.

2.2 Random Walk based approaches

In a random walk we start with a seed node and in each step it moves to its neighbour at random with equal probability. Note that here query is being selected randomly from the current document of the walk instead of selecting from a predefined lexicon. More detail about random walk will be explained in Chapter 3.

Random Walk based PPS sampling

Henzinger et al have proposed Multi thread crawler to estimate various properties of web pages. They do not introduce any new method rather they give some suggestion to improve sampling based on random walk. Instead of using normal crawler they suggest to use the Mercator, a multi threaded web crawler. Here each thread will begin with randomly chosen starting point and for random selection they suggest to make a random jump to the pages which are visited by at least one thread instead of following the hyper links.

Following the idea of Henzinger et al [22], Bar-Yossef et al [4] introduce the Web walker to approximate certain aggregate queries about web pages. They have proposed a new random walk process called Web Walker which

performs a regular undirected random walk and picks pages randomly from its traversed pages. Starting page of the Web Walker is an arbitrary page from strongly connected component of the web. But as it is a PPS sampling estimation is biased.

Rusmevichientong et al.[39] proposed two new algorithms based on approach of Henzinger et al [22] and Bar-Yossef et al. [4]. The first algorithm called Directed-Sample works on the arbitrary directed graph and the other one called Undirected-Sample works on the undirected graph with additional knowledge about inbound links which requires access to the search engine. Both of the algorithms based on weighted random-walk methodology.

Lu et al. [33] have used biased PPS samples obtained by SRW to discover the average degree and population size of the deep web. They have also used the harmonic mean estimator with the biased samples to estimate those properties and shown PPS can outperform real UR when the degree heterogeneity is larger.

Random Walk based UR sampling

In 2006, Bar- Yossef and Gurevich [5] used one of the Monte Carlo simulation methods called Metropolis Hastings algorithm in their random walk approach to obtain uniform random samples from the search engine's index. To overcome the ranking bias they have used those queries which neither overflows nor underflows. The detail of the Metropolis algorithm is explained in Chapter 3.

Using the same approach Gjoka et al [17] have obtained uniform random

samples from the social network Facebook. They also have used the Re-Weighted random walk method, which is similar to our method. Here simple random walk bias is corrected by re-weighting of measured values using Hansen-Hurwitz estimator [20].

A rejection sampling based random walk method have been proposed by Dasgupta et al [14] to obtain uniform random sample from hidden web databases. The authors propose a new algorithm called HIDDEN-DB-SAMPLER, which is based on random walk over the query spaces provided by the public user interface. Three new ideas proposed by the authors are early detection of underflow and valid tuples, random reordering of attributes, boosting acceptance probability via a scaling factor. However, they proposed their method for sampling from a database that is hidden behind a form, structured in a particular way which cannot be compared with the deep web general search interfaces.

In all of those approaches, cost of those sampling techniques is not being studied rigorously. However, Bar- Yossef and Gurevich [7] have presented the theoretical cost analysis in their subsequent work which is represented in terms of upper bound and lower bound. Besides it is specific to their experimental set-up and parameters such as query pool, query cardinality ratios, etc. Hence, there is no specific empirical studies to analyse the cost of these deep web sampling techniques , and the question, which sampling method is better for deep web properties estimation remains unanswered.

Chapter 3

Estimation by Random Walk Sampling

3.1 Deep Web as Graph

A Graph G is an ordered pair $G = (V, E)$ consisting of a set of nodes or vertices (V) and a set of edges (E) which connects a pair of vertices, and $V \cap E = \phi$ [3]. A vertex presents in an edge is called end vertex. Note that a vertex might be present in a graph but may not be in any edges. Degree of a vertex x (deg_x) refers to the number of edges that connects x with other vertices. An undirected graph is an unordered pair $G = (V, E)$ where edges have no direction, means for any two nodes a and b, edges $(a \rightarrow b) = (b \rightarrow a)$. In this research only undirected graph will be considered.

Surface web can be represented using a graph where each web page is a vertex and each hyper link is an edge [25]. In contrast, deep web data source can be represented as a document-term bipartite graph containing two dis-

joint sets of vertices where each edge connects two disjoint sets[36, 45, 47]. The graph G can be represented as $G = (D, T, E)$ and $D \cap T = \emptyset$, where D is the set of documents, T is the set of terms and E is an edge between D and T which represents the presence of a term in a document. The degree of each vertex is the number of its adjacent nodes. More precisely document degree of d_i (deg_i^d) is the number of distinct terms that contained by the document d_i and term degree of t_i (deg_i^t) is the number of documents matched when term t_i is being submitted to the search interface.

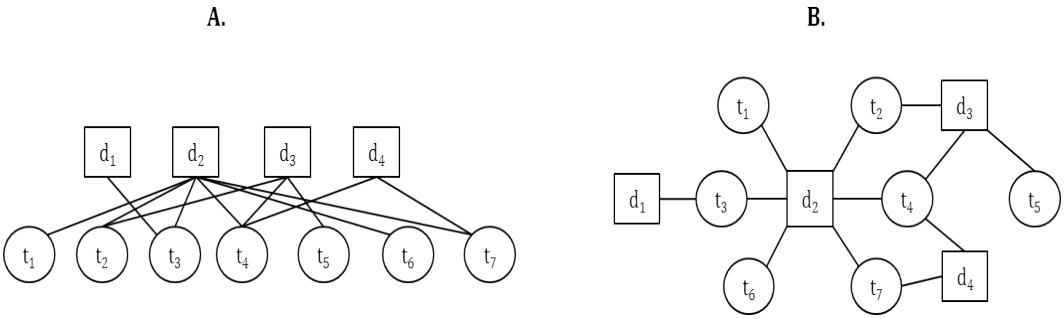

Figure 3.1: A) Deep web source as a bipartite graph B) same graph in spring model

Example 1: Deep Web as bipartite graph A deep web source with 4 documents consisting 7 distinct terms has been depicted in Figure 3.1 where $D = \{d_1, d_2, d_3, d_4\}$ and $T = \{t_1, t_2, ...t_7\}$. An edge $d_1 - t_5$ refers that the term t_5 presents in document d_1. According to the Figure 3.1 degree of documents $deg_1^d = 1$ and $deg_2^d = 6$. Apparently, term degree $deg_1^t = 1$ and $deg_4^t = 3$.

The graph depicted in Figure 3.1 can be represented with an adjacency matrix. An adjacency matrix (A) is a $n \times n$ boolean matrix for n number of nodes where each value of A_{ij} represents the adjacency of nodes i and j. For

a bipartite graph it is called the bi-adjacency matrix which is a $m \times n$ boolean matrix, where m and n represent the number of vertices in two disjoint sets. Bi-adjacency matrix of Figure 3.1 is as following.

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

$|D| \times |T|$ bi-adjacency matrix

A Markov chain is a sequence of nodes or states where the transition of next step is independent of previous or current states. If the current state of a Markov chain is n_i, it will move to an another node n_j with a transition probability p_{ij} which is independent of other nodes[19].

A matrix that represents all transition probabilities is called transition matrix (T) which can be denoted as $T = (p_{ij}) \forall i, j \in V$ where

$$p_{ij} = \begin{cases} \frac{1}{deg_i}, & \text{if ij} \in E \\ 0, & \text{otherwise.} \end{cases} \tag{3.1}$$

Transition matrix of Figure 3.1 will be as following.

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 1/6 \\ 0 & 1/3 & 0 & 1/3 & 1/3 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 1/2 \end{pmatrix}$$

$$|D| \times |T| \text{ transition matrix}$$

A Markov chain is time reversible if the forward and backward edges belongs to same distribution, which means there is a probability distribution π such that $\pi(i)p_{ij} = \pi(j)p_{ji}$ [19]. In terms of uniform distribution transition probability will be equal and Markov chain will be time reversible.

Before proceeding to the Random Walk sampling, some basic definitions and properties of statistics will be explained, which will be helpful for our analysis.

Degree variance σ^2 is the measure of how far all the degrees are spread out from the mean μ, which can be defined as [42]

$$\sigma^2 = \langle deg^2 \rangle - \langle deg \rangle^2 \tag{3.2}$$

For our estimation variance can be defined as following

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (\widehat{\langle deg \rangle}_i - \widehat{\langle deg \rangle})^2 \tag{3.3}$$

Standard error (SE) is the square root of the variance which is as following.

$$SE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\widehat{\langle deg \rangle}_i - \widehat{\langle deg \rangle})^2} \tag{3.4}$$

The coefficient of variation γ is the ratio of the standard deviation and the mean. Standard deviation is nothing but the square root of variance or the SE. The γ can be expressed as following

$$\gamma^2 = \frac{\sigma^2}{\langle deg \rangle^2} = \frac{\langle deg^2 \rangle}{\langle deg \rangle^2} - 1 \tag{3.5}$$

A graph is said to be regular when each vertex has equal degree and when graph is regular $\gamma = 0$.

Bias of an estimation $\hat{x}$ is defined as

$$Bias(\hat{x}) = \mathbb{E}(\hat{x}) - x \tag{3.6}$$

Here $\mathbb{E}$ is the expectation of x, which represents the mean of all possible values of x. When the number of possible value is large it can be approximated using the sample mean. For example if we want to find the bias of estimated average document degree ($\widehat{\langle deg^d \rangle}$), Bias will be as following.

$$\begin{aligned} Bias(\widehat{\langle deg^d \rangle}) &= \mathbb{E}(\widehat{\langle deg^d \rangle}) - \langle deg^d \rangle & (3.7) \\ &= \frac{1}{N} \sum_{i=1}^{N} \widehat{\langle deg_i^d \rangle} - \langle deg^d \rangle. & (3.8) \end{aligned}$$

For evaluation of our estimation we have also used the Relative Rooted MSE (RRMSE) which can be defined as following.

$$RRMSE(\widehat{\langle deg \rangle}) = \frac{1}{\langle deg \rangle} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\langle deg \rangle}_i - \langle deg \rangle\right)^2} \tag{3.9}$$

RRMSE is nothing but the RMSE normalized by the mean and RMSE can be derived from the bias and variance as following.

$$RMSE^2 = Bias^2 + var \tag{3.10}$$

3.2 Random Walk on a Graph

A random walk is a time reversible finite Markov chain [31] which proceeds by stepping forward to a neighbouring node from a current node on a given graph. After n number of successful steps it returns n number of samples which are elements of a Markov chain. A random walk on graph depicted in Figure 3.1 with initial node t_1 can give a sample output as $d_2 - t_2 - d_3 - t_5 - d_3 - t_4$. The

changing of nodes in each steps depends on the mechanism of the random walk. Three different random walk called Simple Random Walk(SRW), Rejection Random Walk (RRW) and Metropolis-Hastings Random Walk (MH-RW) have been explained in the next subsections.

3.3 Simple Random Walk (SRW) sampling

In a Random Walk process, for a given graph (G) and an initial node (n_0), in each $t + 1^{th}$ step one neighbouring node (n_{t+1}) of current node (n_t) is being selected with equal probability $\frac{1}{deg_{n_t}}$. A simple random walk on the document-term bipartite graph can be described as following. First, a valid term t_0 is being selected from a lexicon as a seed query to initiate the random walk. Note that a term is called valid if it is being matched with at least one document while submitted to search interface. Next, one of the neighbouring nodes of t_0 will be selected randomly with equal probability $1/{deg_i}^t$, which will be a document d_i. After that another neighbour of document d_i will be selected randomly with equal probability $1/{deg_i}^d$. Thus these processes will continue until n number of samples are being obtained. Note that, to obtain a neighbouring node of a term we need to submit that term in the search interface and one matched document needs to be taken randomly. On the other hand to obtain a neighbouring node of a document we need to download that document and one term need to be selected randomly from that document. Complete algorithm can be represented as Algorithm 1.

During this random walk only one query need to be submitted to the

Algorithm 1 Simple Random Walk (SRW) sampling

Input : t_0 = seed term, sample size n.
Output: Set of n number of document samples D_s and term samples T_s with their corresponding degrees.

$D_s = T_s$ = empty lists;
$i = 1$;
d_i = select one neighbouring node of t_0 with equal probability;
while $i \leq n$ **do**
    add d_i and its degree ${deg_i}^d$ to D_s;
    t_i = select one neighbouring node of d_i with equal probability;
    add t_i and its degree ${deg_i}^t$ to T_s;
    d_{i+1} = select one neighbouring node of t_i with equal probability;
    $i++$;
end
return D_s and T_s;

search interface and one document need to be downloaded to obtain each sample. For selecting random document from the matched documents we do not required downloading all matched documents. For simplicity we have assumed all matched documents are being returned by the search interface. So, using the search interface we can get all matched documents with their URL. In our algorithm we store all matched documents ID in a list of length m (number of matched documents) and next, we generate a random number r between 1 to m and get the ID of $r-th$ document from the list and download. In our algorithm one document can be visited multiple times. In other words it is a sampling with replacement.

After obtaining samples by SRW we will estimate the average degree using the harmonic mean estimator which has been defined in Equation 1.3. One work through example of whole process is given below.

Example 2: SRW sampling and estimation

according to the graph depicted in Figure 3.1 for a seed term t_2 an output of

a simple random walk sampling can be as following.

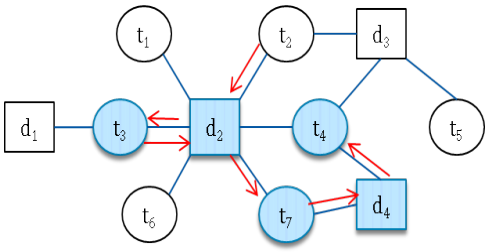

Figure 3.2: An example of SRW sampling

Input: graph depicted in Figure 3.1, $t_0 = t_2$ and $n = 3$

Process: This walk will starts with the seed node t_2. Next, it will select and move to one of its neighbours with equal probability $1/deg_2^t = 1/2$. Assume it selects and moves to d_2. Hence,d_2 will be added as a document sample. In the next step, walk will select and move to another neighbour of d_2 with equal probability 1/6. Assume it selects and moves to t_3. Therefore, it will add t_3 as a term sample. This process will continue until it obtains 3 samples. One possible walk on the given graph is $d_2 - t_3 - d_2 - t_7 - d_4 - t_4$ and has been shown in Figure 3.2. The output of this algorithm can be as following format.

Output:

$T_s = \{(t_3, 2), (t_7, 2), (t_4, 3)\}$

$D_s = \{(d_2, 6), (d_2, 6), (d_4, 2)\}$

After obtaining Document and term samples, our next task is average degree estimation using the harmonic mean estimator. If we consider the document degree, the real average document degree of given graph will be

$$\langle deg^d \rangle = \frac{1+6+3+2}{4} = 3.0 \tag{3.11}$$

If we estimate $\langle deg^d \rangle$ using sample mean which will be biased towards higher degree as following.

$$\widehat{\langle deg^d \rangle}_{SM} = \frac{6+6+2}{3} = 4.66 \tag{3.12}$$

Hence we will estimate the average degree using harmonic mean estimator which basically reduce the bias of higher degree as following.

$$\widehat{\langle deg^d \rangle}_{H} = \frac{3}{\frac{1}{6}+\frac{1}{6}+\frac{1}{2}} = 3.6 \tag{3.13}$$

The advantage of the SRW sampling is all parts of the graph can be traversed regardless of the graph properties such as γ or the graph shape. Because, algorithm always selects and moves to one of the neighbours of current node with equal probability. But, SRW is a PPS sampling.

3.4 Rejection Random Walk (RRW) sampling

RRW is a UR sampling procedure which applies the rejection sampling [43] on random walk. Rejection sampling is the most classical and popular Monte Carlo simulation methods which uses the acceptance-rejection procedure. Assume in a space u, π is the target distribution which is hard to be sampled directly and p is the trial distribution which is easy to be sampled. Note that a sample space refers to all possible outcomes of a random trial or experiment and a probability distribution is a function that specifies the probability of each of the possible outcomes of a random experiment [27]. In this case a Monte Carlo simulation method is a procedure which takes sample from p in order to generate samples from π.

Rejection sampling method requires three main procedures [7]. The first

procedure generate samples from the trial distribution p, such as document $d_1, d_2, d_3...d_n$. In our case trial distribution is the degree distribution from where we can easily obtain sample by submitting query. The two other procedures are used to calculate the unnormalized forms of that particular sample on the target distribution ($\hat{\pi}$) and trial distribution ($\hat{p}$) respectively. An unnormalized form $\hat{p}(x)$ or $\hat{\pi}(x)$ refers to the relative weight which reflects the probability of element x to be sampled from that particular distribution [7]. As π is considered as the uniform distribution, the relative weight will be uniform. Hence, for all $x \in u$ straightforward unnormalized form of π is 1. On the other hand unnormalized form of p is nothing but the $deg(x)$, as in degree distribution the sampling probability is proportional to its degree.

According to the rejection sampling procedure it will repeatedly generates samples from the trial distribution p unless it is being accepted by the acceptance function $a(x)$ as following.

$$a(x) = \frac{\hat{\pi}(x)}{C\hat{p}(x)} \tag{3.14}$$

Here C is a known envelope constant where $\forall x \in supp(p)$, $C \geq max\frac{\hat{\pi}(x)}{\hat{p}(x)}$ and $supp(p) = \{x \in u | p(x) > 0\}$. Note that for UR sampling C can be taken as 1 to satisfy the envelop condition.

Now for obtaining UR samples $\hat{\pi}(x) = 1$, $p(x) = deg(x)$ and $C = 1$, the acceptance function can be simplified as following.

$$a(x) = \frac{1}{deg_x} \tag{3.15}$$

Hence, $a(x)$ is the probability of a document x to be sampled and due to the property of envelope constant, for all $x \in supp(p)$, $a(x) \in [0, 1]$. By using

this acceptance function which is inversely proportional to the degree, is actually reducing the acceptance probability of documents with higher degree and increasing the acceptance probability of documents with lower degree. Eventually, rejection sampling uses the acceptance-rejection procedure to bridge the gap between p and π. The efficiency of this sampling method depends on the similarity between p and π [7]. More similarity between p and π makes less rejection. Also the gap between C and $max\frac{\hat{\pi}(x)}{\hat{p}(x)}$ is crucial for the efficiency. A high value of C makes more rejection and very low value can violates the envelope property.

Algorithm 2 Rejection Random Walk (RRW) sampling

Input : t_0 = seed term, sample size n.
Output: Set of n number of document samples D_s and term samples T_s with their corresponding degrees.

$D_s = T_s =$ empty lists;
$i = 1$;
$d_i =$ select one neighbouring node of t_0 with equal probability;
while $D_s.size < n$ *OR* $T_s.size < n$ **do**
    if *Accept(d_i)* **then**
        add d_i and its degree ${deg_i}^d$ to D_s;
    end
    $t_i =$ select one neighbouring node of d_i with equal probability;
    if *Accept(t_i)* **then**
        add t_i and its degree ${deg_i}^t$ to T_s;
    end
    $d_{i+1} =$ select one neighbouring node of t_i with equal probability;
    $i++$;
end
return D_s and T_s;

A rejection sampling based random walk named Rejection Random Walk (RRW) sampling is given in Algorithm 2. The basic procedure of RRW sampling algorithm is similar to SRW sampling. Likewise SRW, in each step it

Algorithm 3 Accept-RR

Input : d_i OR t_i
Output: $true$ OR $false$
$size$ = degree of the document (${deg_i}^d$) OR term (${deg_i}^t$);
r = one random number between 1 to $size$;
if $r == 1$ **then**
| return $true$;
else
| return $false$;
end

selects and moves to its neighbour with equal probability. But unlike the SRW sampling, it accepts a document or term with an acceptance probability of $\frac{1}{deg_i}$. Hence, documents or terms with higher degree are getting lower probability to be sampled by the acceptance function. Algorithm 3 is simulating that acceptance probability. To simulate the acceptance probability one random number r is being generated between 1 to deg_i and if r comes up 1, document or term is being accepted as a sample. To calculate the document degree particular document needs to be downloaded and for the term degree certain term need to be submitted to the search interface. Likewise SRW, RRW is also a sampling with replacement. One work through example of the whole RRW sampling is given below.

Example 3: RRW sampling and estimation

According to the graph depicted in Figure 3.1 for a seed term t_2 an output of a rejection random walk sampling can be as following.

Input: graph depicted in Figure 3.1, $t_0 = t_2$ and $n = 3$

Process: This walk will be initiated from the seed node t_2 and will select and move to one of its neighbours with equal probability 1/2. Assume it selects and moves to d_2. But, like SRW sampling it will not accept d_2 as a sample

unless it passes the acceptance test with probability $1/deg_2^d$. In the acceptance test a random number r will be generated between 1 to deg_2^d. assume $r = 3$, so it is going to reject d_2 as a sample and continue walk until it accepts n number of samples. So, after t number of steps, algorithm might not accept t number of samples. One possible walk on the given graph can be $d_2 - t_3 - d_2 - t_7 - d_4 - t_4 - d_3 - t_5 - d_3 - t_2 - d_2 - t_4 - d_3$ and has been shown in Figure 3.3. Red sign is indicating the rejection on the same figure.

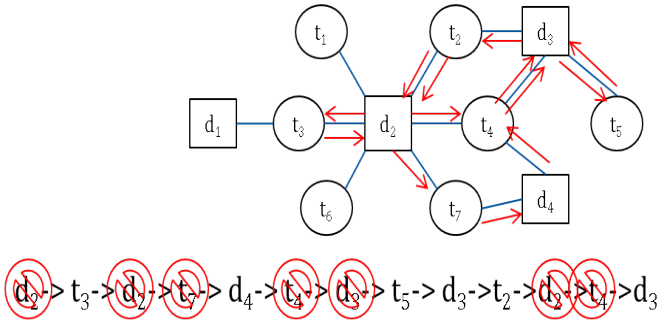

Figure 3.3: An example of RRW sampling

Finally, the output of this algorithm can be as following format.

Output:

$T_s = \{(t_3, 2), (t_5, 1), (t_2, 2)\}$

$D_s = \{(d_4, 2), (d_3, 3), (d_3, 3)\}$

RRW sampling is a UR sampling method. So, for average degree estimation we can use the sample mean estimator as stated in Equation 1.2. The average document degree will be

$$\widehat{\langle deg^d \rangle}_{SM} = \frac{2+3+3}{3} = 2.66 \tag{3.16}$$

In RRW sampling there is no restriction on state transition, rather the

restriction on the accepting of a sample. Hence RRW also able to traverse all parts of the graph regardless of the graph properties such as γ or the graph shape. But, the problem of this algorithm is too many rejection.

After t number of iterations, if this algorithm accepts only a number of samples we can define the sample rejection rate rr_{sample} as following

$$rr_{sample} = \frac{t-a}{a} \tag{3.17}$$

For RRW sampling, sample rejection rate is proportional to the sampling cost. In case of rejection it does not add that document or term as a sample. As a result another iteration is needed to obtain another sample which increase the query cost.

From the Figure 3.3 it also can be observed that for obtaining 3 sample terms and 3 documents algorithm rejected 4 documents and 3 terms. According to this example for documents, samples rejection rate is following

$$rr_{sample} = \frac{7-3}{3} = 1.33 \tag{3.18}$$

From experiments conducted in this research we have found that

$$rr_{sanple} = \langle deg \rangle \tag{3.19}$$

Detail of our experiments has been explained in Chapter 4.

3.5 Metropolis-Hastings Random Walk (MHRW) sampling

Metropolis-Hastings algorithm is a Markov Chain Monte Carlo (MCMC) method which can transform a random walk that converges to a trial dis-

tribution (p), to a new random walk that converges to a target distribution (π) [35, 21]. A MHRW sampler traverses on a Markov Chain to generate samples from a distribution by applying a acceptance-rejection procedure. This acceptance-rejection procedure is being used to determine whether the proposed state will be accepted as a next state of the random walk or not. Eventually this acceptance-rejection procedure transforms the samples from trial distribution (p) to target distribution (π). Note that this is different from the RRW sampling where we apply acceptance-rejection procedure during accepting a sample not in changing state.

The MH algorithm gives best output when the graph is ergodic and supp(p) = supp(π). Note that a graph is called ergodic when it is irreducible or strongly connected and aperiodic. The acceptance function of the MH algorithm is as following.

$$a_{MH}(x, y) = min\{\frac{\pi(y)P(y \to x)}{\pi(x)P(x \to y)}, 1\} \tag{3.20}$$

Here, $\pi(x)$ is the probability of x to be chosen as a sample from the distribution π and $P(x \to y)$ is the transition matrix which can be represented as following.

$$P(x \to y) = \frac{1}{deg_x} \tag{3.21}$$

Hence, after simplification acceptance function will be as following

$$a_{MH}(x, y) = min\{\frac{deg_x}{deg_y}, 1\} \tag{3.22}$$

A Metropolis-Hastings algorithm based random walk is given in Algorithm 4. The foundation of this algorithm is also similar to SRW. But, MHRW differs from the SRW and RRW in state transition. In MHRW, algorithm selects a neighbour of current node with equal probability like SRW and RRW, but

Algorithm 4 Metropolis-Hastings Random Walk (MHRW) sampling

Input : t_0 = seed term, sample size n
Output: Set of n number of document samples D_s and term samples T_s with their corresponding degrees.

$D_s = T_s$ = empty lists;
$current$ = select one neighbouring node of t_0 with equal probability;
while $D_s.size < n$ *OR* $T_s.size < n$ **do**
    $next$ = select one neighbouring node of $current$ with equal probability;
    if $Accept(current, next)$ **then**
        $current = next$;
    end
    add $current$ and its degree $deg_{current}$ to D_s OR T_s;
end
return D_s and T_s;

Algorithm 5 Accept-MH

Input : Two document or term nodes
Output: $true$ OR $false$
$size_1$ = degree of $current$;
$size_2$ = degree of $next$;
r = one random number between 1 to $size_2$;
if $r <= size_1$ **then**
    return $true$;
else
    return $false$;
end

it does not move to that neighbour unless it passes the acceptance test stated earlier. Our Algorithm 5 is simulating the acceptance probability of Equation 3.22. If it passes the test it moves and add that particular neighbour as a sample. Otherwise, it will remain in the same state and will add that current node as a sample. One work through example of whole MHRW for document sampling is given below.

Example 4: MHRW sampling and estimation

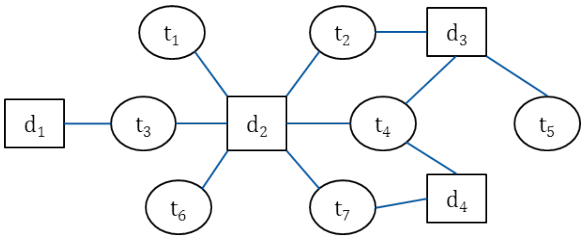

Figure 3.4: Deep web source as a bipartite graph

Given the graph depicted in Figure 3.4 one possible MHRW sampling can be as following

Input: graph depicted in Figure 3.4, $t_0 = t_2$ and $n = 3$

Process: Likewise other two random walks it will start with the seed node t_2 and select one of its neighbours, assume d_2 with equal probability. But it will not move to d_2 unless it passes the acceptance test with probability $min\{1, \frac{deg_2^t}{deg_2^d}\}$. To simulate this probability a random number r will be generated from 1 to deg_2^d. If $r < deg_2^t$ it will move to d_2 and add d_2 as a sample. Otherwise, it will remain in t_2 and add t_2 as a sample. Assume generated $r = 3$, so it will not accept that state and will remain and add t_2 as a sample. This process will continue until 3 document and term samples are being obtained. Note that in MHRW after n number of iterations n number of samples will be obtained regardless of the number of rejection. One possible walk can be $t_2(seed) - t_2 - d_3 - t_5 - t_5 - d_3 - t_4 - d_4$. In this walk consecutive repetition of nodes means rejection of state. Finally, the output of this algorithm can be as following format.

Output:

$D_s = \{(d_3, 3), (d_3, 3), (d_4, 2)\}$

$T_s = \{(t_2, 2), (t_5, 1), (t_5, 1)\}$

As MHRW produces UR sample, we will use the sample mean estimator stated in Equation 1.2 to estimate the average degree. For our considered example average document degree will be as following

$$\widehat{\langle deg^d \rangle}_{SM} = \frac{3+3+2}{3} = 2.66 \tag{3.23}$$

In MHRW whether it accepts or rejects, in each step algorithm will obtain one sample. Therefore according to Equation 3.17 the rejection rate $rr = 0$ which is true in terms of sample acceptance. But, does this rejection of states effects the process? if yes then how?

We have tried to explain this answer using another term named state rejection rate. For n number of iterations, if the algorithm accepts a' number of states, we define the state rejection rate rr_{state} as following

$$rr_{state} = \frac{n - a'}{a'} \tag{3.24}$$

Though, rr_{state} does not effect the query cost but it can effects the sampling

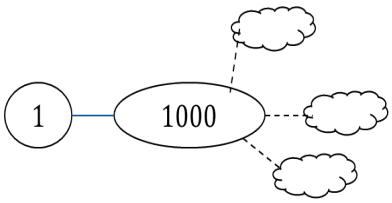

Figure 3.5: An example graph

accuracy. For example let us consider the graph in Figure 3.5 where each node has been labelled with their corresponding degree and cloud represents sub-graph of the whole graph, if we applies MHRW with starting node 1 it will select its neighbouring node 1000 but will move to 1000 with probability

1/1000 which is very low. Hence, even after 500 iterations current state might be the node 1 and all obtained samples during this 500 iterations will be node 1, which might cause a great bias during estimation. So, higher state rejection rate increase the cover time which means expected number of steps to reach every node [31]. As a result the distinct number of nodes traversed by the walk will be small.

In the example walk for obtaining 3 documents samples we need to make 5 iterations. Where the state rejection rate is

$$rr_{state} = \frac{5-3}{3} = 0.66 \tag{3.25}$$

Chapter 4

Experiments and Results

In our estimation process for obtaining UR and PPS samples, our experiment differs from query-based sampling [13] in the following aspects:

- All matched documents or terms are being returned as a query result which means the ranking over the documents is ignored. Hence no query overflow.
- All documents have been indexed with full length. That means no truncation of document size.
- Duplicate and near-duplicate are not in consideration. That means duplicate and near duplicate documents also have the equal probability to be captured.

4.1 Data set

We use 20k Newsgroup data corpus consisting 19,996 xml documents including some empty documents which have been excluded from our experiment, In our experiments we have considered all single alphabetical words as our query term and those are case insensitive. A statistical summary of the data set been given in Table 4.1.

Table 4.1: Statistics of Newsgroup data

Docs	Distinct docs excluding empty files N	11,059
	Average document degree $\langle deg \rangle^d$	190.4528
	Coefficient of variation γ	0.8209
	Minimum document degree deg^t_{min}	37
	Maximum document degree deg^t_{max}	2,353
Terms	Distinct terms N	84,644
	Average term degree $\langle deg \rangle^t$	24.8833
	Coefficient of variation γ	9.0856
	Minimum term degree deg^t_{min}	1
	Maximum term degree deg^t_{max}	11,059

The degree distribution of newsgroup document-term graph is depicted in Figure 4.1. Here we plot the frequency against degree. We can observe that document degrees follow log-normal distribution whereas term degree distribution follows the power law.

4.2 Documents

This section focuses on the estimation for documents. We report the sampling distribution followed by the estimation of average degree using all four

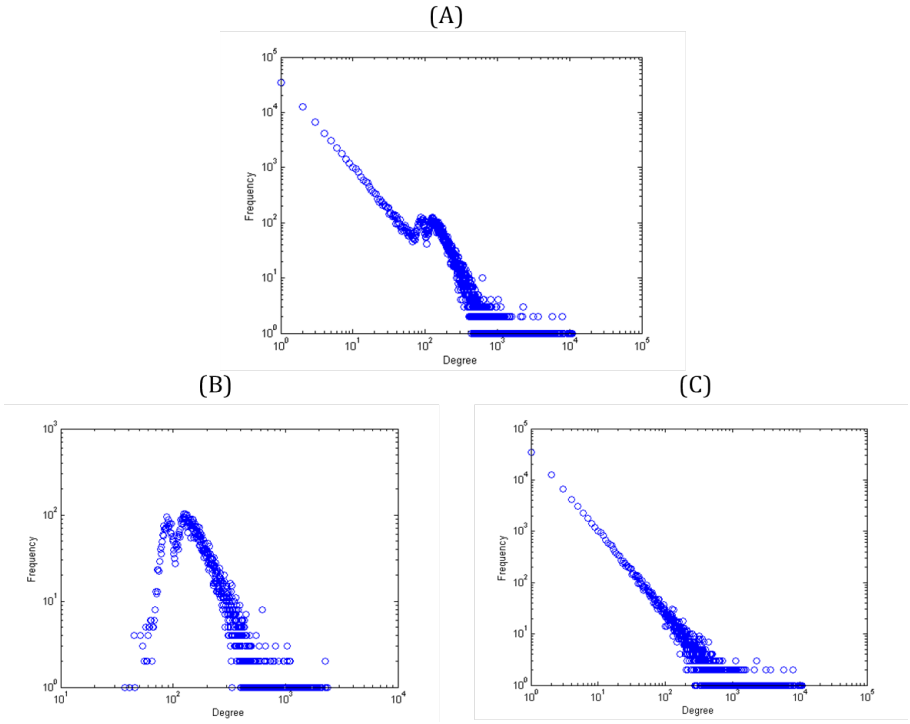

Figure 4.1: (A) Degree Distribution of whole Newsgroup data. (B) Only document degree distribution. (C) Only term degree distribution

methods.

4.2.1 Sampling distribution

We compare the sample distributions that are obtained from four sampling methods including SRW, RRW, MHRW and UR. We obtain 1000 samples and depict in Figure 4.2 which is the the frequency-degree plot. From Figure 4.2 it is observed that SRW sample distribution is different from the real documents distribution. For SRW the tail part is higher compared to the real distribution, which proves that documents with higher degree have higher frequency in the sampled data. Whereas RRW, MHRW and UR samples are uniform

random samples which resemble the real distribution that is log-normal. But, in frequency-degree plot RRW, MHRW and UR seems different because of small sample size. For better observation we depict the corresponding CCDF (Complementary Cumulative Distribution Function) in Figure 4.3. For SRW, CCDF line of sampled data goes upper than the real data as degree increases which means documents with higher degree sample more. In contrast RRW, MHRW and UR samples fits with the real data distribution as those are uniform random.

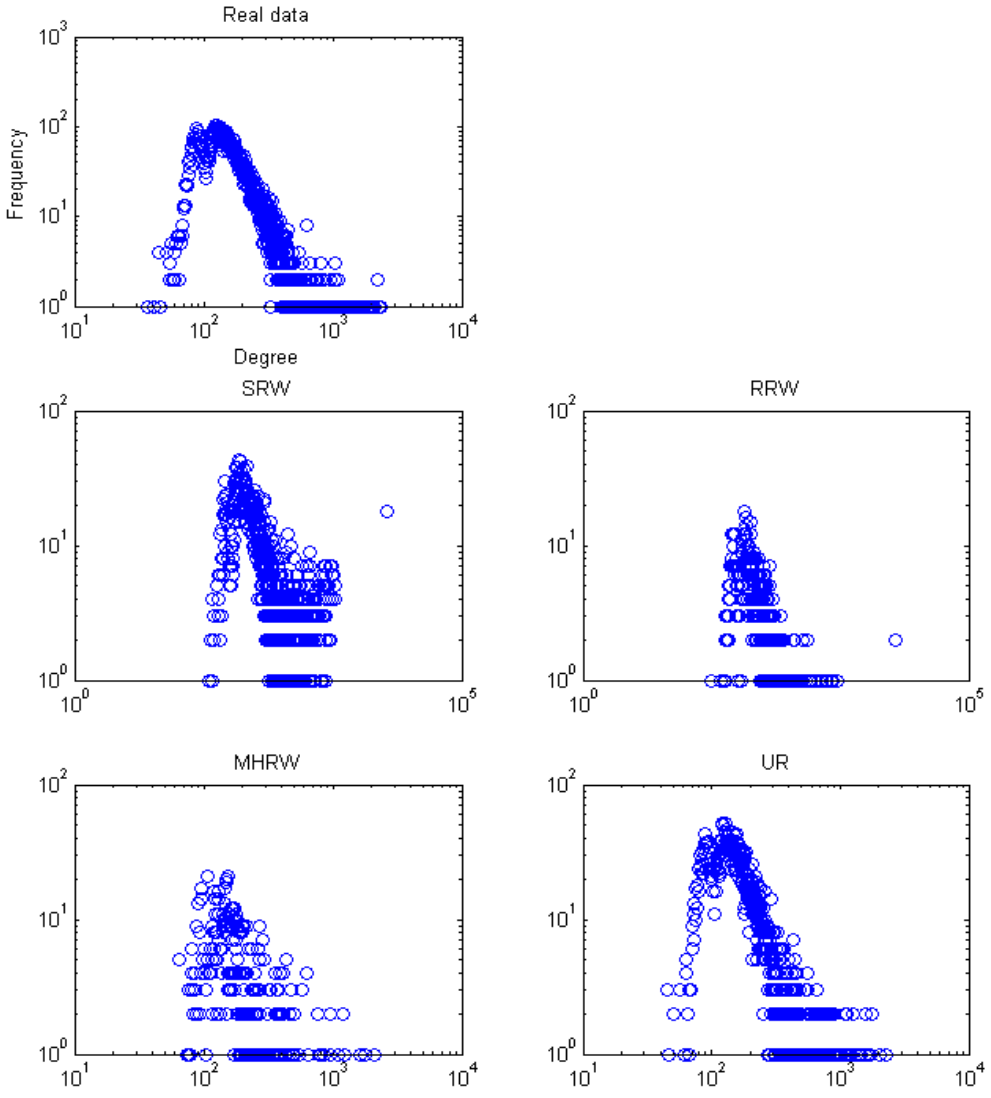

Figure 4.2: Degree distribution of **document** samples of size 1000 using different sampling methods from Newsgroup data.

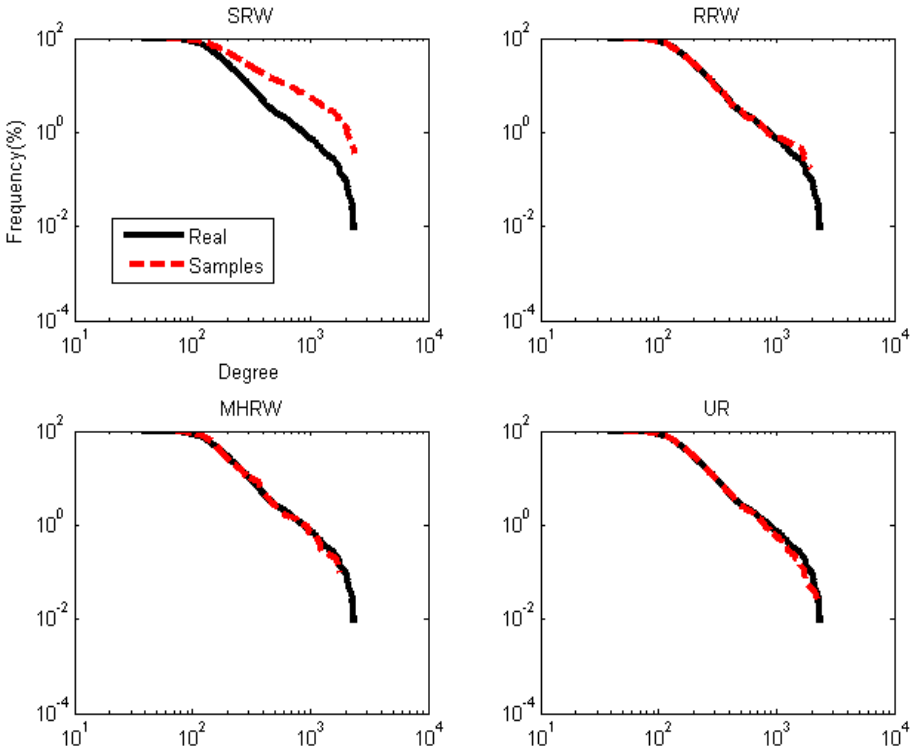

Figure 4.3: CCDF of **document** samples of size 1000 using different sampling methods from Newsgroup data.

4.2.2 Estimate average degree

Comparison in terms of valid sample size

In this section we compare the sampling methods with valid samples only, disregard of the rejected samples during the sampling process. We show that some sampling methods, for instance MHRW, is worse than UR even when only valid samples are considered. Next section we will conduct the comparison based on the actual sampling cost when the rejected samples are also included as sample size.

After obtaining document samples by SRW, RRW, MHRW and real UR sampling, we estimate the average degree of documents using harmonic mean (for SRW) and sample mean (for RRW, MHRW, and UR) estimator stated in Equations 1.3 and 1.2 respectively. We compare these four sampling methods for the estimation of the average degree of documents. Even though UR, RRW, and MHRW all produce uniform random samples in theory, the performances are different because RRW and MHRW obtain uniform random samples only asymptotically.

First we give the box plots for the intuitive understanding of the estimations. In Figure 4.4, for each sampling method we produce 10 box plots for the sample sizes ranging between 1,000 and 10,000. Each box plot is obtained from 200 runs. It shows that MHRW and RRW samplings results in larger variations of the estimations.

Next, we calculate the bias, relative standard error (RSE) and RRMSE of estimated average degree according to the Equation 3.7, 3.4 and 3.9 respec-

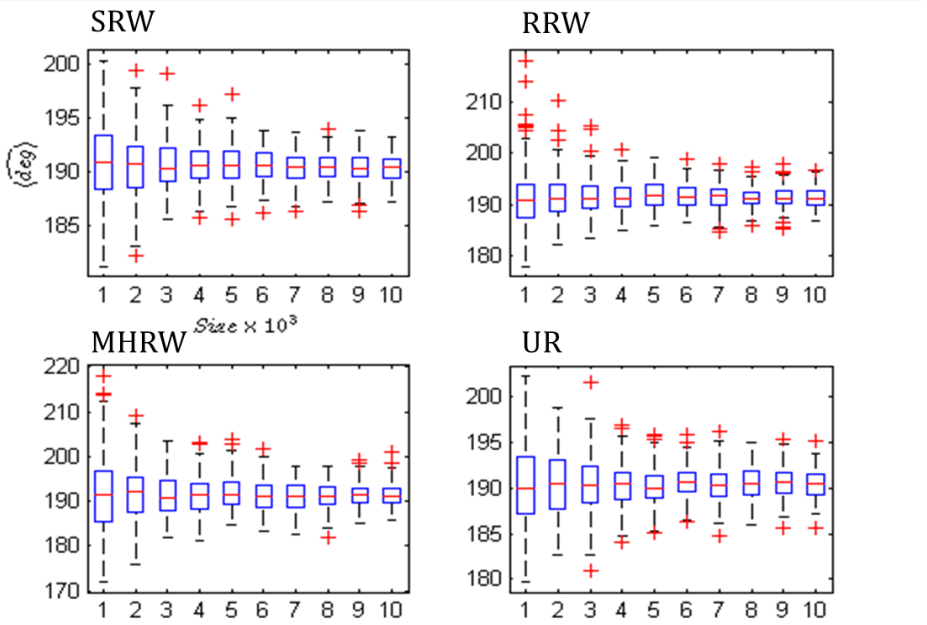

Figure 4.4: Box plots of estimated $\langle deg \rangle$ for **documents** using different sampling methods with different sample size and 200 iteration from Newsgroup data.

tively. RSE is the SE normalized by mean. Here bias evaluates how far the estimation from the real value and RSE determines the variation of estimations. The RRMSE helps us to evaluate the estimation based on both bias and variance together.

In Table 4.2 we reported the bias, RSE and RRMSE of estimation using different sampling methods for documents with different sample size. We also plot these values in Figure 4.5, which gives a more detailed comparison in terms of bias, rse, and rrmse.

First, we notice that UR does not show obvious bias as expected. Other three methods have small positive biases, which may be due to the random

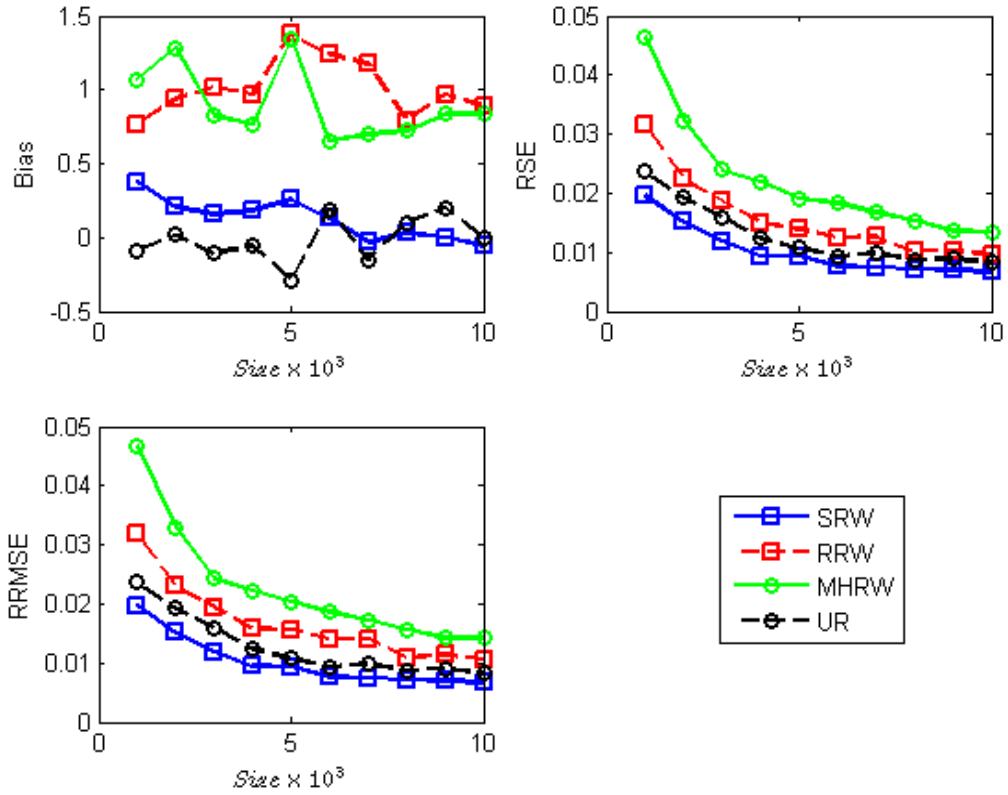

Figure 4.5: Bias, RSE and RRMSE of all **documents** average degree estimation on newsgroup data over 200 runs using different sampling methods.

walk mixing time. The samples are not strictly uniformly at random before random walk mixing. The rse of the four sampling methods are also different. SRW has the smallest variance while MHRW is the worst, because in different iteration MHRW covers a certain part of the graph and reflects in estimation.

Figure 4.5 also shows that the variance dominates the performance of the estimators within this sample size range. Because the bias is rather small compared with the variance, RRMSE is almost the same as RSE.

Table 4.2: Bias, RSE and RRMSE of all **documents** average degree estimation on newsgroup data over 200 runs using different sampling methods with different sample size.

$Size(n)$ $\times 10^3$	Bias				RSE				RRMSE			
	SRW	RRW	MHRW	UR	SRW	RRW	MHRW	UR	SRW	RRW	MHRW	UR
1	0.379	0.769	1.065	-0.091	0.020	0.032	0.047	0.024	0.020	0.032	0.047	0.024
2	0.205	0.936	1.285	0.018	0.015	0.022	0.032	0.019	0.015	0.023	0.033	0.019
3	0.159	1.022	0.826	-0.106	0.012	0.019	0.024	0.016	0.012	0.019	0.024	0.016
4	0.182	0.969	0.766	-0.059	0.009	0.015	0.022	0.012	0.010	0.016	0.022	0.012
5	0.260	1.379	1.338	-0.291	0.009	0.014	0.019	0.011	0.009	0.016	0.020	0.011
6	0.134	1.238	0.657	0.189	0.008	0.013	0.018	0.009	0.008	0.014	0.019	0.009
7	-0.031	1.175	0.696	-0.155	0.007	0.013	0.017	0.010	0.007	0.014	0.017	0.010
8	0.031	0.796	0.724	0.098	0.007	0.010	0.015	0.009	0.007	0.011	0.016	0.009
9	0.000	0.968	0.836	0.201	0.007	0.010	0.014	0.009	0.007	0.011	0.014	0.009
10	-0.057	0.888	0.836	-0.006	0.007	0.010	0.014	0.008	0.007	0.011	0.014	0.008

Comparison in terms of cost

In previous experiments we estimate $\langle deg \rangle$ using valid samples excluding all rejected ones. In this section we conduct similar experiments considering cost. In Figure 4.6 we depict the box plots of all estimations for documents using different sampling methods over 200 runs considering cost or number of steps.

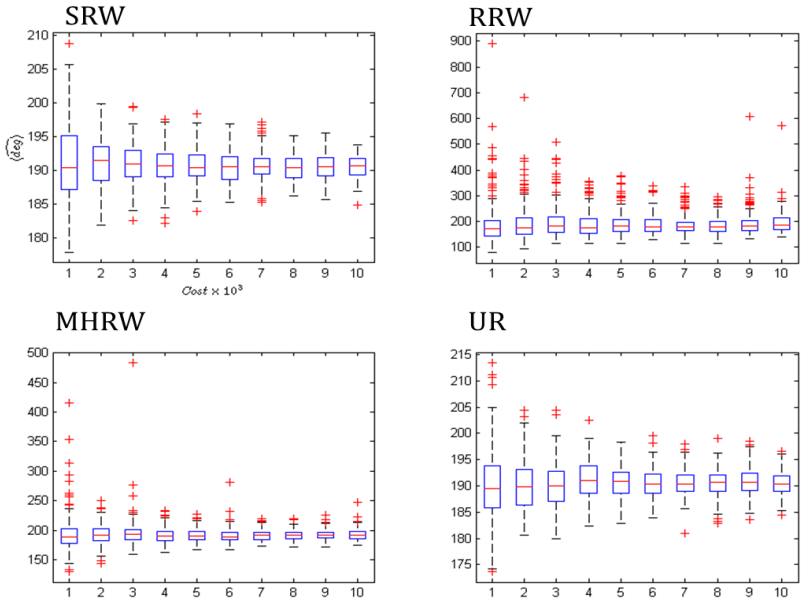

Figure 4.6: Box plots of estimated $\langle deg \rangle$ for **documents** using different sampling methods with different cost and 200 iteration from Newsgroup data.

From the boxplots in Figure 4.6 we observe that SRW and UR are the same as before, since they do not incur extra costs. In SRW, each next random node is taken as a valid sample and in UR sampling, we assume that random nodes can be obtained directly.

In Table 4.3 we report the bias, RSE and RRMSE of estimation using different sampling methods for documents considering cost and plotted these

values in Figure 4.7.

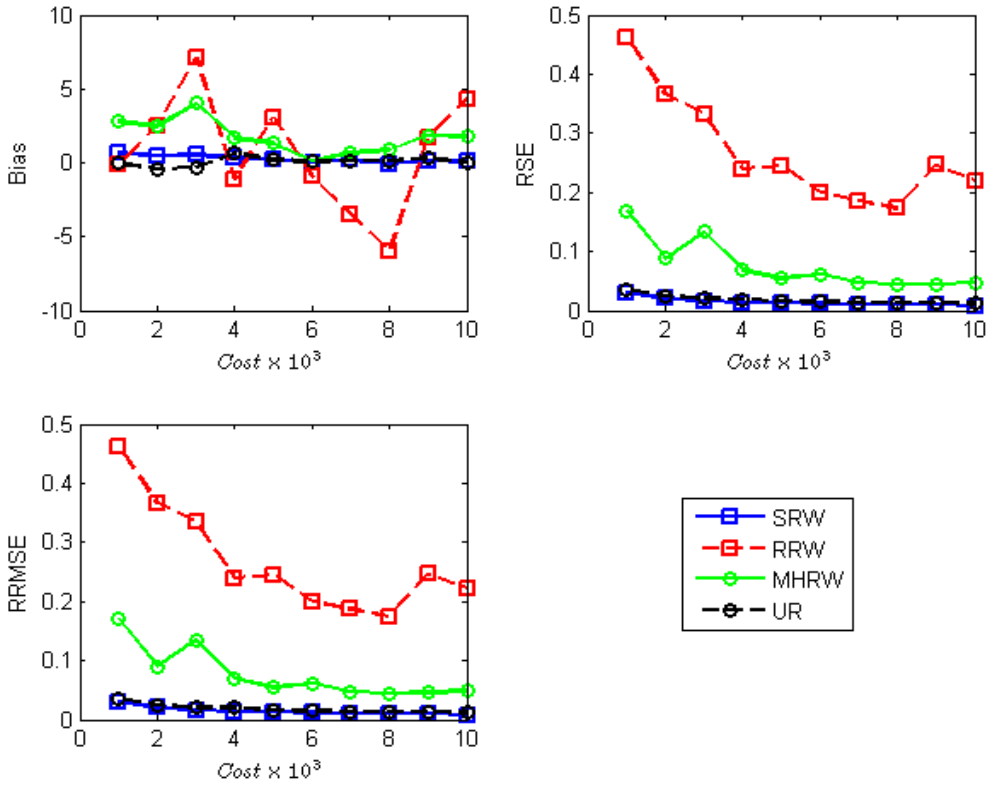

Figure 4.7: Bias, RSE and RRMSE of all **documents** average degree estimation on newsgroup data over 200 runs using different sampling methods with different cost.

Table 4.3: Bias, RSE and RRMSE of all **documents** average degree estimation on newsgroup data over 200 runs using different sampling methods with different cost.

$Size(n)$ $\times 10^3$	Bias				RSE				RRMSE			
	SRW	RRW	MHRW	UR	SRW	RRW	MHRW	UR	SRW	RRW	MHRW	UR
1	0.624	-0.158	2.775	-0.028	0.031	0.463	0.169	0.036	0.031	0.463	0.170	0.036
2	0.463	2.462	2.450	-0.446	0.020	0.366	0.088	0.025	0.020	0.366	0.089	0.025
3	0.542	7.114	4.014	-0.338	0.016	0.332	0.133	0.021	0.016	0.335	0.135	0.021
4	0.308	-1.117	1.691	0.644	0.014	0.240	0.069	0.019	0.014	0.240	0.070	0.020
5	0.245	3.022	1.324	0.214	0.013	0.245	0.054	0.016	0.013	0.245	0.054	0.016
6	0.044	-0.957	0.143	0.081	0.012	0.200	0.062	0.015	0.012	0.200	0.062	0.015
7	0.219	-3.457	0.676	0.131	0.011	0.186	0.048	0.014	0.011	0.187	0.048	0.014
8	-0.112	-6.045	0.901	0.114	0.010	0.173	0.044	0.014	0.010	0.175	0.045	0.014
9	0.158	1.733	1.849	0.331	0.010	0.246	0.045	0.013	0.010	0.246	0.046	0.013
10	0.060	4.319	1.767	0.026	0.008	0.220	0.048	0.012	0.008	0.222	0.049	0.012

For RRW, there are nodes that are accessed but not counted as valid samples. In average RRW rejects $\langle deg \rangle$ number of samples but keeps only one of them as valid. Hence it estimates $\langle deg \rangle$ based on very small number of accepted samples compared to other methods and performs worst. Note that if the real data variance is σ^2, The relation between sample size n and SE is as following.

$$SE = \frac{\sigma}{\sqrt{n}} \tag{4.1}$$

This means more samples result less SE. If n_1 is the number of valid samples and n_2 is the cost including rejected samples where $n_1 > n_2$. The ratio between two RSE will be $\frac{\sqrt{n_1}}{\sqrt{n_1/\langle deg \rangle}} = \sqrt{\langle deg \rangle}$. Hence for documents RSE of estimation in terms of valid sample will be $\sqrt{190.45} = 13.80$ times better than RSE of estimation in terms of cost. In Table 4.4 we report the ratio of both RSE for RRW. We find the average ratio is 17.70 which is higher than 13.80. This might happen because of less number of iterations.

Table 4.4: Ratio of the RSE of estimation in terms of valid sample size **(RSE-valid)** and cost **(RSE-cost)** after 200 iteration using RRW

$Size \times 10^3$	RSE-valid	RSE-cost	Ratio
1	0.463	0.0316	14.639
2	0.366	0.0224	16.343
3	0.332	0.0187	17.775
4	0.240	0.015	16.006
5	0.245	0.0139	17.611
6	0.200	0.0125	15.968
7	0.186	0.0127	14.677
8	0.173	0.0102	16.911
9	0.246	0.0102	24.117
10	0.220	0.0096	22.958
		Average	17.700

We experimented on sample rejection rate for RRW according to the equation 3.17. We calculate the rejection rate to obtain different size of documents and terms samples together. In Table 4.5 we report the sample rejection rate for both documents and terms from newsgroup data using RRW.

Table 4.5: Sample rejection rate rr_{sample} for different iteration by RRW from Newsgroup data

$TotalAccepted$ $\times 10^3$	Documents			Terms		
	Accepted	Rejected	rr_{sample}	Accepted	Rejected	rr_{sample}
1	123	20963	170.43	877	19982	22.78
2	215	42982	199.92	1785	40945	22.94
3	350	66168	189.05	2650	63191	23.85
4	470	88134	187.52	3530	84101	23.82
5	558	109056	195.44	4442	104041	23.42
6	678	132982	196.14	5322	126951	23.85
7	821	153951	187.52	6179	146938	23.78
8	884	173851	196.66	7116	165764	23.29
9	1055	197818	187.51	7945	188763	23.76
10	1150	217803	189.39	8850	207706	23.47
		Average	189.96		Average	23.50

From this table it can be observed that documents sample rejects more than the terms sample. For documents, average sample rejection rate is 189.96 and document average degree is 190.4528. Like wise for terms average sample rejection rate is 23.50 and term average degree is 24.8833. Hence we can infer that sample rejection rate for RRW is equal to average degree.

This relation also can be derived as following. Assume, after N number of steps we obtain N number of samples with following degree

$$d_{x1}, d_{x2}, d_{x3}, ...d_{xN}$$

After rejection procedure we will have only n valid samples as following

$$\frac{1}{d_{x1}}, \frac{1}{d_{x2}}, \frac{1}{d_{x3}}, \dots \frac{1}{d_{xn}}$$

As those samples are uniform random the average of these samples will be same as the real average degree.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{deg_{xi}} = \frac{1}{\langle deg \rangle}$$
$$\text{Therefore, } \sum_{i=1}^{n}\frac{1}{deg_{xi}} = \frac{n}{\langle deg \rangle}$$

Hence, for $\frac{n}{\langle deg \rangle}$ number of samples it will reject n. So, for 1 sample it rejects $\langle deg \rangle$. Hence the sample rejection rate is $\langle deg \rangle$.

We also can observe that MHRW considering cost performs worse than MHRW without cost. The reason is MHRW without cost consider larger sample size than with cost and more samples reduce the RSE and RRMSE. Note that cost 1000 does not provide equal number of documents or terms. There is always a ratio between documents and terms samples. In our experiment term nodes sample almost seven times more than document nodes. Hence estimation perform based on less sample. Therefore, documents sampling with fixed size performs better than documents sampling with cost. We also observe the ratio of obtained documents and terms sample by MHRW in fixed number of cost. In Table 4.6 we report document-term sample ratio in different size of steps. We also depicted the box plots of ratio in Figure 4.8. We can observe that the ratio is almost 0.13 which is the ratio between documents and terms average degree 24.88/190.45 = 0.13.

We also calculate the state rejection rate of MHRW according to the equation 3.24. In Table 4.7 we have reported the state rejection rate from both

Table 4.6: Average Document-term sample ratio in different size of steps by MHRW from Newsgroup data on 200 runs

$Cost \times 10^3$	Average Doc/Term ratio
1	0.1538
2	0.1458
3	0.151
4	0.1443
5	0.1339
6	0.1394
7	0.1375
8	0.138
9	0.1341
10	0.1368

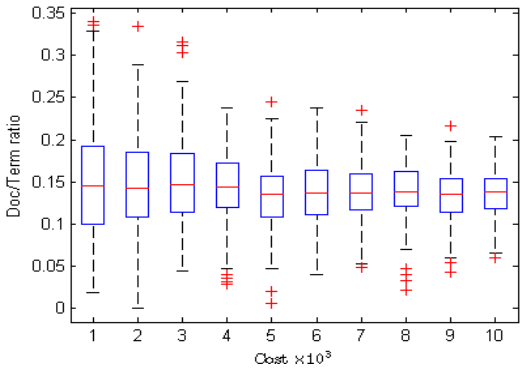

Figure 4.8: Box plot of document-term sample ratio in different size of steps by MHRW from Newsgroup data on 200 runs.

documents and terms using MHRW with different sample size from newsgroup data.

From this table it can be observed that from a term node it rejects more than from a document nodes. Which means it gets stuck more times in a term node compared to a document node. For documents, average state rejection rate is 1.2288 and for terms average state rejection rate is 15.3487.

Table 4.7: State rejection rate rr_{sample} from different iteration by MHRW from Newsgroup data

Accepted	From documents		From terms	
$\times 10^3$	Rejected	rr_{state}	Rejected	rr_{state}
1	1228	1.228	14098	14.098
2	2377	1.1885	34133	17.0665
3	3737	1.2456	45023	15.0076
4	5013	1.2532	60668	15.167
5	5984	1.1968	72452	14.4904
6	7339	1.2231	83908	13.9846
7	8969	1.2812	102558	14.6511
8	9728	1.216	112099	14.0123
9	11211	1.2456	146384	16.2648
10	12106	1.2106	187450	18.745
	Average	1.2288	Average	15.3487

4.3 Terms

This section focuses on the sampling and estimation for terms using all four sampling methods.

4.3.1 Sampling distribution

We also compare terms sample distributions that are obtained from four sampling methods. We obtain 1000 term samples and depict in Figure 4.9 which is the the frequency-degree plot. From Figure 4.9 it is observed that SRW sample distribution is completely different from the real terms distribution. For SRW the tail part is higher compared to the real distribution, which proves that terms with higher degree have higher frequency in the sampled data. Whereas RRW, MHRW and UR samples resemble the real distribution. For better observation we also depict the corresponding CCDF in Figure 4.10. For SRW, CCDF line of sampled data goes upper than the real data as degree increases which means terms with higher degree sample more. Unlikely RRW, MHRW and UR samples fits with the real data distribution as those are uniform random.

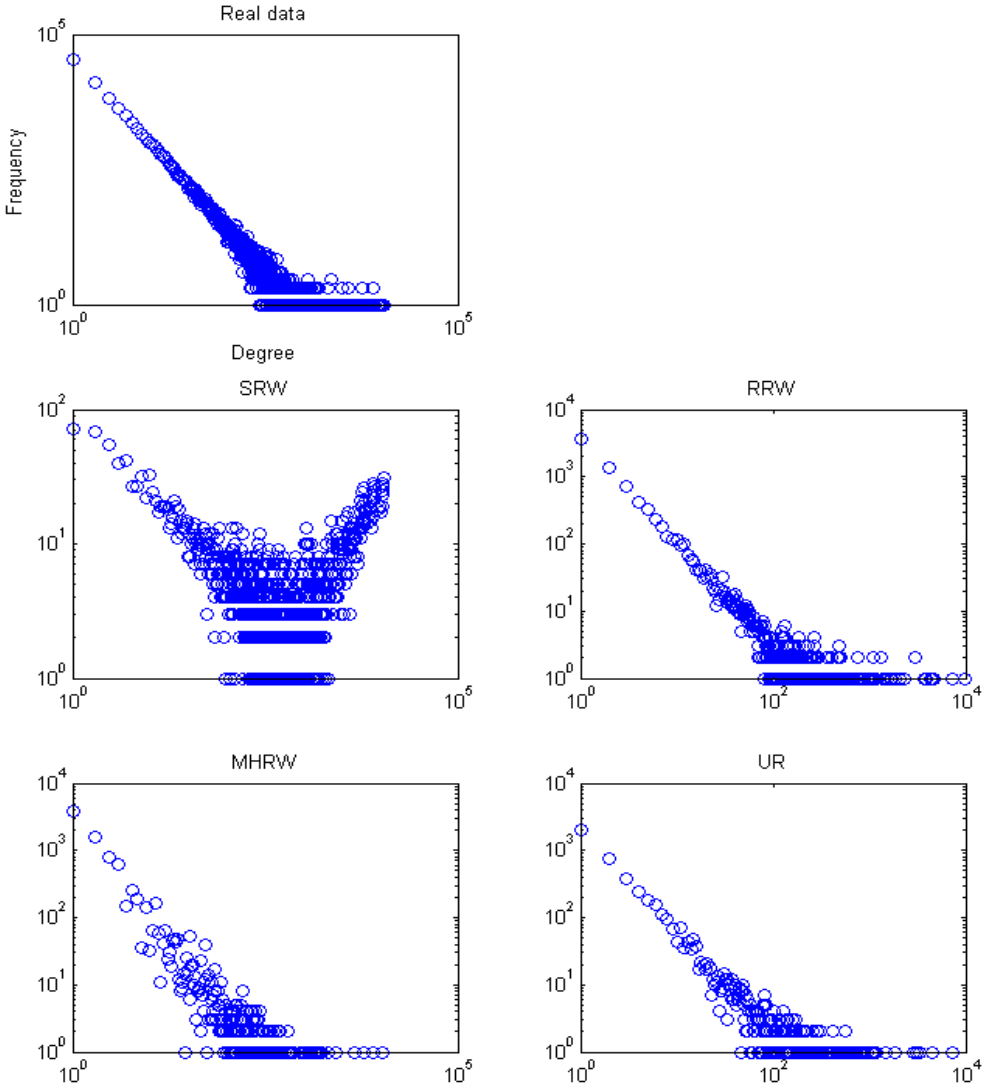

Figure 4.9: Degree distribution of **term** samples of size 1000 using different sampling methods from Newsgroup data.

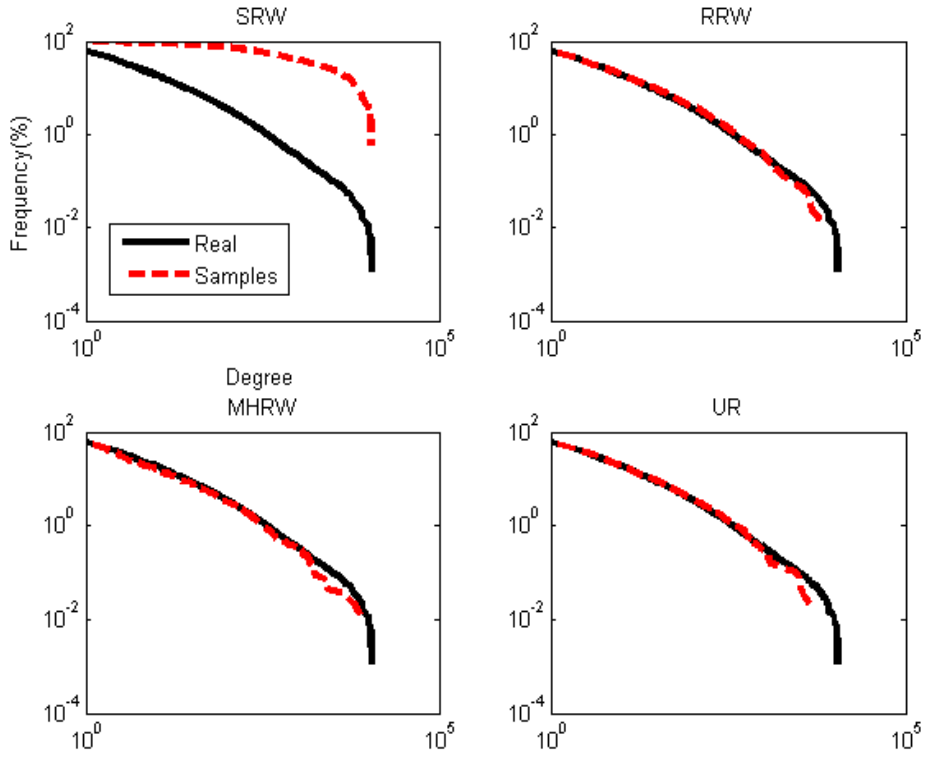

Figure 4.10: CCDF of **term** samples of size 1000 using different sampling methods from Newsgroup data.

4.3.2 Estimate average degree

Comparison in terms of valid sample size

After obtaining term samples by SRW, RRW, MHRW and real UR sampling we also estimate the average degree of terms using harmonic mean and sample mean estimator. In this section we only consider the valid term samples.

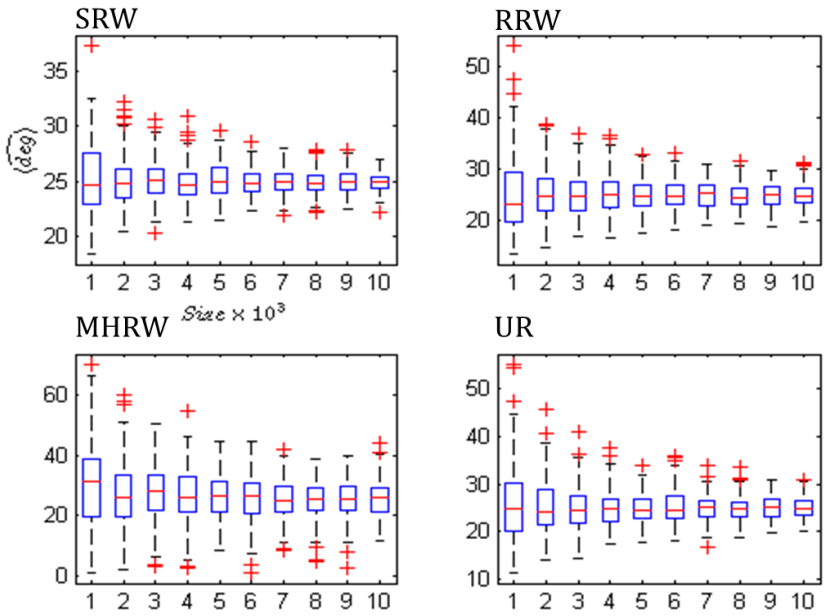

Figure 4.11: Box plots of estimated $\langle deg \rangle$ for **terms** using different sampling methods with different sample size and 200 iteration from Newsgroup data.

In Figure 4.11, we depict the box plots of all estimations for terms using different sampling methods over 200 runs and different sample size, which helps us to visualize the estimation on different runs. We can observe that for terms, MHRW estimation is also worse compared to all other methods as it has large variance on term estimations. SRW performs better than all UR

sampling methods.

Next, we calculate the bias, relative standard error (RSE) and RRMSE of estimated average degree.

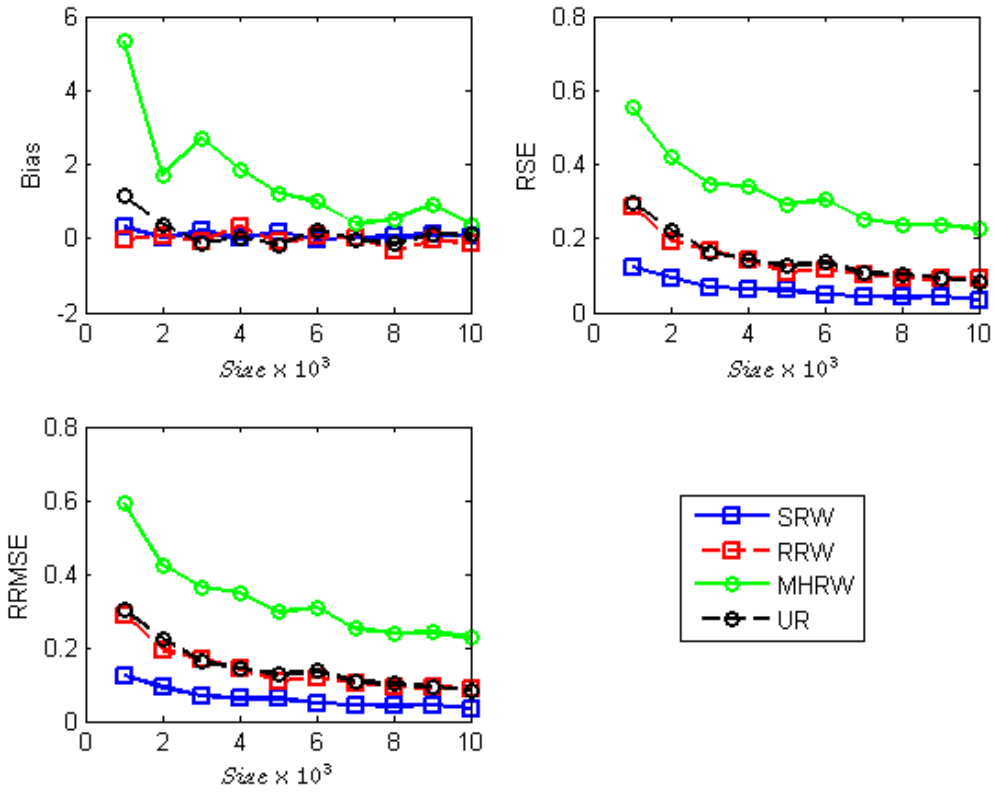

Figure 4.12: Bias and RSE of all **terms** average degree estimation on newsgroup data over 200 runs using different sampling methods.

In Table 4.8 we reported the bias, RSE and RRMSE of estimation using different sampling methods for terms with different sample size. We also plot these values in Figure 4.12. Here we observe that in terms of bias MHRW has large bias when the sample size is small. Because, terms degree follows the

power law which means degree heterogeneity is higher, hence MHRW rejects many states and estimation become biased based on the small part of the graph covered by MHRW. In terms of SE and RRMSE, here again SRW has the minimum and MHRW has the maximum for the degree heterogeneity.

Table 4.8: Bias, RSE and RRMSE of all **terms** average degree estimation on newsgroup data over 200 runs using different sampling methods with different sample size.

$Size(n)$ $\times 10^3$	Bias				RSE				RRMSE			
	SRW	RRW	MHRW	UR	SRW	RRW	MHRW	UR	SRW	RRW	MHRW	UR
1	0.301	-0.034	5.325	1.135	0.122	0.287	0.552	0.298	0.123	0.287	0.592	0.302
2	0.037	0.077	1.724	0.376	0.092	0.192	0.418	0.224	0.092	0.192	0.424	0.224
3	0.210	-0.084	2.714	-0.121	0.068	0.169	0.346	0.162	0.069	0.169	0.363	0.162
4	0.001	0.327	1.853	0.025	0.062	0.142	0.341	0.143	0.062	0.142	0.349	0.143
5	0.178	-0.096	1.229	-0.172	0.060	0.109	0.292	0.126	0.061	0.109	0.296	0.126
6	-0.024	0.043	0.991	0.212	0.048	0.118	0.306	0.136	0.048	0.118	0.308	0.136
7	0.036	0.013	0.410	-0.024	0.045	0.104	0.252	0.107	0.045	0.104	0.252	0.107
8	0.053	-0.326	0.530	-0.140	0.042	0.093	0.237	0.102	0.042	0.094	0.238	0.102
9	0.099	-0.062	0.919	0.112	0.043	0.090	0.238	0.093	0.043	0.090	0.241	0.093
10	0.075	-0.122	0.376	0.111	0.034	0.090	0.225	0.084	0.034	0.090	0.225	0.084

Comparison in terms of cost

In this section we conduct similar experiments for terms considering cost and can observe that RRW performs worst also for terms for its rejection procedure. In Figure 4.13 we depict the box plots of all estimations for terms using different sampling methods over 200 runs considering cost or number of steps.

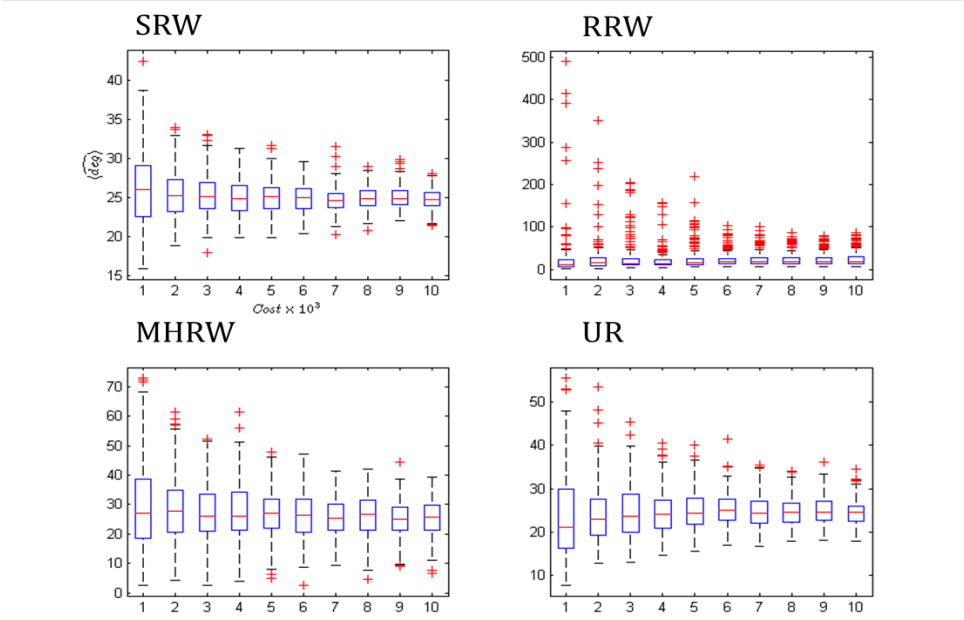

Figure 4.13: Box plots of estimated $\langle deg \rangle$ for **terms** using different sampling methods with different cost and 200 iteration from Newsgroup data.

In Table 4.9 we report the bias, RSE and RRMSE of estimation using different sampling methods for terms considering cost and plotted these values in Figure 4.14. Here also RRW performs 7 times worse in terms of RSE because of its rejection rate as explained before. Note that theoretically it should be $\sqrt{25} = 5$ time worse. SRW outperforms UR as well as MHRW in terms of RSE.

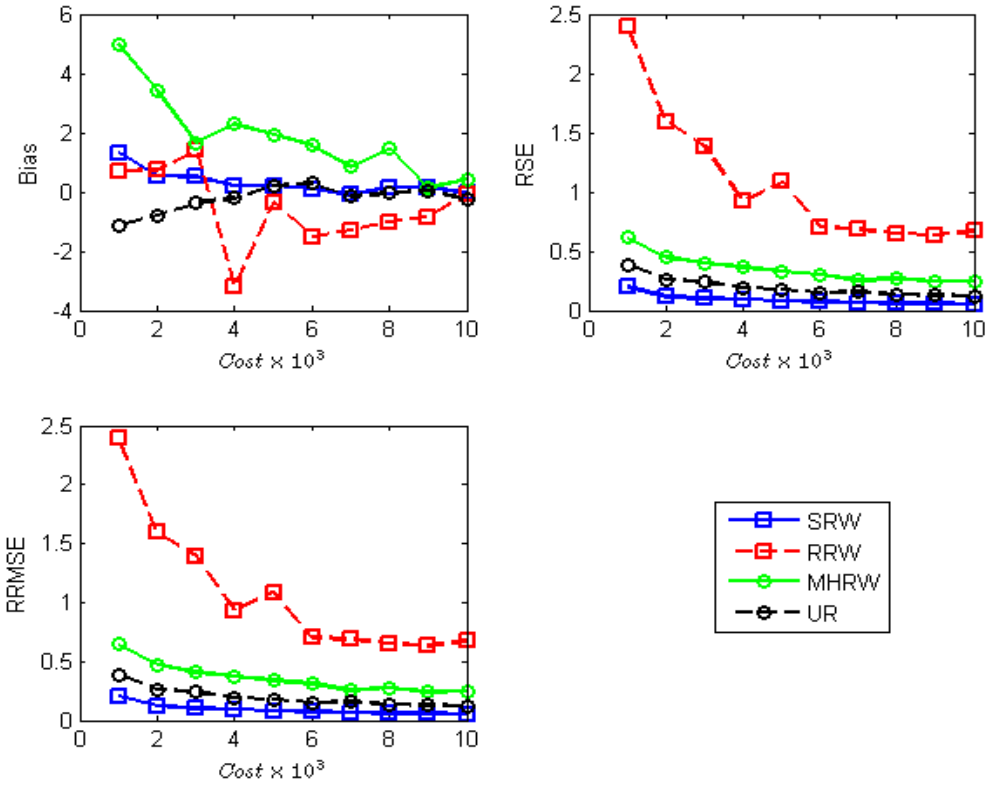

Figure 4.14: Bias, RSE and RRMSE of all **terms** average degree estimation on newsgroup data over 200 runs using different sampling methods with different cost.

Table 4.9: Bias, RSE and RRMSE of all **terms** average degree estimation on newsgroup data over 200 runs using different sampling methods with different cost.

$Size(n)$ $\times 10^3$	Bias				RSE				RRMSE			
	SRW	RRW	MHRW	UR	SRW	RRW	MHRW	UR	SRW	RRW	MHRW	UR
1	1.327	0.713	4.976	-1.124	0.200	2.400	0.614	0.383	0.207	2.400	0.646	0.386
2	0.537	0.735	3.407	-0.790	0.117	1.593	0.450	0.258	0.119	1.593	0.471	0.260
3	0.513	1.410	1.654	-0.378	0.101	1.388	0.401	0.238	0.103	1.389	0.406	0.238
4	0.216	-3.184	2.283	-0.221	0.095	0.921	0.364	0.192	0.096	0.930	0.375	0.192
5	0.216	-0.357	1.934	0.199	0.080	1.084	0.328	0.175	0.080	1.084	0.337	0.175
6	0.087	-1.508	1.563	0.300	0.073	0.703	0.303	0.151	0.073	0.705	0.309	0.151
7	-0.084	-1.280	0.844	-0.140	0.063	0.683	0.256	0.155	0.063	0.685	0.258	0.155
8	0.144	-1.001	1.457	-0.046	0.061	0.649	0.267	0.136	0.061	0.651	0.273	0.136
9	0.153	-0.835	0.137	0.054	0.056	0.632	0.244	0.129	0.057	0.633	0.244	0.129
10	-0.032	-0.084	0.434	-0.260	0.049	0.672	0.247	0.120	0.049	0.672	0.247	0.120

Chapter 5

Conclusions and Future Work

This thesis studies and compares various random walk sampling methods on deep web data sources. The deep web data sources can be modeled as document-term bipartite graphs, thus the sampling methods are reduced to the random walks on bipartite graphs.

We estimate the average degree of documents and terms from the deep web resources. Average degree is an important index of a graph, and also can used in estimation of population size. The highlight of this research is that both RRW and MHRW is worse than the UR samples. MHRW has higher variance and RRMSE in the estimation compared to all other methods. Whereas RRW rejects many samples and become worse in terms of cost.

We shown that PPS sampling such as SRW can outperform the UR sampling method such as RRW and MHRW for both documents and terms in terms of variance. We observe that real UR samples are better when the degree distribution is similar to the log normal or the coefficient of variation γ is low (such as 0.8209 for newsgroup documents). But, for terms whose distri-

bution follows the power law or the γ is higher (such as 9.0856 for newsgroup terms), SRW outperforms the real UR samples as well as RRW and MHRW.

We also show how estimation can differ with and without considering the cost of RRW sampling in terms of sample rejection rate rr_{sample}. We find rr_{sample} is equal to the average degree. When average degree is high, many samples are discarded in RRW.

On the other hand we explained the cost of MHRW sampling in terms of the ratio of document and terms samples, and state rejection rate rr_{state} which increase the cover time of a random walk as it gets stuck in a certain part of the whole distribution. Hence, we also not recommend to use the MHRW which estimation reflects on certain part of the graph.

This thesis can be extended using Random query based approaches. It will be interesting to observe whether simple random query based approach outperforms the SRW or not. For better comparison we have also plan to add one more UR approach called MH algorithm with delayed acceptance (MHDA) which is a modified version of MHRW proposed by Lee et al. [29].

Bibliography

[1] S.C. Amstrup, T.L. McDonald, and B.F.J. Manly. *Handbook of capture-recapture analysis*. Princeton University Press, 2010. [cited at p. 9]

[2] A. Anagnostopoulos, A.Z. Broder, and D. Carmel. Sampling search-engine results. *World Wide Web*, 9(4):397–429, 2006. [cited at p. 7]

[3] R. Balakrishnan and K. Ranganathan. *A textbook of graph theory*. Springer Verlag, 2000. [cited at p. 13]

[4] Z. Bar-Yossef, A. Berg, S. Chien, J. Fakcharoenphol, and D. Weitz. Approximating aggregate queries about web pages via random walks. 2000. [cited at p. 7, 10, 11]

[5] 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. ACM, 2006. [cited at p. 9, 11]

[6] Z. Bar-Yossef and M. Gurevich. Efficient search engine measurements. In *Proceedings of the 16th international conference on World Wide Web*, pages 401–410. ACM, 2007. [cited at p. 8]

[7] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. *Journal of the ACM (JACM)*, 55(5):24, 2008. [cited at p. 3, 4, 7, 12, 21, 22, 23]

[8] Z. Bar-Yossef and M. Gurevich. Efficient search engine measurements. *ACM Transactions on the Web (TWEB)*, 5(4):18, 2011. [cited at p. 4, 7]

[9] M.K. Bergman. White paper: the deep web: surfacing hidden value. *journal of electronic publishing*, 7(1), 2001. [cited at p. 1]

[10] K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. *Computer Networks and ISDN Systems*, 30(1):379–388, 1998. [cited at p. 8]

[11] A. Broder, M. Fontura, V. Josifovski, R. Kumar, R. Motwani, S. Nabar, R. Panigrahy, A. Tomkins, and Y. Xu. Estimating corpus size via queries. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 594–603. ACM, 2006. [cited at p. 4, 7, 9]

[12] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems (TOIS)*, 19(2):97–130, 2001. [cited at p. 1, 3, 8]

[13] J. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. In *ACM SIGMOD Record*, volume 28, pages 479–490. ACM, 1999. [cited at p. 3, 32]

[14] A. Dasgupta, G. Das, and H. Mannila. A random walk approach to sampling hidden databases. In *Proceedings of the 2007 ACM SIGMOD*

international conference on Management of data, pages 629–640. ACM, 2007. [cited at p. 7, 12]

[15] A. Dasgupta, X. Jin, B. Jewell, N. Zhang, and G. Das. Unbiased estimation of size and other aggregates over hidden web databases. In *Proceedings of the 2010 international conference on Management of data*, pages 855–866. ACM, 2010. [cited at p. 7]

[16] S.R.H. Garcia-Molina and S. Raghavan. Crawling the hidden web. In *27th International Conference on Very Large Data Bases*, 2001. [cited at p. 2]

[17] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. A walk in facebook: Uniform sampling of users in online social networks. *arXiv preprint arXiv:0906.0060*, 2009. [cited at p. 4, 7, 11]

[18] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1872–1892, 2011. [cited at p. 4]

[19] C.M. Grinstead and J.L. Snell. *Introduction to probability*. Amer Mathematical Society, 1997. [cited at p. 15, 16]

[20] M.H. Hansen and W.N. Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, pages 333–362, 1943. [cited at p. 4, 12]

[21] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. [cited at p. 27]

[22] M.R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform url sampling. *Computer Networks*, 33(1):295–308, 2000. [cited at p. 4, 10, 11]

[23] P.G. Ipeirotis, L. Gravano, and M. Sahami. Probe, count, and classify: categorizing hidden web databases. *ACM SIGMOD Record*, 30(2):67–78, 2001. [cited at p. 2]

[24] L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. In *Proceedings of the 20th international conference on World wide web*, pages 597–606. ACM, 2011. [cited at p. 7]

[25] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins, and E. Upfal. The web as a graph. In *Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–10. ACM, 2000. [cited at p. 13]

[26] M. Kurant, A. Markopoulou, and P. Thiran. Towards unbiased bfs sampling. *Selected Areas in Communications, IEEE Journal on*, 29(9):1799–1809, 2011. [cited at p. 4]

[27] R.J. Larsen and M.L. Marx. An introduction to mathematical statistics and its applications. 2001. [cited at p. 21]

[28] S. Lawrence and C.L. Giles. Searching the world wide web. *Science*, 280(5360):98–100, 1998. [cited at p. 1, 8]

[29] Chul-Ho Lee, Xin Xu, and Do Young Eun. Beyond random walk and metropolis-hastings samplers: why you should not backtrack for un-

biased graph sampling. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, pages 319–330, New York, NY, USA, 2012. ACM. [cited at p. 61]

[30] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM, 2006. [cited at p. 4]

[31] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46, 1993. [cited at p. 17, 31]

[32] Jianguo Lu and Dingding Li. Sampling online social networks by random walk. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, HotSocial '12, pages 33–40, New York, NY, USA, 2012. ACM. [cited at p. 4]

[33] Jianguo Lu, Yan Wang, and Jie Liang. Surface hidden data source properties by queries. In *Submitted*, 2013. [cited at p. 4, 11]

[34] J. Madhavan, D. Ko, Ł. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy. Google's deep web crawl. *Proceedings of the VLDB Endowment*, 1(2):1241–1252, 2008. [cited at p. 2]

[35] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953. [cited at p. 27]

[36] C. Olston and M. Najork. Web crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010. [cited at p. 14]

[37] M. Papagelis, G. Das, and N. Koudas. Sampling online social networks. *Knowledge and Data Engineering, IEEE Transactions on*, PP(99):1, 2011. [cited at p. 4]

[38] A.H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach. Respondent-driven sampling for characterizing unstructured overlays. In *INFOCOM 2009, IEEE*, pages 2701–2705. IEEE, 2009. [cited at p. 4]

[39] Paat Rusmevichientong, David M. Pennock, Steve Lawrence, and C. Lee Giles. Methods for sampling pages uniformly from the world wide web. In *In AAAI Fall Symposium on Using Uncertainty Within Computation*, pages 121–128, 2001. [cited at p. 11]

[40] M.J. Salganik and D.D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1):193–240, 2004. [cited at p. 4]

[41] M. Shokouhi and L. Si. *Federated search*. Now Publishers Inc, 2011. [cited at p. 1]

[42] S.K. Thompson. *Sampling*. Wiley Series in Probability and Statistics. Wiley, 2012. [cited at p. 2, 16]

[43] John von Neumann. Various Techniques Used in Connection with Random Digits. *J. Res. Nat. Bur. Stand.*, 12:36–38, 1951. [cited at p. 21]

[44] C. Wejnert and D.D. Heckathorn. Web-based network sampling efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods & Research*, 37(1):105–134, 2008. [cited at p. 4]

[45] P. Wu, J.R. Wen, H. Liu, and W.Y. Ma. Query selection techniques for efficient crawling of structured web sources. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 47–47. IEEE, 2006. [cited at p. 14]

[46] S. Ye and S.F. Wu. Estimating the size of online social networks. *International Journal of Social Computing and Cyber-Physical Systems*, 1(2):160–179, 2011. [cited at p. 7]

[47] Mingyang Zhang, Nan Zhang, and Gautam Das. Mining a search engine's corpus: efficient yet unbiased sampling and aggregate estimation. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, SIGMOD '11, pages 793–804, New York, NY, USA, 2011. ACM. [cited at p. 14]

Vita Auctoris

NAME:	Sajib Kumer Sinha
PLACE OF BIRTH:	Comilla, Bangladesh
YEAR OF BIRTH:	1986
EDUCATION:	University of Windsor Windsor, Ontario, Canada 2011-2013 M.Sc.
	North South University Dhaka, Bangladesh. 2005-2010 B.Sc.