Performance Analysis and Comparison of Sampling Algorithms in Online Social Network

Stuti K.
Pranveer Singh Institue of Technolgy y
Kanpur,U.P. (208020)
Dr. A.P.J. Abdul Kalam Technical University,
Lucknow, U.P.

Atul Srivastava
Pranveer Singh Institue of Technolgy
Kanpur,U.P. (208020)
Dr. A.P.J. Abdul Kalam Technical University,
Lucknow, U.P.

ABSTRACT

Graph sampling provides an efficient way by selecting a representative subset of the original graph thus making the graph scale small for improved computations. Random walk graph sampling has been considered as a fundamental tool to collect uniform node samples from a large graph. In this paper, a comprehensive analysis and comparison of four existing sampling algorithms- BFS, NBRW-rw, MHRW and MHDA is presented. The comparison is shown on the basis of the performance of each algorithm on different kinds of datasets. Here, the considered parameters are node-degree distribution and clustering coefficient which effect the performance of an algorithm in generating unbiased samples. The sampling methods as in this study are analysed on the real-network datasets and finally the conclusion says that MHDA performs excellently whereas BFS gives a poor performance.

General Terms

Online Social Network, Sampling Algorithms, BFS, NBRW-rw, MHRW, MHDA

Keywords

Comparison of Sampling Algorithms, Node Degree Distribution, Clustering Coefficient

1. INTRODUCTION

Online social networks have become one of the most popular destinations on the Web. Nowadays there are many social media websites: Google, Yahoo, MySpace, YouTube, Facebook, LinkedIn, Twitter. These sites focus exclusively on connecting users to facilitate the sharing of content using the social networking format and there have been noticeably far fewer studies that have focused on their structural dynamics [1]. Social network offers a platform for sharing knowledge, thoughts, opinions and more often to maintain a social relationship among people. In general, an online social network comprises of three components: users or actors, links or edges and groups.

Social network analysis primarily performs the execution in two phases: data collection and data analysis [2]. The first step includes defining the parameters and the scope of analysis. The second step of SNA is gathering data according to the parameters defined in the first step. Mainly, data gathering is of two types: Elicitation and Registration. Elicitation deals with the questionnaire/survey about specific data while registration deals with the extraction or mining of necessary information from registered entities/information. For example, MEDLINE is one of the large databases that maintain data records of the published papers on the biomedical research and had about two million records according to the statistics from 1995 to 1999 [2].

Before the analysis, the information and the characteristics of the structure of online social networks have to be obtained. This is done by a process known as Crawling. This process continues until some criteria is met for search in the network [3]. The collection of data is an important part of crawling. Crawling techniques can be classified into two categories- (a) Graph traversal techniques and (b) Random walks. In graph traversal techniques, each node in the connected graph is visited exactly once, if the process is running completely. BFS is a basic technique that has been used extensively in the past research for sampling the online social networks including Facebook [4]. BFS is biased towards high degree nodes, which becomes highly non-trivial to be analyzed so far for arbitrary graphs or to be improvised correctly [5]. In Random walks, the re-visiting of nodes is allowed. The application of random walks have been applied to OSNs, such as Twitter and Friendster [4]. Random Walk is biased towards high degree nodes and randomly selects a neighbour node to visit [5].

$$P_{u,w}^{RW} = \frac{1}{k_v}$$

where, k_v is degree of node v

Due to the huge size of the social network graphs, crawling the complete graph is more likely to be impossible, even with the currently available datasets, processing the social network graphs requires expensive clusters as well as large time consuming and computation overhead. An alternative solution to this problem is Graph Sampling that provides an efficient and expensive way by selecting a representative subset of the original graph. This makes the graph scale small while keeping the characteristics of the original social graph and results in improved computations [2].

According to Webster (1985), a sample is a finite part of a statistical population whose properties are studied to obtain information about the whole population. Sampling is defined as the act, process, or technique of selecting a suitable sample, or a representative part of a population for the purpose of a survey and thus determining parameters or characteristics of the whole population [6]. The sampling frame defines the distinction between the population of interest and the actual population. Generalizations can be made only to the actual population. Sampling frame consists of the list or procedure defining the population from which the sample will be drawn. For example- telephone book, voter list, random digit dialling.

Sampling is generally divided into two categories- probability sampling and non-probability sampling. A probability sample is one in which each element of the population has a known non-zero probability of selection [6]. Probability sampling is classified into 5 types- Simple Random Sampling, Systematic Sampling, Stratified Sampling, Cluster sampling, Multistage sampling. In non-probability sampling, sample does not have known probability of being selected. This sampling type is based on human choice rather than random selection. Non-probability sampling is generally classified into 2 types- Convenience Sampling, Quota Sampling [7].

During the sampling process in the online social network graph, the higher degree nodes are given priority over the lower degree nodes and iteratively repeats the visiting of high degree nodes more frequently. This gives rise to the problem of biasing and decreases the performance of the sampled graph. Many sampling algorithms have been proposed to improve this problem based on various parameters. Here in this paper, we mainly focus in analysing the performances of the existing sampling algorithms and comparing them based on some parameters.

1. LITERATURE REVIEW

According to studies, the application of sampling web pages has been taken into consideration by using various methodologies such as purposive, systematic and random sampling. The proposed research seeks to find out whether the sampled pages are representing the whole website using the various methodologies [8]. The previous studies have made it obvious that Breadth First Search (BFS) is very popular and the widely used sampling algorithm for measuring the large social networks, for example WWW or OSNs (Facebook, Twitter, etc) and helps in studying the topological characteristics such as shortest paths, clustering coefficients, node degree distribution of the sampled graphs [9]. As given in the research paper [2], the studies show that BFS has obtained a higher average clustering coefficient and very large normalised mean square error (NMSE) when compared to the other sampling algorithms due to biasing.

A technique of Snowball Sampling as described in [10] is similar to BFS and the authors have proposed a heuristic approach to correct the degree biases with the help of simulations. According to a classic definition by Goodman [10], an n-name "Snowball Sampling is very much identical to BFS, but at every node v, not all kv , but exactly n neighbours are chosen randomly out of all kv neighbours of v". This technique has some limitations that includes community or group bias, vague overall sampling size, lack of definite knowledge in identifying an accurate sample and lack of control over the sampling method. To improve this problem, as described in [9], Respondent-Driven Sampling (RDS) was proposed for penetrating the network of hidden communities.

Random Walk (RW), as described in [11] allows revisiting the same node again and again resulting in the formation of a loop and giving rise to the problem of biasing. Random Walk sampling is inherently biased towards high-degree nodes. In [4] the studies showed that this bias can be evaluated by the analysis of Markov Chain and can be improved by Re-Weighted Random Walk (RWRW) and Metropolis Hastings Random Walk (MHRW) that are unbiased and this study can be justified through properties like assortivity, convergence analysis and estimation. Wang et al. [2] introduced a new sampling method, Frontier Sampling (FS) based on Random Walk (RW) and claimed that FS obtains a very good degree

distribution and clustering coefficient distribution when compared to MHRW and BFS.

The paper [12] showed that Re-Weighted Random Walk (RWRW) sampling algorithm corrects the problem of bias in Random Walk by re-weighting the nodes at the end but suffers from the large deviation problem. RWRW samples are heavily biased and require an appropriate re-weighting procedures for correcting bias but does not stands versatility as distortion is produced due to reweighing and thus it suffers from asymptotic bias. According to [12], a Non-Backtracking Random Walk with reweighting (NBRW-rw) algorithm was introduced that ensured unbiased sampling and guaranteed a smaller asymptotic variance than RWRW resulting in improved node estimation accuracy because the process involves 'never' backtracking to the previous node. The results showed that NBRW algorithm leads to faster mixing rate when compared to RW.

As described in [13], the Metropolis Hastings Random Walk algorithm (MHRW) was introduced as a basic for Markov Chain Monte Carlo technique for sampling from a probability distribution that is generally non-uniform and difficult to sample from directly. In terms of estimation accuracy, MHRW performs very poorly and the clustering coefficient performance depends on the various real-world datasets [4]. The studies in [14] suggests that MHRW can overcome the large deviation problem of RWRW. The number of final samples obtained by MHRW is much smaller than that of RWRW, which results in degrading the performance of MHRW and hence this algorithm suffers from the sample-rejection problem.

The recent studies in [12] proposed an algorithm, Metropolis-Hastings algorithm with Delayed Acceptance (MHDA) as an improvement to MHRW algorithm which guaranteed a smaller asymptotic variance than the previous MHRW algorithm. The process of MHDA includes operating just like the MHRW algorithm at the first step. MHDA, in contrast to MHRW, the actual transition to node k is accepted if k≠i where i is the previous node from which the walker came and on the other hand, the transition to node k is delayed if k=i. Thus, the name is delayed acceptance. The conclusion in the paper [12] says that the walker will spend the same random amount of time at each node as in case of MHRW, followed by reducing the bias towards the previous node while making the transitions to the next node neighbours, allowing us to freely choose the new proposal probabilities as required.

The recent studies in [14] defined an algorithm Maximum-degree random walk (MD) that collects the nodes by running a random walk on a dynamically created regular graph. The algorithm includes the process of modifying the original graph into a regular graph by adding the self-loops on the nodes in order to make the degree of each node equal to the maximum degree of the original graph. The result leads to a conclusion that MD algorithm an overcome the large deviation problem of RWRW but suffers from two drawbacks – repeated samples problem and indefinite knowledge of maximum degree in unbiased graph sampling via crawling. The recent studies in [14] proposed a sampling algorithm, Rejection-Controlled Metropolis-Hastings (RCMH) that introduces a tradeoff between the large deviation problem of RWRW and the sample rejection problem of MHRW and its application to unbiased graph sampling. By setting an appropriate parameter for acceptance function, the large deviation problem of RWRW can be improved and because of the reason that the sample acceptance ratio of RCMH is larger than that of

MHRW, this algorithm also overcomes the sample rejection problem of MHRW.

2. COMPARISON ANALYSIS ON THE BASIS OF PARAMETERS

After going through all the previous algorithms, we selected four existing sampling algorithms namely- BFS, NBRW-rw, MHRW and MHDA. We will compare these sampling algorithms by trying to explore how they perform to maintain the different important properties of the original social graphs. In this study, we will try to give, to the best of my knowledge, the comprehensive and relatively fair comparison among these sampling algorithms.

According to the studies as described in previous and recent research papers, it has been found that these existing sampling algorithms can be evaluated considering many parameters such as- Node Degree Distribution (NDD), Clustering Coefficient (CC), Biasing, Estimation Accuracy, Convergence, Total Variance Distance (TVD) and many more. The measurement studies considering these parameters shows that these algorithms perform very diversely while maintaining the different social graph properties. The different performances are highly correlated with their evaluation procedures on a specific dataset. An algorithm can be seen to perform poorly in some datasets while on the other hand it can perform quite well in another dataset. In this work, we will try to analyze the studies done on these sampling algorithms on different parameters and comparing them.

BFS with MHRW and FS

The process of BFS sampling algorithm [2] is described in the following way. The initial node is selected randomly as a seed node and the further processing of sampling involves formation of two queues: a Sampled queue and a Processed queue. A Sampled queue is used to store the sampled nodes, while a Processed queue is used to store the nodes that have been processed for sampling. The process continues into the looping process until the fixed budget is achieved or the whole graph is explored as per the requirement. The nodes with higher degree in BFS traversal sampling process are said to visit more frequently and thus, the result says that BFS is biased towards high degree nodes [9]. This has been empirically showed by Erdos-Renyi random graphs, that made use of a Poisson degree distribution to show that the observed degree distribution under traceroute sampling follows a power-law, and has been verified by Clauset and Moore, as described in [15].

In the analysis of [2], when considering node degree distribution(NDD) as a parameter, cumulative distribution function (CDF) and normalized mean square error (NMSE) calculations are used to show the performances of different sampling algorithms while keeping the node degrees. It is clearly shown that BFS is biased towards high degree nodes and it is seen that NMSE of BFS is very large when compared with MHRW and FS and hence, converges faster than BFS. For the evaluation of clustering coefficient as a parameter, we need to obtain the network average clustering coefficient (NACC). It is shown in [2] that NACC obtained through BFS is larger than the original one. As clustering coefficient depends strongly on the node degree and BFS is biased towards high degree nodes therefore, the result is larger average clustering coefficient of BFS.

- **NBRW-rw with RWRW**

Chul-Ho Lee et al [12] proposed NBRW-rw sampling algorithm to collect a sequence of samples by crawling the target graph and also applying the re-weighting process at the same time, so that the induced sampling bias can be eliminated from the non-uniform stationary distribution of RWRW. The results have shown that NBRW-rw ensures unbiased graph sampling because of the added function of re-weighting the nodes and eliminating the biasing. NBRW-rw also guaranteed a smaller asymptotic variance when compared to RWRW and it is verified through the calculations provided in [12]. Hence, it provides the improved node estimation accuracy.

The studies of [12] shows the performances of RWRW and NBRW-rw on AS-733 real-world network dataset and depicts that NBRW-rw outperforms RWRW in terms of required number of samples (cost) in order to achieve the same level of estimation error. It has been observed from the above comparison that NBRW-rw is more effective in achieving the higher sampling accuracy than RWRW, even in the case when both the random walks does not start in the stationary regime.

- **MHRW with BFS**

MHRW algorithm is applicable to arbitrary probability distributions and uses the Markov Chain with stationary distribution π as long as the chain is irreducible (i.e. there is a positive probability of visiting all other states) and aperiodic (i.e. no cycles are formed in the chain). This algorithm performs well and is widely being used in modern applications [13]. The Metropolis-Hastings algorithm is used for unbiased graph sampling and constructs a Metropolis-Hastings Random Walk (MHRW) on a graph thus, achieving a uniform stationary distribution i.e. $\pi=u$. The transition probability of MHRW is defined as in [12]-

$$P(i,j) = \begin{cases} \min\left\{\dfrac{1}{d(i)}, \dfrac{1}{d(j)}\right\} & if(i,j) \in \varepsilon, \\ 0 & if(i,j) \notin \varepsilon, i \neq j \end{cases}$$

and, $\qquad P(i,j) = 1 - \sum_{j \neq i} P(i,j)$

When taking the comparison of MHRW with BFS as given in [2], it is clearly seen that MHRW improves the biasing property and has a well-performed node degree distribution. MHRW performs much better than BFS because the cumulative distribution function (CDF) obtained is almost identical to the original dataset and the normalized mean square error (NMSE) calculated is also smaller as compared to BFS and therefore, it converges faster than BFS. In terms of clustering coefficient, MHRW performance greatly depends on the datasets and the connectivity of the graphs. MHRW works better in tightly-connected graphs such as in case of Slashdot0811 dataset and on the other hand performs no good as in case of soc-Epinions1 dataset as it is a loosely-connected graph and shows a much larger difference.

- **MHDA with MHRW and NBRW-rw**

MHDA is an improved process of MHRW algorithm as proposed by Chul-Ho Lee et al. in [12], which theoretically guarantees a smaller asymptotic variance than the generic MHRW algorithm for any given desired distribution π. According to the proposed MHDA algorithm, it results in achieving higher efficiency than MHRW algorithm. One of the additional overhead is keeping in mind the previously visited node from where the random walk was traversed.

It is observed from the analysis of MHDA that if the reversible embedded Markov chain is replaced by a non-reversible Markov chain, which avoids the process of backtracking to the extent possible in contrast to NBRW-rw

while preserving the same stationary distribution then, this transformation will guarantee an unbiased estimator with higher efficiency than the original one. This was the reason for the proposal of MHDA algorithm over the standard MHRW algorithm. According to the analysis of MHDA in [12], the asymptotic variance of MHDA is smaller than the MHRW algorithm.

The comparison analysis demonstrates the different behaviours of BFS, NBRW-rw, MHRW and MHDA sampling algorithms in the social network graph dataset and describes the property metric of graph used in the sampling algorithms. The analysis includes the Uniform Sample taken from the Facebook dataset for analysing the behaviour of the algorithms against the ground-truth uniform sample and the Slashdot0811 dataset considering the parameters node degree distribution and clustering coefficient. This comparison shows the result that which algorithm gives the best performance by generating an unbiased sample.

2.1 Datasets

In order to have a "ground truth" to compare against, the performance of the sampling methods, a uniform sample has been taken from the crawling of the Facebook social graph. It serves as a baseline for comparing the sampling algorithms BFS, NBRW-rw, MHRW and MHDA. A truely uniform sample of Facebook nodes was obtained by generating 32-bit user IDs and discarding the non-existing ones or invalid users. In this study, it is mainly focussed on open/publicly available basic information and do not consider the detailed study of user profiles that are more privacy-sensitive [4]. For implementation, we ran 28 independent crawls for each

algorithm, namely BFS, NBR-rw, MHRW and MHDA, all seeded at the same initial, randomly selected nodes. Each independent crawl continues until exactly 81K samples are collected as described in [4]. This results in a uniform sample (UNI) that is used for the comparison in this study.

Slashdot0811 dataset is a technology-related news website and is known for its specific user community. This network has 77360 nodes and 905468 edges and contains friend/foe links between the users of the dataset. The Slashdot0811 dataset was obtained in November, 2008 [2]. The data is downloaded from the Stanford Large Network Dataset Collection [16] which includes mostly bibliometric, email, online social networks, friendship networks, communication networks, collaboration networks, citation networks, peer to peer networks and graph of linked web pages.

2.2 Node Degree Distribution

The Fig 1 below represents the degree distributions estimated by the sampling algorithms- BFS, NBRW-rw, MHRW and MHDA when considering them accordingly with the behaviour of given uniform sampler (or the dataset). The uniform sample is taken as a 'baseline' for comparing the deviations of the degree distributions of the considered sampling algorithms. It can be observed that MHDA and MHRW performs excellently well and agree almost perfectly with the uniform sample. From these two, MHDA deviates very less from the original sample when compared with the case of MHRW. Thus, MHDA more accurately yields an asymptotically uniform sample with extremely less deviation than the others.
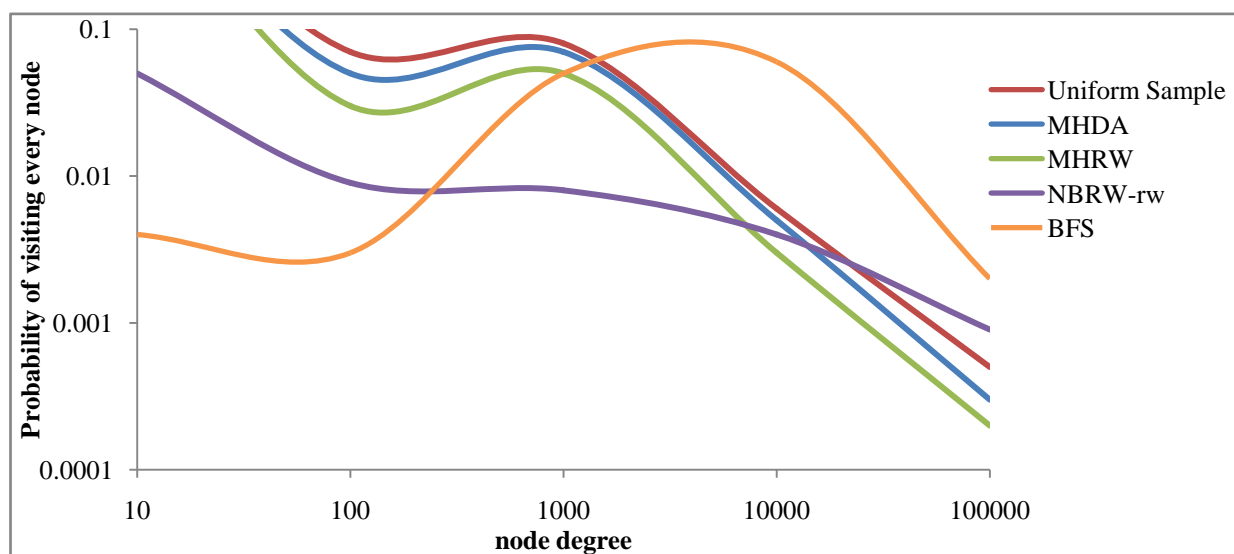

Fig 1: Representing the deviations of the sampling algorithms from the uniform sample

BFS and NBRW-rw gives poor performance and does not lead to the proper results because they deviate significantly more than MHDA and MHRW. BFS is heavily biased to high degree nodes and NBRW-rw samples require use of proper re-weighting procedures to generate correct results.
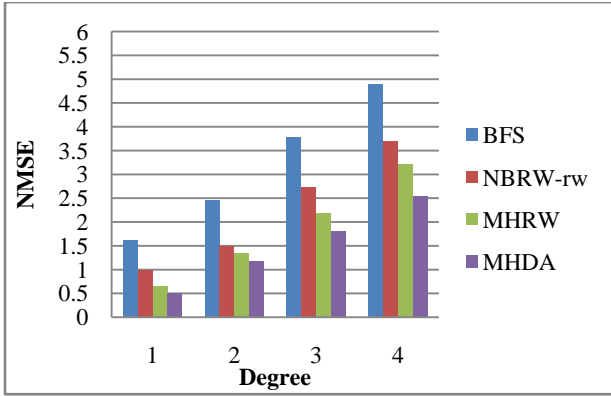
Fig 2: Representing bar graph for NMSE per degree

In Fig 2, the Normalised Mean Square Error (NMSE) for a node degree has been taken as a metric to demonstrate the behaviour of the sampling algorithms on a bar graph. For the estimation of node degree distribution, we need to first calculate the NMSE of node degree k , using the below given formula [2]-

$$NMSE(k) = \frac{\sqrt{E[(\hat{\theta}_k - \theta_k)^2]}}{\theta_k}$$

where $\hat{\theta}_k$ is the estimation of θ_k based on the sampled graph. NMSE(k) metric is defined in order to show the difference between the degree distribution of the sampled graphs and the original one.

The lower the NMSE value of an algorithm, the better is the performance of the sampling algorithm in the social network graph. Thus, it can be observed from the above figure that BFS is biased to high degree nodes and has larger NMSE value than others. In contrary, on the other hand MHDA has smaller NMSE value and hence, the performance of MHDA is excellently well. The smaller NMSE value of MHDA converges faster to the uniform distribution and keeps the node degree distribution well.

2.3 Clustering Coefficient
The given table below shows the Network Average Clustering Coefficient (NACC) for each of the considered sampling algorithms and their deviations (RE) from the original graph.

Table 1. Table captions should be placed above the table

NACC		Deviation (RE)
Slashdot0811 (Original)	0.0555	
BFS	0.106	90.99%
NBRW-rw	0.0498	21.80%
MHRW	0.0504	9.18%
MHDA	0.0519	6.48%

Larger the network average clustering coefficient, poorer is the performance of the algorithm. Clustering coefficient depends on the node degrees. Since, BFS is biased to high degree nodes therefore, the average clustering coefficient of BFS is much larger when compared to NBRW-rw, MHRW

and MHDA sampling algorithms. We need to calculate the NACC for each of the sampling algorithm and the formula [2] is given as-

$$NACC = \frac{1}{n} \sum_{i=1}^{n} C_i$$

where n is the total number of nodes in a graph and, C_i is the local clustering coefficient [2] for a node i in undirected graphs given by the formula-

$$C_i = \begin{cases} \dfrac{2|E_{v,w}|}{k_i(k_i - 1)} & \text{if } k_i > 1 \\ 0 & \text{otherwise} \end{cases}$$

A property metric called as relative error (RE) is defined that quantifies the sampling performance of BFS, NBRW-rw, MHRW and MHDA on different graphs and is expressed by the formula [2]-

$$RE = \frac{|Samples - Original|}{Original}$$

where, Samples represents the NACC metric of the sampled graph and Original represents the NACC metric of the original graph.

An algorithm with less NACC value gives much better performance. MHDA has a smaller NACC value according to the above table. Therefore, MHDA performs far better than BFS, NBRW-rw and MHRW and converges much faster to the uniform stationary distribution.
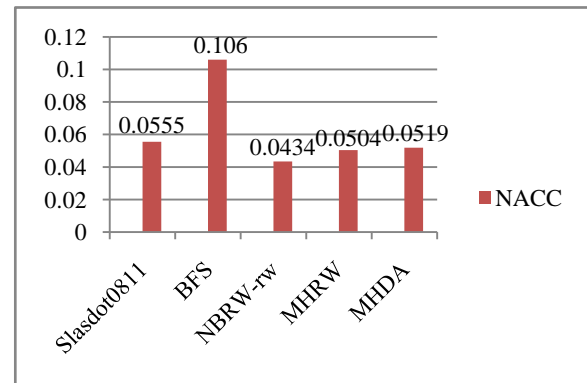

Fig 3: Comparison of Avg. Clustering Coefficients (NACC) of sampling algorithms with respect to the original graph (Slashdot0811).

3. CONCLUSION AND FUTURE SCOPE
This paper shows a comprehensive study on several existing sampling algorithms and the analysis of their performance in maintaining the original properties of the graph. The conclusion says that MHDA gives an unbiased sample and converges faster than MHRW. Therefore, MHDA performs extremely well in terms of both node degree distribution and clustering coefficient when compared to BFS, NBRW-rw and MHRW. BFS and NBRW-rw both gives a poor performance because BFS is biased towards high degree nodes and obtains larger average clustering coefficient and NBRW-rw samples require appropriate re-weighting procedures. Both BFS and NBRW-rw show larger deviations.

The future work can be extended with a parameter Total Variance Distance (TVD) and calculating the estimation accuracy of the existing sampling algorithms and analysing their performances in social network graph.

4. REFERENCES

[1] D. Corlette and F. Shipman, "Capturing On-line Social Network Link Dynamics using Event-Driven Sampling, " International Conference on Computational Science and Engineering, IEEE, 2009.

[2] T. Wang, Y. Chen, Z. Zhang, T. Xu, L. Jin, P. Hui, B. Deng, and X. Li, "Understanding Graph Sampling Algorithms for Social Network Analysis," 31st International Conference on Distributed Computing Systems Workshops (ICDCSW), IEEE, 2011.

[3] C.-I. Wong, K.-Y. Wong, K.-W. Ng, W. Fan, and K.-H. Yeung, "Design of a Crawler for Online Social Network Analysis," WSEAS TRANSACTIONS on COMMUNICATIONS, vol.13, 2014.

[4] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs," in proc. of the IEEE INFOCOM, 2010.

[5] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Practical Recommendations on Crawling Online Social Networks," IEEE Journal on Selected Areas in Communications, vol.29, Issue 9, October 2011.

[6] W. M. K. Trochim, "Sampling," 2002, [Online],Available: http://trochim.human.cornell.edu/tutorial/mugo/tutorial.htm.

[7] M. Doherty, "Probability versus Non-Probability Sampling in Sample Surveys," The New Zealand Statistics Review, pp. 21-28, March 1994.

[8] M. V. Gwetu, "The Application of Sampling to the Design of Structural Analysis Web Crawlers," IJCSI, vol. 7, Issue 4, no. 8, July 2010.

[9] M. Kurant, A. Markopoulou, and P.Thiran, "On the Bias of Breadth First Search (BFS) and of Other Graph Sampling Techniques," International Teletraffic Congress (ITC 22), 2010.

[10] L. Goodman, "Snowball Sampling," Annals of Mathematical Statistics vol.32, pp.148-170, 1961.

[11] S. Dhawan, K. Singh, and K. Saini, "Comparing Crawling Techniques In Social Network," IJIRS, vol.3, Issue 7, July 2014.

[12] C.-H. Lee, X. Xu, and D. Y. Eun, "Beyond Random Walk and Metropolis –Hastings Samplers: Why you should not Backtrack for Unbiased Graph Sampling," SIGMETRICS, 2012.

[13] M. Strand, "Metropolis-Hastings Markov Chain Monte Carlo," May 2009.

[14] R.-H. Li, J. X. Yu, L. Qin, R. Mao, and T. Jin, "On Random Walk Based Graph Sampling," 31st International Conference on Data Engineering (ICDE), IEEE, 2015.

[15] D. Achlioptas, D. Kempe, A. Clauset, and C. Moore, "On the Bias of Traceroute Sampling or, Power-law Degree Distributions in Regular Graphs," in proc. of the 37th ACM Symposium on Theory of Computing (STOC), May 2005.

[16] Stanford Large Network Dataset collection. http://snap.stanford.edu/data/index.html.