SURVEYING ONLINE SOCIAL NETWORKS

RACHITA NAGPAL AND ROOPALI GARG
U.I.E.T, Panjab University, IT Department, Chandigarh, India
Email: rachitanagpal1891@gmail.com, roopali.garg@gmail.com

ABSTRACT: Awareness of online social networking sites is increasing day by day. This has made the topology of networks complex. Analyzing such networks is difficult when we take into consideration a large graph. For such analysis we use graph sampling techniques. These techniques extract a representative sample graph from the original graph and the obtained graph is taken into consideration for analysis. Usage of representative sample graph decreases the complexity of the graph and leads to easy computation. The results can then be generalized on the entire network. In this paper, we discuss various sampling techniques and infer that Metro-polis Hastings Random Walk (MHRW) is an efficient graph sampling algorithm.

KEYWORDS: online social networks, graph sampling, Metro-polis Hastings Random Walk (MHRW)

INTRODUCTION

In today's world social networking site is becoming a part of everyone's life. Social networking refers to the process of making and maintaining relations with people who are similar in nature. Social sites are becoming a platform for individuals for making contacts, expanding business through proper advertisements and also as a source of entertainment. This type of interaction among individuals involves creation of a public profile that consists of the information related to the user. This information is shared among individuals to interact with each other. This type of social sites is often heavy loaded and is difficult to analyse. Therefore, analysis on such sites requires the usage of effective and efficient algorithms. The social networking site can be considered as a large social graph that is to be evaluated. Exploring and evaluating one large graph is difficult, therefore graph sampling algorithms are deployed. In graph sampling, a graph known as the target graph is sampled using the various graph sampling techniques to obtain the required sampled graph known as the representative graph. It is named as representative graph because this graph is used to represent the entire large social graph. The obtained sample graph is said to be similar to target graph when the characteristic properties of the sampled graph matches to the properties of the target graph.

Graph sampling techniques [1] are either based on the nodal properties or the edge properties. Based on the nodal properties graph sampling techniques can be categorized as: Breadth First Search, Random Walk, Random node selection, Random edge selection and Metropolis-Hastings Random Walk. Frontier sampling technique is type of edge based sampling technique. These techniques are discussed later.

The rest of the paper is described as follows: Section II describes the work done related to sampling techniques; Section III defines various graph sampling techniques; Section IV describes MHRW algorithm; Section V discusses the inferences drawn and Section VI derives the conclusions.

Mc-GrawHill Education
GRENZE ID: 02.PEIE.2015.6.506

RELATED WORK

Usage of online social networking sites is growing rapidly during the last decades. People nowadays are becoming aware of the importance of internet. One of the regular usage of internet is for entertainment purposes. Entertainment can be in the form of watching movies, listening to music, chatting with friends and many more. The last source seems to be attractive. Chatting with friends over internet is one of the mediums used by today's generation to remain in contact with the distinct friends in any type of environment. One is able to communicate with more than one friend simultaneously over the network. This type of network is often congested and complex. Many studies have been done to analyze such networks.

Analyzing such networks sounds difficult and complex when seen as a whole. Hence to simplify this type of analytical task, we require the usage of graph sampling in which the target or the original graph is sampled to obtain an efficient sample graph. This resultant sample graph represents the original target graph. Cautions have to be taken for obtaining the sample graph. Hence we obtain the sample graph using the sampling techniques such as Breadth First Search (BFS), Random Walk (RW) and Metro-polis Hastings Random Walk (MHRW).

Kathryn Dempsey, Kanimathi Duraisamy, Hesham Ali and Sanjukta Bhowmick [2] used a parallel graph sampling technique to obtain a chordal graph for making comparison between the properties of the original and the chordal graph and inferred that properties were preserved in chordal graph even when the network size was reduced up to 40%.

Maciej Kurant, Athina Markopoulou and Patrick Thiran [3] explored the biasness of BFS based on arbitrary distribution of degree. Comparison between various sampling techniques was also made.
Graph sampling techniques are also implemented in the arena of data mining [4]. In this area, random area sampling is used to obtain the frequent pattern of sub graphs from the original graph.
Kiran K. Rachuri and C. Siva Ram Murthy stated in [5] that information discovered from networks is efficient when implementation using RW algorithm with level biased steps.

Colin Cooper, Tomasz Radzik and Yiannis Siantos in [6] used RW to evaluate the parameters of the network effectively and efficiently. Also results in [6] showed that it is possible to evaluate the property of the graph even if the crawl is halted prematurely.

Reference [7]-[11] discusses the various applications of RW algorithm in field of OSN and mobile networks. Efficiency of random walk is improved when conductance of network is increased; implementing a path based tool one can overcome with the problem of inverse influence in OSN.

GRAPH SAMPLING TECHNIQUES

The graph sampling techniques categorized in Fig. 1 are described as follows:

Node Sampling

In this type of technique, the nodes of the graph are taken into consideration. Here, the nodes are sampled and the edge connections in the sample graph are the same as between the nodes in the target graph. Mathematically, given a target graph T= (N, E), where N refer to nodes in the target graph and E is the edge between two nodes i and j, where i, j both belong to N. From the above given target graph, a sample graph is obtained with same edge connection as in target graph but

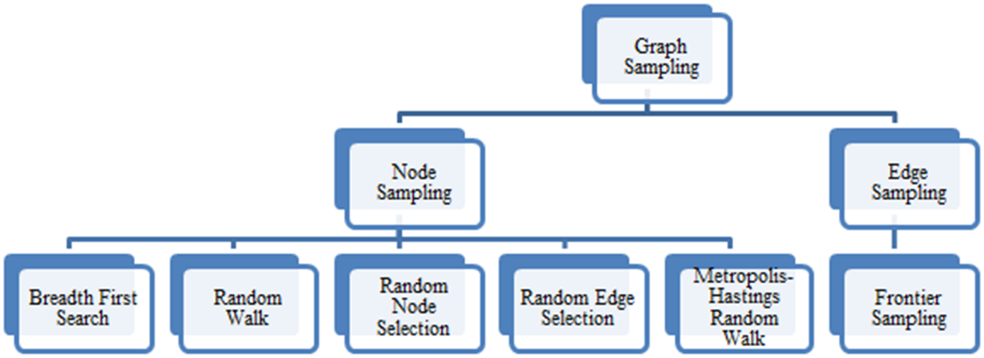

Figure 1. Graph Sampling Techniques

number nodes is less than N in target graph. A sampled graph S= (N’, E) is obtained where N’ depicts the nodes sampled and N’ is the subset of N. E is an edge between the nodes i and j, where i, j belongs to N’. Different types of node sampling techniques are explained as under:

- Breadth First Search [3], [12]: Fig. 2 describes the working of BFS algorithm. In this algorithm, a node is randomly chosen and referred to as the seed node. This seed node is inserted in either of the two sampling queues. The first queue is the sampled queue which consists of the nodes that are sampled and the second queue is the processed queue which consists of nodes that are to be processed. At the initial stage the seed node is in the processed queue and added to the sampled queue to process all of its neighbouring nodes. All the neighbouring node are in the processed queue initially and then added to the sampled queue one by one after completion of the processing of each of the neighbouring node. When nodes in the sampled queue reach the stopping criteria, the algorithm stops and the required representative sampled graph is obtained.

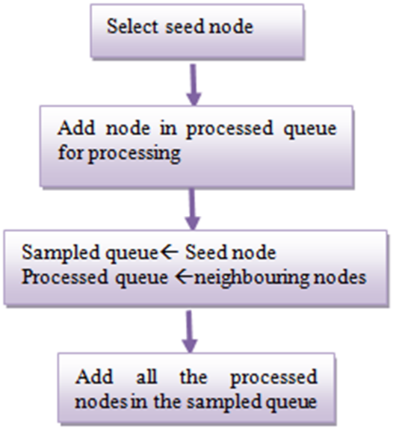
Select seed node
Add node in processed queue for processing
Sampled queue← Seed node
Processed queue ←neighbouring nodes
Add all the processed nodes in the sampled queue

Figure 2. BFS flowchart

- Random Walk [13]: Random walk is one of the graph sampling techniques that forms the basis for analytical purposes in the field of share market, computer science, chemistry and many other fields. Random walk helps in modelling of the processes involving the evaluation of random variables that lead to the invention of a system consisting of random values that may or may not differ with respect to time. In other words, mathematical formulation of paths consisting of various random steps performed in succession is referred to as random walk.
- Random node selection: Using the concept of uniform distribution, nodes are randomly selected to obtain the required sampled graph.
- Random edge selection: This algorithm is similar to random node selection algorithm. In this case, set of edges are selected at random. Selection of edges led to selection of the nodes that are to be included in the sampled graph. The nodes included in the sampled graph are mainly the endpoints of the edges.
- Metropolis-Hastings Random Walk (MHRW) [14], [15], [16]: In this algorithm, sampling is performed on the basis of the proposal function. This proposed function depends on the probability distribution of the degree of nodes. In this algorithm, the proposal function is either accepted or rejected randomly. This leads to the changes in the transition probabilities of the nodes and hence makes the sampled nodes converge to the probability distribution. Detailed algorithm is explained later.

Edge Sampling

In this technique, sampled graph is obtained by applying sampling operation on the edges of the target graph. The edges of the target graph are sampled and the end points of the sampled edges are chosen to be the nodes that are to be included in the required sampled graph.

- Frontier Sampling: A set of nodes is selected at random from the target graph and the nodes in the set of nodes are referred as seed nodes. A seed node ‘v’ belonging to the set of nodes is selected according to the defined probability function and then any outgoing edge (v, u) from ‘v’ is uniformly selected and ‘v’ is then replaced with ‘u’. This process continues until required sample graph of optimal size is obtained.

We will see later in the discussion that MHRW algorithms prove to be an efficient one. The MHRW algorithm is explained in detail in the further section.

METROPOLIS HASTINGS RANDOM WALK

It is one of the efficient node sampling algorithms that obtain the sampled nodes on the basis of probability of the degree distribution of the nodes. These types of sampled nodes are difficult to obtain through direct means of sampling. A proposal function is generated to sample the nodes depending on the nodes degree distribution. This proposal function is either accepted or rejected.

A node ‘n’ is randomly selected as the seed node with degree d. Proposal function is a function of the degree of node defined as P (n) = d (n). In Fig 3, MHRW algorithm randomly chooses a neighbouring node ‘w’ and generates a random number ‘c’ that lies in the range of uniform distribution (0, 1). If c is less than the ratio of P(n) to P(w), then the proposal function is accepted and then node ‘w’ is sampled using MHRW. If the proposal function is rejected, another neighbour of ‘n’ is selected except ‘w’.

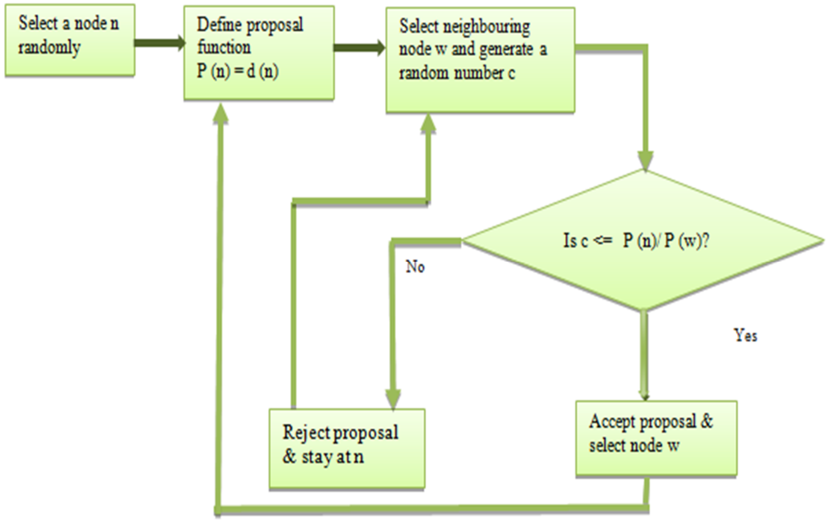
Select a node n randomly
Define proposal function P (n) = d (n)
Select neighbouring node w and generate a random number c
Is c <= P (n)/ P (w)?
No
Yes
Reject proposal & stay at n
Accept proposal & select node w

Figure 3. MHRW flowchart

RESULTS AND DISCUSSIONS

From the related work it is clear that many algorithms have been devised for sampling. Among these sampling algorithms, BFS has been used widely in the previous studies which had already been described in section II. MHRW algorithm is new to the field of sampling.

The reason behind the success of MHRW is that BFS is bias to high degree of nodes. MHRW obtains almost equal degree of distribution. Degree of Normalised Mean Square Error (NMSE) for MHRW is also good than BFS algorithm. From [14] it is clear that NMSE is very small for whole data set considered under sampling. Previous studies shows that MHRW gives a high normalized average clustering coefficient than BFS. From the results shown below it is very clear that BFS is biased for high degree nodes. From the Cumulative Distribution Function (CDF) plot shown in Fig. 4 it can be seen that MHRW and FS shows the same result with original ones[14]. Hence, MHRW obtain high degree of distribution. So the performance of MHRW is better than that of BFS. From Fig. 5 it could also be depicted that NMSE for both algorithms is smaller in tightly connected graphs [14]. This shows that the performance of MHRW goes down in case of loosely coupled nodes.

MHRW can be used in combination with different probability distribution functions providing a very simple and efficient heuristic path for social network over large scale. The average time taken for sampling Facebook using MHRW sampling algorithm was 172s which is very less in comparison to BFS [17].

CONCLUSION

This paper attempts to give a confined description of various graph sampling techniques. These techniques are then compared in terms of CDF and NMSE. We infer from the comparison that

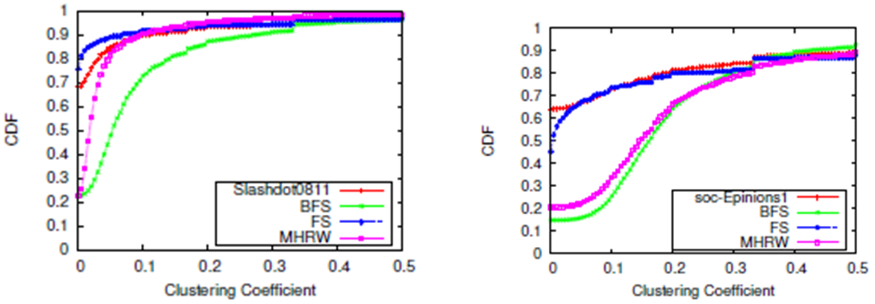

Figure 4. Comparison of Slashdot 0811 and Soc-Epinions1 with BFS, FS and MHRW in terms of clustering coefficient CDF

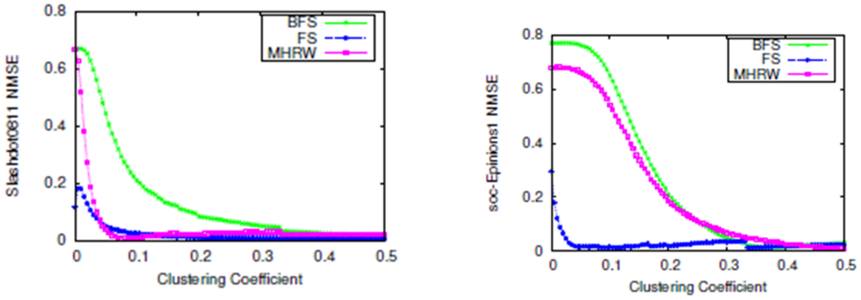

Figure 5. Comparison of Slashdot 0811 Soc-Epinions1 with BFS, FS and MHRW in terms NMSE

BFS and RW show biased nature towards nodes having high degree. MHRW, on the other hand, shows unbiased behaviour. To verify the results we apply these techniques on different data sets and same results were inferred.

We propose to use MHRW algorithm for sampling large graph and then construct a backbone of this sampled graph. Backbone will be constructed in such a way that the properties of energy and connectivity in the sampled graph is the same as that in the original graph. This backbone is constructed using the Connecting Dominating Set (CDS) technology. On the target graph first sampling is done and then backbone is constructed.

ACKNOWLEDGEMENT

I would like to thank my co-author Mrs. Roopali Garg for providing me with her valuable guidance in this field. Her continuous and valuable support has been of great help in completing the research work.

REFERENCES

J. Leskovec and C. Faloutsos, “Sampling from Large Graphs,”ACM, 2006

K.Dempsey, K. Duraisamy, H. Ali and S. Bhowmick, “A parallel graph sampling algorithm for analysing gene correlation networks,” Procedia Computer Science, Elsevier, Volume 4, pp. 136-145, 2011

M. Kurant ; A. Markopoulou, ; P. Thiran, “On the bias of BFS (Breadth First Search),”Teletraffic Congress (ITC),IEEE,pp. 1-8,2010

R Zou and LB Holder , “ Frequent subgraph mining on a single large graph using sampling techniques, ” MLG '10 Proceedings of the Eighth Workshop on Mining and Learning with Graphs, ACM, pp. 171-178, 2010

K.K. Rachuri and C.S.R Murthy, “Energy Efficient Search in Sensor Networks using Simple Random Walks with Level Biased Steps, ”Distributed Computing Systems Workshops, ICDCS Workshops, IEEE pp. 178 – 185, 2009

C. Cooper, T. Radzik, and Y. Siantos, “Estimating network parameters using random walks,” Computational Aspects of Social Networks (CASoN), IEEE, pp.33-40,2012

Zhuojie Zhou , Nan Zhang Zhiguo Gong and G. Das, “Faster random walks by rewiring online social networks on-the-fly,”Data Engineering (ICDE), IEEE, pp. 769-780,2013

Figueiredo and R. Daniel, “Walking around in a changing world: Understandin grandom walks over dynamic graphs,” Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), IEEE, pp.367, 2012

Bo Han , Jian Li ;and A. Srinivasan, “Your Friends Have More Friends Than You Do: IdentifyingInfluential Mobile Users Through Random-Walk Sampling,” Networking, IEEE/ACM Transactions, volume 22 pp. 1389-1400 , 2014

Qingyu Li , Panlong Yang , Yubo Yan and Yue Tao, “Your friends are more powerful than you: Efficient taskoffloading through social contacts,” Communications (ICC), IEEE, pp. 88-93, 2014

Zhaoyan Jin , Quanyuan Wu , Dianxi Shi and Huining Yan , “Random Walk Based Inverse Influence Research in OnlineSocial Networks,”High erformance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), IEEE, pp 2206-2213, 2013

A. Mohaisen, , Pengkui Luo , Yanhua Li , Yongdae Kim and Zhi-Li Zhang, “Measuring bias in the mixing time of social graphs due to graph sampling,” MILITARY COMMUNICATIONS CONFERENCE, IEEE, pp. 1-6, 2012

Rong-Hua Li , J.X. Yu, Xin Huang and Hong Cheng, “Random-walk domination in large graphs” Data Engineering (ICDE), IEEE, pp. 736-747, 2014

Tianyi Wang , Yang Chen , Zengbin Zhang , Tianyin Xu , Long Jin , Pan Hui , Beixing Deng and XingLi , “Understanding Graph Sampling Algorithms for Social Network Analysis,”Distributed Computing Systems Workshops (ICDCSW),IEEE, pp. 123-128 , 2011

M. Gjoka, , M. Kurant, , C.T. Butts and A. Markopoulou, “Walking in Facebook: A Case Study of Unbiased Samplingof OSNs,” INFOCOM,IEEE, pp. 1-9, 2010

M. Gjoka, , C.T. Butts, M. Kurant and A. Markopoulou, “Multigraph Sampling of Online Social Networks,” Selected Areas in Communications, IEEE, pp. 1893-1905, 2011

C.A. Pina-Garcia, and Dongbing Gu , “Collecting Random Samples from Facebook: An EfficientHeuristic for Sampling Large and Undirected Graphs via a Metropolis-Hastings Random Walk,”Systems, Man, and Cybernetics (SMC), IEEE, pp. 2244-2249, 2013