RESEARCH REVIEW JOURNALS
Volume-04
Issue-01
January -2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Social Network Crawling Techniques: A Review

Shweta

Department of Computer science, GJU Hisar, Haryana (India)

ARTICLE DETAILS

Article History
Published Online: 20 January 2019

Keywords
Sampling methods, social network services, Facebook, graph sampling, bias

***Corresponding Author**
Email: chhikarashweta[at]gmail.com

ABSTRACT

In order to crawl online social network such as Facebook, many sampling techniques have been introduced which are based on the undirected Graph sampling methods to produce uniform samples of users. These techniques includes many graph sampling algorithms, trying to extract a snapshot of original graph having almost similar properties. This paper discusses the research that has been done in the area of sampling techniques for crawling OSN. The goal of this paper is to discuss the advantages and disadvantages of currently existing Crawling techniques.

1. Introduction

Online social network (OSN) became very popular now a days, it can be witnessed with huge number of users [13]. Some of popular OSN such as Facebook[16], LinkedIn, Twitter etc. have gathered more than hundreds of millions of users[5][7]. OSN is a very powerful tool for connecting people virtually and provides a mirror image of their real life relationships and society. Facebook reported having 1.2 billion monthly active users on January 2014, globally, therefore it became an interest for researchers. Sociologist want to study FB for understanding the behavior of people at a large level. Marketers want FB for knowing the interest of people so that they can decide the aspects of marketing strategies, technician and researchers show their interest in order to maintain bandwidth and study and implement various properties as well as algorithms.

But crawling the social web [15][10][12] needs more technical challenges than crawling the simple web due to following reasons:

1. Due to high privacy settings of OSN
2. Very huge amount of crawling data which is very difficult to store and handle.

OSNs can be represented as graphs where nodes represent users and edges represent connections. Facebook [16], in particular, is characterized by a simple friendship schema, so as it is possible to represent its structure through a simple, unweighted, undirected graph. If we crawl the whole Facebook graph [13] it come outs to be very complicated and complex. Thus we need a snapshot or sub graph of original graph which can represent the similar properties of original one. Thus we need the sampling techniques for the same.

Crawling OSN composed of many problems such as we do not have computational resources able to mine and work with the whole Facebook graph because it is not minor to challenge large scale mining problems: let's take an example, Gjoka et al. [5] dignified the total amount of overhead included in crawling the complete Facebook graph as 44 Terabytes of data to be downloaded and handled. Moreover, even when such data can be acquired and stored locally, it is non-trivial to devise and implement functions that traverse and visit the graph or even evaluating simple metrics. Hence it is preferred to work on a sample of graph which holds most of the characteristics of the graph. This paper explores the major sampling techniques and analyze those techniques in the certain scenarios. Later, in the next section, comparison of all the techniques is done with their pros and cons.

2. Random selection of nodes

2.1Sampling by random node selection

J. Leskovec and C. Faloutsos [2]introduced sampling algorithm based on random node selection. Simplest method to create a sample of any graph is by randomly selecting a node and then reselecting from the remaining node in same order. A sample that is formed from such a nodes is then a graph induced by the set of nodes N. This algorithm is defined as Random Node (RN) sampling or the Random PageRankNode (RPN) sampling. But the main disadvantage of Random Degree Node (RDN) sampling is that it has more preference towards nodes which have high degree. They set the probability of anode being selected into the sample to be proportional to its PageRank weight. The probability of a node being selected is proportional to its degree.

2.2 Random Walk (RW) sampling

J. Leskovec and C. Faloutsos [2] uniformly at random pick a starting node and then simulate a random walk on the graph. At every step with probability c = 0.15 (the value commonly used in literature) go back to the starting node and re-start the random walk. In random walk, there can come a problem of sucking, for example, if we start with the sink node, or if the node is isolated one or if the number of nodes are small enough. One solution for such problem can be that is if we have large number of nodes then instead of visiting all the nodes so that required sample size can be meet. We can start with another starting node and then repeating the whole procedure again.

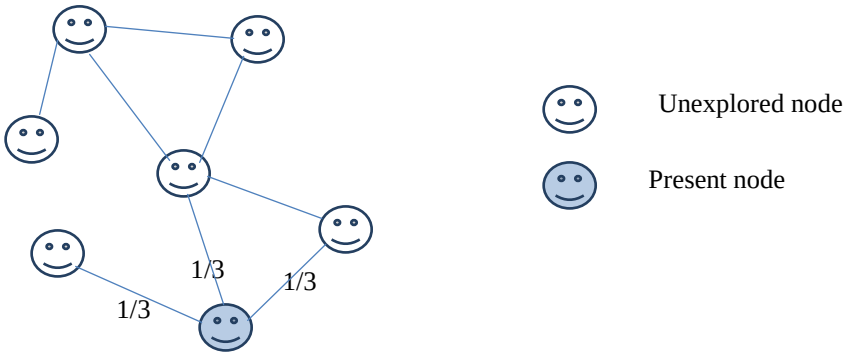

Fig 3. RW approach in social graph [2]

3. Breadth-First-Search Sampling

Breadth-first-search (BFS) is a common traversing algorithm for traversing the graphs. BFS is quite easy to implement and optimal. BFS works well in traversing unweighted and for the undirected graphs. This algorithm starts with the root/seed node and keeps on traversing the neighbors of the seed node, and then putting them into the First in First out Starting (FIFO). Nodes are visited in the arrival order of queue. Hence the Traversing of the graph in BFS looks like expanding wave front. This algorithm, virtually, concludes its execution when all discovered nodes have been visited. If we traverse very large graphs, such as OSNs, BFS terminating condition results in very large computational resources and time. Also, if the BFS is incomplete then sampling can results in biased result [14] [8], mainly towards high degree nodes.

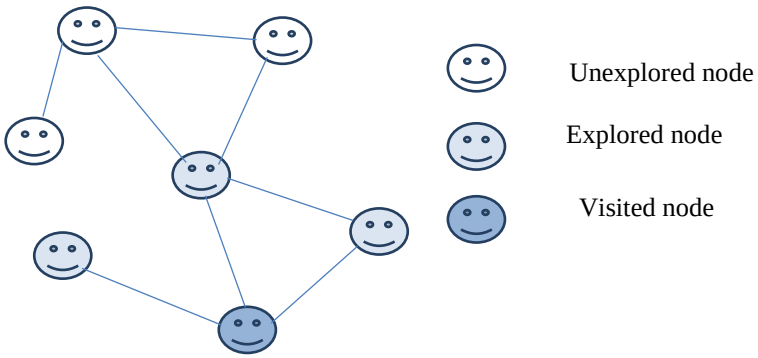

Fig 4. BFS approach in SNS [3]

4. Re-Weighted Random Walk (RWRW)

O. Skare. Introduced RWRW which is essentially a special case of importance sampling [4] reweightedby the stationary distribution of the RW MCMC. In general, Re-Weighted Markov ChainMonte Carlo Walk consists of following two steps: in the first step, an initial MCMC process consisting of a (possibly biased)walk on the graph is used to generate a sample with known asymptotic statistical properties; and in the second step, a re-sampling process is engaged to enable the use of derived sample to uniform the distribution.

Let's consider a random walk which has traversedV = v1....vn distinct nodes. Each node can belong to one of 'm' groups with respect to a property of interest A, which might be the degree, network size or any other discrete-valued node property. Let (A1, A2... Am) be all possible values of A and corresponding groups; U^{m1}Ai = V. E.g., if the property of interest is the node degree, Ai contains all nodes u that have degree ku = i. To esteem ate the probability distribution of A, we need to estimate the proportion of nodes with value Ai, i = 1,..m: [4]

$$\hat{p}(Ai) = \frac{\sum u \in Ai\ 1/ku}{\sum u \in V\ 1/ku}$$

5. Metropolis-Hastings Random Walk (MHRW):

Most of the walks consist of the biasness towards particular nodes. Instead of finding the biasness and the removing it, after the walk, we can modify the traversing style itself. There might be more probability that it converges to the desired uniform sample of the whole graph. One such method is MHRW. The Metropolis-Hastings algorithm [5] is based on the Markov Chain Monte Carlo (MCMC) technique [6]for deriving the sample from a probability distribution μ which is otherwise difficult to sample from directly. Sample nodes from the uniform distribution μv = 1/|V |[5]. This can be achieved by the following transition probability:

$$P_{\upsilon,w}^{MH} = \begin{cases} \frac{1}{k_\upsilon}\min(1,\frac{k_\upsilon}{k_w}) & \text{if } w \text{ neighbor of } \upsilon \\ 1-\sum_{y\neq\upsilon} P_{\upsilon,y}^{MH} & \text{if } w=\upsilon \end{cases}$$

In every iteration of MHRW, at the current node v we randomly select a neighbor w and move there w.p. min (1, kv/kw). In MHRW, next node is selected which has a small degree, while the nodes with the high degree are eliminated. Thus this process succeeded in removing the biasness towards the high degree nodes to some extent.

6. Uniform Sample of User ids (UNI)

UNI sampling method is derived from the technique "Rejection Sampling". This technique guarantees to sample uniformly random userIDs from the allocated Facebook users. This selection is independent of their actual distribution in the userID space, even if they are not allocated evenly or sequentially across the userID space. Minas Gjoka called this method as "UNI" [5]. Every FB user is provided with a unique and random user id. According to Minas Gjoka [5], randomly select the userID from a defined range and validate it weather corresponding user node exists or not. If the nodes exist then that particular node is added to the database of selected nodes otherwise node is directly rejected. This procedure is repeated until the desired sample is extracted. Before starting UNI, a particular range is selected.

7. Comparison of Different Sampling Techniques

Sr .No.	Method	Advantages	Disadvantages
1.	BFS	Fastest and efficient graphs traversing technique.	Leads to biasness towards high degree node[14]. Hence we didn't get unique samples of nodes.
2.	Random walk	Unique nodes are selected every time with a uniformity and most efficient.	Inherently biased towards high degree node. A node with twice degree will be visited twice more often. Hence repetitions of nodes
3.	RWRW	Biasness towards high degree node is corrected using reweighting of nodes.	Biasness is still present in starting, later after crawling reweighting is done to remove the same. Hence more overload and time is needed.
4.	MHRW	Biasness is removed by modifying transition properties which preferred low degree node over high degree node and yield uniform samples.	It require a large number of initial rejection and also lots of computations are required to first calculate probability of each node and then compare, later select the appropriate among them. Also, sometime MHRW suffers from self-loop for a low-degree node which is surrounded by large-degree nodes.
5.	UNI	Unique nodes are sampled by randomly selecting 32 bit user id from a particular range. Hence result in very uniform samples.	Need to know the range of user IDs already. When the range is very high, while nodes are few, then the matrix become sparse matrix which leads to lots of rejection.

8. Conclusion

Facebook graph is a very complex graph and hence taking sample for such graph is itself a challenging task. In this paper, various crawling techniques are explored and analyzed. Also various sampling techniques are compared for crawling the social network. Every techniques works well in one case while there are various limitations attached with them. From the five major techniques we have analyzed that MHRW works significantly well as compared to other techniques. But MHRW still suffers from various disadvantages. Pros and cons of all techniques are explained. A comparison table is framed which state the advantage and disadvantages of all the defined techniques.

References

1. 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.
2. O. Skare, "Improved sampling-importance resampling and reduced bias importance sampling", *Scandin. journ. of stat., theory and apps*, 2003.
3. Metropolis, M. Rosenblut, A. Rosenbluth, A. Teller, and E. Teller, "Equation of state calculation by fast computing machines",*J. Chem. Physics*, 21:1087–1092,1953.
4. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter,"Markov Chain Monte Carlo in Practice",*Chapman and Hall/CRC*, 1996.
5. M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook:A case study of unbiased sampling of OSNs", *in Proceedings of the IEEE Infocom conference*, 2010.
6. W. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, vol. 57, no. 1, pp. 97-109, 1970.
7. Manos Papagelis, Gautam Das, and Nick Koudas, "Sampling Online Social Networks" ,*in IEEE transactions on knowledge and data engineering,* vol. 25, no. 3, march 2013.
8. Maciej Kurant, Athina Markopoulou, Patrick Thiran, "On the bias of BFS (Breadth First Search)", in IEEE.
9. J. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez, "The little engine (s) that could: Scaling online social networks," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 4, pp. 375–386, 2010.
10. S. Agarwal and S. Agarwal, "Social networks as Internet barometers for optimizing content delivery networks," in *Proc. 3rd IEEE Int. Symp. On Advanced Networks and Telecommunication Systems*, New Delhi, India, Dec. 2009.
11. M. Sirivianos, X. Yang, and K. Kim, "SocialFilter: introducing social trust to collaborative spam mitigation," in *Proc. IEEE INFOCOM*, Shanghai, China, 2011.
12. A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. 7th ACM SIGCOMM Conf. on Internet measurement*, San Diego, CA, 2007, pp. 29–42.
13. C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao, "User interactions in social networks and their implications,"

in *Proc. 4th ACM European Conf. on Computer systems*, Nuremberg, Germany, 2009, pp. 205–218.

14. M. Kurant, A. Markopoulou, and P. Thiran, "On the bias of BFS (Breadth First Search)," in *Proc. 22nd Int. Teletraffic Congr., also in arXiv:1004.1729*, 2010.
15. M. Salganik and D. D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, vol. 34, no. 1, pp. 193–240, 2004.
16. B. Viswanath, A. Mislove, M. Cha, and K. Gummadi, "On the evolution of user interaction in facebook," in *Proc. 2nd workshop on Online social networks*, Barcelona, Spain, 2009, pp. 37–42.

