Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 3, March 2016, pg.01 – 05

A SURVEY ON BREADTH FIRST SEARCH (BFS) & METROPOLIS HASTING RANDOM WALK (MHRW)

Ms. Bhagyashree.P[1], Ms. Gayathri.G.S[2]
[1]M.Tech Student, Department of Computer Science and Engineering
New Horizon College of Engineering, Bangalore, India
[2]Assistant Professor, Department of Computer Science and Engineering
New Horizon College of Engineering, Bangalore, India
[1] bhagyashree.pandian@gmail.com; [2] gayathrigummaraj@gmail.com

Abstract--- Graph sampling is a technique of selecting a subset of the original graph making the scale small for improved computations. This technique provides an efficient and yet an inexpensive solution. In this paper, we examine Breadth First Search (BFS) and Metropolis Hasting Random Walk (MHRW) graph sampling algorithm and find out which algorithm performs better than the other.

Keywords--- Graph sampling, Social networks, Social search, BFS, MHRW, OPICS

I. INTRODUCTION

Social networks are popular infrastructures for communication, interaction, and information sharing on the Internet. These social networks provide communication, storage and social applications for hundreds of millions of users. Users join, establish social links to friends, and leverage their social links to share content, organize events, and search for specific users or shared resources. These social networks provide platforms for organizing events, user to user communication, and are among the Internet's most popular destinations. The Internet has spawned different types of information sharing systems, including the Web. Recently, online social networks have gained significant popularity and are now among the most popular sites on the Web. For example, MySpace (over 190 million users), Facebook (over 62 million), Linkedln (over 11 million), and LiveJournal (over 5.5 million) are popular sites built on social networks. Unlike the Web, which is largely organized around content, online social networks are organized around users. Participating users join a network, publish their profile and (optionally) any content, and create links to any other users with whom they associate.

The resulting social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users. An in-depth understanding of the graph structure of online social networks is necessary to evaluate current systems, to design future online social network based systems, and to understand the impact of online social networks on the Internet. For example, understanding the structure of online social networks might lead to algorithms that can detect trusted or influential users, much like the study of the Web graph led to the discovery of algorithms for finding authoritative sources in the Web. Given the large adoption of these networks, there has been increased interest to explore the underlying social structure and data towards social search.

The main idea of social search is to use information collected from a users' social network to improve the accuracy of search results. Social search has recently gained attention as an approach towards personalized search. The utility of social search has been established via experimental user studies. However, for large online social networks, usually consisting of millions of users, a complete crawl of users' extended neighbourhood is infeasible. Therefore, efficient methods are required. An online social network graph is composed of millions of nodes and edges. In order to analyse it we have to store the whole graph in computers memory. Sometimes this is impossible. Even when it is possible it is extremely time consuming only to compute some basic properties. Thus we need to extract a small sample of the graph and analyse it. So, a technique called graph sampling technique was introduced. This technique provided an efficient and inexpensive solution.

Graph sampling is a technique to pick a subset of vertices and/ or edges from original graph. It has a wide spectrum of applications. Example: survey hidden population in sociology, visualize social graph, scale down Internet as graph, graph sparsification, etc. In some scenarios, the whole graph is known and the purpose of sampling is to obtain a smaller graph. In other scenarios, the graph is unknown and sampling is regarded as a way to explore the graph. Commonly used techniques are Vertex sampling, Edge sampling and Traversal Based Sampling. There are few graph sampling algorithms. In this paper we are focussing on Breadth First Search (BFS) and Metropolis Hasting Random Walk (MHRW) graph sampling algorithms.

II. **RELATED RESEARCH WORK**

R.Rajkumar et al, 2015 Behaviour Analysis and Feature Selection in Online Social Network [1]. Establishing an Internet presence permits a lot of people to understand that business exists. Even large companies and news organizations are using social networking websites like Twitter to urge that online presence they have. People prefer to get to understand and check with other people. Social networking permits this to happen on a way larger scale. It proposes three approaches.

1. Rogati and Yang approach: Presented a method for text classification. This approach advised that filter ways, which incorporates the statistics, were systematically higher across classifiers and performance measures.

2. Wrapper approach: This approach uses the induction formula as an area of analysis.

3. Hybrid approach: This approach is bestowed to beat the weakness of filter and wrapper approaches. Several researchers combine each of these approaches to boost the results. This approach is computationally simpler than the above two approaches and it provides higher accuracy than the above two approaches.

Maciej Kurant et al, 2011 Towards Unbiased BFS Sampling [3]. Breadth First Search (BFS) is a widely used approach for large sampling graphs. However, it has been empirically observed that BFS sampling is biased toward high-degree nodes, which may strongly affect the measurement results. This paper focuses to quantify and correct the degree bias of BFS. To quantify the node degree bias of BFS sampling, we calculate the node degree distribution q_k expected to be observed by BFS as a function of the fraction f of covered nodes, in random graph RG (p_k) with a given degree distribution p_k. It is found that for a small sample size, f→0. BFS has the same bias as the simple Random Walk, and with increasing f, the bias monotonically decreases. Based on theoretical analysis, a practical RG (p_k) based procedure is proposed to correct the bias when calculating any node statistics. These techniques are performed well on a broad range of Internet topologies.

Minas Gjoka et al, 2011 Practical recommendations on Crawling Online Social Networks [5]. It proposes a framework for unbiased sampling of users in an Online Social Network by crawling the social graph and provides recommendations for its implementation in practice. This paper has also the following contributions. Comparison of several candidate techniques has been seen in terms of bias (BFS and RW were significantly biased, while MHRW and RWRW provided unbiased samples) and efficiency. It is found that RWRW to be most efficient in practice, while MHRW has the advantage of providing a ready-to-use sample.

Christo Wilson et al, 2009 User Interactions in Social Networks and their Implications [9]. Are social links valid indicators of real user interaction? If not, then how can we quantify these factors to form a more accurate model for evaluating socially enhanced applications? This paper addresses this question through a detailed study of user interactions in the Facebook social network. It proposes the use of interaction graphs to impart meaning to online social links by quantifying user interactions.

1. User Interaction Analysis: It first examines the difference in size between interaction graphs for users in our dataset. Then compute for each user a distribution of the user's interaction events across all users' social links. Then select several points from each distribution (70%, 90%, and 100%) and aggregate across all users the percentage of friends these events involved. The result is a cumulative fraction function plotted.

Alan Misolve et al, 2007 Measurement and Analysis of Online Social Networks [12]. It presents a large-scale measurement study and analysis of the structure of multiple online social networks. The data gathered from four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut are examined. It proposes the following approach.

1. Power-law Nodes Degree: It examines the graph structure by considering the node degree distribution. The degree distribution of many complex networks, including offline social networks, has been to conform to power-laws. The best power-law coefficients approximate the distribution very well for Flickr, LiveJournal, and YouTube, the Orkut data deviates significantly.

Yong Yeol Ahn et al, 2007 Analysis of Topological Characteristics of Huge Online Social Networking Services [13]. Social networking services are fast growing business in the Internet. However, it is unknown if online relationships and their growth patterns are the same as in real life social networks. This paper compares the structure of three online social networking services: Cyworld, MySpace, and Orkut, each with more than 10 million users respectively. It has analysed the complete network of an online social network service, Cyworld.

In addition, it has also analysed sample networks from Cyworld, Orkut, and MySpace in terms of degree distribution, clustering coefficient, degree correlation, and average path length. There is a report of a multi-scaling behaviour in Cyworld's degree distribution and have sustained the claim that heterogeneous types of users are the force behind the behaviour with detailed analysis of the clustering coefficient distribution, assortivity (or dissassortivity), and the historical evolution of the network size, the average path length, and the effective diameter.

Jure Leskovec et al, 2006 Sampling from Large Graphs [14]. Given a huge real graph, how can we derive a representative sample? This paper addresses this question. There are many known algorithms to compute interesting measures (shortest paths, centrality, etc.), but several of them become impractical for large graphs. Thus graph sampling is essential. Several graph sampling methods are considered, novel methods are proposed to check the goodness of sampling, and develop a set of scaling laws that describe relations between the properties of the original and the sample. Overall, best performing methods are the ones based on random walks and "forest fire". They match very accurately both static as well as evolutionary graph patterns, with sample sizes down to about 15% of the original graph. Back-in-time goal was proposed.

1. The Back-in-time goal: This approach was proposed for sampling. This is better than the Scale-down sampling goal. A second approach to sampling is the Back-in- time sampling goal. Here we do not match the properties of the final target graph G, but rather match G as it grew and evolved. This means that we compare the patterns from sample graph S on nodes with the target graph G, when it was of the same size as S.

Luca Becchetti et al, 2006 A Comparison of Sampling techniques for Web Graph Characterization [16]. The main contributions of this paper are to compare different methods for Web sampling by sub-sampling a large Web collection. The following are different sampling methods.

1. Uniform random sampling: Pages are chosen uniformly at random with a certain probability. This sampling strategy is actually not possible for a standard Web crawler that must discover pages by following links but we used it as a baseline for the comparison.

2. Sampling by selecting entire sites: Sites are chosen uniformly at random with a certain probability and all of the pages inside a site are included in the sample. We continued this process until we have a predefined fraction of the nodes in the graph. This is feasible in practice and the crawler must be instructed not to follow links outside the sampled sites.

3. Sampling by breadth-first search (BFS): All the initial pages of sites (the starting or home page, located in the root directory of the site and typically named "/index.*" or just*/") were sampled. We consider those pages to be at depth equal to 1. All of the pages that are linked by those pages are considered to have depth equal to 2 and so on. This strategy simulates a BFS search that stops when a given threshold of nodes is reached.

4. Sampling by OPIC: The OPIC algorithm (Online Page Importance Computation) was introduced by Abiteboul as an algorithm for ranking pages while discovering them. It can be seen as a biased breadth first search in which the pages that are highly linked are more likely to be chosen. To implement this algorithm in external memory, we approximated it by recalculating page importance 20 times during the simulated crawl (instead of after inserting every code).

After analyzing these results the conclusion seems robust. For many measures the BFS and OPIC strategies perform much better sampling by sites. This seems to indicate that many characteristics of the connectivity of the Web arise from the interaction among many different sites, presumably under the control of different Web site administrators.

Colin Cooper et al, Sampling Regular Graphs and a Peer-to-Peer Network [18]. This paper has two parts. In the first part it considers a simple Markov chain for 'd' regular graphs on 'n' vertices, where d = d(n) may grow with ‘n’. It is shown that the mixing time of this Markov chain is bounded above by a polynomial in ‘n’ and ‘d’. In the second part of the paper, a related Markov chain for ‘d’ regular graphs on a varying number of vertices is introduced, for even constant ‘d’. This is a model for a certain peer-to-peer network. It is proved that the related chain has mixing time which is bounded above by a polynomial in ‘N’, the expected number of vertices, provided certain assumptions are met about the rate of arrival and departure of vertices.

A peer-to-peer (P2P) network is decentralised, dynamic network for sharing data and computing resources among the vertices (participants) of the network. The network is dynamic in the sense that the vertices join and leave the network. The participants must follow the protocol of the network, and if they do, then the network should maintain some desirable properties. The properties which the designers of protocols would like to ensure include connectivity, low degree and small diameter of the network. In addition, the protocol should ideally be simple, local (not requiring global knowledge of the network) and robust.

Cooper, Klasing and Radzic proposed a decentralized protocol based on random walks. Each joining vertex ‘v’ donates a fixed number of tokens with its address. These tokens perform a random walk on the network and can be used by any vertex which requires them. Similarly, a vertex which has lost a connection can obtain a new one using the next token to visit it. The network is robust under adversarial deletion of vertices and edges and actively reconnects itself.

III. **CONCLUSION**

After effective scrutiny and thorough study of the above cited papers, all the observations have been offered thoroughly.

Breadth First Search (BFS) is one of the simplest algorithms for searching a graph. Given a graph and an eminent source vertex, breadth first search explores and examines the edges of the graph to find every vertex reachable from source. It calculates the distance (with the fewest number of edges) from source to all reachable vertices. This algorithm uses first-in-first-out (FIFO) queue for the vertices that are to be visited next.

Metropolis Hasting Random Walk (MHRW) is an algorithm which works by generating a sequence of samples value in such a way that, as more and more sample values are produced, the distribution of values more closely approximates the desired distribution, P(x). These sample values are iteratively produced, with the distribution of the next sample being dependent only on the current sample value. Specifically, at every iteration, the algorithm picks a node for the next sample value based on the current sample value. Then, with some probability, the node is either accepted or rejected.

Therefore, MHRW has better performance than BFS based on time, cumulative distribution frequency, normalized mean square error, and efficiency.

REFERENCES

[1] R. Rajkumar ,” Behaviour Analysis and feature selection in Online Social Network”,In Proc. Of ACM, 2015.

[2] M. Kurant, M. Gjoka, Y. Wang, Z. Almquist, C. Butts, andA. Markoloulou, “Coarse-grained topology estimation via graph sampling, In Workshop on OSNs, Helsinki, Norway, August 2012.

[3] M. Kurant, “Towards Unbiased BFS Sampling”, vol.29, NO.9, October 2011.

[4] A. S. Maiya and T. Y. Berger-Wolf, Benefits of Bias: Towards Better Characterization of Network Sampling. In Proc. of KDD, San Diego, CA, USA, August 2011.

[5] M. Gjoka, “ Practical Recommendations on Crawling Online Social Networks”, vol. 29, NO. 9, October 2011.

[6] M. Gjoka, M. Kurant, C.T Butts, and A. Markopoulou, “Walikng in Facebook: A Case Study of Unbiased Sampling of OSNs,” In Proc.of IEEE INFOCOM, 2010.

[7] S. Ye, J. Lang, and F. Wu, “Crawling Online Social Graphs,” in Proc. *12th Asia-Pacific Web Conference*, Busan, Korea, 2010, pp. 236-242.

[8] N. Ahmed, J. Neville, and R. Kompella, “Reconsidering the foundations of network sampling”. In Proc. of the 2nd workshop on information in Networks, New York, September 2010.

[9] C. Wilson, B. Boe, A. Sala, K.P.N. Puttaswamy, and B.Y. Shao, “User interactions in social Networks and their Implications,” In Proc. Of ACM EuroSys, 2009.

[10] D. Achiloptas, A. Clauset, D. Kempe, and C.Moore,”On the Bias of Trace route Sampling: Or, Power-law degree distributions in regular Graphs, J.ACM, 56(4), July 2009.

[11] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen and W. Willinger. “On unbiased sampling for unstructured peer-to-peer networks”. IEEE/ACM Transactions on Networking, 16(6):377-390, April 2008.

[12] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee, “Measurement and Analysis of Online Social Networks,” In Proc. Of ACM IMC, 2007.

[13] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong,”Analysis of Topological Characteristics of Huge Online Social Networking Services”, In Proc. Of WWW, 2007.

[14] Jure Leskovec and Christos Faloutsos, “ Sampling from Large Graphs”, KDD’06, August 2006.

[15] S.H. Lee, P.J. Kim, Jeong, “Statistical Properties of Sampled Networks,” *Physical Review E*, vol. 73, p. 16102, 2006.

[16] L. Becchetti, C. Castillo, D. Donato, A. fazzone, and I. Rome, “A Comparison of Sampling Techniques for Web graph Characterization,” in *Proc. Workshop on Link Analysis,* Philadelphia, PA, 2006.

[17] M.P.H.Stumpf and C. Wiuf, “Sampling properties of random graphs”: The degree distribution, Phys.Ref.E, 72:036118, Sep 2005.

[18] Colin Cooper, Martin Dyer, and Catherine Greenhill, “Sampling Regular Graphs and a peer-to-peer Network”, Research supported by the UNSW Faculty Research Grants Scheme.