IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. I (May-Jun. 2014), PP 78-82
www.iosrjournals.org

Low selectivity problem using the concept of sampling

[1]Mr. Y. P. Murumakar, [2]Prof. Y.B.Gurav,
[1]Student PV Ptt, Bavdhan, Pune
[2]Ast. Professor, PV Ptt, Bavdhan, Pune

***Abstract*:** *There is an increase in the interest to utilize various available network structures and also the available information on social peers for improving the information needs of a user or node, this is because of the birth of online social networks. In this paper, the focus is on improvement of the performance of collecting information from the neighborhood of a user or a node in a dynamic social network. To explore user's or node's social network correctly we have introduced sampling based algorithms by keeping in mind the structure of social network and to approximate the quantities of interest in short time. By showing correlations across our samples we have introduced and also analyzed variants of basic sampling scheme. Here, the models of distributed and centralized network are considered. Assuming that information for each user or node is available we have showed that our algorithms can be utilized to rank nodes which are neighbor to the user. We demonstrate the working of our algorithms for approximation of various quantities of interest and validate analysis results. This is done using real and synthetic data. The methods we describe can be possibly easily adopted in verities of strategies which aim to collect information efficiently from a social graph.*
***Index Terms*:** *Protecting privacy; secure communication; online social networking.*

I. Introduction

There are various changes in the use of web technology which aims to add to interconnectivity, self-expression and information moving on the web have led to the birth of online social network services. This can be seen by the multitude of activity and social interaction which takes place in web sites like Facebook, Myspace, and Twitter.

Also the need t o connect and interact evolves far beyond centralized social networking sites and takes the form of ad hoc social networks formed by instant

messaging clients, VoIP software, or mobile geosocial networks. Even though interactions with people who are not in one's contact list is currently not possible (e.g., via query capabilities), the implicit social networking structure is in there. As these networks are being adopted in large quantity, there hasbeen an increased interest in exploring the available social structure and information in order to improve on information retrieval tasks of social peers. The tasks like these are in the core of many application domains. To further add to our research, we discuss in more detail the case of social search. Social search engine is a type of

search method which tries to determine the relation of search results by considering contributions of various

users. We can improve the accuracy of search results by collecting information from user's social network. For example we can consider the following search scenario:

1. User a submits query to search engine.
2. Using ranking algorithm search engine calculates ordered list L of most relevant results.
3. Search engine gets the information that is present for the node near to a and compares it to the results in L.
4. Search engine uses this information to rearrange the list L to a new list L' that is given to a.

The use of social search has been developed via experimental study of users. For example, in [1], Mislove et al.report improved result accuracy for searching of web asurlsfor a query are not ranked based on some global ranking criteria, but based on the number of times people in the same social environment endorsed them. To realize online social search number of ideas have been suggested. This is done basically from human search engines which utilize humans to filter the results of search and help users to clarify their requests for search. Because of this, the pages which this person visits will be ranked higher. If we consider any case, the important goal is to give end users a limited number of related results told by human judgment which is against traditional search engines that very frequently return very big number of results which may not be relevant. This information can be used to improve the quality of search available for users. Generally, many algorithms and tools which exist for analysis of networks, they only focus on analyzing various properties of the structure of the network instead of focusing on the contents of the node. They operate on whole graphs instead of any user specific graph.(i.e. nodes near to the particular user).Hence it would be beneficial to design algorithms which operate on a particular or a single node. We can have an example like starting from a particular user in the

network algorithms should crawl its neighbors and get the information that is present on close peers. But to crawl completely all social peers is not feasible because those networks may consist of thousands of nodes and also their structure might not be static. That's why very efficient methods are required. Our focus is on improving the performance for collecting information from the near of node or user in various social networks. We try to make some contributions, some of which are the following:

-To quickly obtain a near-uniform random sample of users or nodes in the neighbor. We introduce an algorithm for this. This algorithm can be applied to fast

the approximation of number of users in a particular user's neighbor which have endorsed an item.

-Variants of these basic sampling schemes are introduced and analyzed that aim to decrease the total number of users or nodes in the network visited by exploring correlations across samples.

-Using real and synthetic data we have evaluated our algorithms in terms of accuracy as well as efficiency. Also demonstrated the utility of our approach.

-Also, we concluded that basic sampling schemes can be utilized for many strategies which aim to rank its items in the network. This is done by assuming that information for each user is available.

- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting
- Data which can be selected

The remaining paper is organized as given below: Section 3 defines interested problems and

pros and cons. Basic ideas of sampling are presented in section 4 and details of algorithms with how they can be implemented are described in section5.These algorithms are evaluated experimentally in section 4 and in section 5 we review related work and concluded in section 6.

II. Related Work

Our work is related for working on sampling large graphs using randomwalks.Generating a uniform random subset of nodes of a graph using random walks is a very known problem.It generally arises in the analysis of convergence properties of Markov chains (e.g., see [5], [6], [7], [8]) or for the problem of doing the sampling of a search engine's index [6], [7]. The basic idea behind it is to get started from any specific node, say a, and initiate a random walk by proceeding to neighbor nodes which are selected at random at each iteration. Assume the probability of reaching any node u after k steps of this walk is p (u). We know that if k is suitably large where values of k variable depend on the topological properties of the graph, given probability distribution is stationary i.e. it does not depend on the starting node.But, this stationary distribution cannot be an uniform distribution; the probability associated with each node or user is related inversely to its degree in the graph. It is possible to make this stationary distribution uniform using techniques like Metropolis Hastings algorithm (see [7]), or using t e c h n i q u e r e j e c t i o n s a m p l i n g . This process can be iterated for obtaining various random samples of a required size. Approaches similar to this have been applied in [8] where Hastings describes sampling-based methods to correctly collect information from nodes in a social graph and in [3] where sampling techniques are used to collect unbiased This research targets the improvements in the performance using sampling. Statistical sampling, In statistics, quality assurance, & survey methodology, sampling is related to the selection of a subset of individuals considering a statistical population to calculate various functionalities of the whole population. To determine whether a production lot of material meets the governing specifications the technique used here is called as Acceptance sampling. Lower cost and fast data collection are the two advantages of sampling.

The sampling process comprises several stages:

Defining the population of concern

Specifying a sampling frame, a set of items or events possible to measure

Specifying a sampling method for selecting items or events from the frame

samples of Facebook. Similarly, in [10] Katzir et al.

Design algorithms for calculating the number of users in large social networks using biased sampling, and in [10] sampling methods are given to approximate community structures in a social network. Our research presents ways to make improvement on these generic random walk methods for graphs by knowing the fact that we need to sample from the neighbor of a node v (i.e., it may be a few links away from v).

A. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs

With more than 300 million active users [1], Facebook (FB) is currently one of the most important online social networks. Our aim in this paper is to obtain a representative (unbiased) sample of Facebook users by crawling its graph which is social. Two approaches that are found to perform well are the Metropolis-Hasting random walk (MHRW) and a re-weighted random walk (RWRW). Both have pros and cons, which we demonstrate through a comparison to each other and to the"ground-truth". In contrast, the traditional Breadth-First-Search and Random Walk (without re-weighting) perform quite poorly, producing substantially biased results. Also,we introduce online formal convergence diagnostics to assess sample quality during the data collection process. We show how these can be used to correctively determine when a random walk sample is of adequate size and quality for

subsequent use. Unbiased sample of Facebookis considered using these methods. Finally, we use one of our

representative datasets, collected through MHRW, to characterize various key properties of Facebook.

B. A Chernoff Bound for Random Walks on Expander Graphs

Here we consider a finite random walk on a weighted graph Gand show that the sample average of visits to a set of vertices A converges to the stationary probability $\pi(A)$ with error probability exponentially small in the length of the random walk and the square of the size of the deviation from $\pi(A)$. The exponential bound is in terms of the expansion of G.Itimproves previous results. We show that the method of taking the sample average from one trajectory is a more efficient estimate of $\pi(A)$ than the standard method of generating independent sample points from several trajectories. Using this good sampling method,algorithms of Jerrum and Sinclair (1989)are improved which are used for approximating the number of perfect matchings in a dense graph as well as for approximating the partition function of anIsing system.

C. Deterministic Simulation in LOGSPACE

A wide class of probabilistic algorithms can be simulated by deterministic algorithms,this is proved in this paper.To name

a few if there is a test in LOGSPACE so that a random sequence of length (log n)2/ log log n,passes the test with probability at least 1/n then a deterministic sequence can be constructed in LOGSPACE which also passes the test.Theimportant thing is that the machine performing the test should get each bit ofthe sequence

only once. The sequence that. we construct does not, really depend on the test. This family is the same even if the test is

allowed to use an oracle of polynomial size, and this can be constructed in LOGSPACE (without, using an oracle).

D. Efficient Sampling of Information in Social Networks

In this paper, focus is dynamic social network for improving the performance of information collection from the

neighborhood of a user.For this sampling based.

algorithms are introduced. Models of centralized and distributed social networks are considered. It is shown that our algorithms can be utilized to rank items in the neighborhood of a user or node by assuming that information for each user in the network is available.Real and synthetic data sets are used to validate the results of analysis and demonstrationofthe efficiency of algorithms for approximating quantities of interest.The methods described are general and can probably be easily adopted in a variety of strategies targeting to efficiently collect information from a social graph.

D. How to recycle random bits

In this paper it is shown that modified versions of the linear congruential generator and the shift register generator

are provably sufficient for filtering the correctness of a probabilistic algorithm. More precisely, if r random bitsare needed for a BPP algorithm to be correct with probability at least 2/3, then $O(r+k^2)$ bits are neededto improve this probability to $1-2^{-k}$.A different pseudorandom generator which is optimal, upto some constant factor, in this relation is also given.It uses only O(r+k) bits to improve the probability to $1-2^{-k}$.

This generator is based on random walks on expanders. The unproven assumptions are not used for results. It is shown that themodified versions of the shift register and linear congruential generators can be used.

III. Implementation Details:

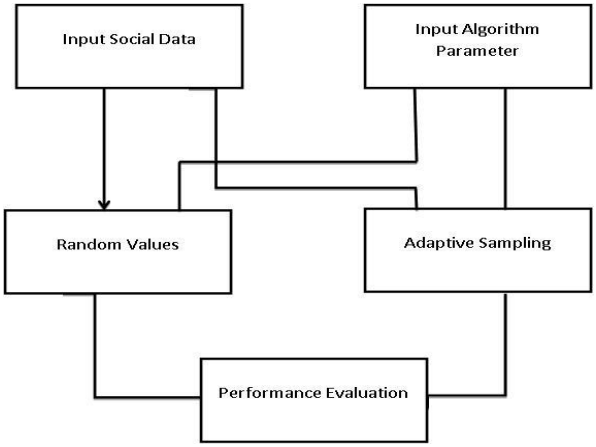

Fig 1 : Basic Blocks for analysis

Our objectives:

1.Compare our algorithm with base paper algorithm. 2.Measure the performance improvements.Identify and conclude with reasoning. We use objects and links among objects to form clusters.It also can identify relations among clusters.We used k-means clustering algorithm for analysis.The k-means algorithm aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean,serving as a prototype of the cluster.

IV. Math

Algorithm: Cluster Analysis
Step 1: Assignment of variables
$S_i^t = \{ xp \mid xp - m_i^i \mid \leq \mid xp - m_j^t \text{ for all I where } 1 \leq j \leq k \}$

Where each xp is assigned to S^t

Step 2:Update and calculate the mean
$M_i^{(t+1)} = 1/\mid S_i^t \mid \sum_{xj \in Si(t)}^{n} xj$

Where (x1,x2,……..xn) are observations|properties in d-dimensional vector of social network
Using k-means algorithm we partition it into k number of sets
S = (s1, s2, ……..sn)

Above algorithm assigns to nearest cluster by design.
Random walk Markova chain

V. Result Set And Data Set

In this experiment we analyzed different algorithms and compared those algorithms. Also we concentrated on low selectivity problem. Our main concern will be low selectivity problem. We try to solve low selectivity problem with our algorithm. We are expecting the results in our experiments.

W_k is the node of the graph selected at kth wave.
$a_{ij} = 1$ indicates a link from node i to node j.

$\{W_0, W_1, W_2, \dots\}$ is a Markov chain with
$P(W_{k+1} = j \mid W_k = i) = a_{ij}/a_i.$

Q is the transition matrix of the chain,
$q_{ij} = P(W_{k+1} = j \mid W_k = i).$

The stationary probabilities $(\pi_1, \dots, \pi_N)$ satisfy $\pi_j = \sum \pi_i q_{ij}$
for $j = 1, \dots, N$.

VI. Conclusion:

Our research suggests methods which collect information in no time from the neighborhood of a node in a dynamic social network where knowledge of its structure is not available or else is limited. Methods given are based on sampling. Using sampling it is not necessary to visit all

nodes in the area of user and hence get improved performance. The real and synthetic data is used to run experiments. The Support is provided to ranking

Algorithms and strategies. We studied what is low selectivity problem and tried to find the solution for the same. This problem arises during answering aggregation queries in sampling. In our case data stored at each node is changing rapidly, so this method is not directly applicable. This paper assumes that information for each user in the network for example web history logs is available. Privacy concerns could serve as major point of our algorithms. The systems using our algorithms should follow clear approach to designing social systems which carry a balance of visibility and awareness and accountability.

Acknowledgment

We would like to express our gratitude and sincere regards to the following people to whom we are grateful for their support and help. We would like to thank Dr**. Y.V. Chavan**, **Principal, and P.V.P.I.T.** for providing us with excellent facilities and valuable guidance. We would like to express my profound and sincere gratitude to D r . A . M . D i x i t , **Prof. Y. B. Gurav HOD**, Computer Department who has given all co-operation and help to complete this paper. We would like to express my profound and sincere gratitude to **Prof. N. D. Kale Coordinator**, Computer Department who has given all his co- operation and help to complete this paper.

References

[1]. C. Gkantsidis, M. Mihail, and A. Saberi,"Random Walks in Peer-to-Peer Networks: Algorithms and Evaluation,"Performance Evaluation, vol. 63, n o . 3, pp. 241-263, 2006.
[2]. D. Gillman, "A Chernoff Bound for Random Walks on Expander Graphs," SIAM J. Computing, vol. 27, no. 4,pp. 1203-1220, 1998.
[3]. G. Das, N. Koudas, M. Papagelis, and S.Puttaswamy, "Efficient Sampling of Information in Social Networks," Proc. ACM Workshop Search in Social Media (SSM), 2008.
[4]. L. Katzir, E. Liberty, and O. Somekh,"Estimating Sizes of Social Networks via Biased Sampling," Proc. 20th Int'l Conf. World Wide Web (WWW), 2011.
[5]. M. Ajtai, J. Komlos, and E. Szemeredi "Deterministic Simulation inLogspace," Proc. 19th Ann. ACM Symp. Theory of Computing (STOC), 1987.
[6]. M. Gjoka, M. Kurant, C.T. Butts, and A.Markopoulou, "Walking in Fa c eb ook : A C a s e S t u d y of Un b i a s ed Sampling of Osns," Proc. INFOCOM, 2010.
[7]. R. Impagliazzo and D. Zuckerman, "How to Recycle Random Bits," Proc. 30th Ann. Symp. Foundations of Computer Science (FOCS), 1989.
[8]. W. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," Biometrika,vol. 57, no. 1, pp. 97-109, 1970.
[9]. Z. Bar-Yossef and M. Gurevich,"Random Sampling from a Search Engine's Index," Proc.15th Int'l Conf.World Wide Web (WWW),2006.
[10]. Z . Bar-Yossef, A.Berg, S.Chien, J . Fakcharoenphol, and D. Weitz, "Approximating Aggregate Queries About Web Pages via Random Walks," Proc. 26th Int'l Conf. Very Large Data Bases (VLDB), 2000.