Unbiased Sampling in Directed Social Graph

Tianyi Wang
Department of Electronic Engineering
Tsinghua University, Beijing, China
tsinghuawty@gmail.com

Yang Chen
Institute of Computer Science
University of Goettingen, Goettingen, Germany
chenyang03@gmail.com

Zengbin Zhang
Department of Computer Science
University of California, Santa Barbara, USA
zhangzengbin@gmail.com

Peng Sun
Department of Electronic Engineering
Tsinghua University, Beijing, China
sunp1988@gmail.com

Beixing Deng
Department of Electronic Engineering
Tsinghua University, Beijing, China
dengbx@mail.tsinghua.edu.cn

Xing Li
Department of Electronic Engineering
Tsinghua University, Beijing, China
xing@cernet.edu.cn

ABSTRACT

Microblogging services, such as Twitter, are among the most important online social networks(OSNs). Different from OSNs such as Facebook, the topology of microblogging service is a directed graph instead of an undirected graph. Recently, due to the explosive increase of population size, graph sampling has started to play a critical role in measurement and characterization studies of such OSNs. However, previous studies have only focused on the unbiased sampling of *undirected* social graphs. In this paper, we study the unbiased sampling algorithm for *directed* social graphs. Based on the traditional Metropolis-Hasting Random Walk (MHRW) algorithm, we propose an unbiased sampling method for *directed* social graphs(USDSG). Using this method, we get the first, to the best of our knowledge, unbiased sample of directed social graphs. Through extensive experiments comparing with the "ground truth" (UNI, obtained through uniform sampling of directed graph nodes), we show that our method can achieve excellent performance in directed graph sampling and the error to UNI is less than 10%.

Categories and Subject Descriptors

J.4 [**Computer Application**]: Social and behaviorial sciences

General Terms

Human Factors,Measurement

Keywords

Online social network

1. INTRODUCTION

In recent years, the population of Online Social Networks (OSNs) has experienced an explosive increase. Twitter, for example, has attracted over 100 million users by April 2010. The world-wide spreading of OSNs has motivated a large amount of studies from research community to measure and analyze the characteristics of social graphs. The data that these studies use are either complete datesets from network operators, which are commonly not publicly accessible, or self-crawled datasets, which are normally incomplete or biased to high-degree nodes sometimes. Thus, lots of attention has been put on how to obtain a representative or unbiased dataset from a large social graph using graph sampling techniques.

Breadth-First-Search [1] and Random Walk (without distinguishing the degree of neighboring nodes) are the most popular ways to sample the social graphs. However, previous study [2] has shown that both of them are biased towards high-degree nodes. In [2], Metropolis-Hasting Random Walk(MHRW) is proposed to obtain samples from an *undirected* graph, such as Facebook. This algorithm can guarantee the unbiasedness of the sampling procedure, thus can keep all the statistical properties of *undirected* social graphs. However, unlike Facebook, microblogging networks such as Twitter do not require reciprocation in the relationship: one can follow anyone without being followed. In other words, Twitter is a *directed* other than *undirected* social graph, and thus can not be sampled using the previous methods.

In this paper, we propose the solution to unbiasedly sample a directed graph. Our contributions are twofold. Firstly, based on the unique properties of directed graphs, we propose a sampling algorithm that is unbiased. Secondly, through comparison with "ground truth" (global uniform sampling) by extensive experiments, we proved that this method can achieve excellent performance in directed graph sampling.

2. DESIGN FRAMEWORK

Metropolis-Hasting Random Walk(MHRW) is a Markov Chain Monte Carlo(MCMC) algorithm to obtain random samples from a probability distribution for which direct sampling is difficult. However, MHRW is not applicable in directed graphs because there is a probability that we 'walk' to a node whose out degree is 0. It means, once we 'walk' to this node, we can never go to other nodes through pure random walk. An intuitive approach to solve the problem is to randomly choose a neighboring node (in degree neigh-

SIGCOMM'10, August 30–September 3, 2010, New Delhi, India.
ACM 978-1-4503-0201-2/10/08.

bors) as next node. But it would be biased to high-degree nodes. For each source node, the Markov chain is not long enough to converge to the target probability distribution.

In our method, USDSG, we consider all unidirectional edges as bidirectional edges to solve the problem. In this case, after choosing a well-connected initial node (we don't want nodes with no edges connecting with), we can reach all the other nodes.

Furthermore, a new proposal function is needed in USDSG, which depends on current state to generate a new proposal sample. In previous study[2], node degree k_v is used as proposal function to obtain unbiased samples in undirected graphs. The proposal function changes the transition probabilities and modifies the bias towards high-degree nodes. As a result, the sample converges to uniform distribution. In a directed graph, neither in degree nor out degree can form the proposal function by itself, because the properties relevant to the other would get lost.

As we have mentioned above, all edges are considered as bidirectional edges. In this condition, we can simply use the number of connected neighbors of each node as proposal function. It is the same as node degree in undirected graphs.

USDSG algorithm works in the following way. First, we obtain a random node v as current state, and the proposal function is $Q(v)$. A node w is then chosen from node v's connected nodes as the next proposal sample. In the next step, we generate α from uniform distribution $U(0,1)$: if $\alpha \leqslant \frac{Q(v)}{Q(w)}$, we say a proposal is accepted and w is taken as the next sample; else, v remains to be the sample. It can be proved that if we have enough steps, the sampling will converge to an uniform distribution.

To conclude, by changing directed graphs into undirected graphs and then applying a new proposal function, USDSG can leverage MHRW to do unbiased sampling in the resulted graphs. This is theoretically correct because MHRW guarantees that we can get unbiased sample from an undirected graph regardless of the topology.

3. EVALUATION

In this section we conduct experiments on real datasets to evaluate our method. The datasets we use are available at [4], which collects extensive large network datasets of social graphs. The datasets we use are soc-Epinions1 with 75,879 nodes and 508,837 edges, soc-Slashdot0811 with 77,360 nodes and 905,468 edges and soc-Slashdot0922 with 82,168 nodes and 948,464 edges.

For each dataset, we calculate node degree and record connected nodes for each node. In this way we change directed graphs into undirected graphs. We omit those nodes whose degree is 0 because these nodes cannot be used as initial nodes. It won't affect the result because once we start random walk, we won't walk to these nodes. Then we use MHRW to obtain samples and analyze the average degree of the samples.

From table 1 we can see the result of our method is almost the same as that of UNI. We define error as follows:

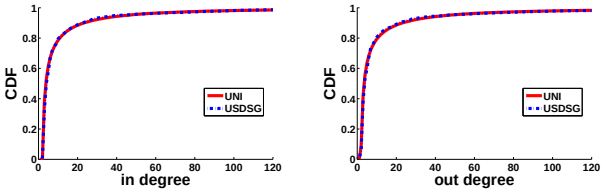
$$error = \frac{|USDSG - UNI|}{UNI}$$

All errors are less than 10%. This shows USDSG performs very well in unbiased sampling.

To further study the unbiased estimation, we consider the CDFs of node degree distribution. We obtain both in degree and out degree of the sampled nodes and compute the distribution function. For all the three different datasets, we get similar results. Due to the space limitation, we will only present the results from Slashdot0811. As we have mentioned, the variation is very large, some nodes even have in degree or out degree that is more than 1,000. In the figures we only plot the CDFs when in degree or out degree is less than 120. From the figure we can see that the sample we obtain through our method is almost identical to the UNI. This demonstrates that USDSG can get unbiased samples.

dataset		UNI	USDSG	error
Slashdot0811	in degree	11.7046	11.4208	2.4%
	out degree	11.566	12.1501	5.5%
Slashdot0902	in degree	11.5431	11.6202	0.67%
	out degree	11.5829	11.9697	3.34%
soc-Epinions1	in degree	6.70512	6.9976	4.36%
	out degree	6.94397	7.5262	8.38%

Table 1: Average Degree of Samples



(a) In Degree CDF (b) Out Degree CDF

Figure 1: Degree distributions

4. CONCLUSIONS

In this paper, we obtained the first, to the best of our knowledge, unbiased (i.e., uniform) sampling method of directed social graphs, USDSG based on traditional MHRW method. We also compare the result with UNI, and the results show that USDSG can performs sampling almost identical with UNI.

5. REFERENCES

[1] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy and Ben Y. Zhao. User Interactions in Social Networks and their Implications. In Proc. of ACM EuroSys 2009.

[2] Minas Gjoka, Maciej Kurant, Carter T Butts, Athina Markopoulou. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. In Proc. of IEEE Infocom, 2010.

[3] Haewoon Kwak, Changhyun Lee, Hosung Park, Sue Moon. What is Twitter, a Social Network or a News Media?. In Proc. of WWW, 2010.

[4] Stanford large network dataset collection:http://snap.stanford.edu/data/index.html