Unbiased Sampling Method Analysis on Online Social Network

Siyao Wang [a], Bo Liu, Jiajun Zhou, Guangpeng Li
College of Computer Science, Nation University of Defense Technology, Changsha, China
[a] 627188726@qq.com

Abstract. The study of social graph structure has become extremely popular with the development of the Online Social Network (OSN). The main bottleneck is that the large account of social data makes it difficult to obtain and analyze, which consume extensive bandwidth, storage and computing resources. Thus unbiased sampling of OSN makes it possible to get accurate and representative properties of OSN graph. The widely used algorithm, Breadth-First Sampling (BFS)and Random Walking (RW) both are proved that there exists substantial bias towards high-degree nodes. By contrast the Metropolis-Hasting random walking (MHRW), re-weighted random walking (RWRW) and the unbiased sampling with reduced self-loop (USRS)which are all based on Markov Chain Monte Carlo(MCMC) method could produce approximate uniform samples. In this paper, we analyze the similarities and differences among the four algorithms, and show the performance of unbiased estimation and crawling efficient on the data set of Facebook. In addition, we provide formal convergence test to determine when the crawling process attain an equilibrium state and the number of nodes should be discarded.

Keywords: OSN, Unbiased Sampling, online convergence test.

1. Introduction

Online social networks(OSNs) have gradually become prevalent in daily life.By 2016 there are more than 2.2 billion registered users in Facebook.In the second quarter of 2017, there are more than 2 billion active users in Facebook.This global phenomenon has attracted many researchers to conduct research on OSN.

Though by means of complete data set provided by the OSN companies can lead to optimal result, the complete data set is usually not open to public and, like most OSNs, Facebook dose not share the private data set of users [1]. So, we need to acquire small-scale but representative data set so as to analysis OSN structure properties. Through sampling OSN, keep the social relationships among users, so as to ensure that there exists a great similarity between origin graph and sampled graph.

There are some questions to be solved:

What is a good sampling method and how to evaluate them? Random select nodes by user id or by some algorithms?

Use which measurements to evaluate the effectiveness of sampling?and determine when the sampling reach the convergence state

BFS and RW are considered as the most common crawling methods.However, studies have shown that these sampling techniques introduce the bias towards high degree nodes, resulting in neglecting other nodes. Such bias leads to the sampled graph failing to reflect the topology of OSN graph correctly and completely. Notice that bias introduced by RW could be quantified by Markov Chain Monte Carlo method and corrected by Hasen-Hurwitz estimation which could generate a uniform stationary distribution of graph properties [2]. The Metropolis-Hasting Random Walk (MHRW) method correct the bias through yielding continuous sample iterations on low-degree nodes [3]. Unbiased sampling to reduce self-loops (USRS) could reduce the probabilities of self-loops introduced by MHRW [7] and still keep unbiased.

In this paper, the main works are as followed:

We sampled the OSN graph using the traditional sampling strategy, RW, and the unbiased sampling method MHRW, RWRW, and USRS to compare the performance among them.

We apply the formal convergence diagnostics to assess sample quality. This method (evolved from MCMC applications)enable us to determine when a sample is adequate for use and how many nodes should be discarded before attaining desired result.

In addition, we compare the effectiveness of each algorithm in terms
of link coverage rate as well as node coverage rate with the same number of iterations.

2. Related Work

In this section,we review the previous study about large-scale OSN sampling.The sampling efficiency of BFS and RW are presented in [4].Although these two methods are commonly used in OSN sampling, the structure properties generated form these sampling methods are deviated from origin graph.

Gjoka et al. propose MHRW [3] method to sample nodes uniformly in case of not knowing the total graph to correct the bias towards high-degree nodes; Gjoka also used RWRW method proposed in [5]to correct the bias introduced by Random Walking, applying this method on a large sample (over 1M nodes) of the Facebook Graph. However, the original assumption of MHRW is that social graph is well connected [6], which results in MHRW not proper for sampling disconnected or loosely connected graph. In addition, sampling method based on MHRW introduce relatively massive self-loops which make it difficult to find more nodes for crawler. Wang [7] proposed USRS which reduce the probability of self-loops to increase the transition probabilities of new unseen neighbours.

3. Sampling Method

3.1 Metropolis-Hasting Random Walk(MHRW)

The Metropolis-Hasting Random Walk, a variant of MCMC method, is a classic and unbiased method for sampling from a probability distribution μ that is difficult to yield directly [9].

The method begins from a randomly selected seed node and iteratively accesses neighbours of seed node .At each sampling step,MHRW selects a neighbour v of the current node u at a case of probability.The transition probability from u to v $P_{u,v}^{MH}$ is defined in formula(1):

$$P_{u,v}^{MH} = \begin{cases} \min(\frac{1}{k_u}, \frac{1}{k_v}) & if \quad v \quad is \quad a \quad neighbour \quad of \quad u, \\ 1 - \sum_{y \neq u} P_{u,y}^{MH} & if \quad w = v, \\ 0 & otherwise \end{cases} \tag{1}$$

If the neighbor node $v^{'}$ degree k_v is larger than the current node $u^{'}$ degree k_u ,then the probability is $\frac{1}{k_v}$,otherwise, the probability is $\frac{1}{k_u}$ .If the transition probabilities for moving to all the neighbors is lower than 1,then u will stay at the current node with the remaining probabilities.

We can see from formula(1) for any two adjacent nodes u and v, $P_{u,v} = P_{v,u}$ .In MHRW since both of the two probabilities for moving to neighbors are $1/\max\{k_u, k_v\}$ .So its transition probability matrix is symmetrical.It can be proved that the stationary probability distribution is $\pi_v^{MH} = \frac{1}{|V|}$,which is a constant without relevance to node degree.

3.2 Random Walking(RM)

In the typical random walk process, crawler starts from the current node i, randomly select a neighbor as the sampled node at the next step by the equal probability. P_{ij} is defined as formula (2):

$$P_{ij} = \begin{cases} \frac{1}{k_i}, & j \ \ is \ \ a \ \ neighbor \ \ of \ \ i, \\ 0, & otherwise \end{cases} \tag{2}$$

Where i is the current node where crawler stays, P_{ij} is the transition probability of sampling from i to j, k_i is the degree of i.

Given a connected graph, the probability of being at the particular node v converge to the stationary distribution $\pi_v^{RW} = \frac{k_v}{2 \cdot |E|}$ .Since the number of edges |E| is a constant, the result distribution is positively correlated with node degree.

3.3 Re-weighted Random Walk(RWRW)

Correcting the bias introduced by Random Walk could be by means of Hansen-Hurwitz estimator [8] and it is was also later applied in [10]. Suppose a typical random walk that has accessed node set $V = v_1, v_2 ... v_n$. Every node could attribute to one of n groups in relative to a property of interest A, which could be degree diameter or other node properties with discrete value. Let $(A_1, A_2, ..., A_n)$ be the range of valuein terms of interest A; $\bigcup_1^m A_i = V$ .E.g. if the property of interest is the node degree, then A_i includes all nodes u whose degree is i.In order to assess the probability distribution of A ,we need to assess the proportion of nodes with A_i ,i=1,2,...,m:

$$p(A_i) = \frac{\sum_{u \in A_i} 1/k_u}{\sum_{u \in V} 1/k_u} \tag{3}$$

3.4 Unbiased Sampling with Reduced Self-Loop(USRS)

The basic idea of USRS sampling is that take the self-loop L_i , L_j of current node i and its neighbors into consideration based on transition probability computed by MHRW.If part of the L_i and L_j transfers to P_{ij} and P_{ji} ,then reduce the self-loop of node itself and increase the probability of transition between nodes. For current node i, USRS calculate the self-loop P_{ii} , P_{jj} for node i and all its neighbors j.The transition probability of the current node i can be calculated in formula(4)

$$P_{ij} = \begin{cases} \frac{1}{k_i} \times \min\left\{1, \frac{k_i}{k_j}\right\} + \frac{\Delta r_i}{N_i}, j \in R_i, \\ 1 - \sum_{y \neq i} P_{iy} - \Delta r_i, j = i, \\ \\ 0, \qquad\qquad otherwise, \end{cases} \tag{4}$$

Where R_i represents the set of the neighbors of node i; N_i represents the number of nodes whose self-loop is greater than 0 calculated by formula(3)among the neighbors of node i. Δr_i is the modifying factor of self-loop,computed by formula(5)

$$\Delta r_i = \min\{P'_{ii}, N_i \times \{P'_{ii}, \bigcup_{j \in R_j \& P'_{jj} > 0} P'_{jj}\}\} \tag{5}$$

Where P'_{ii} is the self-loop of node i computed by formula(3). Δr_i is defined by formula (5), which is to ensure that as much as possible to reduce P'_{ii} ,and the self-loop of node i and all neighbor nodes does not drop to a negative value.USRS sampling reduces the self-loop introduced by MHRW,at the same time,ensure the symmetry of the transition probability matrix($\forall i, j \in V, P_{ij} = P_{ji}$).Thus, it is easy to prove that the USRS sampling algorithm, like MHRW, satisfies the condition of unbiased sampling

3.5 Convergence

Here we conduct a normal convergence test, Geweke diagnostic [9]. And we apply the test on node degree during crawling. Geweke [11]proposed a convergence diagnostic based on normal time-series techniques.Let X be a single sequence of a certain metric of graph interest.For each interest, the chain is separated into two parts containing the first X_a (typically the first 10%)and its end X_b (typically the last 50%).According to X_a and X_b ,we produce the z-statistics

$$z = \frac{E(X_a) - E(X_b)}{\sqrt{Var(X_a) + Var(X_b)}} \tag{6}$$

If the whole chain is static, the means value of the first and end chain should be alike.With the increasing of iterations,the value of z will fall into the interval of [-1,1]

4. Result

We have crawled more than 50,000 Facebook active users and their social information, which contains about 10million friend's relationship. From April 2017 to October 2017, parallel crawlers were used to complete the above work. As a result of the relationship between friends is mutual, the edges in the graph are all un-directed.

4.1 Unbiased Estimation

We used RW, MHRW, RWRW and USRS to sample the social graphs respectively, and estimate metrics of interest in OSN structure including degree distribution, clustering coefficient and assortativity .

Degree distribution

In Fig.1 we present the node degree distribution estimated by RW, MHRW, RWRW and USRS. It can be shown from graph that MHRW, RWRW and USRS avoid or correct the bias towards high degree node introduced by RW to a certain content. The results of MHRW, USRS and RWRW are similar. Compared with MHRW, the proportion of low degree nodes in USRS is relatively large, because the introduction of self-loop enables the crawler traverse more nodes.

Clustering Coefficient

We compute the the clustering coefficient of sampled graph by above algorithm ans Fig.2shows the change of clustering coefficient accompanied by the increase of crawled nodes.As can be shown from figure that the clustering coefficient in RW is significantly lower than the other three algorithms. The reason might be the bias toward high degree nodes reduces the probability of forming edges between the neighbors of high degree nodes. Whats more,the clustering coefficient estimated by the other three algorithms reach nearly 0.145, slightly less than that reported in [12]

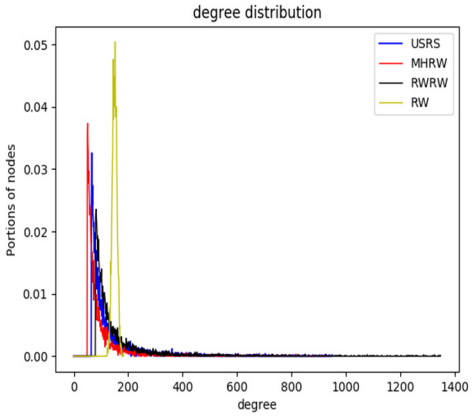

Fig.1 degree distribution

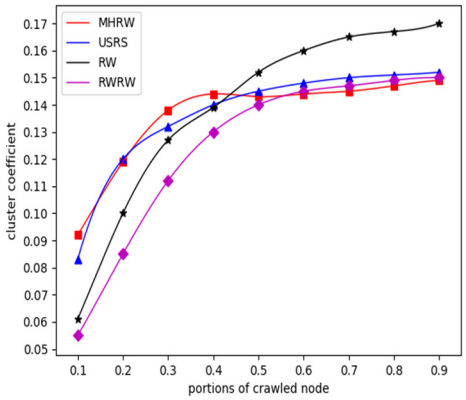

Fig.2 cluster Coefficient

Assortative

From the table 1,we can see a positive correlation for all of the four algorithm,which is in agreement with previous studies on OSN.The results of MHRW and RWRW are similar, and the estimate of USRS is slightly higher than that of UNI.Through calculating the Pearson correlation for five crawlers.The value 0.16 is similar to the $r^{'} = 0.17$ reported in[13].

Table 1. assortativity coefficient

Sampling algorithm	assortativity coefficient
MHRW	0.18
RWRW	0.16
USRS	0.16
RW	0.17

4.2 Crawling Efficiency

We assess the crawling efficiency by means of node coverage(NC)and link coverage(LC) [14],which are defined by formula(7),in which$|V_{seen}|$ and$|E_{seen}|$are numbers of nodes and links found by the crawlers.V and E are the total numbers of nodes and edges in the origin graph.

$$NC = \frac{|V_{seen}|}{V} \qquad LC = \frac{|E_{seen}|}{E} \tag{7}$$

From Fig.3 and Fig.4,we can see that USRS finds more nodes than other algorithms in the case of the same number of iterations .The reason is that USRS reduce the self-loop probabilities compared with MHRW and RW.Thus,USRS enables the crawler move to new nodes more easily and quickly instead of getting stuck in a local area for a long time.In addition, NC and LC attained by RWRW grow faster in the first 40,000 iterations,but the proportion is lowest in the end.The reason might be that compared with the other two algorithms RWRW has no self-loop rate, with the increasing of iterations,RW may easily get stuck in some local area and then can not traverse other areas of the graph.

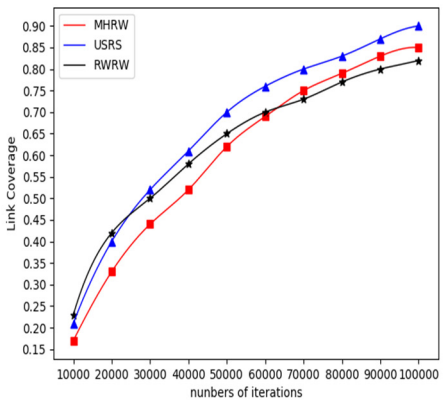

Fig.3 link coverage

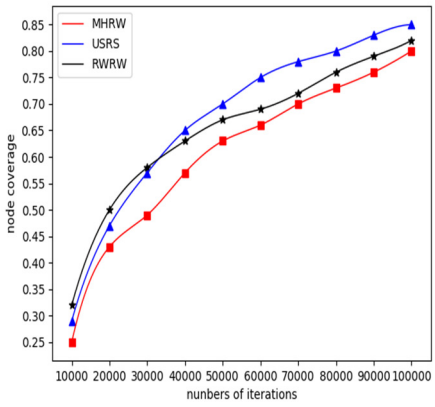

Fig.4 Node Coverage

4.3 Convergence Analysis

In order to get a rough estimate of the sample quality during the crawling process,we used Geweke diagnostic to identity convergence. In order to answer the question proposed before: when the sampling is reaching equilibrium state.We choose node degree as the parameter of z-score calculation.

Fig.5,6,7provides the trace plot for the metric of node degree, presenting the z-score value against the numbers of iteration.We declare convergence when all the values fall into the[-1,1] interval. We can see that after 4000 iterations, USRS and MHRW crawler respectively reach convergence state. However, the convergence process of RWRW is relatively slow, after the 6000 iterations the crawler achieve the desired effect. In general, after ten experiments, in each independent chain we conservatively discard 5K nodes out of 50k nodes.

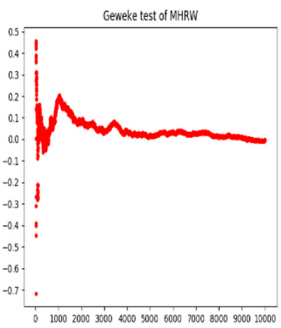

Fig.5 MHRW

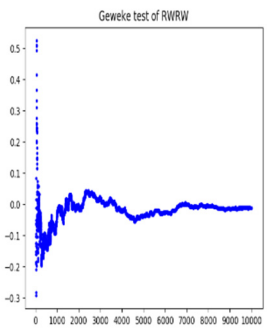

Fig.6 RWRW

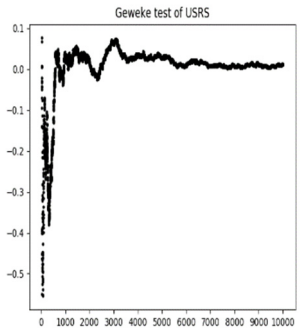

Fig.7 USRS

5. Conclusion

In this paper, we evaluated experimentally four sampling algorithms: MHRW, RWRW, USRS and RW. We picked up three metrics of interests as the indicator of unbiased estimation (including degree distribution, cluster coefficient, and assortative coefficient) from large-scale social networks like Facebook. In addition, we also analyze the performance of four algorithms on sampling efficiency and convergence.

We demonstrated that three principled algorithm: MHRW,RWRW and USRS perform better than traditional sampling one(RW) both on unbiased estimation and sampling efficiency. We also determine the numbers of nodes should be discarded in the burn-in period.

Future research would focus on further analyzing the features of each algorithm and theirs performance on sampling different types of ONS graph.A different direction is sampling with weighted edge instead of non-weighted graph.

References

[1]. Leskovec J, Faloutsos C. Sampling from large graphs[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006:631-636.

[2]. Gjoka M, Kurant M, Butts C T, et al. Practical Recommendations on Crawling Online Social Networks[J]. IEEE Journal on Selected Areas in Communications, 2011, 29(9):1872-1892.

[3]. Gjoka M, Kurant M, Butts C T, et al. Walking in Facebook: A Case Study of Unbiased Sampling of OSNs[C]// INFOCOM, 2010 Proceedings IEEE. IEEE, 2010:1-9.

[4]. Ye S, Lang J, Wu F. Crawling Online Social Graphs[C]// Web Conference. IEEE, 2010:236-242.

[5]. Heckathorn D D. Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations[J]. Social Problems, 1997, 44(2):174-199.

[6]. Gjoka M. Measurement of online social networks[M]. California State University at Long Beach, 2010.

[7]. Wang D, Li Z, Xie G. Towards Unbiased Sampling of Online Social Networks[C]// IEEE International Conference on Communications. IEEE, 2011:1-5.

[8]. Hansen M H, Hurwitz W N. On the Theory of Sampling from Finite Populations[J]. Annals of the Rheumatic Diseases, 1943, 14(12):2111-2118.

[9]. Mauro Gasparini. Markov Chain Monte Carlo in Practice[J]. Technometrics, 1997, 2(3):9236–9240.

[10]. Rasti A H, Torkjazi M, Rejaie R, et al. Respondent-Driven Sampling for Characterizing Unstructured Overlays[C]// INFOCOM. IEEE, 2009:2701-2705.

[11]. Geweke J. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments[J]. Staff Report, 1991, 4:169-193.

[12]. Lovász L, Lov L, Erdos O P. Random Walks on Graphs: A Survey[J]. Combinatorics, 1993, 8(4):1-46.

[13]. Wilson C, Boe B, Sala A, et al. User interactions in social networks and their implications[C]// ACM European Conference on Computer Systems. ACM, 2009:205-218.

[14]. Korolova A, Motwani R, Nabar S U, et al. Link privacy in social networks[C]// IEEE, International Conference on Data Engineering. IEEE, 2008:1355-1357.