Physica A 424 (2015) 254–268

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Sampling social networks using shortest paths

Alireza Rezvanian *, Mohammad Reza Meybodi

Soft computing laboratory, Computer Engineering and Information Technology Department, Amirkabir University of Technology, Tehran, Iran

HIGHLIGHTS

- We propose to use the concept of shortest path for sampling social networks.
- The proposed algorithm is studied on several well-known synthetic and real networks.
- The proposed algorithm is compared with other well-known sampling methods in terms of RE, NMSE, and KS test.
- The experimental results show that proposed sampling method is a proper method for sampling social networks.

ARTICLE INFO

Article history:
Received 15 June 2014
Received in revised form 11 October 2014
Available online 13 January 2015

Keywords:
Online social networks
Social network analysis
Network sampling
Shortest path

ABSTRACT

In recent years, online social networks (OSN) have emerged as a platform of sharing variety of information about people, and their interests, activities, events and news from real worlds. Due to the large scale and access limitations (e.g., privacy policies) of online social network services such as *Facebook* and *Twitter*, it is difficult to access the whole public network in a limited amount of time. For this reason researchers try to study and characterize OSN by taking appropriate and reliable samples from the network. In this paper, we propose to use the concept of shortest path for sampling social networks. The proposed sampling method first finds the shortest paths between several pairs of nodes selected according to some criteria. Then the edges in these shortest paths are ranked according to the number of times that each edge has appeared in the set of found shortest paths. The sampled network is then computed as a subgraph of the social network which contains a percentage of highly ranked edges. In order to investigate the performance of the proposed sampling method, we provide a number of experiments on synthetic and real networks. Experimental results show that the proposed sampling method outperforms the existing method such as random edge sampling, random node sampling, random walk sampling and Metropolis–Hastings random walk sampling in terms of relative error (RE), normalized root mean square error (NMSE), and Kolmogorov–Smirnov (KS) test.

1. Introduction

In recent years, online social networks (OSN) have emerged as a platform for sharing variety of information about users, their activities, events and news in their real worlds. OSN similar to many real world networks modeled and represented as a graph with a set of nodes (e.g., users of OSN) and edges (a certain type of relationship between users of OSN). It is shown that for different OSNs there are common fascinating properties such as small-world and scale-free properties [1,2]. The nature of OSN is large, dynamic, complex, and also some part of these networks is not fully accessible due to computational or privacy settings and thus most of the time the direct access on these networks is not feasible [3]. Therefore, these

* Corresponding author.
E-mail address: a.rezvanian@aut.ac.ir (A. Rezvanian).

http://dx.doi.org/10.1016/j.physa.2015.01.030
0378-4371/

networks [4–6] are studied and characterized via metrics (i.e., centrality measures) or techniques (i.e., sampling methods) instead of access to the whole network [7–10]. In practice, researchers via sampling from networks can process information in a reasonable amount of time and lower computational effort. From the sampled network which contains only a portion of the whole data, one can reveal some hidden important properties of networks such as user age distribution, user activities, user connectivity, and many more [11,12].

Let $G = \langle V, E\rangle$ be the input graph where V is the set of nodes with size $n = |V|$ and E is the set of edges. Sampling is a function $f : G \to G_s$ from graph G to sampled graph $G_s = \langle V_s, E_s\rangle$ such that $V_s \subset V$, $E_s \subset E$, and $|V_s| = \phi n$, where $0 < \phi < 1$ denotes the sampling rate. Sampling methods play an important role in preprocessing, characterizing, and studying real networks [7,9,13]. Sampling can be used to study a small part of networks while preserving main features of the original network. In this paper, a sampling method for sampling social networks will be presented using the concept of shortest path. In the proposed sampling method, we first find the shortest paths between several pairs of nodes which are selected according to some criteria. Then the edges in these shortest paths are ranked according to the number of times that each edge has appeared in the set of found shortest paths. The sampled network is then created as a subgraph of the social network which contains only a percentage of highly ranked edges/nodes. The proposed algorithm could be simply implemented using repeated shortest path algorithm. In order to study the performance of the proposed sampling method, a number of experiments on synthetic and real networks are provided. Experimental results show that the proposed sampling method outperforms the existing method such as random edge sampling, random node sampling, random walk sampling, and Metropolis–Hastings random walk sampling in terms of relative error (RE), normalized root mean square error (NMSE), and Kolmogorov–Smirnov (KS) test [14]. The rest of this paper is organized as follows. Section 2 introduces sampling methods in brief. In Section 3, the proposed sampling method and some of its improvements are described. The performance of the proposed sampling method is studied through simulations whose results are reported in Section 4. Section 5 concludes the paper.

2. Related works

There are a limited number of recent researches on studying, characterizing and estimating the properties of online social networks via sampling. Several sampling methods have been proposed for sampling networks that can be categorized into three main strategies in terms of collecting samples: random sampling, crawling based sampling, and coarse graining based sampling.

- **Random sampling:** In random sampling, random selection is done based on either nodes or edges without considering its topological structure [15]. Random sampling including two main simple techniques: random edge sampling and random nodes sampling. In random edge sampling (RES), an edge is selected at random and two nodes incident to the edge collected for sampled network. In random node sampling (RNS), each node is selected uniform randomly to form sampled network. RES and RNS due to simplicity are applicable for theoretical investigation. Also, due to simplicity of these methods, some improved methods based on RES and RNS are developed for sampling networks in recent years by researchers [16–18].
- **Crawling based sampling:** Crawling based sampling (also called topology based sampling or traversal based sampling) such as breadth-first-search (BFS) [19], depth-first-search (DFS) [20], forest fire sampling (FFS) [21], snowball sampling (SBS) [22], and random walk sampling (RWS) [23] have been used for collecting samples from networks. RWS iteratively selects the next node uniformly at random among all adjacent of that node. In RWS, a node with more edges will have higher probability of being sampled. Therefore, the sample mean tends to overestimate the average degree. It is noted that due to the simplicity and efficiency of RWS a variety of random walks such as Metropolis–Hastings Random Walk (MHRW) [24], Weighted Random Walk (WRW) [25], Stratified Weighted Random Walk (SWRW) [25], Respondent Driven Sampling (RDS) [26], and Distributed Learning Automata based Sampling method (DLAS) [27] are improved and widely used in literature.
- **Coarse graining based sampling**: In coarse graining based sampling, the goal is to reduce the scale of networks, which mainly used for network visualization applications to display in a limited screen [28]. In this method, several approaches are presented such as clustering methods [29], k-core methods [30], and fractal based methods [31].

There are some special complexities and challenges in sampling from real networks which can be still discussed as new research fields [15]. A good study for sampling from complex networks presented by *Leskovec* et al. [15]. They proposed sampling method from large graph by introducing several basic methods with two goals: back in time and down scale. The study has shown that RNS and RES do not provide appropriate results. In Ref. [19] *Lee* et al. presented that RNS performs better than RES with respect to estimating the clustering coefficient of networks. *Lu* et al. studied sampling on *Twitter* data and this research reveals that results of RWS is much better than results of RNS or RES methods [32]. An analytical comparison between RWS and BFS sampling has been presented by *Kurant* et al. [8] to sampling from network. Their study indicates that the degree of graph is overestimated by the BFS, while it is underestimated by the RW sampling. Therefore, they suggested analytical solutions to correct the biasness of estimation. A practical framework for uniform sampling from Facebook users has been developed based on crawling in Ref. [33]. In this study, the advantage of unbiased estimation of MHRW and ReWRW (RWRW) over random sampling and BFS has been addressed with comparing various approaches. *Rejaie* et al. tried to estimate the number of users for MySpace and Twitter by generating sequential user id [34], but this technique is failed for those online social networks where the user id is randomly generated. RDS was analyzed in Ref. [35] to reduce the biases