Physica A 396 (2014) 224–234

Contents lists available at ScienceDirect

Physica A

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

Sampling from complex networks using distributed learning automata

Alireza Rezvanian [a,*], Mohammad Rahmati [b], Mohammad Reza Meybodi [a]

[a] Soft Computing Laboratory, Computer Engineering and Information Technology Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Hafez Ave., 424, Iran

[b] Image Processing and Pattern Recognition Laboratory, Computer Engineering and Information Technology Department, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Hafez Ave., 424, Iran

HIGHLIGHTS

- We propose a distributed learning automata based algorithm for sampling from complex networks.
- The proposed algorithm is studied on 9 popular complex networks.
- The proposed algorithm is compared with well-known sampling methods.
- The experimental results show that the proposed algorithm is a viable approach for sampling from complex networks.

ARTICLE INFO

Article history:
Received 8 July 2013
Received in revised form 6 September 2013
Available online 18 November 2013

Keywords:
Complex networks
Social networks
Network sampling
Distributed learning automata

ABSTRACT

A complex network provides a framework for modeling many real-world phenomena in the form of a network. In general, a complex network is considered as a graph of real world phenomena such as biological networks, ecological networks, technological networks, information networks and particularly social networks. Recently, major studies are reported for the characterization of social networks due to a growing trend in analysis of online social networks as dynamic complex large-scale graphs. Due to the large scale and limited access of real networks, the network model is characterized using an appropriate part of a network by sampling approaches. In this paper, a new sampling algorithm based on distributed learning automata has been proposed for sampling from complex networks. In the proposed algorithm, a set of distributed learning automata cooperate with each other in order to take appropriate samples from the given network. To investigate the performance of the proposed algorithm, several simulation experiments are conducted on well-known complex networks. Experimental results are compared with several sampling methods in terms of different measures. The experimental results demonstrate the superiority of the proposed algorithm over the others.

1. Introduction

Many real world systems such as biological, social, ecological, technological and information systems are modeled as networks, which are represented as graphs with a set of vertices (e.g. users in social networks) and edges (relationship between users in social networks). It is shown that there are universal common features in different real world networks such as small-world and scale-free properties [1,2]. In recent years, complex networks have attracted an great deal of attraction by researchers in many applications [3–5]. In Ref. [6], information in the form of periodic orbits (cycles) exists

* Corresponding author. Tel.: +98 21 6454 5120.
E-mail address: a.rezvanian@aut.ac.ir (A. Rezvanian).

0378-4371/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.physa.2013.11.015

in the network modeled as complex networks. Moreover, complex networks have provided a suitable framework for real world systems such as software systems [7]. Due to the dynamicity, complexity, and large scale of complex networks, the conventional techniques of problem solving might experience some difficulties facing real networks [8]. Therefore, a network is characterized using metrics (e.g. centrality measures) or techniques (e.g. sampling), instead of access to the whole network [9–11]. In case of online social networks, in addition to the large scale of the network, limited access due to privacy settings has arisen. So, it is impossible to fully access the whole network. In practice, researchers via sampling from networks can estimate the characteristics of given large networks [12].

A sample from graph $G = \langle V, E\rangle$, where $V(G)$ is the set of $|V| = n$ vertices and $E(G)$ denotes the edge set including $|E| = m$ edges, is a function $f : G \to S$ from the given graph G to the sample-set $S = \langle V', E'\rangle$ such that $V' \subset V$ and $E' \subset E$. It is noted that the main goal of the sampling technique is to select a smaller subgraph from G. Sampling methods [9,11,13] play a significant role in preprocessing, characterizing, and studying networks. Sampling can be used to study a small part of the networks while preserving features of the initial network. For example in online social networks, due to privacy restrictions, it could not access the whole network information, therefore, the networks are characterized by estimated small samples from a significant part of the networks. Taking into account complexity, large scale and dynamic properties of the complex networks, most traditional methods were not scalable and thus failed. Moreover, the policies of some social network websites have restricted accessibility to the whole network simultaneously [14], because of computational or privacy restrictions. Hence, there are some special complexities and challenges in sampling from complex networks which can still be discussed as new research fields [15].

In general, for studying social networks such as Facebook/Twitter, in fact there is no assumption about the probability distribution of particular measures e.g. degree distributions, whereas some researchers have to address considering special assumptions of the network *a priori* [10] for computing the biasness. In p2p networks which are highly dynamic, vertices/edges are continuously added or removed from the network, while the sampling process covers some other parts of the network [16]. Moreover, the weights of edges considered in the network are usually assumed constant, though these weights are variables over time [17]. In some cases, sampling methods not only are limited to using smaller scales in order to reduce storage space and computational complexity, but also they are used to estimate, characterize and study networks. For example, the latest estimation of indexed web pages reports 3.6 billion pages in the internet [18].

The various methods reported in the literature for sampling from graphs can be categorized into three general approaches for the sampling techniques. In the first approach, the scale of the initial graph is reduced, which is mainly used for visualizing the graph applications [19] and describing the initial graphs. In this approach, several techniques can be utilized such as clustering [20], coarse graining [21], k-core [22] and fractal based methods [23]. In the second category, the random selection is done based on either nodes or edges which often does not provide proper results since it models just a small part of the graph without considering its topological structure [15]. The third category includes crawling methods or topology based sampling methods which has attracted a great deal of attention in the literature. As the topological structure is provided relatively by the crawling methods such as Breadth-First-Search (BFS) [24], Depth-First-Search (DFS) [25], Forest Fire Sampling (FF) [26], Snowball [27], Random Walk (RW) [28], Metropolis–Hastings Random Walk (MHRW) [29], Weighted Random Walk (WRW) [30], Stratified Weighted Random Walk (SWRW) [30], and Respondent Driven Sampling (RDS) [31], they offer more effective results in comparison with random selection methods. All of these approaches are similar in their basics with their general difference being the selection strategy for a part of the graph [9].

Leskovec et al. [15] proposed a sampling method from a large graph with two goals of back in time and down scale, while several methods have been introduced and compared. The study has demonstrated that vertex selection and edge selection techniques do not provide desirable results regarding their ignorance of the correlations between selected vertices/edges. An analytical comparison between RW and BFS sampling has been presented by Kurant et al. [10] to sample from a network. Their study reveals that the degree of the graph is overstated by the BFS, while it is understated by the RW sampling. Therefore, they suggested analytical solutions to correct the bias of estimation. A practical framework for uniform sampling from users of the social network Facebook has been developed based on crawling in Ref. [16]. In this research, the advantage of unbiased estimation of MHRW and Re-WRW (RWRW) over random sampling and breadth-first-search has been addressed by comparing various approaches. RDS was analyzed in Ref. [32] to reduce the biases associated with chain referral sampling of hidden populations. And then later, sampling from Twitter using the RDS method has been reported to characterize it in Ref. [33]. They have shown, through experimental examinations on Twitter, the lower error in RDS than that of MHRW. Analysis of the random walk method has been developed by Cooper et al. [34], where the authors have tried to sample from the high degree vertices and similar graphs regarding the power law distribution. Cumulative distribution of degrees is estimated via sampling based on tracerouting and some methods were studied for eliminating bias of the high degrees [35]. Ribeiro et al. proposed a sampling method, called Frontier sampling. It is developed from the traditional random walk which used several dependent random walks. The frontier sampling outperforms the conventional random walk and generates small errors in sparse graphs. According to the diversity communication between users of the social networks, a multi-graph has been introduced using random walk [11] and the results of its simulation indicate improvement of the proposed method by Gjoka. Random jump in MHRW for unbiased estimation has been proposed and it also prevents being trapped in local structures in Ref. [36] by Jin et al.

Variety in social network models makes various kinds of modeling possible. A modular structure has been studied in complex networks with sampling from the network being implemented based on identifying the communities by Maiya et al. [37]. In other research, structure of a bipartite graph has been considered in Ref. [38] for some social networks, then