Engineering Applications of Artificial Intelligence 59 (2017) 244–259

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Irregular cellular learning automata-based algorithm for sampling social networks

Mina Ghavipour, Mohammad Reza Meybodi*

Department of Computer Engineering and IT, Amirkabir University of Technology, Tehran, Iran

ARTICLE INFO

Keywords:
Complex networks
Social networks
Network sampling
Graph mining
Cellular learning automata

ABSTRACT

Since online social networks usually have quite huge size and limited access, smaller subgraphs of them are often produced and analysed as the representative samples of original graphs. Sampling algorithms proposed so far are categorized into three main classes: node sampling, edge sampling, and topology-based sampling. Classic node sampling algorithm, despite its simplicity, performs surprisingly well in many situations. But the problem with node sampling is that the connectivity in sampled subgraph is less likely to be preserved. This paper proposes a topology–based node sampling algorithm using irregular cellular learning automata (ICLA), called ICLA-NS. In this algorithm, at first an initial sample subgraph of the input graph is generated using the node sampling method and then an ICLA isomorphic to the input graph is utilized to improve the sample in such a way that the connectivity of the sample is ensured and at the same time the high degree nodes are also included in the sample. Experimental results on real–world social networks indicate that the proposed sampling algorithm ICLA-NS preserves more accurately the underlying properties of the original graph compared to existing sampling methods in terms of Kolmogorov-Smirnov (KS) test.

1. Introduction

Social network and the analysis of it is an inherently interdisciplinary field which is emerged from computer science, sociology, social psychology, statistical physics, and graph theory (Yang et al., 2013; Li et al., 2013; So and Long, 2013). The research on social network offers a framework for analysing the structure of whole network graph, identifying local and global patterns in these structures and studying dynamical properties on the network. Despite the importance of studying real–world social networks, it would be difficult to capture the structural properties of the whole network since we often face large scale networks with access limitations. To deal with this problem, many sampling methods have been reported in literature. Generally, the main goal of a sampling method is to produce a representative subgraph from the original network which can be used for studying characteristics of the larger network. The term "representative subgraph sampling" defined by Leskovec and Faloutsos (2006) refers to producing a small sample of the original network, whose characteristics represent as accurately as possible the entire network. There exist many characteristics which describe a network structure such as degree, clustering coefficient, and k –core distributions. Authors in (Leskovec and Faloutsos, 2006) proposed a set of empirical rules by which the measurements of the sample can be scaled up, to recover estimates for the original graph. Work in (Ebbes et al., 2013) investigated the ability of nine different sampling methods in preserving the structural properties such as degree, clustering coefficient, betweenness centrality, and closeness centrality of social networks. Lee et al. (2006) exploited three sampling algorithms and investigated the statistical properties of the samples taken by them. They focused on the topological properties such as degree distribution, average path length, assortativity, clustering coefficient, and betweenness centrality distribution.

Existing sampling methods can be categorized into three main classes: node, edge and topology–based sampling. Despite its simplicity, classic node sampling method performs surprisingly well in many situations (Leskovec and Faloutsos, 2006). The problem with node sampling is that connectivity in sampled subgraph is less likely to be preserved (Lee et al., 2006). Considering this in mind, this paper propose a topology–based node sampling algorithm, called ICLA-NS, that utilizes an irregular cellular learning automata (ICLA) to produce representative subgraphs by ensuring the connectivity of sampled subgraphs and sampling the nodes with high degree. ICLA-NS first constructs an initial sample subgraph of the input graph using the node sampling method, and then uses an ICLA isomorphic to the input graph to improve the sample by repeatedly replacing nodes in the sample with the nodes found by exploring the input graph. In order to

[*] Corresponding author.
E-mail address: mmeybodi@aut.ac.ir (M.R. Meybodi).

http://dx.doi.org/10.1016/j.engappai.2017.01.004
Received 17 March 2016; Received in revised form 30 November 2016; Accepted 4 January 2017
Available online 14 January 2017

evaluate the performance of the proposed sampling algorithm, we conduct a number of experiments on real–world networks. Based on our experimental results, the proposed sampling algorithm outperforms the existing sampling algorithms such as node sampling, Random Walk sampling and Forest Fire sampling in terms of Kolmogorov-Smirnov (KS) test for degree, clustering coefficient, and k–core distributions.

The rest of the paper is organized as follows. The next section presents notations and related works on network sampling. In Section 3, learning automata and cellular learning automata are introduced. Section 4 describes the proposed sampling algorithm ICLA-NS. Experimental results are given in Section 5. Section 6 concludes the paper.

2. Foundations of graph sampling

While the explicit goal of graph sampling algorithms is to produce a smaller subgraph from the original graph, there often exist other implicit goals for a sampling process. Three possible goals of graph sampling algorithms are: *Scale-down sampling*, *Back-in-time sampling*, and *Supervis*ed *sampling*. The Scale-down sampling aims to sample a representative subgraph that have similar (or scaled-down) topological properties to those of the original graph (Leskovec and Faloutsos, 2006). In Back-in-time sampling, the sampled subgraph matches temporal evolution of the original graph (Leskovec and Faloutsos, 2006). That is, the sampled subgraph G_s is similar to what the original graph G looked like when it was of the same size asG_s. Finally, the goal of supervised sampling is to identify nodes belonging to a specific category (Fang et al., 2013, 2015, 2016). For this purpose, a biased sampling is done to sample a subgraph under the requirements related to that category.

In this paper, we focus on the goal of sampling a representative subgraph (Scale-down sampling). This section provides some common notations and a formal definition of the graph sampling problem considered in this paper. In this section, several taxonomies of sampling algorithms reported in the literatures are also given.

2.1. Notations and definitions

Let$G(V, E)$ be an unweighted and undirected graph with the node set $V = \{v_1, v_2, ..., v_n\}$ and the set of edges $E = \{e_{ij} \mid v_i \in V, v_j \in V\}$, such that $|V| = n$ denotes the number of nodes, and $|E| = m$ denotes the number of edges. The neighbourhood of node v_i is defined as $N(v_i) = \{v_j \mid e_{ij} \in E, v_j \in V\}$, such that $d(v_i) = |N(v_i)|$ is the degree of nodev_i.

In this paper, we consider the following graph sampling problem. Given an input graph $G(V, E)$ and a sampling fraction f, a sampling algorithm samples a subgraph $G_s(V_s, E_s)$ with a subset of the nodes $V_s \subset V$ and a subset of the edges$E_s \subset \{e_{ij} \mid v_i \in V_s, v_j \in V_s\}$, such that $|V_s| = fn$. The goal is to ensure that the sampled subgraph G_s preserves the topological properties of the original graph G.

2.2. Related works

Graph sampling algorithms can be classified in several ways. We present three such classifications, namely random versus topology – based sampling, static versus streaming graph sampling, and simple versus extended sampling.

2.2.1. Random versus topology – based sampling

In random sampling methods, a sample subgraph is constructed by the random selection of either nodes or edges, and so these methods can be categorized into two main subclasses: node, and edge sampling. Classic node sampling (NS) (Leskovec and Faloutsos, 2006) chooses nodes uniformly at random from the original graph G. For a required fraction f of nodes, each node is independently sampled with a probability of f. The sampled subgraph G_s consists of the chosen nodesV_s as well as all the edges among them (E_s). Despite its simplicity, classic NS performs surprisingly well in many situations (Leskovec and Faloutsos, 2006). Authors in (Stumpf et al., 2005) indicated that classic node sampling does not accurately retain properties for graphs with power–law degree distributions. Although node sampling includes all the edges related to the sampled node set V_s, Lee et al. (2006) shown that the original level of connectivity is less likely to be preserved. Many other variations of NS have been developed in recent years (Krishnamurthy et al., 2005; Leskovec and Faloutsos, 2006; Ahmed et al., 2010, 2013).

Classic edge sampling (ES) chooses edges (rather than nodes) independently and uniformly at random from the original graph$G$. For each edge chosen (and added to E_s), both incident nodes are added to V_s. So, the sampled subgraph G_s is constructed by including the sampled edges in E_s and their incident nodes in V_s. Since classic edge sampling is biased towards high degree nodes and samples both incident nodes of chosen edges, it can accurately preserve the path length distributions (Ahmed et al., 2010, 2013). However, ES is less likely to capture the original clustering and connectivity, since it samples the edges independently (Lee et al., 2006). Classic edge sampling generally produces sparse subgraphs. Some improved variations of ES have been proposed so far (Krishnamurthy et al., 2005; Leskovec and Faloutsos, 2006; Ahmed et al., 2010, 2013).

There also exist many sampling methods based on topological structure of graph. The common idea in this class of sampling methods is to explore the neighbouring nodes of a given node. These methods can be categorized into two subclasses: *random walks* and *graph traversals*. In the category *random walks*, sampling is performed with replacement, i.e. nodes can be revisited. This category includes classic Random Walk (RW) (Lovász, 1993; Yoon et al., 2007; Lu and Li, 2012) and its variations (Henzinger et al., 2000; Gkantsidis et al., 2006; Stutzbach et al., 2009; Rasti et al., 2009; Ribeiro and Towsley, 2010; Avrachenkov et al., 2010; Kurant et al., 2011; Lee et al., 2012; Rezvanian et al., 2014). In the category *graph traversals*, each node is visited at most once (sampling without replacement). Methods in this category differ in the order in which they visit the nodes. Examples are Breadth –First Search (BFS) (Lee et al., 2006), Depth–First Search (DFS) (Even, 2011), Forest Fire (FF) (Leskovec and Faloutsos, 2006), Snowball Sampling (SBS) (Goodman, 1961; Newman, 2003b; Illenberger et al., 2011), Respondent–Driven Sampling (RDS) (Heckathorn, 1997; Goel and Salganik, 2010), and Expansion Sampling (Maiya and Berger-Wolf, 2010). The sampled subgraph G_s in topology–based sampling methods consists of the explored nodes and edges. Sampling methods based on topology outperform simple methods such as NS and ES (Leskovec and Faloutsos, 2006).

2.2.2. Static versus streaming graph sampling

The sampling algorithms based on the assumption of a static graph (Goodman, 1961; Lovász, 1993; Heckathorn, 1997; Leskovec and Faloutsos, 2006; Lee et al., 2006; Maiya and Berger-Wolf, 2010; Even, 2011) consider the input graph only at one point in time and assume that it is of moderate size which can fit in the main memory. However, many real–world networks are too large to fit in the memory, and evolve continuously over time and thus are not fully observable at any point in time. Activity networks (e.g. email), social media (e.g. Twitter), and content sharing (e.g. Facebook, YouTube) are the examples of such large dynamic networks. Analysing these networks, called streaming graphs, is increasingly important for identifying patterns of interactions among individuals and investigating how the network structure evolves over time. As a result, the streaming graph sampling has received more attention in recent years. Researchers have developed algorithms for sampling from streaming graphs (Ahmed et al., 2010, 2013). Authors in (Ahmed et al., 2013) outlined a spectrum of computational models for designing sampling algorithms, going from static to streaming graphs. They presented the streaming