Computer Communications 000 (2015) 1–12

Contents lists available at ScienceDirect

Computer Communications

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

More than topology: Joint topology and attribute sampling and generation of social network graphs

Michael Seufert*, Stanislav Lange, Tobias Hoßfeld[1]

Institute of Computer Science, University of Würzburg, Würzburg, Germany

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Social networks
Attributes
Graph sampling
Graph similarity
Graph generation

ABSTRACT

Graph sampling refers to the process of deriving a small subset of nodes from a possibly huge graph in order to estimate properties of the whole graph from examining the sample. Whereas topological properties can already be obtained accurately by sampling, current approaches do not take possibly hidden dependencies between node topology and attributes into account. Especially in the context of online social networks, node attributes are of importance as they correspond to properties of the social network's users. Therefore, existing sampling algorithms can be extended to attribute sampling, but still lack the capturing of structural properties. Analyzing topology (e.g., node degree and clustering coefficient) and attribute properties (e.g., age and location) jointly can provide valuable insights into the social network and allows for a better understanding of social processes. As major contribution, this work proposes a novel sampling algorithm which provides unbiased and reliable estimates of joint topological and attribute based graph properties in a resource efficient fashion. Furthermore, the obtained samples allow for the generation of synthetic graphs, which show high similarity to the original graph with respect to topology and attributes. The proposed sampling and generation algorithms are evaluated on real world social network graphs, for which they demonstrate to be effective.

1. Introduction

With steadily increasing size and popularity, online social networks (OSNs) such as Facebook, Twitter, and Google+ have drawn the interest of the scientific community. Analyzing the structure and properties of these networks allows research in various fields. By studying social behavior and finding patterns in the networks' structure, it is possible to acquire knowledge about requirements and parameters for future networks and applications. Additionally, OSNs have a high impact on today's users' choice and consumption of online media. Coupled with widespread availability of mobile Internet, these phenomena give rise to the scientific field of socially aware traffic management [29]. The main idea in this discipline is to utilize social information about Internet users in order to enhance existing traffic management strategies. For example, information from OSNs about users' interests allows for improved caching solutions. However, complete social networks are seldom available due to privacy restrictions and the OSN providers' reluctance to publish the core of their business data. Furthermore, running complex algorithms on social network graphs in the order of magnitude of hundreds of millions of nodes is usually infeasible due to time and resource constraints.

Graph sampling techniques address the latter issue by examining only a representative subset of a given graph. Formally, the task of deriving a node sample from a given graph $G = (V, E)$ with node set V of size n and edge set E can be defined as finding a subset of nodes $V_S \subseteq V$ whose topological information can be used in order to reliably estimate various properties of G. There are two main quality requirements for sampling strategies. First, the generated sample has to be unbiased. That is, the expected value of the sampled data and the actual value of the estimated parameter are equal. Second, the minimum amount of samples required for reliable results should be low.

Unfortunately, state-of-the-art graph sampling techniques are limited to the topological analysis of huge graphs whereas socially aware traffic management requires additional information on user attributes like interests, geographic location, and age. Retrofitting these algorithms with attribute sampling capabilities implicitly assumes the independence of attributes and topology and can provide inferior results. Therefore, this work transfers ideas from graph sampling to graphs with node attributes by proposing a joint sampling of structure and attributes, which takes possible dependencies between

* Corresponding author. Tel.: +49 931 3188475; fax: +49 931 3186632.
E-mail addresses: seufert@informatik.uni-wuerzburg.de (M. Seufert), stanislav.lange@informatik.uni-wuerzburg.de (S. Lange), tobias.hossfeld@uni-due.de (T. Hoßfeld).
[1] Now at: University of Duisburg-Essen, Chair of Modeling of Adaptive Systems, Essen, Germany.

http://dx.doi.org/10.1016/j.comcom.2015.07.023
0140-3664/

topology and attributes into account. The resulting sampling mechanism provides unbiased and reliable estimates of joint topological and attribute based properties of social network graphs in a resource efficient fashion. As an application of this sampling approach, a graph generation method [8] is augmented to use a collected sample for generating synthetic social network graphs, which show joint structure and attribute properties similar to the original graph. In order to quantify this similarity, measures were developed, which assess similarity not only with respect to topology, but also take attributes into account.

Thus, the contribution of this work is threefold:

- Existing sampling algorithms are extended to node attributes
- A novel sampling algorithm is proposed which allows for joint capturing of structure and attribute characteristics
- A graph generation method is presented that reproduces topology and attribute related properties of the original graph based on sampling

Therefore, this work is structured as follows. Section 2 covers relevant related work on attribute sampling and graph generation, and describes the used social network data sets. Section 3 introduces the sampling mechanism and presents results. Topological graph similarity measures are extended to additionally assess attribute related similarity in Section 4. A method to generate synthetic social network graphs from a node sample with attributes is proposed in Section 5. The performance of the algorithm is evaluated for different social network graphs and attributes. The results are discussed and an outlook on future work is given in Section 6.

2. Related work

2.1. Graph sampling

Numerous approaches have emerged since graph sampling became a relevant scientific topic. We revisit state-of-the-art graph sampling algorithms that are classified into three categories, namely, Uniform Node Sampling (UNI), Breadth First Search (BFS), and random walks (RW). Though primarily focused on sampling topological graph properties, these algorithms provide a solid foundation for the design of novel sampling algorithms. Due to extensive research and several performance benchmarks [17–19], they have proven properties and behavior and are also well-established in practice.

In the context of UNI, a given amount of n_S nodes is drawn at random from the original graph's set of nodes V. While this procedure guarantees an unbiased sample, it is not practical in most situations due to several possible restrictions. These include sparse ID spaces where multiple queries may be required in order to obtain a single sample, or even completely unknown ID spaces where no information about the domain of user IDs is available. Thus, UNI is considered as reference for the theoretical best case. Although BFS and related methods like depth first search [17] and snowball sampling [12] were used for various sampling tasks in the past [24,25], current research suggests avoiding these methods due to a bias towards nodes with high degree [17,18]. Additionally, this bias is graph specific and no mechanism for correcting this bias has been developed yet. Recent graph sampling mechanisms rely on random walks [11,20,27], a family of algorithms that require only the basic operation of querying a node for its set of neighbors. With the possibility to exactly quantify the node degree bias encountered in random walks, techniques for correcting this bias have been developed and allow collecting samples that are unbiased with respect to topology. Most commonly used representatives include the Metropolis Hastings Random Walk (MHRW) and the Re-Weighted Random Walk (RWRW). Based on the Metropolis Hastings algorithm [23], MHRW is a rejection sampling technique that corrects the bias on the fly. Its applications range from the analysis of P2P networks [26,31] to that of directed [34] and undirected [7] OSNs. In contrast, RWRW first performs a biased RW and then applies the Hansen–Hurwitz estimator [10] to the degree distribution observed in the sample. By dampening the occurrence probabilities of high degree nodes, this yields an unbiased distribution. The main advantage of RWRW over MHRW is that RWRW avoids spending a large portion of its sampling budget on rejections. In the case of MHRW, around 55% of iterations are rejections [7]. Therefore, on average, MHRW's resulting set of sampled nodes not only consists of fewer unique nodes, but also stems from a shorter walk. A generalization of the RWRW algorithm is presented in Section 3 and is the basis of the developed sampling algorithm. It allows for estimation of the two-dimensional distribution of node degrees and attribute values. While the literature also offers techniques for sampling from dynamic, time dependent graph streams [1], our proposed method works with static graphs. The main reason for this behavior is that typical methods involving dynamic graphs require operations like drawing a graph's edges uniformly at random which is not possible in real world OSN graphs. Furthermore, algorithms for estimating a graph's size have been proposed [13]. However, we assume that the graph's size is part of the input.

2.2. Graph generation

Modeling real networks is an important branch of science with many applications, including the analysis of biological and social systems. When a model is able to use a real graph to consistently generate synthetic graphs that capture the majority of the original graph's properties, its resulting graphs can be used as input for algorithm benchmarks and simulations, or as a means of anonymizing crawled data before publication. This section outlines three state-of-the-art models that can be used to generate synthetic graphs given either the full input graph or even just a sample of its node set.

Exponential Random Graph Models (ERGMs) [16,30,35] constitute a family of statistical models whose goal is to reveal dependencies in the process of edge creation in networks. This is achieved by quantifying the importance of various graph statistics that summarize structural patterns in the graph. ERGMs are often used for characterizing social networks [4,6,28]. Additionally, these models allow generating synthetic graphs once model parameters have been estimated. However, ERGMs require complete information on the input graph and current implementations' time complexity prohibits the analysis of graphs whose size exceeds a few thousand nodes [36], thus excluding most real world OSNs.

In [14], the Multiplicative Attribute Graph Model (MAG), a generative model for graphs with categorical node attributes, is proposed. The basic idea is that the probability of two nodes being involved in an edge depends on the nodes' pairwise combinations of attribute values. By quantifying the probability of edge formation for each possible combination of attribute values, various relationships like homophily, heterophily, or the tendency to seek connections to a specific attribute value can be expressed. Given the original graph, model parameters representing the aforementioned probabilities can be estimated. Based on these parameters, the model is capable of generating synthetic graphs with realistic topological and attribute related properties [15]. Unfortunately, MAGs also require the full input graph G for parameter estimation. As in the case of ERGMs, this is the factor that makes the approach unsuitable for this work's goals of analyzing real world OSNs and generating similar networks based on samples.

The authors of [8] present a method for generating a topology similar to the one of a real world graph without requiring full knowledge of that graph. The input consists of a node sample collected during a random walk on the graph of interest. Main aspects to reproduce are the joint degree distribution (JDD), basically an edge count for each type of degree–degree combination, as well as the average clustering coefficient per node degree. For this purpose, values of the aforementioned measures are derived from the collected sample and are

Please cite this article as: M. Seufert et al., More than Topology: Joint topology and attribute sampling and generation of social network graphs, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.07.023