Discrete Applied Mathematics 206 (2016) 109–114

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Improving the characterization of P-stability for applications in network privacy

Julián Salas [a,*], Vicenç Torra [b]

[a] Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Tarragona, Spain
[b] School of Informatics University of Skövde, Skövde, Sweden

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 15 January 2016
Accepted 24 January 2016
Available online 20 February 2016

Keywords:
P-stability
k-anonymity
Graphic sequence
Degree sequence
FPRAS
Rapidly mixing Markov chain
Fully polynomial-time randomized approximation scheme

ABSTRACT

Recently, we have found that the concept of P-stability has interesting applications in network privacy. In the context of Online Social Networks it may be used for obtaining a fully polynomial randomized approximation scheme for graph masking and measuring disclosure risk. Also by using the characterization for P-stable sequences from Jerrum, McKay and Sinclair (1992) it is possible to obtain optimal approximations for the problem of *k*-degree anonymity. In this paper, we present results on P-stability considering the additional restriction that the degree sequence must not intersect the edges of an excluded graph *X*, improving earlier results on P-stability. As a consequence we extend the P-stable classes of scale-free networks from Torra et al. (2015), obtain an optimal solution for *k*-anonymity and prove that all the known conditions for P-stability are sufficient for sequences to be graphic.

1. Introduction

With the increasing use of social networks, researchers and enterprises have found that valuable data can be obtained from their analysis, this has fostered the development of data mining algorithms for extracting relevant information from the data, cf. [1]. The extraction of information has implicit the possibility of revealing private attributes or relationships from the participants in the networks. This can be attenuated by the development of privacy protection methods.

Formerly, with the objective of protecting data from census and surveys the fields of statistical disclosure control and privacy preserving data mining developed methods and measures for the evaluation of disclosure risk and information loss of protected data, cf. [10]. More recently these methods have been extended to Social Networks. Social networks can be modelled by a graph in which the vertices have personal information attached to them and the edges represent the structure of their relations.

Sometimes the structural properties and node attributes are considered together, e.g. [2,26]. However, they are more commonly studied separately: the information attached to the vertices represented by a database, and the relational structure represented by a graph.

Several protection methods have been developed, and many techniques from database protection have been inherited by graphs when possible, such as *k*-anonymity [18,22].

* Corresponding author.
E-mail addresses: julian.salas@urv.cat (J. Salas), vtorra@his.se (V. Torra).

http://dx.doi.org/10.1016/j.dam.2016.01.025
0166-218X/

In the case of k-anonymity, there are many different interpretations that originate different definitions for graphs, e.g. k-degree anonymity [14], k-neighbourhood anonymity [27], in general all of them can be resumed as k-candidate anonymity [9] or $k-\mathcal{P}$-anonymity [3], i.e. for a given structural property $\mathcal{P}$ and a vertex in the graph G there are at least $k-1$ other vertices with the same property $\mathcal{P}$.

Note that the most restrictive of all the structural properties, in the sense that it implies all the others, is when $\mathcal{P}$ is the neighbourhood, cf. [21]. On the other hand, the most basic is the degree, i.e. all these definitions imply k-degree anonymity, therefore the minimum number of edge modifications needed to obtain k-degree anonymity could be used as a lower bound for all the other properties.

Another technique for anonymization is randomization. In [9] a random deletion and insertion of m edges is performed to protect a graph, with m taking the values of %5, %10, %100 of the total edges, the effects of perturbation are compared for graph measures such as the degree, diameter, path length, closeness, betweenness and clustering coefficient for real world datasets.

As it is pointed out in [25] the generation of graphs with a given degree sequence does not guarantee that the generated graphs preserve the topological features of the real graph. The authors implement an algorithm using a Markov chain that preserves various features of the original graphs. They state that it has been shown empirically for many networks that the number of steps k for the Markov chain to approach stationarity is around $k = 10m$, cf. [16].

However, theoretical bounds are also important and in many cases lacking. Hence we focus on this problem improving the theoretical condition (P-stability) for the Markov chain for generating the graphs with degree sequence $\mathbf{d}$ from [12] to be rapidly mixing. Note that [12] uses P-stability to prove that there exist a fully polynomial almost uniform generator for $\mathcal{G}(\mathbf{d})$ the set of graphs with given degree sequence $\mathbf{d}$, and a fully polynomial randomized approximation scheme for $\mathcal{N}(\mathbf{d})$ the number of such graphs.

In [11], the authors study further conditions in order to characterize P-stable sequences in terms of the maximum and minimum degrees of a sequence $\mathbf{d}$.

This paper broadens the class of P-stable sequences considering excluded graphs and in many cases improves the characterizations of [11] and [12], that are stated in Eqs. (1)–(4). See [15], for a survey of recent results on excluded graphs. Also, we use a condition obtained from Theorem 2.4 for a degree sequence $\mathbf{d}$ to be P-stable to prove that it is graphic. So it must not be assumed in advance that the sequences are graphic as in [12,11], i.e., that $\mathcal{G}(\mathbf{d}) \neq \emptyset$. The study of conditions for a degree sequence to be graphic is a relevant problem by itself. There are well known conditions for a degree sequence to be graphic, cf. [8,5,6]. More recent conditions can be found in [4,19,24].

As an application of our results, we extend the class of scale-free networks from [23] and give optimal solutions for k-anonymity when the information loss metric is the distance δ_2, the Euclidean distance for degree sequences.

It should be noted that for applications certain features (e.g., path length, clustering coefficient, betweenness, closeness, centrality, modularity, spectral measurements) must remain barely modified in order to preserve the utility of the original network after anonymization. However, a measure that is optimal for one application is often sub-optimal for a different application, and the utility for a given algorithm for anonymization should be calculated by applying it to a real or synthetic graph. We would like to remark that this aspect of the anonymization is not treated in this article, nevertheless the Euclidean distance gives a numerical value that is independent of the application and therefore can be used as general criterion for anonymization.

1.1. Definitions

We consider only undirected, simple graphs (without loops or multiple edges). All the graphs have vertex set $V = \{v_1, \ldots, v_n\}$ and edge set E. The *degree* d_i of a vertex v_i, is the total number of edges incident at v_i. We say that a degree sequence is graphic if it can be realized as the degree sequence of a simple graph.

Suppose that $\mathbf{d} = (d_1, d_2 \ldots, d_n)$ is a degree sequence and X is a graph to be avoided. Let $\mathcal{G}(\mathbf{d}, X)$ denote the set of graphs $G \in \mathcal{G}(\mathbf{d})$ such that the edge sets of G and X are disjoint. We will call X an *excluded graph.*

Let $\mathcal{G}'(\mathbf{d}, X)$ be the union $\cup'_{\mathbf{d}}\, \mathcal{G}(\mathbf{d}', X)$, where $\mathbf{d}'$ ranges over all sequences $\mathbf{d}' = (d'_1, d'_2 \ldots, d'_n)$ that satisfy $\sum_{i=1}^n d_i = \sum_{i=1}^n d'_i$ and $\sum_{i=1}^n |d_i - d'_i| = 2$. We say that a class of degree sequence/excluded graph pairs $(\mathbf{d}, X)$ is *P-stable* if there exists a polynomial $p(n)$ such that $\frac{|\mathcal{G}'(\mathbf{d},X)|}{|\mathcal{G}(\mathbf{d},X)|} \leq p(n)$ for every $(\mathbf{d}, X)$ in the class.

Observe that as an application of degree sequence/excluded graph, the probability that a random graph with degree sequence $\mathbf{d}$ has X as a subgraph can be calculated as $P_{\mathcal{G}(\mathbf{d})}(\mathbf{X}) = \frac{|\mathcal{G}(\mathbf{d}-\mathbf{x},\mathbf{X})|}{|\mathcal{G}(\mathbf{d})|}$, where $\mathbf{x}$ denotes the degree sequence of X. Notice, that when the excluded graph is empty $X = \emptyset$, we have the usual definition for P-stability.

In [12], the authors proved that P-stability is a sufficient condition to guarantee that there exists a *fully polynomial almost uniform generator* for $\mathcal{G}(s)$ and a *fully polynomial randomized approximation scheme* for $|\mathcal{G}(s)|$.

An *almost uniform generator* for Y is a probabilistic algorithm which, given an instance x and a positive real bias ϵ, outputs an element of $Y(x)$ such that the probability of each element appearing approximates $|Y(x)|$ within ratio $(1+\epsilon)$. The generator is *fully polynomial* if its execution time is bounded by a polynomial in $log\epsilon^{-1}$ and the size of x, cf. [13,20].

If u and v are distinct vertices with $uv \notin E(G)$, then u and v are *non-adjacent* and the pair u, v is a *non-edge* in G. Define an *alternating path* of length l in G to be a sequence of vertices $v_0, v_1, \ldots, v_l$ such that $v_i v_{i+1}$ is an edge when i is even and a non-edge when i is odd.