Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Practical anonymity models on protecting private weighted graphs

Yidong Li [a,*], Hong Shen [b], Congyan Lang [a], Hairong Dong [c]

[a] School of Computer and Information Technology, Beijing Jiaotong University, China
[b] School of Information Science and Technology, Sun Yat-sen University, China
[c] State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form
21 July 2016
Accepted 27 August 2016
Communicated by: Steven Hoi

Keywords:
Anonymity
Weighted graph
Privacy preserving graph mining
Weight anonymization

ABSTRACT

Identity disclosure control (IDC) on graph data has attracted increasing interest in security and database communities. Most existing work focuses on preventing identity disclosure derivable from certain structural information in unweighted graphs. In weighted graphs, when the weight of an edge implying relevance/association between its adjacency vertices is taken into account, this problem becomes more complex due to the diversity of weight-related information which may expose to many types of background knowledge attacks and hence significantly increases the time complexity for preventing privacy breaches. This paper systematically studies IDC in weighted graphs, which has no known solution to our knowledge, by employing *elementary weight invariants* as background knowledge. We propose a general anonymity model against weight-related attacks, and introduce a new utility metric based on spectral graph theory. Then we distinguish two types of practical breaches, namely *volume* and *histogram* attack, which the adversary has the knowledge of the sum and the set of adjacent weights for each vertex respectively. We propose an efficient method for volume anonymization, and a heuristic scheme for histogram anonymization which we show to be NP-hard. We show how to construct the graph under these anonymized properties to protect a graph from both attacks. Our approaches are effective in terms of efficiency and data utility preservation: run in near-quadratic time on graph size, and preserve a similar utility as the original graph. The performances of the algorithms have been validated by extensive experiments on both synthetic and real-world datasets.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

As more and more real-world graphs modelling complex systems such as social networks are released publicly, there is a growing concern about privacy breaches for the entities involved in these graphs. As an integral part of data security, *identity disclosure (ID)* is to reveal the identification of an entity according to the related information in a released database. Then the adversary can acquire the corresponding sensitive information of the entity. A popular attack model for such a breach concerns with background knowledge of an individual that contains sufficient information for an adversary to identify the individual. A rich body of research on this subject focuses on modelling different types of background knowledge attacks for relational data [26,17,14,1] and simple graphs (i.e. undirected, unweighted and acyclic) [12,2].

This paper takes a further step to discuss the ID problem on weighted graphs of which most studies mentioned above steer clear. The motivation is that, many graphs are intrinsically weighted as has long been appreciated, which introduce more information than its simple version and makes them more vulnerable to privacy breaches. In a social network there may be stronger or weaker social ties between individuals. In a transportation network there may be longer or shorter distances between stations. In a communication network there may be more or less bandwidth or data flow between routers/clients.

Intuitively, weight-related properties (or *weight properties* in short), whose quantitative measures are called *weight invariants*, embody richer information leading to more effective attacks on a graph. This is true even for those elementary invariants such as the sum and the set of adjacent weights for a vertex. In contrast to the breaches with weight-independent invariants such as degree or neighborhood structure, attacks with elementary weight invariants show the effectiveness referring to the following three aspects.

First, weight properties bring in more distinct values increasing the possibility of disclosing individuals in real-world datasets. Let us consider a coauthorship network used in our experiments, named NetSci, where vertices (with real labels removed) denote researchers and weights on edges represent the numbers of co-written papers. Our experimental result shows that this graph contains 1% of the total 1589 nodes with exclusive values for

* Corresponding author.
E-mail address: ydli@bjtu.edu.cn (Y. Li).

http://dx.doi.org/10.1016/j.neucom.2016.08.084
0925-2312/© 2016 Elsevier B.V. All rights reserved.

Please cite this article as: Y. Li, et al., Practical anonymity models on protecting private weighted graphs, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.08.084

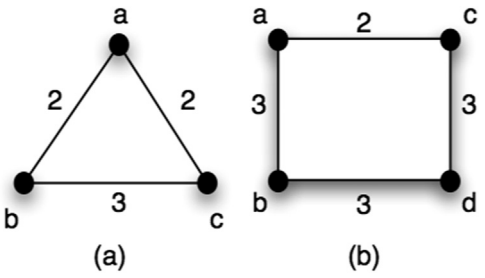

Fig. 1. Weighted graphs with degree anonymity.

degree but around 6% for the so-called volume (the sum of adjacent weights for a vertex). Moreover, such weight invariants are quite easily known by public as background knowledge under no awareness to information owners.

Secondly, most 'already-secured' graphs prevented from non-weighted attacks, however, have a high risk to be breached with weight properties. As a consequence, many preserving algorithms for simple graphs may not be extended or adapted to their weighted version. For instance, Fig. 1(a) states a 3-degree anonymous graph [15], in which none of the vertices can be distinguished from the other two with the degree value (all with degree 2). However an adversary can determine the real entity represented by vertex v_1 if he has prior knowledge about either the sum (4) or the set ($\{2, 2\}$) of adjacent weights of the entity, which are both unique information in the released graph.

In addition, an adversary usually obtains weight invariants as background knowledge quite easily under no awareness to information owners. Consider a coauthorship network of computer scientists released publicly, where vertices (with real labels removed) denote researchers and weights on edges represent the numbers of co-written papers. Assuming the information that there is an author, say Bob, in the network is known publicly, an adversary can identify Bob easily with the volume representing the total number of published papers, which is the public information provided in the resume on Bob's personal website.

Although a number of privacy models and preservation techniques have been proposed for the ID problem on graphs, most existing studies assume that the adversary has knowledge of certain structural information based on unweighted graphs. The problem of introducing weight is challenging due to the diversity of weight-related information, which leads to many types of background knowledge attacks, and would significantly increases the time complexity if applying the existing preservation methods to against these attacks. Hence, as discussed later, in neither models nor techniques the existing work can be easily extended to guarantee privacy in weighted graphs. For example, the model of k-degree anonymity [15] is shown impractical in real-world scenarios and the extension of its anonymization methods only maintains stable and acceptable data utility with the assumption of non-negative integer weights.

In this paper, we systematically demonstrate prevention of a graph from weight-related attacks with elementary weight invariants. We show how easily the values of these invariants can be obtained by an adversary in the real-world scenarios. Further, we identify two concrete attacks, and provide anonymization algorithms for each case based on a two-step framework. The contribution of this paper are summarized as follows:

- This paper proposes a general model for *weighted graph anonymization (WGA)*, in order to address the ID problem in weighted graphs with elementary weight invariants as background knowledge. Two invariants are considered: (1) volume: the sum of adjacent weights for a vertex; and (2) histogram: the neighborhood weight distribution of a vertex, and we show empirically how high the disclosure risk is while an adversary intents to breach real-world graphs with some values of these invariants *a priori*.
- The paper also proves the histogram anonymization problem is NP-hard in general case, and presents an efficient heuristic algorithm for this problem running in near-quadratic time on graph size. We also show theoretically and empirically how the proposed methods perform on both data privacy and utility.
- Finally the paper provides graph construction methods for both anonymization problems based on the perturbed volume sequence and histogram set. In addition, we propose a spectrum-based metric to quantify the information loss incurred in graph anonymization, and theoretically justify the impact of weight modification on the proposed metric.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of the literature on the graph anonymization problem. In Section 3, we formally define a general model for weighted graph anonymization and provide two concrete cases of weight-related attacks. We also discuss the use of graph spectrum as the information loss during anonymization in this part. Sections 4 and 5 focus on methods against volume attack and histogram attack respectively. We present the experimental results in Section 6 and conclude this paper in Section 7.

2. Related work

In the past decade, privacy preserving data publishing on tabular data has been extensively studied [26,17,14]. With the great development of new data applications and services, such as social networks and bioinformatics, research on privacy preserving graph publishing has attracted increasing interest in security and database communities [2,30,15,20,32,29].

As one of the pioneer work, [2] describes a family of attacks on a social network whose labels for vertices are replaced with meaningless unique identifiers. The basic idea behind these attacks is to create or find unique subgraphs embedded in an arbitrary network. Then, adversaries can learn whether edges exist or not between specific targeted pairs of nodes. The similar problem is also discussed in [12], which points out the risk that simply removing the identifiers (or label) of the nodes does not always guarantee privacy. The authors study a spectrum of adversary external information and its power to re-identify individuals in a social network. In details, two types of adversary knowledge are formalized: (1) vertex refinement queries, which reveal the structure of a graph around a vertex. For a node v, such information includes its label, degree, the list of its neighbors' degree, and so on. (2) Subgraph knowledge queries, which investigate the uniqueness of a subgraph around the target node.

The above two directions are extended by following studies [15,33] accordingly. [15] studies a specific graph-anonymity model named *k-degree anonymity*, which prevents the re-identification of individuals by adversaries with a priori knowledge of the degrees of certain nodes. The work in [33] identifies the *k-neighborhood anonymity* problem: if an adversary has some knowledge about the neighbors of a target node and the relationship among them, it is possible to re-identify the node in the network. The authors propose an anonymization approach based on neighborhood component coding technique, but admit that the algorithm will be a serious challenge in computation as the neighborhood size increased. Zheleva and Getoor [31] consider the problem of protecting sensitive relationships among individuals in anonymized social networks. This is closely related to the link-prediction problem that has been widely studied in the link mining community.

Please cite this article as: Y. Li, et al., Practical anonymity models on protecting private weighted graphs, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.08.084

Download English Version:

https://daneshyari.com/en/article/4948169

Download Persian Version:

https://daneshyari.com/article/4948169

Daneshyari.com