Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

Utility-aware social network graph anonymization

Mohd Izuan Hafez Ninggal, Jemal H. Abawajy

Parallel and Distributed Computing Lab, School of Information Technology, Deakin University, Victoria, Australia

ARTICLE INFO

Article history:
Received 15 January 2014
Received in revised form
2 December 2014
Accepted 14 May 2015
Available online 9 June 2015

Keywords:
Social network
Data publishing
Anonymization
Data utility
Privacy
K-anonymization

ABSTRACT

As the need for social network data publishing continues to increase, how to preserve the privacy of the social network data before publishing is becoming an important and challenging issue. A common approach to address this issue is through anonymization of the social network structure. The problem with altering the structure of the links relationship in social network data is how to balance between the gain of privacy and the loss of information (data utility). In this paper, we address this problem. We propose a utility-aware social network graph anonymization. The approach is based on a new metric that calculates the utility impact of social network link modification. The metric utilizes the shortest path length and the neighborhood overlap as the utility value. The value is then used as a weight factor in preserving structural integrity in the social network graph anonymization. For any modification made to the social network links, the proposed approach guarantees that the distance between vertices in the modified social network stays as close as the original social network graph prior to the modification. Experimental evaluation shows that the proposed metric improves the utility preservation as compared to the number-of-change metric.

1. Introduction

The number of social network users to communicate with family, friends, and colleagues is exponentially increasing. For example, Facebook boosts 1.15 total users around the world as of 2013. The social network data is a collection of entities and connections (links) among them. The links represents relationships (e.g., friends, family, etc.),financial exchange, web links, etc. The usefulness of these data in capturing online social activities promises benefits in many fields such as economics, sociology and information science. Thus, the rapidly increasing data generated by online social media services offers great opportunities for new information discovery.

It has been shown that social network data makes commerce much more profitable (Swamynathan et al., 2008). As a result, network operators are increasingly sharing social network graphs with advertising partners to enable better social targeting of advertisements (Clauset, 2005; Newman and Girvan, 2003). Since social network data contains sensitive information about the individual users, publishing social network data raises privacy concerns. In social network graph, the vertex attributes and relationships between vertices may be sensitive information that could be exploited to breach the privacy of the individuals. For instance, in a sexual-relationship network with gender information attached to vertices, it can reveal sexual orientation. Thus any information released may be subject to privacy implications for the involved individuals. To ameliorate the privacy concerns of social network data publishing, various social network data anonymization techniques have been proposed (Zhou and Pei, 2010; Wu et al., 2010; Liu and Terzi, 2008; Zhou and Pei 2008; Zou et al., 2009; Cheng et al., 2010; Tai et al., 2011; Xiaoyun et al., 2009). Since the cost of the anonymization is quantified by the utility loss, the utility of the anonymized social network data is an important factor in publishing social network data. In this paper, we focus on the *re-identification attack* (Liu and Terzi, 2008) as this attack is one of the most serious privacy problems in social network platform. Identity re-identification attack using semi-identifiable attribute has been well studied in micro-data privacy-preserving area (Fung et al., 2010). However, besides personal attributes (identifiable and semi-identifiable), social network data also consists of relationship information.

Ideally, the anonymized social network should preserve the privacy of the individuals with the smallest utility loss such that an analysis derived from the anonymized social network data are very similar to those from the original one (Wang et al., 2014). Thus, the problem of maintaining high data utility in social network data anonymization is paramount importance. By high data utility, we meant that the pure information carried by the original social network data is highly preserved. The information that has distorted too much from its original will likely produce unreliable analysis outcomes. However, most of the existing anonymization techniques fail to generate anonymized social

E-mail addresses: mninggal@deakin.edu.au (M.I.H. Ninggal), jemal@deakin.edu.au (J.H. Abawajy).

http://dx.doi.org/10.1016/j.jnca.2015.05.013

network data with high utility (Wang et al., 2014). Thus, it is important to understand and model utility of social network data being published through utility-aware metrics (Watanabe et al., 2011).

In this paper, we address the problem of anonymizing the relationship information (i.e., social links) with minimal changes to the social network graph with the aim to protect the identity of each individual involved. It is very challenging to maintain high utility of the data when the link structure of a social network graph is modified to pursue anonymity. To this end, we propose a metric to capture the impact of changes on the structural properties of social network data. Currently, the amount of changes made to the social graph is the common metric used to control the utility distortion in structural-based anonymization (Zhou and Pei, 2010, 2008; Liu and Terzi, 2008; Zou et al., 2009; Cheng et al., 2010). However, this metric does not consider the impacts on the social links structure, which has serious impact on the graph properties. In contrast, the proposed metric leverages the shortest path length and the neighborhood overlap to weigh the connection edge that subject to be modified. Based on the weight value, the anonymization algorithm then heuristically performs edge perturbation. The proposed mechanism guarantees, for any perturbation made to the social links, the distance between vertices is as close as the original network. This is beneficial in preserving the relative importance of vertices in social network data. In summary, we make the following main contributions:

a. We identify the ineffectiveness of the common utility measurement adopted by many existing k-anonymization algorithms on preserving the structural features (i.e., utility) of the published social networks;
b. We propose a new metric to determine the impact on structural integrity from anonymization operation. We propose to build the utility loss measurement based on the community-based graph model instead of the simple degree sequence-based graph model.
c. We compare the performance of the proposed approach with the popular existing approach.
d. We conduct extensive experiments to verify the effectiveness and efficiency of our proposed approach. The results show that our scheme achieves significant improvement on the utility of the anonymized social networks compared with the existing anonymization algorithms. In most cases, our algorithm generates anonymized social networks with utility loss less than 1% on many important network statistics.

The remainder of the paper is organized as follows. Section 2 briefly reviews related works about social network data anonymization with focused on utility control mechanism used in anonymization. Section 3 provides the background knowledge of the problem to be solved. Section 4 discusses about our proposed metric to preserve utility in anonymization. We report the experiment results in Section 5. Finally, Section 6 concludes the paper.

2. Background

In this section, we will introduce background information. We will present the models for the social network and adversary. We also present the problem overview and the related work. Table 1 shows the parameters used throughout this paper.

2.1. Conceptual model

A social network is generally modeled as a graph consisting of a set of entities and the connections between them. As previous studies (Zhou and Pei, 2010, 2008; Liu and Terzi, 2008; Zou et al., 2009; Cheng et al., 2010; Xiaoyun et al., 2009; Wang et al., 2011; Masoumzadeh and Joshi, 2012), we model a social network data as a graph $G(V,E)$ where $V=\{v_1,v_2,\ldots,v_n\}$ is a set of n unlabeled vertices representing individuals in the social network data and $E=\{e_1,e_2,\ldots,e_m\}$ is a set of m unlabeled edges describing the relationship (e.g., friendship and collaboration) between the individuals. Fig. 1 shows an example of a social network graph. There are ten vertices $v=\{A,\ B,C,D,E,F,G,H,I,J\}$ and each vertex is connected to at least one other vertex.

Given an original social network graph $G(V,\ E)$, it is published version $\overline{G}(\overline{V},\ \overline{E})$ is generated by removing all the vertex identity information of G as well as possible structural modifications (e.g., edge and/or vertex insertion and/or deletion). How $\overline{G}(\overline{V},\ \overline{E})$ is derived from $G(V,\ E)$ will be discussed through out this paper. To demonstrate our utility-loss control technique, we employ degree-based anonymization model called k-degree anonymity (Liu and Terzi 2008).

Definition 1. (**Vertex degree):** *The degree of a vertex $v_i \in V$, denoted as $d(v_i)$, is the number of connections or edges v_i has to other vertices in G. For example, the $d(J)$ in Fig. 1 is three.*

Definition 2. (***k*-degree anonymity**)**:** *A graph $G(V,\ E)$ is said to be k-degree anonymous if for every vertex $v_i \in V$, there exist at least k other vertices in G with the same degree as v_i.*

As in (Watanabe et al., 2011), we assume that a query type for data manipulation operations applied to the published social network datasets. We define structural query as follows:

Definition 3. (**Structural Query):** *Given a social network $G(V,E)$ and a target individual $t \in V$, a query q over G based on the structural information about t in G returns a set of vertices $\mathcal{W} \subset V$ with cardinality $C \mid 0 \leq C \leq n$.*

In this paper, we consider an adversary $\mathcal{A}$ with *a priori* knowledge β of the degree of certain vertices in the social network graph $G(V,\ E)$. Specifically, an adversary $\mathcal{A}$ aims to identify some target individuals as vertices in the published social network graph $\overline{G}(\overline{V},\ \overline{E})$ by using structural query over the published social network data and the background knowledge (β). The structure-based vertex re-identification attack is formally defined as follows:

Definition 4. (***Re-identification Attack*** (Liu and Terzi, 2008))***:*** *Let $G(V,\ E)$ be an original social network graph and $\overline{G}(\overline{V},\ \overline{E})$ be the*

Table 1
Parameter description.

Symbol	Description
$G(V,E)$	Original social network graph with a set of V vertices and E edges.
$\overline{G}(\overline{V},\ \overline{E})$	Anonymized version of $G(V,E)$ such as $V=\overline{V}$ and $E\neq\overline{E}$.
$d(v)$	The degree of vertex v
$U(G)$	Utility of the original social network graph
$U(\overline{G})$	Utility of the anonymized social network graph
$\mathcal{A}$	An adversary
β	An adversary background knowledge
k	Anonymization value

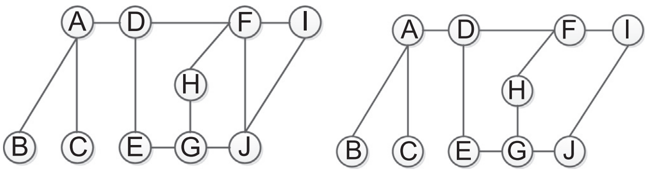

Fig. 1. Example of a social network graph $G(V,\ E)$ (left) and derived $\overline{G}(\overline{V},\ \overline{E})$ (right).