Available online at www.sciencedirect.com

ScienceDirect

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

ELSEVIER

Computers & Security

A graph-based multifold model for anonymizing data with attributes of multiple types

Li-E. Wang, Xianxian Li *

Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, PR China

ARTICLE INFO

Article history:
Received 20 February 2017
Received in revised form 15 June 2017
Accepted 8 September 2017
Available online 22 September 2017

Keywords:
Privacy protection
Data publishing
Transactional data
Uncertain graph
High-dimensional data
Anonymization

ABSTRACT

Transactional data with attributes of multiple types may be extremely useful to secondary analysis (e.g., learning models and finding patterns). However, anonymization of such data is challenging because it contains multiple types of attributes (e.g., relational and set-valued attributes). Existing privacy-preserving techniques are not applicable to address this problem. In this paper, we propose a novel graph-based multifold model to anonymize data with attributes of multiple types. Under this model, such data are modelled as a graph, and multifold privacy is guaranteed through fuzzing on sensitive attributes and converting associations among items into an uncertain form. Specifically, we define a multi-objective attack model in a graph and devise a safety parameter and algorithm to prevent such attacks. Experiments have been performed on real-life data sets to evaluate the performance.

1. Introduction

In general, transactional data are considered set-valued data, as such data consist of sets of items, for which the privacy-preserving model is usually the extensional k-anonymization model. However, transactional data are usually much richer in structure, involving objects of multiple types that are related to each other, such as customers and products in an e-commerce system. Every transaction of these datasets refers to attributes of multiple types, called transactional datasets with attributes of multiple types, *AMT-datasets* for short. *AMT-datasets* have attracted more and more attention in recent years due to their high impact on various important applications, such as recommending systems (Amatriain and Pujol, 2015; Chang et al., 2010; Cho et al., 2015), data mining (Amatriain and Pujol, 2015; Cho et al., 2015; Lee et al., 2013) and other research purposes. Because *AMT-dataset* contains multifold sensitive information, such as individuals' private information, the privacy of the correlated objects and their relations must be preserved. The anonymizing approaches for publishing such data need to thwart *sensitive attributes*, *association disclosures* and *identity disclosure*. This means that *AMT-datasets* present a great challenge for anonymizing techniques.

AMT-dataset enables us to associate single values (e.g., Age, Sex) and set values (e.g., Purchased-product, Disease) and analyze their relationships. Single values are called relational attributes, and set values are called transaction attributes (Poulis et al., 2013). To anonymize *AMT-datasets* for publishing, in general, the datasets would be divided into multiple parts

* *Corresponding author.*
E-mail address: lixx@gxnu.edu.cn (X. Li).

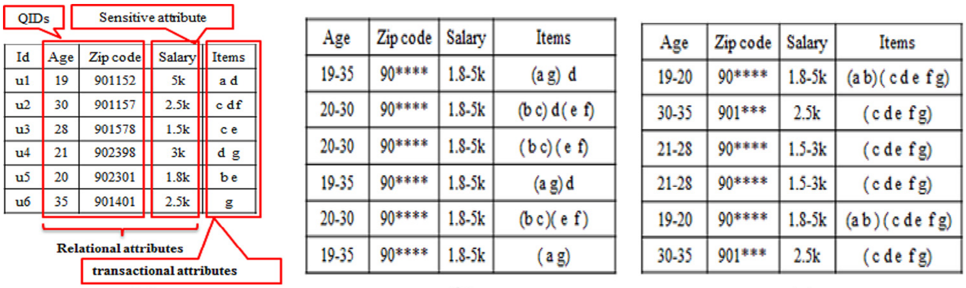
https://doi.org/10.1016/j.cose.2017.09.003

such that every part has a single type of data structure for using existing privacy-preserving techniques. There are two streams of relevant studies that target different types of privacy requirements. Most of the existing works focus on anonymizing based on relational attributes, such as k-anonymity (Samarati, 2001; Sweeney, 2002), l-diversity (Machanavajjhala et al., 2007), and t-closeness (Li et al., 2007), while other studies aim to anonymize based on transaction attributes which is considered set-valued data. Suppression and Generalization has been proposed as a way to address this problem. Suppression removes sensitive items directly to guarantee privacy (Xu et al., 2008), and Generalization maps the original items to generalized items (He and Naughton, 2009; Liu and Wang, 2010; Terrovitis et al., 2008, 2011). Due to its features of high dimensionality and sparsity, Suppression and Generalization results in considerable information loss. To address the problem, a straightforward approach is Bucketization and Perturbation. Bucketization operates by separating sensitive items from the QID (Ghinita et al., 2008, 2011), and Perturbation operates by adding or altering items from the individuals' transactions (Chen et al., 2009; Fung et al., 2010).

However, *AMT-datasets* cannot be independently anonymized using the two aforementioned principles and algorithms for two reasons. First, existing data anonymizations only consider a small number of possible transformations to anonymize a single type of data, and simple transformations will not be as efficient when sensitive information about an entity has many types. Second, the two types of data, relational attributes and transaction attributes, are anonymized and published separately, which will lose the correlations among different types of data items. Thus, Poulis et al. (2013) proposed (k, k^m)-anonymity for anonymizing data with relational and transaction attributes. They enforce (k, k^m)-anonymity to offer a privacy guarantee, with a bounded information loss in one attribute type and minimal information loss in the other. Takahashi et al. (2013) proposed an anonymization approach via integrating recordings for single-valued attributes and set-valued attributes into a whole top-down anonymization. Although these approaches (Poulis et al., 2013; Takahashi et al., 2013) are concerned with both relational and transaction attributes, they assume that the relational attributes are a *quasi-identifier* (QID) and only defend the privacy of the transaction attributes. However, the relational attributes in *AMT-datasets* also contain sensitive information, such as *occupation* and *salary*, which should be protected before publishing. As verified in Wang and Li (2014b), the transaction attributes of *AMT-datasets* involve sensitive and insensitive items, so the anonymizing approaches proposed in Poulis et al. (2013) and Takahashi et al. (2013) are not suitable for *AMT-datasets*. On the other hand, these applications for supporting secondary analysis (e.g., learning models and finding patterns) require acquiring the associations between the two attributes. Figuratively, examples include *at what age is susceptible to suffering from some type of disease* and *which areas tend to purchase certain products*. However, generalization for anonymizing data with attributes of multiple types, which is mainly adopted in Poulis et al. (2013) and Takahashi et al. (2013), is considered an ill-advised practice due to incurring excessive information loss, even making the data useless. To make the point clearer, we give an example as below.

Example 1. *Anonymizing on AMT-datasets*

In this paper, the *AMT-dataset* exemplified by e-commerce data consists of relational attributes (e.g., attribute information of customers, such as *age*, *zip code* and *salary*) and transaction attributes (e.g., market basket data), as shown in Fig. 1(a). When publishing these data, we must anonymize both the relational and transaction attributes while masking the relations between them, which make it more challenging. In this example, both *salary* and *items* are sensitive information, and the relations between them are also private. As analyzed above, the existing data anonymizations, which are used to preserve the privacy of datasets containing only relational attributes (Li et al., 2007; Machanavajjhala et al., 2007; Samarati, 2001; Sweeney, 2002) or only transaction attributes (Cao et al., 2010; Chen et al., 2009; Cormode et al., 2010; Fung et al., 2010; Ghinita et al., 2008, 2011; Gkoulalas-Divanis and Loukides, 2012; He and Naughton, 2009; Liu and Wang, 2010; Loukides et al., 2010, 2011, 2013; Terrovitis et al., 2008, 2011; Wang and Li, 2014a, 2014b; Xiao and Tao, 2006; Xu et al., 2008; Xue et al., 2012), are not enough. We attempt to apply existing approaches from Poulis et al. (2013) and Takahashi et al. (2013) to anonymize *AMT-datasets* in Fig. 1(b) and 1(c). The table in Fig. 1(b) is produced from the table in Fig. 1(a) by applying the method of T_{FIRST} in Poulis et al. (2013), which minimizes the information loss incurred by the generalization on transaction attributes with a bounded information loss in the other. Observe that all values of *zip code* and *salary* are replaced by the same generalized range. That is, the associations between the two attributes are lost. For example, using the table in Fig. 1(b), we will no longer be


QIDs
Sensitive attribute
Relational attributes
transactional attributes
(a)
(b)
(c)

Id	Age	Zip code	Salary	Items
u1	19	901152	5k	a d
u2	30	901157	2.5k	c df
u3	28	901578	1.5k	c e
u4	21	902398	3k	d g
u5	20	902301	1.8k	b e
u6	35	901401	2.5k	g

Age	Zip code	Salary	Items
19-35	90****	1.8-5k	(a g) d
20-30	90****	1.8-5k	(b c) d(e f)
20-30	90****	1.8-5k	(b c) (e f)
19-35	90****	1.8-5k	(a g) d
20-30	90****	1.8-5k	(b c)(e f)
19-35	90****	1.8-5k	(a g)

Age	Zip code	Salary	Items
19-20	90****	1.8-5k	(a b)(c d e f g)
30-35	901***	2.5k	(c d e f g)
21-28	90****	1.5-3k	(c d e f g)
21-28	90****	1.5-3k	(c d e f g)
19-20	90****	1.8-5k	(a b) (c d e f g)
30-35	901***	2.5k	(c d e f g)

Fig. 1 – Attempting to apply existing anonymization to *AMT-datasets*: (a) *AMT-dataset*; (b) (k, k^m)-anonymity by applying T_{FIRST} in Poulis et al. (2013); (c) Multi-dimensional k-anonymization in Takahashi et al. (2013).