Information Sciences 367–368 (2016) 324–336

Contents lists available at ScienceDirect

Information Sciences

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

A hybrid approach to prevent composition attacks for independent data releases

Jiuyong Li[a,*], Muzammil M. Baig[a], A.H.M. Sarowar Sattar[a], Xiaofeng Ding[a,b], Jixue Liu[a], Millist W. Vincent[a]

[a] School of Information Technology and Mathematical Science, University of South Australia, Mawson Lakes, SA-5095, Australia
[b] School of Computer Science and Technology, Huazhong University of Science and Technology, China

ARTICLE INFO

Article history:
Received 10 December 2014
Revised 2 October 2015
Accepted 11 May 2016
Available online 9 June 2016

Keywords:
Privacy
Anonymization
Composition attack

ABSTRACT

Data anonymization is one of the main techniques used in privacy preserving data publishing, and many methods have been proposed to anonymize both individual data sets and multiple data sets. In real life, a data set is rarely isolated and two data sets published by different organizations may contain records pertaining to the same individual. For example, some patients might have visited two hospitals for the same disease, and their records are independently anonymized and published by the two hospitals. Although each published data set alone might pose a small privacy risk, the combination of two data sets may severely compromise the privacy of the individuals common to both data sets. An attack on individual privacy which uses independent data sets is called a *composition attack*. The topic of how to anonymize data sets to prevent a composition attack using independent data releases has not been widely investigated. In this paper, we propose a new principle to protect data sets from composition attacks. We propose a hybrid algorithm, which combines sampling, perturbation and generalization to protect data privacy from composition attacks. We experimentally demonstrate that the proposed anonymization technique significantly reduces the risk of composition attacks and also preserves good data utility.

1. Introduction

Privacy is at risk in a process of data sharing, which is crucial for social, medical and various scientific research. In the past decades, several instances of data privacy breaches [2,30], due to sharing of data, have resulted in financial and reputation losses for enterprises. Privacy preserving data publishing techniques address this problem by anonymizing data such that individual privacy is preserved while allowing organizations to release/share anonymized data.

Anonymization is a major technique for data privacy in data publication. The idea behind the anonymization is one in the crowd, which ensures that an individual cannot be distinguished from a number of other people. Anonymization techniques have been widely discussed in literature and well known schemes include *generalization* [25,27,34,41] and *perturbation* [7,18,36]. Generalization coarsens the values of a set of attributes called *quasi-identifier* that potentially identify an individual, such as Gender, Age and Postcode, so that a group of individuals appear to have the same quasi-identifier values and they are indistinguishable in a published data set. Such a group is called an *equivalence class*, and an example of two equivalences is given in Table 1. Perturbation randomly adds noises to or swamp quasi-identifier values in a data set. It

* Corresponding author. Tel.: +61883023898.
E-mail address: Jiuyong.Li@unisa.edu.au (J. Li).

http://dx.doi.org/10.1016/j.ins.2016.05.009
0020-0255/

Table 1
An illustrative example of a composition attack.

G	A	PC	SV	G	A	PC	SV
m	21–25	5095	A	m	21–25	5095	C
m	21–25	5095	B	m	21–25	5095	D
m	21–25	5095	C	m	21–25	5095	E

(a) published data set 1 (b) published data set 2 G: Gender; A: Age; PC: Postcode; SV: Sensitive value.

reduces the confidence of an adversary to find the record of an individual based on the quasi-identifier values. Again, the higher level of generalization/perturbation, the better privacy protection, and normally the lower data utility.

Existing anonymization techniques mainly focus on one-time publication [27,34], where a data owner anonymizes a data set without considering other published data sets. In many cases, multiple views of a data set [44,45] are published or a series of data sets in different time stamps are published [38,40,42]. An example of the former is the publications of data with different generalization schemes for different purposes, and an example of the latter is quarterly publications of hospital data. Both examples are multiple data publications but are not independent data publications since the data sets are controlled by the same data owner. We will explain multiple independent data publications late. When the information of an individual resides in multiple data sets, an adversary may intersect a number of anonymized data sets to reveal the privacy of the individual, which is preserved in each single publication [15,38].

Multiple publications are vulnerable to a *composition attack*, which uses intersection of a number of published data sets to infer the sensitive values of individuals whose records are in multiple data sets. For example, suppose that a victim has the following personal information, (Gender = male, Age = 22, Postcode = 5095), known to an adversary. The adversary also knows that the victim's records are in two data sets. Table 1 lists data segments from two published data sets containing victim's records. In each data set, the adversary could not find the victim's sensitive information since there are three possible sensitive values. However, the intersection of two data segments contains only one sensitive value, thus the privacy of the victim is breached.

Multiple independent data publishing poses a new challenge for privacy preserving data publication. In multiple independent data publications, a data owner does not know which published data set will be used for a composition attack. Multiple independent data publication is different from normal multiple data publications, such as multiple view data publication [44,45] and series data publication [38,40,42], where a data publisher knows all data sets (different views or previous versions of the current data set) that can be used for composition attacks and can use information in the known data sets to anonymize the current data set. There is not any communication or information sharing between data owners in multiple independent data publications, and hence collaborative privacy preserving data publication techniques [19,20,28,43] could not be used for protecting the privacy in multiple independent data publications.

Multiple independent data publishing is a situation that occurs in practice. For example, a patient may visit two hospitals for the treatment of the same disease and the two hospitals then publish their data sets independently without co-ordination. The reasons for not co-ordination may be due to a regulation that hospitals are not allowed to share their raw data, or infeasibility because a hospital may share common patients with many other hospitals. An adversary who has background knowledge that the victim's records are in two published data sets can then conduct a composition attack. The adversary may obtain such background knowledge from many sources, such as personal acquaintance or information revealed in a social network.

A differential privacy based data randomization method can protect data from the composition attacks [15]. A randomized mechanism on a data set satisfies differential privacy if the removal or inclusion of a single record from the data set has only a small effect on the output of the randomization mechanism. When two data sets are processed with differential privacy with the privacy budgets ϵ_1 and ϵ_2 (the smaller a privacy budget, the higher privacy). A composition attack could not completely compromise the privacy of an individual since the privacy is ensured in the privacy budget $\epsilon_1 + \epsilon_2$.

In contrast, generalization methods alone do not protect data from composition attacks in multiple independent publications [4].

Although differential privacy can protect against a composition attack, a drawback with the technique is the loss of utility. The loss of utility in general cases has been observed in [9,31], and in the following we explain it in our case. Most of the work in differential privacy [10] is based on the interactive setting, whereby the user access a data set via a numerical query and the anonymization technique adds noise to the query result. However, the interactive setting is not always applicable since in some situations data sets are required to be published publicly, and so the differential privacy principle has been extended to non-interactive data access [29]. In this approach, values within the quasi-identifier are grouped into equivalence classes where all values in an equivalence class are the same, so individuals in an equivalence class are indistinguishable. To ensure good data utility, the size of such a group should not be large since otherwise detailed information about the quasi-identifier attributes is lost. Then, Laplacian noise is added to the count of every sensitive value in an each equivalence class. Since the size of an equivalence class is small, the counts of sensitive values in an equivalence class are also small. When the counts are small, the addition of Laplacian noise to achieve differential privacy results in low data utility.