Knowledge-Based Systems 67 (2014) 361–372

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments

A.H.M. Sarowar Sattar [a,*], Jiuyong Li [a,*], Jixue Liu [a], Raymond Heatherly [b], Bradley Malin [b,c]

[a] School of Information Technology and Mathematical Science, University of South Australia, Mawson Lakes SA-5095, Australia
[b] Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
[c] Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA

ARTICLE INFO

Article history:
Received 18 November 2013
Received in revised form 4 April 2014
Accepted 10 April 2014
Available online 24 April 2014

Keywords:
Databases
Data publication
Privacy
Composition attack
Anonymization

ABSTRACT

Organizations share data about individuals to drive business and comply with law and regulation. However, an adversary may expose confidential information by tracking an individual across disparate data publications using quasi-identifying attributes (e.g., age, geocode and sex) associated with the records. Various studies have shown that well-established privacy protection models (e.g., *k*-anonymity and its extensions) fail to protect an individual's privacy against this "composition attack". This type of attack can be thwarted when organizations coordinate prior to data publication, but such a practice is not always feasible. In this paper, we introduce a probabilistic model called (d, α)-linkable, which mitigates composition attack without coordination. The model ensures that d confidential values are associated with a quasi-identifying group with a likelihood of α. We realize this model through an efficient extension to *k*-anonymization and use extensive experiments to show our strategy significantly reduces the likelihood of a successful composition attack and can preserve more utility than alternative privacy models, such as differential privacy.

1. Introduction

The increasing collection of large quantities of person-specific information has created tremendous opportunities for knowledge-based decision making in a number of domains [24]. To fully maximize the knowledge that can be learned, the data needs to be made available beyond the organizations that performed the initial collection [13]. Data sharing, however, must be accomplished in a manner that respects the privacy of the individuals from which the data was gathered [31]. There are a wide variety of computational protection models that have been suggested [7], but, the majority fail to consider situations in which records on the same individual occur in multiple organizations' data sets. This is a concern because, in certain environments, an individual's information will be collected and published by disparate organizations [20]. And when such a situation arises, an adversary may invoke a *composition attack* [9] on the published data sets to compromise the privacy afforded by traditional protection models.

The composition attack will be formalized below, but here, we take a moment to illustrate how such a problem transpires to provide context. Imagine two healthcare organizations, Organization-A and Organization-B, collect demographics and confidential health information as shown in Table 1(a) and (b), respectively. Notice that Alice, a 22 year-old female living in ZIP code 5095, was diagnosed with 'Diabetes' at both organizations. The organizations choose to publish versions of their data sets as depicted in Table 2(a) and (b). These adhere to a traditional formal privacy model called *k*-anonymity, which classifies attributes as explicit identifiers (that is, information that allows for direct communication with an individual such as *name* and *Social Security Number*), quasi-identifiers (that is, in combination, can uniquely characterize an individual and be leveraged for identification purposes, such as *age*, *sex*, and *ZIP code* of residence) and confidential attributes (e.g., diagnoses). In a published data set, explicit identifiers are suppressed, quasi-identifiers are masked, and confidential attributes are retained in their original form. To mask quasi-identifiers, their values are often generalized to less specific concepts. Alice's information, for instance, has been generalized to an age range of 15–25 and a ZIP code of 50** in one table and 10–30 and 50** in the other table. Yet, when an adversary knows that Alice visited both institutions, they may learn her health status because there is only one common confidential value in the sets of records that could possibly correspond to Alice.

* Corresponding authors. Tel.: +61 420863356 (A.H.M.S. Sattar).
E-mail addresses: satay003@mymail.unisa.edu.au (A.H.M.S. Sattar), Jiuyong.Li@unisa.edu.au (J. Li).

http://dx.doi.org/10.1016/j.knosys.2014.04.019
0950-7051/

Table 1
The data managed by (a) Organization-A (b) Organization-B in their private collections.

Name	Age	Sex	ZIP code	Diagnosis
(a)				
Emu	25	M	5095	Cough
Alex	24	M	5085	Flu
Clark	20	M	5001	Diabetes
Hafiz	23	M	5005	Flu
Alice	22	F	5095	Diabetes
Mina	25	F	5001	Fever
Sofia	20	F	5002	Diabetes
Anju	21	F	5087	Fever
(b)				
Emu	25	M	5095	Cough
Michel	24	M	5085	Fever
Bokul	20	M	5031	Diabetes
Safiq	23	M	5025	Flu
Alice	22	F	5095	Diabetes
Lima	25	F	5065	Cough
Nima	20	F	5002	Diabetes
Fa mi	21	F	5077	Cough

Table 2
Publications of data sets from (a) Organization-A (b) Organization-B after the application of *k*-anonymization.

Age	Sex	ZIP code	Diagnosis
(a)			
15–25	M	50**	Flu
15–25	M	50**	Flu
15–25	M	50**	Cough
15–25	M	50**	Diabetes
15–25	F	50**	Fever
15–25	F	50**	Fever
15–25	F	50**	Diabetes
15–25	F	50**	Diabetes
(b)			
10–30	M	50**	Cough
10–30	M	50**	Fever
10–30	M	50**	Diabetes
10–30	M	50**	Flu
10–30	*	50**	Diabetes
10–30	*	50**	Diabetes
10–30	*	50**	Cough
10–30	*	50**	Cough

The *composition attack* can be thwarted when organizations coordinate during the *k*-anonymization process. Specifically, such coordination can take place by sharing their data sets in the clear (e.g., [18]) or computing over encrypted transformations (e.g., [11,19]) prior to publication to discover and address potential violations. However, such coordination is not always possible and may even be prohibited by law [2]. Moreover, in some countries, such as the United States, healthcare is decentralized. As a result, it is not uncommon for a patient to be seen at multiple hospitals that do not coordinate with one another as discussed in [20]. We refer to this setting as a *non-coordinated environment*.

In non-coordinated environments, privacy enhancing methods based on randomization can be applied to limit the detection of an individual in a data set. In particular, differential privacy (which is discussed in further detail in the following section), can prevent the composition attack [22]. However, the utility of such data sets may be too low for practical [22]. Tables 2,3 depict examples of *k*-anonymized and differentially private data sets, respectively.[1] The composition attack is successful for the pair of *k*-anonymized data sets because an adversary can restrict their focus to only one confidential value, namely 'Diabetes', to link with Alice's record. In contrast, for the pair of differentially private data sets, the adversary has all values from the confidential attribute's domain to link with Alice's record.[2]

[1] The hypothesized tables in Table 3 have been created by following the differential privacy mechanism in [22].

Table 3
Published data sets from (a) Organization-A (b) Organization-B after the application of a differential privacy mechanism.

Age	Sex	ZIP code	Diagnosis
(a)			
15–25	M	50**	Cough (5) Fever (10) Flu (O) Diabetes (8) Hepatitis (4) Heart disease (6)
15–25	F	50**	Cough (7) Fever (4) Flu (10) Diabetes (5) Hepatitis (ll) Heart disease (5)
(b)			
10–30	M	50**	Cough (3) Fever (7) Flu (12) Diabetes (0) Hepatitis (6) Heart disease (2)
10–30	*	50**	Cough (9) Fever (4) Flu (7) Diabetes (5) Hepatitis (8) Heart disease (2)

Thus, the goal of our current work is to develop generalization-based strategies to protect individuals' privacy whose records are disclosed by disparate organizations when coordination is not permitted. We propose a protection model that is designed to increase the likelihood that an adversary will have multiple confidential values to link with an individual's record after combining disparate *k*-anonymized data sets. Specifically, the contributions of this paper are as follows.

- First, we propose a novel model to reduce the risk of the composition attack. Our model is applicable to each publisher's data set independently and without coordination. This model uses statistical information regarding the quasi-identifying and confidential attributes of the underlying population to simulate a *k*-anonymized data set published by another organization.
- Second, we design an efficient algorithm to achieve the proposed protection model. The algorithm is implemented as a post-processing method applied to partition-based *k*-anonymization [16,17,30] approaches.[3] Note that in the publications from different independent organizations, the privacy of the records is preserved by the *k*-anonymity model. Thus, when applying the post-processing method on top of *k*-anonymization, we retain this privacy guarantee.
- Third, we provide an extensive empirical evaluation of our method on publicly avaialble data sets from the U.S. Census Bureau. We compare our method with a strategy based upon differential privacy [6] and show that our method can preserve better utility, with a negligible effect on data quality.

[2] In this example there are six confidential values {'Cough', 'Diabetes', 'Flu', 'Fever', 'Hepatitis', 'Heart disease'} in the confidential attribute's domain.

[3] We acknowledge that even though there are many of *k*-anonymity methods based on generalization, there are other methods like microaggregation, that replace equivalence classes by averaged values [4,21,5,1]. However, in this paper, we only consider generalization because our algorithm is implemented as a post-processing step for partition-based anonymization.