Information Sciences 298 (2015) 344–372

Contents lists available at ScienceDirect

Information Sciences

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

Privacy by diversity in sequential releases of databases

Erez Shmueli [a,*], Tamir Tassa [b]

[a] Department of Industrial Engineering, Tel Aviv University, Tel Aviv, Israel
[b] Department of Mathematics and Computer Science, The Open University, Ra'anana, Israel

ARTICLE INFO

Article history:
Received 30 September 2013
Received in revised form 2 October 2014
Accepted 1 November 2014
Available online 11 November 2014

Keywords:
Privacy preserving data publishing
Anonymization
Diversity
Sequential release
Continuous data publishing
Multipartite graphs

ABSTRACT

We study the problem of privacy preservation in sequential releases of databases. In that scenario, several releases of the same table are published over a period of time, where each release contains a different set of the table attributes, as dictated by the purposes of the release. The goal is to protect the private information from adversaries who examine the entire sequential release. That scenario was studied in [32] and was further investigated in [28]. We revisit their privacy definitions, and suggest a significantly stronger adversarial assumption and privacy definition. We then present a sequential anonymization algorithm that achieves ℓ-diversity. The algorithm exploits the fact that different releases may include different attributes in order to reduce the information loss that the anonymization entails. Unlike the previous algorithms, ours is perfectly scalable as the runtime to compute the anonymization of each release is independent of the number of previous releases. In addition, we consider here the fully dynamic setting in which the different releases differ in the set of attributes as well as in the set of tuples. The advantages of our approach are demonstrated by extensive experimentation.

1. Introduction

1.1. Overview

Large organizations regularly collect personal data, such as medical records, marketing information, or census data, in order to perform on it data mining for the purpose of revealing trends and patterns in the general population. However, the use of data containing personal information has to be restricted in order to protect individual privacy. Although identifying attributes like ID numbers and names are never released for data mining purposes, sensitive information might still leak due to linking attacks, whereby an attacker may uncover hidden identities or sensitive information by joining the released data attributes with other publicly available data. The attributes that can be efficiently used to create such links, such as gender, zipcode, and age, are called quasi-identifiers. The problem of anonymity calls for the modification of those attributes in order to thwart such attacks, while maintaining as much as possible of the utility of the released data.

The first model of privacy-preserving data publication was k-anonymity [27,29]. That model suggests to generalize the values of the quasi-identifiers so that each of the released records becomes indistinguishable from at least $k-1$ other records, when projected on those attributes. As a consequence, each individual may be linked to sets of records of size at least k in the released anonymized table, whence privacy is protected to some extent. While k-anonymity aims at preventing identity disclosure, the later models of ℓ-diversity [20,34] and t-closeness [19] aim at preventing sensitive attribute

* Corresponding author.

http://dx.doi.org/10.1016/j.ins.2014.11.005
0020-0255/

Table 1
A table (a) and two corresponding releases (b) and (c).

	name	age	gender	disease
(a)	Alice	20	female	measles
	Bob	20	male	hepatitis
	Carol	30	female	flu
	David	30	male	angina

	age	disease
(b)	20	measles
	20	hepatitis
	30	flu
	30	angina

	gender	disease
(c)	female	measles
	male	hepatitis
	female	flu
	male	angina

disclosure by imposing conditions on the distribution of the sensitive values within each subset of records that are indistinguishable with respect to their quasi-identifiers.

In all those models, the values of the database are typically modified via the operation of generalization, while keeping them consistent with the original ones. A cost function is used to measure the amount of information that is lost by the generalization process. The objective is to modify the table entries in order to respect the underlying privacy condition while minimizing the information loss.

Most of the studies thus far concentrated on scenarios of a single release, in which the underlying table is released just once in an anonymized manner. However, there are scenarios in which the same table is released more than once, for example, when new records are added to the table, or when partial views of the data have to be released to different clients. In such scenarios, it is imperative to consider the potential threats that may be caused by joining information from the different views.

Example 1.1. Consider the table with two quasi-identifiers, `age` and `gender`, and one sensitive attribute, `disease`, that is given in Table 1, alongside with two releases of it. The first release includes the `age` and `disease` attributes, while the second release includes the `gender` and `disease` attributes. Each of the two releases satisfies 2-diversity, since it can be used to link each individual to two sensitive values with equal probabilities. For instance, if an adversary wishes to find sensitive information about Alice, a female of age 20, then the first release allows him to infer that she has either measles or hepatitis, while the second release reveals that she has either measles or flu. However, the combination of the two releases discloses with certainty that Alice has measles. Therefore, in order to achieve 2-diversity, one of the two releases would have to be further generalized.

The two main scenarios of multiple releases of a given table are the following (see [9]): (a) The scenario of sequential release publishing [28,32], where different (vertical) projections of a given table on different subsets of attributes are released in a sequential manner; and (b) Continuous data publishing [1,3,10,25,28,35,37], in which the underlying table changes over time (e.g., tuples are added, removed, or updated), and updated snapshots of the table are released over time.

In both scenarios, several releases of partial views of the same basic table are published in a sequential manner, where already published releases cannot be modified. The goal is to anonymize the next release so that the combination of information from all releases does not lead to a privacy breach. In the scenario which is called above "sequential release publishing", the set of tuples (rows) is fixed, while the set of attributes (columns) changes from one release to another. On the other hand, in the scenario of so called "continuous data publishing", the set of attributes is fixed while the set of tuples is dynamic.

In the lion's part of this paper we study the scenario of sequential release publishing (in which the set of attributes changes between releases, but the set of tuples is fixed). The approach that we propose here differs significantly from the one proposed in [32] and then further developed in [28]. Then, in Section 5, we explain how to extend our approach to handle also the case of dynamically changing tables, where tuples can be added from time to time; in such cases, the set of attributes as well as the set of tuples may change from one release to the next one.

1.2. Related work on privacy-preservation in sequential release publishing

Wang and Fung [32] were the first to study the privacy problem in sequential releases and they developed an algorithm for anonymizing such a release. They focused on the case in which only one previous release of the underlying table was published, and then the problem is to generalize a second release so that a given privacy goal is met. They (as well as the