Data & Knowledge Engineering xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

# Hiding outliers into crowd: Privacy-preserving data publishing with outliers

Hui (Wendy) Wang *, Ruilin Liu

*Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA*

ARTICLE INFO

*Article history:*
Received 16 August 2013
Received in revised form 8 May 2015
Accepted 30 June 2015
Available online xxxx

*Keywords:*
Security
Integrity and protection
Data sharing
Data anonymization
Outliers

ABSTRACT

In recent years, many organizations publish their data in non-aggregated format for research purpose. However, publishing non-aggregated data raises serious concerns in data privacy. One of the concerns is that when outliers exist in the dataset, they are easier to be distinguished from the crowd and their privacy is prone to be compromised. In this paper, we study the problem of privacy-preserving publishing datasets that contain outliers. We define the *distinguishability-based attack* by which the adversary can identify outliers and reveal their private information from an anonymized dataset. We show that the existing syntactic privacy models (e.g., *k*-anonymity and $\ell$-diversity) cannot defend against the distinguishability-based attack. We define the *plain $\ell$-diversity* to provide privacy guarantee to outliers against the distinguishability-based attack, and design efficient algorithms to anonymize the dataset to achieve plain $\ell$-diversity with low information loss. We extend our anonymization approach to deal with continuous release of a series of datasets that contain outliers. Our experiments demonstrate the efficiency and effectiveness of our approaches.



## 1. Introduction

Recent years have witnessed increasing volume of released microdata (i.e., data in raw, non-aggregated format) for ad-hoc analysis in a variety of domains. However, releasing data to the public raises serious concerns of revealing private information of individuals. To protect privacy of individuals, the datasets must be anonymized before being released to the public. Previous study [1] has shown that simply removing explicit identifiers, e.g., name and social security number (SSN), from the microdata, is insufficient to protect privacy. This is because the *quasi-identifiers* (QI-attributes), e.g., zipcode, gender and date of birth, can jointly identify individuals uniquely. Therefore, the identity of individuals as well as their sensitive information can be easily revealed if the released dataset is joined with any external public dataset (e.g., voting registration lists) that contains the QI-attributes.

In general, there are two types of privacy: *attribute privacy* and *row privacy* [2]. Attribute privacy protects a number of "sensitive" attributes even when the other attributes are known to an attacker, while row privacy protects an entire row of the database even given leakage of other rows. An extensive body of research has been developed to protect both types of privacy. The attribute privacy can be protected via the notion of *k-anonymity* [3,1] and its variants [4–6], while the row privacy can be protected by the well-known *differential privacy* [7,8] and its offsprings [9,10]. Informally, *k*-anonymity requires that when certain attributes, known as quasi-identifiers (*QIDs*), are accessible to the attacker, each individual is not identifiable from a group of at least *k* individuals according to his/her *QIDs* [3]. Hence the probability of associating any individual with his/her sensitive values is no larger than 1/*k*. Differential privacy tries to ensure that the removal or addition of any individual record in the database will not affect the final output of the

* Corresponding author.
*E-mail addresses:* Hui.Wang@stevens.edu (H.(W.) Wang), rliu3@stevens.edu (R. Liu).

http://dx.doi.org/10.1016/j.datak.2015.06.012
0169-023X/

Please cite this article as: H.(W.) Wang, R. Liu, Hiding outliers into crowd: Privacy-preserving data publishing with outliers, Data Knowl. Eng. (2015), http://dx.doi.org/10.1016/j.datak.2015.06.012

query significantly. In the literature, *k*-anonymity and differential privacy have been considered as two different privacy models: *k*-anonymity provides *syntactic* privacy guarantee, while differential privacy provides *semantic* privacy protection.

Many application domains naturally produce correlated data. In general, there may exist data correlations in many real-world datasets. The correlations can be classified into two main categories:

- *Attribute correlations* that exist on two or more attributes. A typical example of attribute correlations is the *functional dependency* that catches the relationship between attributes. Informally, a functional dependency occurs when a set of attributes uniquely determine another set of attributes. For example, the attribute *Zipcode* determines the attribute *AddressCity*, meaning that all tuples of the same zipcode values must have the same *AddressCity* values.
- *Row correlations* that exist among rows. Row correlations include row closeness (e.g., the rows are involved in similar patterns) and row outlierness (e.g., the rows that do not comply with the patterns of the majority). An example of outliers is the multi-billionaires whose income is much higher than the U.S. average.

In this paper, we mainly consider *row outlierness* as the data correlations. As an example, consider the microdata in Table 1 (a). It contains two outliers: *Bill* whose income is much higher than the whole population in the dataset, and *Justin* who earns much more

**Table 1**
An Example of base table and bad anonymization.

(a) Base table $D$

| ID | Quasi-identifiers (QI) | | | Sensitive |
|---|---|---|---|---|
| Name | Age | Sex | Zipcode | Income |
| Alice | 20 | F | 06006 | 20 K |
| Bob | 20 | M | 06011 | 25 K |
| **Justin** | **20** | **M** | **06013** | **1 M** |
| Carol | 30 | F | 06001 | 30 K |
| Allan | 30 | M | 06010 | 50 K |
| **Bill** | **30** | **M** | **06022** | **2B** |
| Ben | 40 | M | 06004 | 1.1 M |
| Susan | 40 | F | 06002 | 1.2 M |
| David | 40 | M | 06003 | 1.3 M |

(b) A bad anonymization scheme $D_1^*$ that cannot hide Justin's record

| Age | Sex | Zipcode | Income |
|---|---|---|---|
| 20 | * | [06006, 06013] | 20 K |
| 20 | * | [06006, 06013] | 25 K |
| **20** | * | [06006, 06013] | **1 M** |
| 30 | * | [06001, 06022] | 30 K |
| 30 | * | [06001, 06022] | 50 K |
| **30** | * | [06001, 06022] | **2B** |
| 40 | * | [06002, 06004] | 1.1 M |
| 40 | * | [06002, 06004] | 1.2 M |
| 40 | * | [06002, 06004] | 1.3 M |

(c) A bad anonymization scheme $D_2^*$ that cannot hide Justin's record

| Age | Sex | Zipcode | Income |
|---|---|---|---|
| [20, 30] | * | [06001, 06011] | 20 K |
| [20, 30] | * | [06001, 06011] | 25 K |
| [20, 30] | * | [06001, 06011] | 30 K |
| [20, 30] | * | [06001, 06011] | 50 K |
| [20, 40] | * | [06002, 06013] | 1 M |
| [20, 40] | * | [06002, 06013] | 1.1 M |
| [20, 40] | * | [06002, 06013] | 1.2 M |
| [20, 40] | * | [06002, 06013] | 1.3 M |

(d) A good anonymization scheme $D_3^*$ that can hide Justin's record

| Age | Gender | Zipcode | Income |
|---|---|---|---|
| [20, 40] | * | [06002, 06010] | 20 K |
| [20, 40] | * | [06002, 06010] | 50 K |
| [20, 40] | * | [06002, 06010] | 1.2 M |
| [20, 40] | * | [06002, 06010] | 1.3 M |
| [20, 40] | * | [06001, 06013] | 25 K |
| [20, 40] | * | [06001, 06013] | 30 K |
| [20, 40] | * | [06001, 06013] | **1 M** |
| [20, 40] | * | [06001, 06013] | 1.1 M |

Please cite this article as: H.(W.) Wang, R. Liu, Hiding outliers into crowd: Privacy-preserving data publishing with outliers, Data Knowl. Eng. (2015), http://dx.doi.org/10.1016/j.datak.2015.06.012