Future Generation Computer Systems 27 (2011) 348–356

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A family of enhanced (L, α)-diversity models for privacy preserving data publishing

Xiaoxun Sun [a,*], Min Li [b], Hua Wang [b]

[a] Australian Council for Educational Research, Australia

[b] Department of Mathematics & Computing, University of Southern Queensland, Australia

ARTICLE INFO

Article history:
Received 22 December 2009
Received in revised form
9 May 2010
Accepted 18 July 2010
Available online 24 July 2010

ABSTRACT

Privacy preservation is an important issue in the release of data for mining purposes. Recently, a novel l-diversity privacy model was proposed. However, even an l-diverse data set may have some severe problems leading to the revelation of individual sensitive information. In this paper, we remedy the problem by introducing distinct (l, α)-diversity, which, intuitively, demands that the total weight of the sensitive values in a given QI-group is at least α, where the weight is controlled by a pre-defined recursive metric system. We provide a thorough analysis of the distinct (l, α)-diversity and prove that the optimal distinct (l, α)-diversity problem with its two variants entropy (l, α)-diversity and recursive (c, l, α)-diversity are NP-hard, and propose a top-down anonymization approach to solve the distinct (l, α)-diversity problem with its variants. We show in the extensive experimental evaluations that the proposed methods are practical in terms of utility measurements and can be implemented efficiently.

1. Introduction

Many data holders publish their microdata for different purposes. However, they have difficulties in releasing information such that no privacy is compromised. The traditional approach of releasing the data tables without breaching the privacy of individuals in the table is to de-identify records by removing the identifying fields such as name, address, and social security number. However, joining this de-identified table with a publicly available database (like the voters database) on attributes like race, age, and Zipcode (usually called quasi-identifiers) can be used to identify individuals. For example, Sweeney reported in [1] that 87% of the population of the United States can be uniquely identified by the combinations of attributes: gender, date of birth, and 5-digit Zipcode.

In order to protect privacy, Sweeney [1] proposed the k-anonymity model, where some of the quasi-identifier fields are suppressed or generalized so that, for each record in the modified table, there are at least $(k - 1)$ other records in the modified table that are identical to it along the quasi-identifier attributes. In the literature of k-anonymity problem, there are two main models. One model is global recoding [2–5] while the other is local recoding [6,5]. Here, we assume that each attribute has a corresponding conceptual generalization hierarchy or taxonomy tree. A lower level domain in the hierarchy provides more details than a higher level domain. For example, Zipcode 14248 is a lower level domain and Zipcode 142$**$ is a higher level domain. We assume such hierarchies for numerical attributes too. In particular, we have a hierarchical structure defined with {value, interval, $*$}, where value is the raw numerical data, interval is the range of the raw data and $*$ is a symbol representing any values. Generalization replaces lower level domain values with higher level domain values. For example, Ages 27 and 28 in the lower level can be replaced by the interval (27–28) in the higher level.

1.1. Motivation

When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosures have been identified in the literature [7,8]: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure happens when the new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before releasing the data. Although k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. Several models such as p-sensitive k-anonymity [9], l-diversity [10] and t-closeness [11] were proposed. However, depending on the nature of the sensitive attributes, even these enhanced properties still permit the information to be disclosed.

p-sensitive k-anonymity principle: The purpose of p-sensitive k-anonymity is to protect against attribute disclosure by requiring that there be at least p different values for each sensitive attribute within the records sharing a combination of quasi-identifiers.

[*] Corresponding author.
E-mail addresses: sun@acer.edu.au (X. Sun), limin@usq.edu.au (M. Li), wang@usq.edu.au (H. Wang).

0167-739X/$ – see front matter Crown Copyright © 2010 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2010.07.007

This approach has the limitation of implicitly assuming that each sensitive attribute takes values uniformly over its domain; that is, the frequencies of the various values of a sensitive attribute are similar. When this is not the case, achieving the required level of privacy may cause a huge data utility loss.

l-diversity principle: The l-diversity model protects against sensitive attribute disclosure by considering the distribution of the attributes. The approach requires l "well-represented"[1] values in each combination of quasi-identifiers. This may be difficult to achieve and, like p-sensitive k-anonymity, may result in a large data utility loss. Further, as we shall discuss in Section 2, l-diversity is insufficient to prevent similarity attack.

t-closeness principle: The t-closeness model protects against sensitive attributes disclosure by defining semantic distance among sensitive attributes. The approach requires the distance between the distribution of the sensitive attribute in the group and the distribution of the attribute in the whole data set to be no more than a threshold t. Whereas Li et al. [11] elaborate on several ways to check t-closeness, no computational procedure to enforce this property is given. If such a procedure was available, it would greatly damage the utility of data because enforcing t-closeness destroys the correlations between quasi-identifier attributes and sensitive attributes.

Faced with these limitations, we intend to enhance the current privacy paradigms to make them preserve the better tradeoff between data quality and privacy. The work presented in this paper is greatly inspired by [10]. The main contribution of [10] is to introduce the basic l-diversity property, which provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary. In this paper, we propose a family of enhanced (l, α)-diversity models, where l is an integer and α is a real number. In addition to l-diversity, we further require that the total weight of sensitive values in any QI-group should be at least α after modification. We also propose an efficient anonymization method to tackle our problems.

2. Preliminaries

Let T be the initial microdata and T' be the released microdata. T' consists of a set of tuples over an attribute set. The attributes characterizing microdata are classified into the following three categories.

- *Identifier attributes* can be used to identify a record such as Name and Medicare card.
- *Quasi-identifier (QI) attributes* may be known by an intruder, such as Zipcode and Age. QI attributes are presented in the released microdata T' as well as in T.
- *Sensitive attributes* are assumed to be unknown to an intruder and need to be protected, such as Disease or ICD-9 Code.[2] Sensitive attributes are presented both in T and T'.

In what follows we assume that the identifier attributes have been removed and the quasi-identifier and sensitive attributes are usually kept in the released and initial microdata table. Another assumption is that the values of the sensitive attributes are not available from any external source. This assumption guarantees that an intruder cannot use the sensitive attributes to increase the chances of disclosure. Unfortunately, an intruder may use record linkage techniques [12] between quasi-identifier attributes and external available information to glean the identity of individuals from the modified microdata. To avoid this possibility of privacy disclosure, one frequently used solution is to modify the initial microdata, more specifically the quasi-identifier attributes values (a minimal set Q of attributes in T that can be joined with external information to re-identify individual records), in order to enforce the k-anonymity property.

Table 1
The raw microdata.

ID	Age	Country	Zipcode	Disease
1	27	USA	14248	HIV
2	28	Canada	14207	HIV
3	26	USA	14206	Cancer
4	25	Canada	14249	Cancer
5	41	China	13053	Hepatitis
6	48	Japan	13074	Phthisis
7	45	India	13064	Asthma
8	42	India	13062	Obesity
9	33	USA	14242	Flu
10	37	Canada	14204	Flu
11	36	Canada	14205	Flu
12	35	USA	14248	Indigestion

Table 2
2-anonymous microdata.

ID	Age	Country	Zipcode	Disease
1	(27–28)	America	142∗∗	HIV
2	(27–28)	America	142∗∗	HIV
3	(25–26)	America	142∗∗	Cancer
4	(25–26)	America	142∗∗	Cancer
5	>40	Asia	130∗∗	Hepatitis
6	>40	Asia	130∗∗	Phthisis
7	>40	Asia	130∗∗	Asthma
8	>40	Asia	130∗∗	Obesity
9	(33–35)	America	142∗∗	Flu
10	(36–37)	America	142∗∗	Flu
11	(36–37)	America	142∗∗	Flu
12	(33–35)	America	142∗∗	Indigestion

Definition 1 (*k-Anonymity*)**.** T' is said to satisfy k-anonymity if and only if each combination of quasi-identifier attributes in T' occurs at least k times.

A QI-group in the modified microdata T' is the set of all records in the table containing identical values for the QI attributes. There is no consensus in the literature over the term used to denote a QI-group. This term was not defined when k-anonymity was introduced [4,1]. More recent papers use different terminologies such as equivalence class [13] and QI-cluster [14].

For example, let the set {Age, Country, Zipcode} be the quasi-identifier of Table 1. Table 2 is one 2-anonymous view of Table 1 since there are five QI-groups and the size of each QI-group is at least 2. So k-anonymity can ensure that even though an intruder knows that a particular individual is in the k-anonymous microdata table T, she/he cannot infer which record in T corresponds to the individual with a probability greater than $1/k$.

The k-anonymity property ensures protection against identity disclosure, i.e. the identification of an entity (person, institution). However, as we will show next, it does not protect the data against attribute disclosure, which occurs when the intruder finds something new about a target entity. Consider Table 2, where the set of quasi-identifiers is composed of {Age, Country, Zipcode} and Disease is the sensitive attribute. As we discussed above, identity disclosure does not happen in this modified microdata. However, assuming that external information in Table 3 is available, attribute disclosure can take place. If the intruder knows that in Table 2 the Age attribute was modified to '(25–26)', she/he can deduce that both Rick and Rudy have Cancer, even he does not know which record, 3 or 4, is corresponding to which person. This example shows that even if k-anonymity can protect identity disclosure,

[1] The interpretation of the term "well-represented" can be found in [10].

[2] International Statistical Classification of Diseases and Related Health Problems: ICD-9 — provides multiple external links for looking up ICD codes. Available at http://icd9~cm.chrisendres.com/.