Computer Communications 34 (2011) 526–535

Contents lists available at ScienceDirect

Computer Communications

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

Extended *k*-anonymity models against sensitive attribute disclosure

Xiaoxun Sun *, Lili Sun, Hua Wang

Department of Mathematics & Computing, University of Southern Queensland, Australia

ARTICLE INFO

Article history:
Received 30 November 2009
Received in revised form 25 February 2010
Accepted 18 March 2010
Available online 27 March 2010

Keywords:
k-Anonymity
NP-hard
Attribute disclosure
Algorithm
Experiments

ABSTRACT

p-Sensitive *k*-anonymity model has been recently defined as a sophistication of *k*-anonymity. This new property requires that there be at least *p* distinct values for each sensitive attribute within the records sharing a set of quasi-identifier attributes. In this paper, we identify the situations when the *p*-sensitive *k*-anonymity property is not enough for the sensitive attributes protection. To overcome the shortcoming of the *p*-sensitive *k*-anonymity principle, we propose two new enhanced privacy requirements, namely p^+-sensitive *k*-anonymity and (p,α)-sensitive *k*-anonymity properties. These two new introduced models target at different perspectives. Instead of focusing on the specific values of sensitive attributes, p^+-sensitive *k*-anonymity model concerns more about the categories that the values belong to. Although (p,α)-sensitive *k*-anonymity model still put the point on the specific values, it includes an ordinal metric system to measure how much the specific sensitive attribute values contribute to each QI-group. We make a thorough theoretical analysis of hardness in computing the data set that satisfies either p^+-sensitive *k*-anonymity or (p,α)-sensitive *k*-anonymity. We devise a set of algorithms using the idea of top-down specification, which is clearly illustrated in the paper. We implement our algorithms on two real-world data sets and show in the comprehensive experimental evaluations that the two new introduced models are superior to the previous method in terms of effectiveness and efficiency.

1. Introduction

Agencies and other organizations often need to publish microdata, e.g. medical data or census data, for research and other purposes. However, if individuals can be uniquely identified in the microdata then their private information (such as their medical condition) would be disclosed, and this is unacceptable. To avoid the identification of records in microdata, uniquely identifying information like names and social security numbers are removed from the table. However, this traditional method still does not ensure the privacy of individuals in the data. A recent study estimated that 87% of the population of the United States can be uniquely identified by "linking attack" using the seemingly innocuous attributes gender, date of birth, and 5-digit zip code [28]. To avoid linking attacks, Samarati and Sweeney [23,30] proposed a definition of privacy called *k*-anonymity. A table satisfies *k*-anonymity if every record in the table is indistinguishable from at least $k-1$ other records with respect to every set of quasi-identifier attributes; such a table is called a *k*-anonymous table. This ensures that individuals cannot be uniquely identified by linking attacks. For example, Table 2 is a 2-anonymous view of Table 1. The sensitive attributes (Disease) is retained without change in this example.

In recent years, numerous algorithms have been proposed for implementing *k*-anonymity via generalization and suppression. Samarati [23] presents an algorithm that exploits a binary search on the domain generalization hierarchy to find minimal *k*-anonymous table. Sun et al. [24] recently improve his algorithm by integrating the hash-based technique. Bayardo and Agrawal [3] presents an optimal algorithm that starts from a fully generalized table and specializes the dataset in a minimal *k*-anonymous table, exploiting ad hoc pruning techniques. LeFevre et al. [13] describes an algorithm that uses a bottom-up technique and a priori computation. Fung et al. [7] present a top-down heuristic to make a table to be released *k*-anonymous. As to the theoretical results, Meyerson and Williams [20] and Aggarwal et al. [1,2] proved the optimal *k*-anonymity is NP-hard (based on the number of cells and number of attributes that are generalized and suppressed) and describe approximation algorithms for optimal *k*-anonymity. Sun et al. [25] proved that *k*-anonymity problem is also NP-hard even in the restricted cases, which could imply the results in [1,2,20] as well.

In the literature of *k*-anonymity problem, there are two main models. One model is global recoding [7,13,29,23,24,26] while the other is local recoding [1,29,25]. Here, we assume that each attribute has a corresponding conceptual generalization hierarchy or taxonomy tree. A lower level domain in the hierarchy provides

* Corresponding author. Tel.: +61 0746315530.
E-mail addresses: sunx@usq.edu.au (X. Sun), sun@usq.edu.au (L. Sun), wang@usq.edu.au (H. Wang).

0140-3664/$ - see front matter Crown Copyright © 2010 Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2010.03.020

Table 1
Raw microdata.

ID	Age	Country	Zip Code	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	14248	Flu
10	37	Canada	14204	Flu
11	36	Canada	14205	Flu
12	35	USA	14248	Indigestion

more details than a higher level domain. For example, Zip Code 14248 is a lower level domain and Zip Code 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, Age 27, 28 in the lower level can be replaced by the interval (27–28) in the higher level (See Table 2).

1.1. Motivation

When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in the literature [5,12]: *identity disclosure* and *attribute disclosure*. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when 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 the data release. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. Several models such as p-sensitive k-anonymity [31], l-diversity [19] and t-closeness [16] were proposed. However, depending on the nature of the sensitive attributes, even these enhanced properties still permit the information to be disclosed or have other limitations.

Limitation of p-sensitive k-anonymity: The purpose of p-sensitive k-anonymity [31] is to protect against attribute disclosure by requiring that there should be at least p different values for each sensitive attribute within the records sharing a combination of quasi-identifier. This approach has the limitation of implicitly assuming that each sensitive attribute takes values uniformly over its domain; that is, that 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.

Table 2
2-anonymous microdata.

ID	Age	Country	Zip Code	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
12	(33–35)	America	142**	Indigestion
10	(36–37)	America	142**	Flu
11	(36–37)	America	142**	Flu

Limitation of l-diversity: The l-diversity model [19] 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, l-diversity is insufficient to prevent similarity attack.

Limitation of t-closeness: The t-closeness model [16] 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. [16] 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.

Facing with these limitations, we intend to enhance the current privacy principles to make them preserve good data quality and data privacy. In this paper, we identify situations when p-sensitive k-anonymity property is not enough for privacy protection and study two solutions to overcome this identified problem. Our comprehensive experimental results show that the enhanced privacy models are better than the previous one in terms of data quality and utility.

2. Preliminaries

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

- *Identifier attributes* that can be used to identify a record, such as Name and Social Security Number. Since our objective is to prevent sensitive information from being linked to specific respondents, we will assume in what follows that *identifier attributes* in the microdata have been removed or encrypted in a pre-processing step.
- *Quasi-identifier (QI) attributes* are those, such as Zip Code and Age, that in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in the microdata belong. Unlike *identifier attributes*, *QI attributes* cannot be removed from the microdata, because any attribute is potentially a *QI attribute*.
- *Sensitive attributes* that are assumed to be unknown to an intruder and need to be protected, such as Disease or ICD-9 Code.[2]

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 value for 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 [35] between quasi-identifier attributes and external available information to glean the identity of individ-

[1] The interpretation of the term "well-represented" can be found in [19].

[2] International Statistical Classification of Diseases and Related Health Problems: ICD-9, which provides multiple external links for looking up ICD codes. Available http://www.icd9cm.chrisendres.com/.

Table 3
External available information.

Name	Age	Country	Zip Code
Rick	26	USA	14246
Hassen	45	India	13064
Rudy	25	Canada	14249
Yamazaki	48	Japan	13074

Table 4
Categories of disease.

Category ID	Sensitive values	Sensitivity
One	HIV, cancer	Top secret
Two	Phthisis, hepatitis	Secret
Three	Obesity, asthma	Less secret
Four	Flu, indigestion	Non secret

uals 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, in order to enforce the *k*-anonymity property.

Definition 1 (*k-anonymity*). The modified microdata table 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 [23,30]. More recent papers use different terminologies such as equivalence class [19,16] and QI-cluster [31].

For example, let {Age, Country, Zip Code} be the set of quasi-identifier attributes 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. The *k*-anonymity property ensures protection against identity disclosure, however, as we shall 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, in which the set of quasi-identifier is {Age, Country, Zip Code} 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 the modified table (Table 2) the Age attribute was modified to '(25–26)', he can deduce that both Rick and Rudy have Cancer, even he does not know which record, 3 or 4, corresponds to which person. This example shows that even if *k*-anonymity can well protect identity disclosure, sometimes it fails to protect against sensitive attribute disclosure. To deal with this problem in privacy breach, the *p*-sensitive *k*-anonymity model was introduced in [31].

Definition 2 (*p-sensitive k-anonymity*). The modified microdata table T' satisfies *p*-sensitive *k*-anonymity property if it satisfies *k*-anonymity, and for each QI-group in T', the number of distinct values for each sensitive attribute is at least *p* within the same QI-group.

Although the *p*-sensitive *k*-anonymity principle represents an important step beyond *k*-anonymity in protecting against attribute disclosure, it still has some shortcomings. Sometimes, the domain of the sensitive attributes, especially the categorical ones, can be partitioned into categories according to the sensitivity of attributes. For example, in medical datasets Table 1, the Disease attribute can be classified into four categories (see Table 4). The different types of diseases are organized in a category domain. The attribute values are very specific, for example they can represent HIV or Cancer, which are both Top Secret information of the individuals. In the case that the initial microdata contains specific sensitive attributes like Disease, the data owner can be interested in protecting not only these most specific values, but also the category that the sensitive values belong to. For example, the information of a person who affected with Top Secret needs to be protected, no matter whether it is HIV or Cancer. If we modify the microdata to satisfy *p*-sensitive *k*-anonymity property, it is possible that in a QI-group with *p* distinct sensitive attribute values, all of them belong to the same pre-defined confidential category. For instance, the values {HIV, HIV, Cancer, Cancer} of one QI-group in Table 5 all belong to Top Secret category. To avoid such situations, we introduce two new enhanced privacy protection models, namely, p^+-sensitive *k*-anonymity model and (p,α)-sensitive *k*-anonymity model, which are aware of not only protecting specific sensitive values. Such attack is known as *Similarity Attack*, which refers to the situation when the sensitive attribute values in a QI-group have distinct but similar sensitivity, an adversary can learn important information.

3. New privacy protection models

Let *S* be a categorical sensitive attribute we want to protect against attribute disclosure. All of the concepts in this paper are easily explained in the single sensitive attribute setting, but can also be generalized to multiple sensitive attributes. First, we sort the values of *S* according to their sensitivity, forming an ordered value domain *D*, and then partition the attribute domain into *m*-categories $(S_1, S_2, \ldots, S_m)$, such that $S = \cup_{i=1}^{m} S_i, S_i \cap S_j = \emptyset$ (for $i \neq j$) and $S_i \leqslant S_{i+1}$ (for $i = 1, \ldots, m$), where $S_i \leqslant S_j$ means that S_i is more sensitive than the S_j (for $1 \leqslant i \leqslant j \leqslant m$). For example, consider the Disease S = {HIV, Cancer, Phthisis, Hepatitis, Obesity, Asthma, Flu, Indigestion} in Table 1, it has been partitioned into four categories according to the sensitivity of the diseases (Table 4), where S_1 (Top Secret) is the most sensitive and S_4 (Non Secret) is the least one.

Definition 3 (*p^+-sensitive k-anonymity*). The modified Microdata table T' satisfies p^+-sensitive *k*-anonymity property if it satisfies *k*-anonymity, and for each QI-group in T', the number of distinct categories for each sensitive attribute is at least *p* within the same QI-group.

Table 6 is a 2^+-sensitive 4-anonymous view of Table 1. The first four records in Table 6 corresponds to the records 1, 4, 9 and 12 in Table 1 after anonymization. As you can see, for example, in Table 6, the first four records belong to one QI-group in which

Table 5
2-Sensitive 4-anonymous microdata.

ID	Age	Country	Zip Code	Disease
1	<30	America	142**	HIV
2	<30	America	142**	HIV
3	<30	America	142**	Cancer
4	<30	America	142**	Cancer
5	>40	Asia	130**	Hepatitis
6	>40	Asia	130**	Phthisis
7	>40	Asia	130**	Asthma
8	>40	Asia	130**	Obesity
9	3*	America	142**	Flu
10	3*	America	142**	Flu
11	3*	America	142**	Flu
12	3*	America	142**	Indigestion

Table 6
2^+-sensitive 4-anonymous microdata.

Age	Country	ZipCode	Disease	Category
<40	America	1424*	HIV	One
<40	America	1424*	Cancer	One
<40	America	1424*	Flu	Four
<40	America	1424*	Indigestion	Four
>40	Asia	130**	Hepatitis	Two
>40	Asia	130**	Phthisis	Two
>40	Asia	130**	Asthma	Three
>40	Asia	130**	Obesity	Three
<40	America	1420*	HIV	One
<40	America	1420*	Cancer	One
<40	America	1420*	Flu	Four
<40	America	1420*	Flu	Four

the Disease is not that easy to be referred since they belong to two different categories defined in Table 4. Compared with the previous anonymous solution shown in Table 5, this new model could overcome the shortcomings of previous models and reduce the possibility of leaking privacy. Before introducing our next enhanced (p,α)-sensitive k-anonymity model, we first define an ordinal weight for each category, which captures the degree that each specific sensitive value contributes to the QI-group.

Let $D(S) = \{S_1, S_2, \ldots, S_m\}$ denote a partition of categorical domain of an attribute S and let $weight(S_i)$ denote the weight of category S_i. Then,

$$\begin{cases} weight(S_i) = \frac{i-1}{m-1}; & 1 \leqslant i < m \\ weight(S_m) = 1, \end{cases} \quad (1)$$

Note that the weight of the specific sensitive value is equal to the weight of the category that the specific value belongs to. The weight of the QI-group is the total weight of each specific sensitive values that the QI-group contains.

We illustrate these concepts by taking Table 6 as an example. Given the partition of sensitive attributes as shown in Table 4 and four corresponding values set A = {Cancer, Phthisis, Asthma, Flu}. According to formula (1), $weight(S_1) = 0$, $weight(S_2) = 1/3$ and $weight(Asthma) = 2/3$, $weight(Flu) = 1$, the total weight of A is 0 + 1/3 + 2/3 + 1 = 2. Our next enhanced privacy principle is defined as follows:

Definition 4 ((p,α)-*sensitive k-anonymity*). The modified microdata table T' satisfies (p,α)-sensitive k-anonymity property if it satisfies k-anonymity, and each QI-group has at least p distinct sensitive attribute values with its total weight at least α.

For instance, Table 7 is a (3,1)-sensitive 4-anonymous view of Table 1. Since there are at least three different values in each QI-group and the least total weight of the QI-group is 1. We can easily see that the (p,α)-sensitive k-anonymity model can well protect sensitive information disclosure as well when compared with previous p-sensitive k-anonymity model.

Table 7
(3, 1)-Sensitive 4-anonymous microdata.

Age	Country	ZipCode	Disease	Weight	Total
<40	America	142**	HIV	0	1
<40	America	142**	HIV	0	
<40	America	142**	Cancer	0	
<40	America	142**	Flu	1	
>40	Asia	130**	Hepatitis	1/3	2
>40	Asia	130**	Phthisis	1/3	
>40	Asia	130**	Asthma	2/3	
>40	Asia	130**	Obesity	2/3	
<40	America	14***	Cancer	0	3
<40	America	14***	Flu	1	
<40	America	14***	Flu	1	
<40	America	14***	Indigestion	1	

These two new introduced models focus on different perspectives in protecting sensitive attributes disclosures. Instead of focusing on the specific values of sensitive attributes, p^+-sensitive k-anonymity model cares more about the categories that the values belong to. Although (p,α)-sensitive k-anonymity model still put the point on the specific values, it includes an ordinal metric system to measure how much the specific sensitive attribute values contribute to each QI-group. In the next section, we theoretically prove that both p^+-sensitive k-anonymity and (p,α)-sensitive k-anonymity are NP-hard. We use different approaches to derive the hardness results. For the computing harness of p^+-sensitive k-anonymity, it can be proved directly as a deduction from the known results in [31], while to prove the hardness of the optimal (p,α)-sensitive k-anonymity problem, it takes a standard procedure by reducing it to a well-known NP-hard problem.

4. Hardness results

The optimal p-sensitive k-anonymity problem is NP-hard as discussed in [31]. It is easy to deduce that the optimal p^+-sensitive k-anonymity model is also NP-hard. Recall that the difference between the p^+-sensitive k-anonymity and p-sensitive k-anonymity principles is that the former requires p distinct categories, while the later enforces p different values. Consider the situation when each pre-defined category contains only one sensitive value, then the p^+-sensitive k-anonymity could be reduced to the p-sensitive k-anonymity principle. Because the optimal p-sensitive k-anonymity problem is NP-hard [31], it is easy to obtain that computing the optimal p^+-sensitive k-anonymity is NP-hard as well. Next, we show that the optimal (p,α)-sensitive k-anonymity problem is NP-hard.

Theorem 1. (p,α)-*sensitive k-anonymity is NP-hard for a binary alphabet* ($\sum = \{0,1\}$).

Proof. The proof is by transforming the problem of *edge partition into 4-cliques* to the (p,α)-sensitive k-anonymity problem. □

Edge partition into 4-cliques [8]: Given Given a simple graph $G = (V,E)$, with $|E| = 6m$ for some integer m, can the edges of G be partitioned into m edge-disjoint 4-cliques?

Given an instance of edge partition into 4-cliques. Set $p = 2$, $\alpha = 6$ and $k = 12$. For each vertex $v \in V$, construct a non-sensitive attribute. For each edge $e \in E$, where $e = (v_1, v_2)$, create a pair of records $r_{v1,v2}$ and $\tilde{r}_{v1,v2}$, where the two records have the attribute values of both v_1 and v_2 equal to 1 and all other non-sensitive attribute values equal to 0, but one record $r_{v1,v2}$ has the sensitive attribute equal to 1 and the other record $\tilde{r}_{v1,v2}$ has the sensitive attribute equal to 0.

We define the cost of the (2,6)-sensitive 12-anonymity to be the number of suppressions applied in the data set. We show that the cost of the (2,6)-sensitive 12-anonymity is at most $48m$ if and only if E can be partitioned into a collection of m edge-disjoint 4-cliques.

Suppose E can be partitioned into a collection of m disjoint 4-cliques. Consider a 4-clique C with vertices v_1, v_2, v_3 and v_4. If we suppress the attributes v_1, v_2, v_3 and v_4 in the 12 records corresponding to the edges in C, then a cluster of these 12 records are formed where each modified record has four *s. Note that the (p,α)-sensitive requirement can be satisfied as the frequency of the sensitive attribute value 1 is equal to 6. The cost of the (2,6)-sensitive 12-anonymity is equal to $12 \times 4 \times m = 48m$.

Suppose the cost of the (2,6)-sensitive 12-anonymity is at most $48m$. As G is a simple graph, any twelve records should have at

least four attributes different. So, each record should have at least four *s in the solution of the (2,6)-sensitive 12-anonymity. Then, the cost of the (2,6)-sensitive 12-anonymity is at least $12 \times 4 \times m = 48m$. Combining with the proposition that the cost is at most $48m$, we obtain the cost is exactly equal to $48m$ and thus each record should have exactly four *s in the solution. Each cluster should have exactly 12 records (where six have sensitive value 1 and the other six have sensitive value 0). Suppose the twelve modified records contain four *s in attributes v_1, v_2, v_3 and v_4, the records contain 0s in all other non-sensitive attributes. This corresponds to a 4-clique with vertices v_1, v_2, v_3 and v_4. Thus, we conclude that the solution corresponds to a partition into a collection of m edge-disjoint 4-cliques.

5. Utility measurements

In this section, we discuss three generic utility metrics for measuring the quality of anonymized data.

There are a number of quality measurements presented in previous studies. Many metrics are utility based, for example, model accuracy [7,15] and query quality [14,37]. They are associated with some specific applications. Three generic metrics have been used in a number of recent works.

Discernability metric (DM): The Discernability metric was proposed by Bayardo et al. [3] and has been used in [14,37]. It is defined in the following:

$$\mathrm{DM} = \sum_{\mathrm{QI-group}\ G} |G|^2$$

where $|G|$ is the size of the QI-group G. The cost of anonymisation is determined by the size of the QI-group. An optimization objective is to minimize discernability cost.

Normalized average QI-group (CAVG): Normalized average QI-group size was proposed by LeFevre et al. [14] and has been used in [37]. It is defined as the following:

$$\mathrm{CAVG} = \left(\frac{\text{total records}}{\text{total QI} - \text{groups}}\right)/(k)$$

The quality of k-anonymisation is measured by the average size of QI-groups produced. An objective is to reduce the normalized average QI-group size.

These measurements are mathematically sound, but are not intuitive to reflect changes being made to an anonymized data set. In this chapter, we use the most generic criterion, called *distortion ratio*, which measures changes caused by the operation of data generalisation.

Distortion ratio: Suppose the value of the attribute in a tuple (record) has not been generalized, there will be no distortion. However, if the value of the attribute in a tuple is generalized to a more general value in the taxonomy tree or the conceptual generalization hierarchy, there is a distortion of the attribute of the tuple associated with the operation of the generalization. If the value is generalized more (i.e. the original value is updated to a value at the node of the taxonomy near to the root), the distortion will be greater. Thus, the distortion of this value is defined in terms of the height of the value generalized. For example, if the value has not been generalized, the height of the value generalized is equal to 0. If the value has been generalized one level up in the taxonomy, the height of the value generalized is equal to 1.

Let $h_{i,j}$ be the height of the value generalized of attribute A_i of the tuple t_j. The distortion of the whole data set is equal to the sum of the distortions of all values in the generalized data set. That is, distortion = $\sum_{i,j} h_{i,j}$. *Distortion ratio* is equal to the distortion of the generalized data set divided by the distortion of the fully generalized data set, where the fully generalized data set is the one with all values of the attributes are generalized to the root of the taxonomy tree.

6. The anonymization algorithms

In this section, we propose a set of algorithms for achieving new enhanced privacy principles, p^+-sensitive k-anonymity and p-sensitive k-anonymity principles. We adopt the local recording mechanism, since it produces less distortion than global recoding model. We first describe the idea of developing the local recoding algorithms, and then use a simple example to illustrate how the algorithm works.

The idea of the algorithm is to first generalize all tuples completely so that, initially, all tuples are generalized into one QI-group. Then, tuples are specialized in iterations. During the specialization, we must maintain p^+-and (p,α)-sensitive k-anonymity properties. The process continues until we cannot specialize the tuples any more (*Algorithm 1*). For ease of illustration, we present how the algorithm works for (p,α)-sensitive k-anonymity for a set of quasi-identifier attributes with size 1.

Let us illustrate it with an example in Table 8(a). Suppose the QI contains Zip Code only. Because there are only two sensitive values, so we assume that $\alpha = 1, p, k = 2$. Initially, we generalize all four tuples completely to a most generalized value Zip Code=***** (Fig. 1(a)). Then, we specialize each tuple one level down in the generalization hierarchy. We obtain the branch with Zip Code = 1**** in Fig. 1(b). In the next iteration, we obtain the branch with Zip Code = 14***, the branch with Zip Code = 142** and the branch with Zip Code = 1424* in Fig. 1(c)–(e), respectively. Next, we can further specialize the tuples into the two branches as shown Fig. 1(f). Hence the specialization processing can be seen as the growth of a tree.

Algorithm 1. The top-down local recoding algorithm (*Localpk*(p,k))

1. Fully generalize all tuples such that all tuples are equal.
2. Let P be a set containing all these generalized tuples
3. $S \leftarrow \{P\}$; $O \leftarrow \emptyset$.
4. Repeat
5. $S' \leftarrow \emptyset$
6. For all $P \in S$ do
7. Specialize all tuples in P one level down in generalization hierarchy forming a number of specialized child nodes.
8. Un-specialize the nodes which do not satisfy p-sensitive k-anonymity by moving the tuples back to the parent node.
9. If the parent P does not satisfy p-sensitive k-anonymity then.
10. Un-specialize some tuples in the remaining child nodes so that the parent P satisfies p-sensitive k-anonymity.
11. For all non-empty branches B of P, do $S' \leftarrow S' \cup \{B\}$
12. $S \leftarrow S'$
13. If P is non-empty then $O \leftarrow O \cup \{P\}$
14. Until $S = \emptyset$
15. Return O.

If each leaf node satisfies (p,α)-sensitive k-anonymity, then the specialization will be successful. However, we may encounter some problematic leaf nodes that do not satisfy (p,α)-sensitive k-anonymity. Then, all tuples in such leaf nodes will be pushed upwards in the generalization hierarchy. In other words, those tuples cannot be specialized in this process. They should be kept unspe-

Table 8
(a): Sample Data from Table 1; (b): Original projected table; (c): Generalized projected table.

Age	Zip Code	Disease
(a)		
27	14248	HIV
35	14248	Indigestion
33	14248	Flu
25	14247	Cancer
(b)		
1	14248	HIV
2	14248	Indigestion
3	14248	Flu
4	14247	Cancer
(c)		
1	14248	HIV
2	14248	Indigestion
3	1424*	Flu
4	1424*	Cancer

cialized in the parent node. For example, in Fig. 1(f), the leaf node with Zip Code = 14247 contains only one tuple, which violates (p,α)-sensitive k-anonymity. Thus, we have to move this tuple back to the parent node with Zip Code = 1424*. See Fig. 1(g).

After the previous step, we move all tuples in problematic leaf nodes to the parent node. However, if the collected tuples in the parent node do not satisfy (p,α)-sensitive k-anonymity, we should further move some tuples from other leaf nodes L to the parent node so that the parent node can satisfy (p,α)-sensitive k-anonymity while L also maintain the (p,α)-sensitive k-anonymity. For instance, in Fig. 1(g), the parent node with Zip Code = 1424* violates (p,α)-sensitive k-anonymity. Thus, we should move one tuples upwards in the node B with Zip Code = 14248 (which satisfies (p,α)-sensitive k-anonymity). In this example, we move tuple 3 upwards to the parent node so that both the parent node and the node B satisfy the (p,α)-sensitive k-anonymity.

Finally, in Fig. 1(h), we obtain a data set where the Zip Code of tuples 3 and 4 are generalized to 1424* and the Zip Code of tuples 1 and 2 remains 14248. So the final allocation of tuples in Fig. 1(h) is the final distribution of tuples after the specialization. The results can be found in Table 8(c).

7. Proof-of-concept experiments

We performed two sets of experiments on our proposed p^+-sensitive k-anonymity and (p,α)-sensitive k-anonymity models with real-world data sets to show their effectiveness and efficiency.

In the first set of experimental studies, we use Adult database, publicly available at the UC Irvine Machine Learning Repository[3] to evaluate the algorithms of the proposed two enhanced k-anonymity models in terms of *Similarity Attack*, *Effectiveness*, *Efficiency* and *Distortion Ratio*. The experimental results show that sensitive attribute disclosures can be significantly reduced under our new proposed models, therefore, emphasizing the need to protect the data against attribute disclosures beyond the previous p-sensitive k-anonymity model. Our second set of experiment deploys a real database CENSUS database[4] commonly used in the literature [38–40], and we compare our proposed algorithms with a clustering method, *Clustering* [32] in terms of three data quality measures, distortion ratio, discernability (DM) and normalized average QI-group size (CAVG). Both sets of experiments show that the proposed enhanced k-anonymity models are efficient and effective for privacy protection in real-world data publication.

[3] www.ics.uci.edu/-mlearn/MLRepository.html.
[4] downloadable at http://www.ipums.org

7.1. First set of experiments

7.1.1. Experiment setup

Data Sets. In this set of experiments, we adopted the publicly available Adult Database [21], which has become the benchmark of this field and was adopted by [13,7,16,31,19]. For Adult database, we used a configuration similar to [13]. We eliminated the records with unknown values. The resulting data set contains 45222 tuples. Seven of the attributes were chosen as the set of quasi-identifier attributes. We add a column with sensitive values called "Health Condition" consisting of {HIV, Cancer, Phthisis, Hepatitis, Obesity, Asthma, Flu, Indigestion} to the whole data set and randomly assign one sensitive value to each record of the Adult data set. Table 9 provides a brief description of the modified data set including the attributes we used, the type of each attribute, the number of distinct values for each attribute, and the height of the generalization hierarchy for each attribute.

On default, we set $\alpha = 1,\ p = 2$ and $k = 3$. We denote the previous p-sensitive k-anonymity model as Model 1, p^+-sensitive k-anonymity model as Model 2 and (p,α)-sensitive k-anonymity model as Model 3. We modified the Incognito algorithm [13] so that it produces p^+- and (p,α)-sensitive k-anonymous data sets as well. All the experiments are run on top of Windows XP on a machine with a 2.0 GHz Pentium 4 processor and 1 GB RAM.

7.1.2. Experimental results

We evaluate the proposed models in terms of the *similarity attack*, *effectiveness*, *execution time* and *distortion ratio*, and summarize the experimental results as follows.

7.1.2.1. Similarity attack. We use the first 7 attributes in Table 9 as the quasi-identifier attributes and treat "Health Condition" as the sensitive attribute. We divide the eight values of the Health Condition attribute into four pre-defined equal-size categories, based on the confidentiality of the values (See Table 4). Any QI-group that has all values falling in one category is viewed as vulnerable to the similarity attack. We first generate all 2-sensitive 2-anonymous tables. In total, there are 21 minimal data sets and 13 of them suffer from the similarity attack (13/21 = 0.62%). In one such anonymized table, a total of 916 records can be inferred about their sensitive value class. We then use generate all 30 minimal (2,1)-sensitive 2-anonymous tables, and found that only 4 of which are vulnerable to the similarity attack (4/30 = 13%). Similar results are obtained with p^+-sensitive k-anonymity model. To summarize, both p^+-sensitive k-anonymity and (p,α)-sensitive k-anonymity models could significantly reduce the chance of similarity attacks.

7.1.2.2. Effectiveness. Table 10 shows that even under two new enhanced p^+-sensitive k-anonymity and (p,α)-sensitive k-anonymity models, disclosure channels still exits so that the Health Condition can be inferred. This is because of the nature of sensitive attributes. However, compared with the previous p-sensitive k-anonymity model, our new enhanced models could significantly reduce the number of sensitive attribute disclosures, which help achieve better privacy protection.

7.1.2.3. Efficiency. We compare the efficiency among three privacy measures: (1) p-sensitive k-anonymity; (2) p^+-sensitive k-anonymity; (3) (p,α)-sensitive k-anonymity. Results of efficiency experiments are shown in Fig. 2. The running times for p^+-sensitive k-anonymity and p-sensitive k-anonymity are similar, which makes p^+-sensitive k-anonymity usable in practice. Fig. 2(a) shows the running times with fixed $p = 4,\ \alpha = 4$ while varying the size s of the quasi-identifier attributes, where $2 \leqslant s \leqslant 7$. A set of quasi-identifier attributes with size s consists of the first s attributes listed in Table 9. Fig. 2(b) shows the running times of three privacy

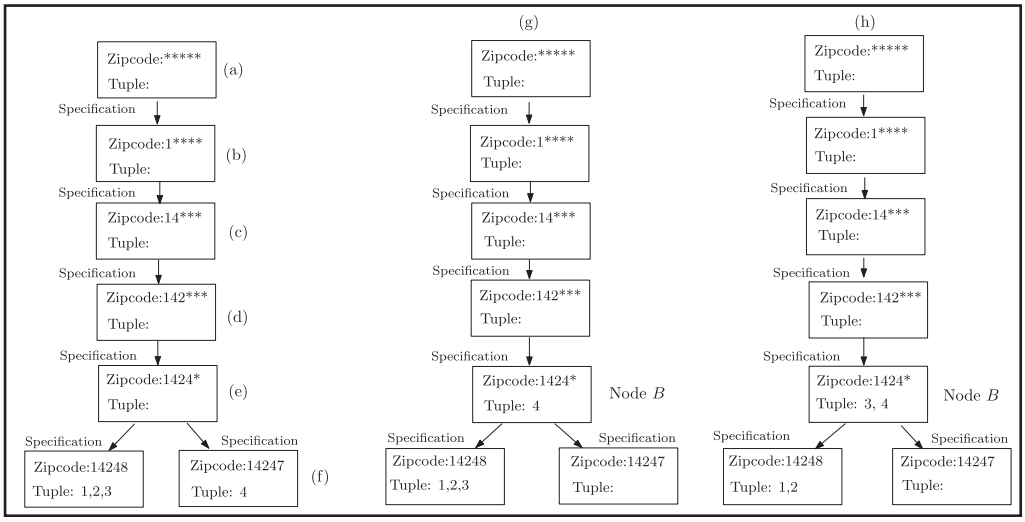

Fig. 1. Algorithm illustration for QI={Zip Code}.

Table 9
Features of quasi-identifier attributes.

Attribute	Type	Height
Age	Numeric	5
Workclass	Categorical	3
Education	Categorical	4
Country	Categorical	3
Marital status	Categorical	3
Race	Categorical	3
Gender	Categorical	2
Health condition	Sensitive	–

Table 10
Attribute disclosures comparisons.

k, p	Number of attribute disclosures		
$k = 3,\ p = 2$	Model 1	Model 2	Model 3
	25	3	2
$k = 4,\ p = 2$	Model 1	Model 2	Model 3
	30	4	6
$k = 3,\ p = 3$	Model 1	Model 2	Model 3
	15	2	3
$k = 4,\ p = 3$	Model 1	Model 2	Model 3
	21	1	2

measures with the same set of quasi-identifier attributes but with different parameters settings of p and α. As shown in the figures, p^+-sensitive k-anonymity run faster than the (p, α)-sensitive k-anonymity; the difference gets larger when α increases.

7.1.2.4. Distortion ratio. Results of distortion ratio are shown in Fig. 3. From Fig. 3(a), the distortion ratio almost increases as the size of the quasi-identifier attributes grows. This is because when the set of quasi-identifier attributes contains more attributes, there is more chance that two tuples are different with respect to the set of the quasi-identifier attributes. In other words, there is more chance that the tuples will be generalized. Thus, the distortion ratio is greater. On average, the distortion ratio of Model 3 is greater than Model 2, since Model 3 require a more strict privacy requirement causing more data generalization. In Fig. 3(b), when α increases, the distortion ratio decreases. Intuitively, if α is larger, meaning that there is less requirement of metric α, it yields fewer operations of generalization of the values in the data set. Thus, the distortion ratio is smaller.

7.2. Second set of experiments

7.2.1. Data sets

Our experimentation deploys a real database CENSUS commonly used in the literature [38–40]. It contains 500 k tuples, each of which describes the personal information of an American. The CENSUS data set includes four numerical attributes Age, Birthplace, Occupation and Income, whose domains are [16,93], [1,710],

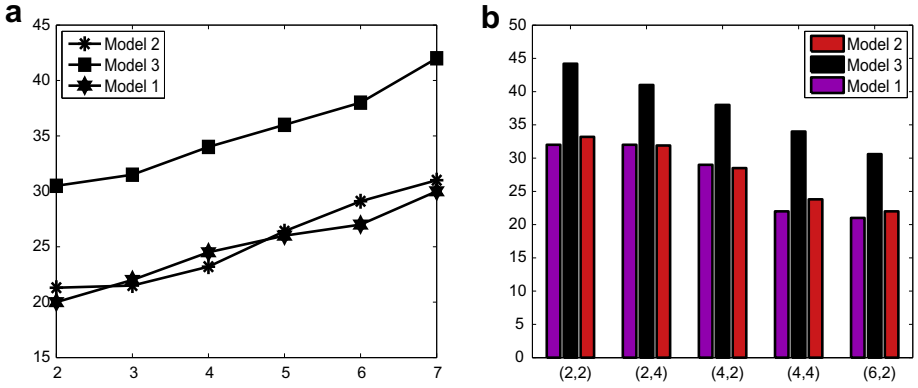

Fig. 2. Execution time vs. three privacy measures.

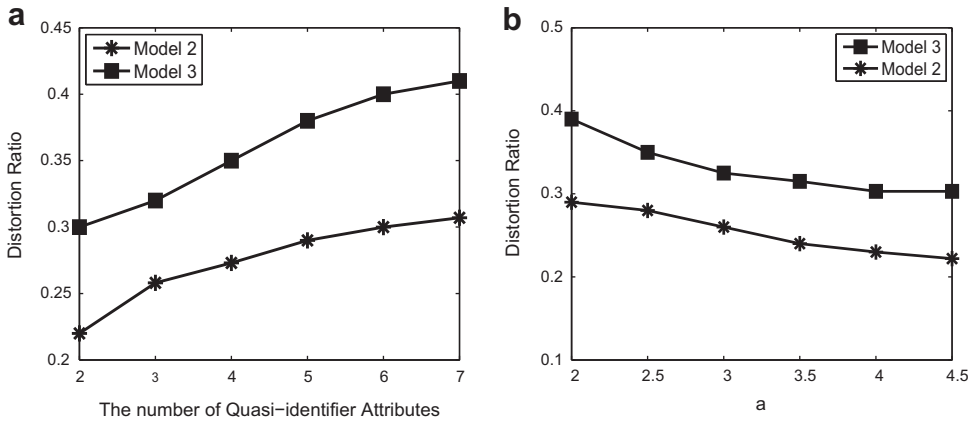

Fig. 3. Distortion ratio vs. two enhanced privacy measures.

[1,983] and [1k,100k], respectively. We treat the first three columns as the set of quasi-identifier attributes, and Income as the sensitive attribute. We further divide the attribute Income into four categories shown in Table 11. By default, we set $\alpha = 1,\ p = 2$ and $k = 3$. We denote the clustering algorithm used in [32] as *Clustering*, the local recoding algorithm for Model 2 and Model 3 as *localpk2* and *localpk3*. We run comparisons throughout three quality measures, distortion ratio, discernability (DM) and normalized average QI-group size (CAVG).

In Fig. 4, we compare three local recoding algorithms in terms of the distortion ratio, DM and CAVG. Based on distortion ratio, our proposed models are superior to the previous one. Specifically, both *localpk2* and *localpk3* perform consistently better than the clustering method. This shows that our defined α metric could significantly reduce the distortion ratio. For other two measures, our models are better than the previous one as well. In comparison to big difference of CAVG among different algorithms, differences of an algorithm in variant k are negligible.

Fig. 5 shows the graphs of the execution time against k and α when $p = 2$. In Fig. 5(a), when k varies, the execution time of all algorithms increases with k. This is because, when k increases, the number of candidates (representing the generalization domain) increases, and thus the execution time increases. In Fig. 5(b), when α varies, different algorithms change differently. The execution time of local recoding algorithms decreases when α increases. In the local recoding algorithms, we may have to unspecialize some tuples in the branches satisfying (p,α)-sensitive k-anonymity so that the parent satisfies (p,α)-sensitive k-anonymity. When α is small, it is more likely that the parent cannot satisfy (p,α)-sensitive k-anonymity, triggering this step of un-specialization. As the un-specialization step is more complex, the execution time is larger when α is smaller.

Table 11
Categories of Income.

Category ID	Income	Sensitivity
One	[1k,20k]	Lower income
Two	(20 k,40k]	Average income
Three	(40k,70k]	Above average
Four	(70k,100k]	Higher income

8. Related work

Several types of information disclosure in microdata publishing have been identified in the literature [5,12]. An important type of information disclosure is attribute disclosure. Attribute disclosure occurs when a sensitive attribute value is associated with an individual. This is different from both identity disclosure (i.e., linking an individual to a record in the database) and membership disclosure [6,22] (i.e., learning whether an individual is included in the database). As in [4,19,16,17,26,31,32,36], this paper focuses on how to limit attribute disclosures in data publishing.

k-anonymity [23,28] (requiring each QI-group contains at least k records) aims at preventing identity disclosure. Because identity disclosure leads to attribute disclosure (once the record is identified, its sensitive value is immediately revealed), k-anonymity can partly prevent attribute disclosure. But because attribute

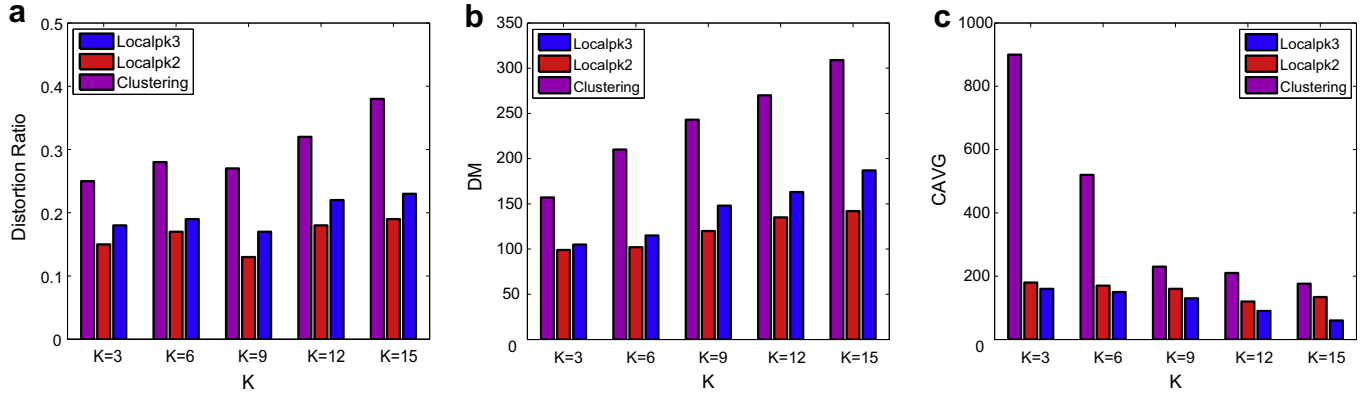

Fig. 4. Performance of different local recoding algorithms with varying k: (a) distortion ratio, (b) discernability and (c) normalized average QI-group size.

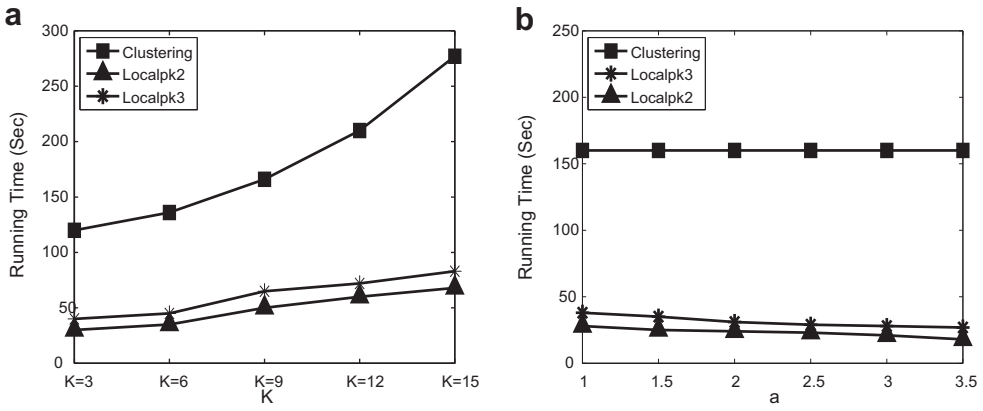

Fig. 5. Running time comparison of different local recoding algorithms.

disclosure can occur without identity disclosure [19,36,31,32] (for example, when all records in the equivalence class have the same sensitive value), *k*-anonymity does not prevent attribute disclosure. *l*-diversity [19] remedies the above limitations of *k*-anonymity by requiring that in any equivalence class, each sensitive value can occur with a frequency of at most $1/l$. While there are several other definitions of *l*-diversity such as recursive (c, l)-diversity, the above probabilistic interpretation is the most widely used one in the literature. Similar privacy requirements are the (α, k)-anonymity [36] and *p*-sensitive *k*-anonymity [31]. *l*-diversity ensures that the probability of inferring the sensitive value is bounded by $1/l$. However, this confidence bound may be too strong for some sensitive values (e.g., a common form of disease) and too weak for some other sensitive values (e.g., a rare form of cancer). *t*-closeness [16] remedies the limitations of *l*-diversity, by requiring the sensitive attribute distribution in QI-group to be close to that in the overall data. A closely-related privacy requirement is the template-based privacy [34] where the probability of each sensitive value is bounded separately.

Two popular anonymization techniques are generalization and bucketization. Generalization [23,28,29,33] replaces a value with a less-specific but semantically consistent value. Three types of encoding schemes have been proposed for generalization: global recoding, regional recoding, and local recoding. Global recoding has the property that multiple occurrences of the same value are always replaced by the same generalized value. Regional record [14] is also called multi-dimensional recoding (the Mondrian algorithm) which partitions the domain space into non-intersect regions and data points in the same region are represented by the region they are in. Local recoding does not have the above constraints and allows different occurrences of the same value to be generalized differently. Bucketization [10,18,39] first partitions tuples in the table into buckets and then separates the quasi-identifiers with the sensitive attribute by randomly permuting the sensitive attribute values in each bucket. The anonymized data consists of a set of buckets with permuted sensitive attribute values. In particular, bucketization has been used for anonymizing high-dimensional data [9]. In this paper, we used the generalization technique with local recoding schemes to ensure the less information loss.

It is important that the anonymized data can be used for data analysis or data mining tasks. Because of this, most utility measures are workload-independent, i.e., they do not consider any particular data mining workload. For example, the utility of the anonymized data has been measured by the number of generalization steps, the average size of the QI-goup (CAVG) [14,37,27], the discernibility metric (DM) [3,27] which sums up the squares of QI-group sizes, the *KL*-divergence between the reconstructed distribution and the true distribution for all possible quasi-identifier values [11] and the distortion ratio, which is equal to the distortion of the generalized data set divided by the distortion of the fully generalized data set. In this paper, we used the average size of the QI-goup (CAVG), the discernibility metric (DM) and the distortion ratio to evaluate the utility of the anonymized data.

9. Conclusion and future work

p-sensitive *k*-anonymity is a novel property that, when satisfied by microdata sets, can help increase the privacy of the respondents whose data is being used. However, as shown in the paper, to some extent, this property is not enough for protecting sensitive attributes. In this paper, we proposed two new models, called p^+-sensitive *k*-anonymity and (p, α)-sensitive *k*-anonymity models to enhance the previous *p*-sensitive *k*-anonymity model. Our experimental results show that our proposed models have advantages in terms of effectiveness, efficiency and distortion ratio.

The ordinal metric discussed in this paper to enhance *p*-sensitive *k*-anonymity property can also be extended to other privacy paradigms, like *l*-diversity and *t*-closeness. As we stated before, both privacy principles have their own limitations, and we could adopt this new introduced ordinal metric to amend these limitations and to achieve a better balance between data quality and privacy level.

Acknowledgements

This research is supported by Australian Research Council (ARC) Grants DP0774450 and DP0663414.

References

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, Anonymizing tables, in: Proceedings of the 10th International Conference on Database Theory (ICDT'05), Edinburgh, Scotland, pp. 246–258.
[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, A. Zhu, Approximation algorithms for *k*-anonymity. Journal of Privacy Technology, paper number 20051120001.
[3] R. Bayardo, R. Agrawal, Data privacy through optimal *k*-anonymity, in: Proceedings of the 21st International Conference on Data Engineering (ICDE), 2005.
[4] J. Brickell, V. Shmatikov, The cost of privacy: destruction of data-mining utility in anonymized data publishing, In KDD (2008) 70–78.
[5] G.T. Duncan, D. Lambert, Disclosure-limited data dissemination, Journal of the American Statistical Association 18 (393) (1986) 10–28.
[6] C. Dwork, Differential privacy, In ICALP (2006) 1–12.
[7] B. Fung, K. Wang, P. Yu, Top-down specialization for information and privacy preservation, in: Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan.
[8] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, Freeman, San Francisco, 1979.

[9] G. Ghinita, Y. Tao, P. Kalnis, On the anonymization of sparse high-dimensional data, In ICDE (2008) 715–724.
[10] N. Koudas, D. Srivastava, T. Yu, Q. Zhang, Aggregate query answering on anonymized tables, In ICDE (2007) 116–125.
[11] D. Kifer, J. Gehrke, Injecting utility into anonymized datasets, In SIGMOD (2006) 217–228.
[12] D. Lambert, Measure of disclosure risk and harm, Journal of Official Statistics 9 (1993) 313–331.
[13] K. LeFevre, D. DeWitt, R. Ramakrishnan, Incognito: efficient full-domain k-anonymity, in: ACM SIGMOD International Conference on Management of Data, June 2005.
[14] K. LeFevre, D.J. DeWitt, R. Ramakrishnan. Mondrian multidimensional *k*-anonymity, in: ICDE'06: Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), IEEE Computer Society, Washington, DC, USA, 2006, p. 25.
[15] K. LeFevre, D.J. DeWitt, R. Ramakrishnan, Workload-aware anonymization, in: KDD 06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 2006, ACM Press, pp. 277–286.
[16] N. Li, T. Li, S, Venkatasubramanian *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity, ICDE 2007, pp. 106–115.
[17] T. Li, N. Li, On the tradeoff between privacy and utility in data publishing, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), 2009.
[18] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, J.Y. Halpern, Worst-case background knowledge for privacy-preserving data publishing, In ICDE (2007) 126–135.
[19] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, *l*-Diversity: privacy beyond *k*-anonymity, In ICDE (2006).
[20] A. Meyerson, R. Williams, On the complexity of optimal *k*-anonymity, in: Proceedings 23rd ACM-SIGMOD-SIGACT-SIGART Symposium on the Principles of Database Systems, Paris, France, 2004, pp. 223–228.
[21] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases. Available at <www.ics.uci.edu/-mlearn/MLRepository.html, University of Califonia, Irvine, 1998>.
[22] M.E. Nergiz, M. Atzori, C. Clifton, Hiding the presence of individuals from shared databases, in: SIGMOD 665C676, 2007.
[23] P. Samarati, Protecting respondents' identities in microdata release, IEEE Transactions on Knowledge and Data Engineering 13 (6) (2001) 1010–1027.
[24] X. Sun, M. Li, H. Wang, A. Plank, An efficient hash-based algorithm for minimal *k*-anonymity problem, in: Thirty-First Australasian Computer Science Conference (ACSC, 2008), Wollongong, Australia, pp. 101–107.
[25] X. Sun, H. Wang, J. Li. On the complexity of restricted k-anonymity problem, in: 10th Asia Pacific Web Conference (APWEB, 2008), Shenyang, China, pp. 287–296.
[26] X. Sun, H. Wang, J. Li, *l*-diversity based updating technique for large time-evolving microdata, in: 21st Australasian Joint Conference on Artificial Intelligence (AusAI, 2008), 3–5 December 2008, Auckland, New Zealand, pp. 461–469.
[27] X. Sun, H. Wang, J. Li, Injecting purposes and trust into data anonymization, in: The 18th ACM Conference on Information and Knowledge Management (CIKM 2009), Hong Kong, China.
[28] L. Sweeney, Uniqueness of simple demographics in the U.S. Population, Technical Report, Carnegie Mellon University, 2000.
[29] L. Sweeney, Achieving *k*-anonymity privacy protection using generalization and suppression, International Journal of Uncertainty, Fuzziness and Knowledge-Based System 10 (5) (2002) 571–588.
[30] L. Sweeney, *k*-anonymity: a model for protecting privacy, International Journal on Uncertainty Fuzziness Knowledge-based Systems 10 (5) (2002) 557–570.
[31] T.M. Traian, V. Bindu, Privacy protection: *p*-sensitive *k*-anonymity property, in: The 22th International Conference of Data Engineering (ICDE), Atlanta, 2006.
[32] T.M. Truta, A. Campan, P. Meyer, Generating microdata with *p*-sensitive *k*-anonymity property, SDM 2007, pp. 124–141.
[33] K. Wang, P.S. Yu, S. Chakraborty, Bottom-up generalization: a data mining solution to privacy protection, in: The Fourth IEEE International Conference on Data Mining, (ICDM 2004), pp. 249–256.
[34] K. Wang, B.C.M. Fung, P.S. Yu, Template-based privacy preservation in classification problems, in: ICDM, 2005, p. 466C473.
[35] W.E. Winkler, Advanced methods for record linkage, proceedings of the section on survey research methods, American Statistical Society (2002) 467–472.
[36] R. Wong, J. Li, A. Fu, K. Wang, (α, k)-anonymity: an enhanced *k*-anonymity model for privacy preserving data publishing, *KDD 2006*, pp. 754–759.
[37] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A.W.C. Fu, Utility-based anonymization using local recoding, in: KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 2006, ACM Press, pp. 785–790.
[38] X. Xiao, Y. Tao, Personalized privacy preservation, in: SIGMOD'06: Proceedings of the 2006 ACM SIGMOD International Conference on Management of data, 2006.
[39] X. Xiao, Y. Tao, Anatomy: simple and effective privacy preservation, In VLDB (2006) 139–150.
[40] X. Xiao, Y. Tao, M-invariance: towards privacy preserving re-publication of dynamic datasets, in: SIGMOD Conference, 2007, pp. 689–700.