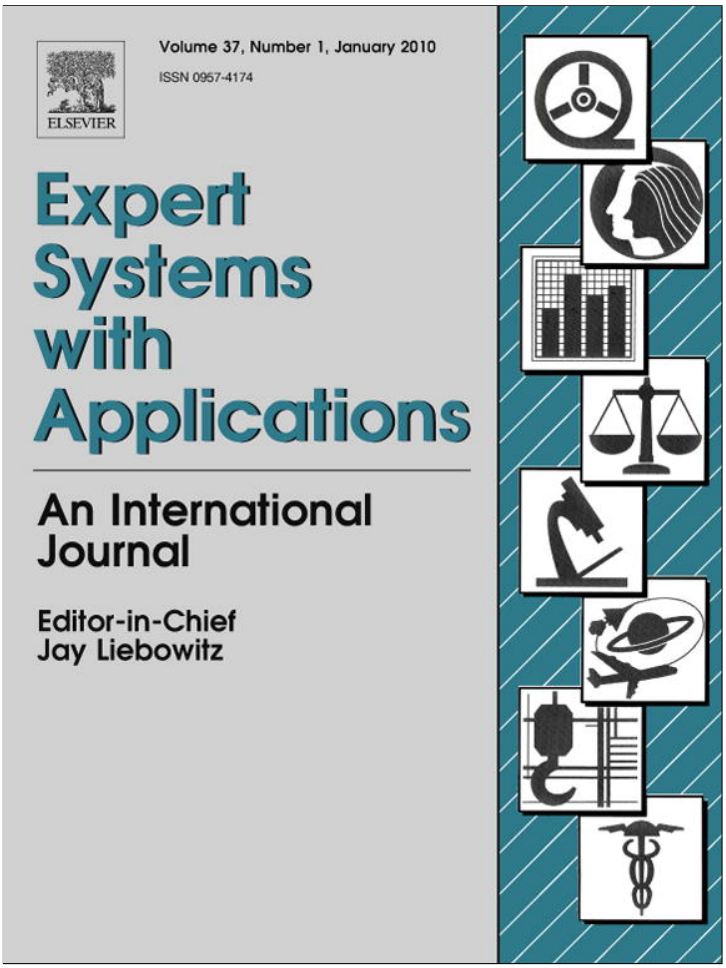

Expert Systems with Applications 37 (2010) 756–766

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A novel anonymization algorithm: Privacy protection and knowledge preservation

Weijia Yang [a,*], Sanzheng Qiao [b]

[a] Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200030, China
[b] Department of Computing and Software, McMaster University, Hamilton, Ont., Canada L8S 4K1

ARTICLE INFO

Keywords:
Data mining
Privacy protection
Data anonymization
Knowledge preservation

ABSTRACT

In data mining and knowledge discovery, there are two conflicting goals: privacy protection and knowledge preservation. On the one hand, we anonymize data to protect privacy; on the other hand, we allow miners to discover useful knowledge from anonymized data. In this paper, we present an anonymization method which provides both privacy protection and knowledge preservation. Unlike most anonymization methods, where data are generalized or permuted, our method anonymizes data by randomly breaking links among attribute values in records. By data randomization, our method maintains statistical relations among data to preserve knowledge, whereas in most anonymization methods, knowledge is lost. Thus the data anonymized by our method maintains useful knowledge for statistical study. Furthermore, we propose an enhanced algorithm for extra privacy protection to tackle the situation where the user's prior knowledge of original data may cause privacy leakage. The privacy levels and the accuracy of knowledge preservation of our method, along with their relations to the parameters in the method are analyzed. Experiment results demonstrate that our method is effective on both privacy protection and knowledge preservation comparing with existing methods.

1. Introduction

Privacy protection is an important issue in data processing. Suppose a data set shown in Fig. 1 is used in a medical research article. Before the article is published, the data must be anonymized to protect privacy. For example, if Michael knows that Hannah is 32 working as a clerk in UK, then he can infer that Hannah has diabetes, even the data set is published without names. A group of attributes, such as {Age, Job, Country} in the above example, which can be used to infer patients is commonly called *quasi-identifier* (Q-I). The attribute like disease is called *sensitive* attribute. In medical research, the associations such as "clerk → hypertension" and "[50–60] → diabetes" are often studied. This kind of associations compose *non-sensitive knowledge*, if individual names cannot be inferred from the associations. Thus, it is important that a data set is protected in such a way that the miner can hardly infer the sensitive data of an individual and, at the same time, discover non-sensitive knowledge.

In recent years, many methods have been proposed to protect the data privacy in data mining (Agrawal & Aggarwal, 2001; Agrawal & Srikant, 2000; Aggarwal & Yu, 2008; Clifton, Kantarcioglu, Vaidya, Lin, & Zhu, 2002; Du & Zhan, 2003; Lindell & Pinkas, 2000). Among these methods, data anonymization (Ciriani, Vimercati, Foresti, & Samarati, 2008; Machanavajjhala, Gehrke, Kifer, & Venkitasubramaniam, 2006; Sweeney, 2002) provides an effective yet simple way of preventing the user from learning sensitive data. Referring to the *k*-anonymity method (Sweeney, 2002), any individual is indistinguishable from at least $k-1$ other ones in the anonymized data set. Machanavajjhala et al. (2006), however, has pointed out that the user may guess the sensitive values with high confidence when the sensitive data is lack of diversity, and introduced the *l*-diversity method. Both *k*-anonymization and *l*-diversity have drawbacks. Most of the related works (Kifer & Gehrke, 2006; LeFevre, DeWitt, & Ramakrishnan, 2005, 2006; Li, Li, & Venkatasubramanian, 2007; Meyerson & Williams, 2004) generalize the data values. Some attributes in the quasi-identifier may even be totally suppressed. Both the attribute generalization and suppression lead to loss of data details (Aggarwal, 2005). As a result, the useful data knowledge such as the non-sensitive knowledge and data distributions can only be obtained in general forms, which may be of little value for the miner.

Recently, the permutation method has been proposed by Xiao and Tao (2006) and Koudas et al. (2007). Instead of generalizing the values of attributes, it randomly permutes the sensitive values in record groups. In this way, the user can hardly match the sensitive values with the right individuals. However, it is also difficult to discover the data knowledge from the permuted data.

In this paper, we propose a novel anonymization method. Instead of generalizing or permuting attribute values as in most anonymization methods, our method randomly breaks the links

* Corresponding author. Tel.: +86 21 5474 0000.
E-mail addresses: weijia.yang@yahoo.com.cn (W. Yang), qiao@cas.mcmaster.ca (S. Qiao).

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.05.097

	Quasi-identifier			Sensitive Data
Name	Age	Job	Country	Disease
Christopher	[50–60]	Doctor	USA	Hypertension
William	[40–50]	Clerk	USA	Hypertension
Jacob	[30–40]	Clerk	USA	Hypertension
Isabella	[30–40]	Clerk	Germany	Hypertension
Michael	[50–60]	Trader	USA	Diabetes
Hannah	[30–40]	Clerk	UK	Diabetes
Olivia	[50–60]	Engineer	USA	Diabetes
Madison	[20–30]	Trader	UK	Heart Disease
Matthew	[60–70]	Banker	USA	Cancer
Andrew	[30–40]	Banker	India	Cancer

Fig. 1. A sample of original confidential data.

between the value combinations of the quasi-identifier and the values of the sensitive data, while maintaining statistical relations between quasi-identifier data and sensitive data, thus preserving the non-sensitive knowledge for statistical study. It protects privacy and, at the same time, allows the miner to discover non-sensitive knowledge with high confidence. Furthermore, an enhanced method is proposed for preventing the privacy leakage when the user has some prior knowledge of the original data associations.

The paper is organized as follows: in Section 2, we introduce our privacy definition after presenting the preliminary knowledge of data anonymization. Our random anonymization (RA) algorithm is presented in Section 3. Then we analyze its level of privacy protection in Section 4. The discussion of the accuracy of the knowledge preservation is given in Section 5. Afterwards, we present in Section 6 an enhanced RA algorithm to limit the extra privacy leakage. Finally, we demonstrate our experiments in Section 7.

2. Data privacy

Since our method provides a way of anonymizing data, it is necessary to first introduce some preliminary definitions of data anonymization in Section 2.1. The privacy is disclosed when the user is able to correctly link the individuals with their sensitive values. In Section 2.2, we introduce our new definition of privacy by data association.

2.1. Preliminary

In this section, we adopt several definitions commonly used in the k-anonymization and l-diversity methods from Sweeney (2002) and Machanavajjhala et al. (2006). First, we present the formal definition of quasi-identifier.

Definition 1 (*Quasi-Identifier (Q-I)*). Given a data set $D(A_1, A_2, \ldots, A_m)$ and an external table D_E. For all records $r_i \in D$, if the value combination $r_i(A_j, \ldots, A_k), j, k \leqslant m, \{A_j, \ldots, A_k\}$ contains no identifiers, can be uniquely located in D_E, we call the set of attributes $\{A_j, \ldots, A_k\}$ a quasi-identifier.

For example, in the data set in Fig. 1, the external table D_E would consist of Name, Age, Job, and Country and a quasi-identifier would be $\{\text{Age}, \text{Job}, \text{Country}\}$.

The k-anonymization method uses the generalization technique, formally defined by:

Definition 2 (*Generalization*). Suppose that a domain M consists of disjoint partitions $\{P_i\}, i = 1 \ldots n$, and $\cup P_i = M$. On a given value combination v, we call the generalization process as returning the only partition P_i containing v.

By generalizing the quasi-identifier, each individual in a k-anonymous table is identical to at least $k-1$ other ones with respect to the quasi-identifier:

Definition 3 (*k-Anonymous*). Given a data set $D(A_1, A_2, \ldots, A_m)$ and its quasi-identifier QI. If for any subset $C \subseteq QI$ and for any record $r_i \in D$, there exist at least $k-1$ other records sharing the same values with r_i on the attribute set C, then data set D is k-anonymous.

When the data set is k-anonymous, we can group together the records with the same value combinations of the quasi-identifier:

Definition 4 (*Q-I Group*). Given a data set D and its quasi-identifier QI. We define the Q-I group as the set of all the records with the same values on QI.

Fig. 2 is a 2-anonymous data set generalized from Fig. 1, where the data set is partitioned into three groups. As shown in Fig. 2, the data of Age are grouped into wider intervals and Jobs are clustered. We denote by symbol "$*$" a wildcard in attributes. As a result, each record is indistinguishable from at least one other record by the quasi-identifier.

However, k-anonymity has drawbacks. When an individual belongs to the last group in Fig. 2, the user can infer that the individual has hypertension with more than 66% confidence, since two out of three in the group have hypertension. Moreover, since quite a few data values have been generalized in the anonymous data set, non-sensitive associations such as "[50–60] $\rightarrow$ diabetes" cannot be obtained.

While the k-anonymization method only focuses on hiding the Q-I information of the individuals, the l-diversity model (Machanavajjhala et al., 2006) pays attention to the relations between the Q-I and sensitive data:

Definition 5 (*l-diversity*). A Q-I group is said to satisfy l-diversity if it contains at least l "well-represented" values for the sensitive attribute. A data set is said to have l-diversity if every Q-I group of it satisfies l-diversity.

In particular, for each Q-I group, if the entropy of the sensitive attribute is greater than $\ln l$, then the data set is said to satisfy

Quasi-identifier			Sensitive Data
Age	Job	Country	Disease
[40–70]	*	USA	Hypertension
[40–70]	*	USA	Hypertension
[40–70]	*	USA	Diabetes
[40–70]	*	USA	Diabetes
[40–70]	*	USA	Cancer
[20–40]	[Trader,Banker]	*	Cancer
[20–40]	[Trader,Banker]	*	Heart Disease
[20–40]	Clerk	*	Hypertension
[20–40]	Clerk	*	Hypertension
[20–40]	Clerk	*	Diabetes

Fig. 2. A 2-anonymous data set by generalizing Fig. 1.

Quasi-identifier			Sensitive Data
Age	Job	Country	Disease
[40–70]	*	USA	Hypertension
[40–70]	*	USA	Hypertension
[40–70]	*	USA	Diabetes
[40–70]	*	USA	Diabetes
[40–70]	*	USA	Cancer
[20–40]	*	*	Cancer
[20–40]	*	*	Heart Disease
[20–40]	*	*	Hypertension
[20–40]	*	*	Hypertension
[20–40]	*	*	Diabetes

Fig. 3. A 2-diversity data set derived from Fig. 1.

"entropy *l*-diversity". When a user tries to guess the sensitivity of an individual from an anonymized data, the diversity quantifies the average number of the possible sensitive values. Fig. 3, also derived from Fig. 1, shows an anonymized data set satisfying more than (entropy) 2-diversity. From this data set, the user cannot infer the sensitive value of any individual with more than 50% confidence. However, Q-I values have been generalized further, thus the non-sensitive knowledge is hard to obtain. Moreover, the diversity of the sensitivity in each Q-I group is limited by the distribution of sensitive data in the whole data set.

To keep the data details, the permutation method permutes the sensitive values within each Q-I group. As shown in Fig. 4, the data set in Fig. 1 are permuted so that each group satisfies at least (entropy) 2-diversity. The permuted data set is actually divided into two views: the view of the quasi-identifier and the view of the sensitivity. During data mining, the user will join these two views for each record group. Although the data details are preserved, the non-sensitive associations can hardly be discovered from Fig. 4, because their confidences decrease dramatically.

Quasi-identifier				Sensitive Data
Age	Job	Country		Disease
[50–60]	Doctor	USA		Heart Disease
[50–60]	Trader	USA		Hypertension
[20–30]	Trader	UK	–	Hypertension
[30–40]	Clerk	Germany		Cancer
[60–70]	Banker	USA		Diabetes
[50–60]	Engineer	USA		Hypertension
[30–40]	Banker	India		Diabetes
[30–40]	Clerk	UK	–	Diabetes
[40–50]	Clerk	USA		Hypertension
[30–40]	Clerk	USA		Cancer

Fig. 4. An anonymous data set by permuting Fig. 1.

2.2. Measuring data privacy

Most *k*-anonymization methods (Aggarwal, 2005; Kifer & Gehrke, 2006; LeFevre et al., 2005, LeFevre, DeWitt, & Ramakrishnan, 2006; Meyerson & Williams, 2004; Sweeney, 2002) emphasize the protection of the quasi-identification information. The minimal size of the Q-I groups is used to quantify the privacy level of the anonymized data. The larger the minimal size is, the more difficult it is to identify an individual from a value combination of Q-I. Similarly, it is difficult to identify an individual from a value of sensitive data alone. However, if the user can establish a strong association between a value combination of Q-I and a value of sensitive data, then privacy can be compromised. Thus we propose that privacy is measured by the association between the value combinations of Q-I and the values of sensitive data. A probabilistic measurement of privacy or anonymity is given in Section 4. Also, the anti-monotone property of *k*-anonymity shows that if the quasi-identifier is *k*-anonymous, any subset of it is also *k*-anonymous. The value combinations of the subset do not disclose more privacy than those of the quasi-identifier. Thus, unlike the *k*-anonymization methods, we do not group the value combinations of Q-I. How do we prevent the user from getting these important associations? We will present our anonymization algorithm in the next section.

3. Anonymization algorithm

To break the associations between the value combinations of Q-I and the values of sensitive data, it is not necessary to anonymize all the attributes in Q-I. Our main idea is to randomly replace part of the Q-I data for each record by using the distributions of the original values of an attribute in the Q-I. In this way, no new information is added to the anonymized data, the associations between Q-I and sensitive data are broken, and the original data distribution is preserved. Consequently, the associations in individual records are broken, whereas the statistics of associations in the whole data set is preserved. We present the anonymization process in Algorithm 1.

Algorithm 1. The RA Algorithm

1: **Input** : the original data set D, the Q-I attributes Q, and the probability distribution $\{p_1, \ldots, p_m\}$, where $m = |Q|$, the number of attributes in Q-I.
2: **Output**: the original data set D is overwritten by an anonymized one
3: **begin**
4: $n := |D|$;
5: $Dist := \emptyset$;
6: **for** i: = 1 **to** m **do**
7: **begin**
8: $Dist_i$:= the distribution of the values of Q_i;
9: **end**
10: **for** j: = 1 **to** n **do**
11: **begin**
12: Randomly select an attribute Q_k in Q-I of the jth record with probability p_k;
13: Randomly generate a new value for Q_k based on $Dist_k$;
14: Replace the value of Q_k with the new value;
15: **end**
16: **end**

For example, in the data set in Fig. 1, the Q-I attributes $Q = \{Q_1, Q_2, Q_3\} = \{\text{Age}, \text{Job}, \text{Country}\}$. The values of Q_i and their corresponding distributions are:

Values of Q_1	[20–30]	[30–40]	[40–50]	[50–60]	[60–70]
$Dist_1$	1/10	4/10	1/10	3/10	1/10

Values of Q_2	Doctor	Clerk	Trader	Engineer	Banker
$Dist_2$	1/10	4/10	2/10	1/10	2/10

Values of Q_3	USA	Germany	UK	India
$Dist_3$	6/10	1/10	2/10	1/10

Setting $p_1 = p_2 = p_3 = 1/3$, for every record in Fig. 1, we equally likely select an attribute Q_k from the three attributes in the Q-I. Then we randomly generate a value for the selected Q_k based on $Dist_k$ and replace the original value of Q_k with the new value. Fig. 5 shows an anonymized data set produced by Algorithm 1.

How do we choose p_i in this algorithm? How does this algorithm protect privacy while preserving non-sensitive knowledge? We will address these issues in the following two sections.

4. Privacy analysis

Since our RA algorithm, unlike the *k*-anonymity and *l*-diversity methods, anonymizes a data set by randomly breaking the associations between its Q-I and sensitive data, we first define a statistical measurement of anonymity in Section 4.1. Then we use this definition to analyze the anonymity of our method. Moreover, since we randomize the values of the selected attributes, for comparison, we also use the measurement in Evfimievski, Gehrke, & Srikant (2003) to analyze the "privacy breaches" in our method in Section 4.2.1. In addition, in Section 4.2.2, we further discuss the privacy leakage when the user has some prior knowledge of the data distribution.

4.1. Probabilistic anonymity

For a good anonymization, it should be very unlikely that the user can infer the original associations from the corresponding associations in an anonymized data set. Thus we propose the following definition of anonymity.

Quasi-identifier			Sensitive Data
Age	Job	Country	Disease
[50–60]	Doctor	USA	Hypertension
[40–50]	Clerk	USA	Hypertension
[20–30]	Clerk	USA	Hypertension
[30–40]	Clerk	USA	Hypertension
[50–60]	Trader	USA	Diabetes
[40–50]	Clerk	UK	Diabetes
[50–60]	Engineer	Germany	Diabetes
[20–30]	Clerk	UK	Heart Disease
[60–70]	Trader	USA	Cancer
[30–40]	Banker	India	Cancer

Fig. 5. An anonymized data set produced by Algorithm 1 from Fig. 1.

Definition 6 (*Probabilistic anonymity*). Suppose that a data set D is anonymized to D'. Let r be a record in D and $r' \in D'$ be its anonymized form. Denote $r(QI)$ as the value combination of the quasi-identifier in r. The probabilistic anonymity of data set D' is defined by $1/P(r(QI)|r'(QI))$, where $P(r(QI)|r'(QI))$ is the probability that $r(QI)$ (for all $r \in D$) may be inferred given $r'(QI)$.

The probabilistic anonymity gives a measurement of how unlikely the user can infer original associations. The greater the probabilistic anonymity, the less probable the user can guess the original data.

Now that we have a measurement of anonymity, we will show how to determine p_i in Algorithm 1 to maximize the anonymity of the data set produced by the algorithm.

Proposition 7. *Let $Q_i, i = 1, \ldots, m$ be the ith Q-I attribute (category attribute) in a data set D and Entropy(Q_i) be the value of the entropy of Q_i. Then the probabilistic anonymity of D', the anonymized form of D, reaches the maximal value when each p_i in Algorithm 1 is directly proportional to the value of $e^{Entropy(Q_i)}$.*

Proof 1. Let c_{ij} be the jth category value in Q_i and $Freq(c_{ij})$ be the frequency of c_{ij} in Q_i. We calculate $Pa(D')$, the probabilistic anonymity of D', by:

$$\begin{aligned}\ln Pa(D') &= -\sum_{i=1}^{m}\sum_{j=1}^{J_i}[p_i Freq(c_{ij}) \ln(p_i Freq(c_{ij}))] \\ &= \sum_{i=1}^{m}\sum_{j=1}^{J_i}\left[p_i Freq(c_{ij}) \ln\frac{1}{p_i}\right] \\ &\quad + \sum_{i=1}^{m}\sum_{j=1}^{J_i}\left[p_i Freq(c_{ij}) \ln\frac{1}{Freq(c_{ij})}\right],\end{aligned}$$

where J_i is the number of the category values in Q_i. Since $\sum_{j=1}^{J_i} Freq(c_{ij}) = 1$, we can also derive:

$$\ln Pa(D') = \sum_{i=1}^{m} p_i(-\ln p_i + Entropy(Q_i)) \quad (1)$$

Next, we use the method of Lagrange Multipliers (Chen, Jin, Zhu, & Ouyang, 1983) to find the maximal value of $Pa(D')$. We incorporate the constraint $\sum_{i=1}^{m} p_i = 1$ into (1):

$$F(p_1, \ldots, p_m, \mu) = \sum_{i=1}^{m}[p_i(-\ln p_i + Entropy(Q_i))] + \mu\left(\sum_{i=1}^{m} p_i - 1\right)$$

where μ is an unknown scalar. Setting $\nabla_{p_1,\ldots,p_m} F(p_1, \ldots, p_m, \mu) = 0$, we have:

$$\begin{cases} \frac{\partial F}{\partial p_1} = Entropy(Q_1) - \ln p_1 - 1 + \mu = 0 \\ \cdots \\ \frac{\partial F}{\partial p_m} = Entropy(Q_m) - \ln p_m - 1 + \mu = 0. \end{cases}$$

Thus, for all $i, j \in [1, m]$, we have

$$\ln\frac{p_i}{p_j} = Entropy(Q_i) - Entropy(Q_j),$$

implying that

$$\frac{p_i}{p_j} = \frac{e^{Entropy(Q_i)}}{e^{Entropy(Q_j)}}. \quad (2)$$

From (2), when $p_i = \frac{e^{Entropy(Q_i)}}{\sum_{j=1}^{m} e^{Entropy(Q_j)}}$, $Pa(D')$ reaches its maximal value. □

The above proposition also shows a general method for calculating the probabilistic anonymity. When $p_i = \frac{1}{m}, i = 1, \ldots, m$, we can have a quick estimation of the scaled $Pa(D')$ by taking the geometric mean of the diversities of all Q-I attributes:

$$\ln Pa(D') = \ln m + \sum_{i=1}^{m} \left(\frac{1}{m} Entropy(Q_i) \right) = \ln \left(m \left(\prod_{i=1}^{m} Diversity_i \right)^{\frac{1}{m}} \right), \tag{3}$$

where $Diversity_i = e^{Entropy(Q_i)}$ is the entropy diversity of Q_i (Machanavajjhala et al., 2006). For an arbitrary record, the probability that the user may guess its original Q-I values, given an anonymized data set, is $1/Pa(D')$. This is also the user's confidence in associating a sensitive value with an individual. From (3), $Pa(D')$ is usually greater than the geometric mean of the diversities of all Q-I attributes. Similarly, $Pa(D')$ is usually greater than the diversity of the sensitive attribute. Even when the user is sure that an individual is in the data set, the user's maximal confidence in inferring the corresponding sensitivity is $1/Diversity_s$, where $Diversity_s$ is the diversity of the sensitive attribute. In comparison, in the l-diversity method, the user's confidence in guessing the data privacy is $1/l$, which is usually greater than $1/Diversity_s$. Therefore, by disassociating the relationships between the Q-I and sensitive values, the RA method provides a better protection of the privacy than the l-diversity method does.

4.2. Privacy breaches

The measurement of probabilistic anonymity evaluates the general level of privacy protection. Although the privacy is protected on average, certain privacy breaches (Evfimievski et al., 2003) may still take place.

4.2.1. Common breaches

Suppose $q \in D$ is one of the original value combinations of the quasi-identifier. And $q' \in D'$ is its anonymized form. The privacy breaches can be represented by:

$$\text{there exists } q \in D \text{ and } q' \in D' :$$
$$P(q \subset t | q' \subset t') \gg P(o \subset t | q' \subset t'), \quad \text{for all } o \neq q,$$

where $t \in D$ is the original record and $t' \in D'$ is its anonymized form. By the above definition, if the user is able to get the original q with high confidence once q' appears in the anonymized data, then there are privacy breaches. Privacy breaches are quantified by the γ-amplification (Evfimievski et al., 2003):

Definition 8 (γ-*Amplification*). Suppose a data set D is anonymized to D'. For some value $v' \in D'$, if for all $v_1, v_2 \in D, \frac{P(v_1 \to v')}{P(v_2 \to v')} \leqslant \gamma$, then the anonymization is at most γ-amplifying for v'. The anonymization is at most γ-amplifying if it is at most γ-amplifying for all $v' \in D'$.

By the definition, once the anonymized value v' is revealed, when all candidate values are equally possible, i.e., they share a common probability, then the probability that the user may obtain its original value is at most γ times the common probability. Thus, the closer the value of γ approaches 1, the more difficult it is for the user to infer the original data. Note that $\gamma \geqslant 1$.

In the RA algorithm, each record is anonymized without referring to its original Q-I data. It seems that we have $\gamma = 1$. But the original data distribution is used to generate the new values for the selected attributes. How does this affect γ? In the following, we analyze the privacy breaches in our method.

Suppose a data set D is anonymized to D' by Algorithm 1. Let $q' = \{q'_1, q'_2, \ldots, q'_m\}$ be a value combination of the quasi-identifier in D' where q'_i is the value of the ith Q-I attribute Q_i. We calculate the probability $P(q \to q')$ for each possible value combination $q \in D$:

$$\begin{cases} P(q = \{*, q'_2, q'_3, \ldots, q'_m\} \to q') = p_1 P(Q_1 = q'_1) \\ P(q = \{q'_1, *, q'_3, \ldots, q'_m\} \to q') = p_2 P(Q_2 = q'_2) \\ \ldots \\ P(q = \{q'_1, \ldots, q'_{m-1}, *\} \to q') = p_m P(Q_m = q'_m) \end{cases}$$

If there exists a k such that $p_k P(Q_k = q'_k) \geqslant p_i P(Q_i = q'_i)$ for all $i = 1, \ldots, m$, then there is a privacy breach by Definition 8. Although the user can infer with high confidence that q' is generated by anonymizing the kth attribute, he is unable to find the original value of the kth attribute, since each value is anonymized independently from the original one. Thus the user can hardly leverage the breaches in our method to identify the individuals or their sensitive data. However, things will be different when the user has some prior knowledge of the distributions of the Q-I data, especially the support of the value combinations of the quasi-identifier. This is probable when data set D contains all the individuals in the public data. In the next subsection, we analyze the data privacy that our method preserves in this situation.

4.2.2. Privacy breaches by prior knowledge

Let $t' \in D'$ be an anonymized record and $q' \subseteq t'$ be its Q-I value combination, then t' may be anonymized from one of the following two types of records in D:

- The records which contain q' both before and after the anonymization.
- The records whose Q-I values have intersection with q' of size $m - 1$ and are anonymized to q'.

Definition 9 (*Support*). The support of a value combination q in D, denoted by $supp_D(q)$, is the percentage of the records in D which contain all the attribute values in q. The support of a record t, denoted by $supp_D(t)$, is the support of its value combination.

A support gives the frequency of the occurrence of a value combination in D. Assuming $p_i = 1/m, i = 1, \ldots, m$, we have the expectation of the support for the first type of records in D':

$$\begin{aligned} &E(supp_{D'}(\{t' | q' \subset t \& q' \subset t'\})) \\ &\quad = supp_D(q') P(q' \text{is anonymized to itself}) \\ &\quad = supp_D(q') \frac{1}{m} \sum_{i=1}^{m} P(Q_i = q'_i). \end{aligned} \tag{4}$$

For the second type, $|q' \cap q| = m - 1$, assuming $p_i = 1/m, i = 1, \ldots, m$, the expectation of the support is:

$$\begin{aligned} &E(supp_{D'}(\{t' | q' \subset t' \& q \subset t\})) \\ &\quad = \sum_{q} (supp_D(q) P(q \text{ is anonymized to } q')) \\ &\quad = \sum_{q} \left(supp_D(q) \frac{1}{m} P\left(Q_{w(q',q)} = q'_{w(q',q)} \right) \right), \end{aligned} \tag{5}$$

where $w(q', q)$ is the index of the Q-I attribute in which q' and q differ. This time, by the prior knowledge, a breach for the user to guess the original value of the anonymized attribute may occur. Combining (4) and (5), we can derive the γ-amplification for the breach:

$$\gamma_{q'} = \max_{q} \left(\frac{supp_D(q') \cdot \sum_{i=1}^{m} P(Q_i = q'_i)}{supp_D(q) \cdot P(Q_{w(q',q)} = q'_{w(q',q)})} \right)$$

If $\gamma_{q'} < 1$, we set $\gamma_{q'} = 1/\gamma_{q'}$. The value of $\gamma_{q'}$ varies with the distributions of different data sets. It is desirable to have a constant mea-

surement. Thus, we derive an upper bound for the maximal probability that the user may infer the original q from q'. Let *maxSupp* be the larger of

$$\max_{|q\cap q'|=m-1}(supp_D(q)P\left(Q_{w(q',q)}=q'_{w(q',q)}\right) \quad \text{and}$$

$$supp_D(q')\sum_{i=1}^{m}P(Q_i=q'_i)$$

then the maximal probability is

$$\frac{maxSupp}{\sum_q\left(supp_D(q)P(Q_{w(q',q)}=q'_{w(q',q)})\right)+supp_D(q')\sum_{i=1}^{m}P(Q_i=q'_i)} \tag{6}$$

The following proposition gives an upper bound for the maximal probability.

Proposition 10. *Suppose $p_i = 1/m$, for $i = 1,\ldots,m$. Then the maximal probability that the user may guess the original data is bounded above by*

$$\max\left(\frac{supp_D(q')}{\min_i(supp_D(q'\ominus q'_i))},\frac{P(Q_{w(q',q)}=q'_{w(q',q)})}{\sum_{i=1}^{m}P(Q_i=q'_i)}\right).$$

Proof 2. We consider two cases based on the *maxSupp* in (6). In the first case when $maxSupp = supp_D(q')\sum_{i=1}^{m}P(Q_i=q'_i)$, the probability that the user may guess the data is

$$\begin{aligned}
&P(q\in D|q'\in D')\\
&\quad=\frac{supp_D(q')\sum_{i=1}^{m}P(Q_i=q'_i)}{\sum_{i=1}^{m}\left(P(Q_i=q'_i)\left(supp_D(q')+\sum_{w(q',q)=i}supp_D(q)\right)\right)}\\
&\quad=\frac{supp_D(q')\sum_{i=1}^{m}P(Q_i=q'_i)}{\sum_{i=1}^{m}\left(P(Q_i=q'_i)\cdot supp_D(q'\ominus q'_i)\right)}\\
&\quad\leqslant\frac{supp_D(q')}{\min_i(supp_D(q'\ominus q'_i))},
\end{aligned}\tag{7}$$

where $q'\ominus q'_i$ is the value combination obtained by deleting q'_i from q. If $q=\{q_i\}$, then $supp_D(q'\ominus q'_i)=1$. In the second case, $maxSupp = supp_D(q)P(Q_{w(q',q)}=q'_{w(q',q)})$ for some q, we have

$$\begin{aligned}
P(q\in D|q'\in D')&=\frac{supp_D(q)P(Q_{w(q',q)}=q'_{w(q',q)})}{\sum_{i=1}^{m}\left(P(Q_i=q'_i)supp_D(q'\ominus q'_i)\right)}\\
&\leqslant\frac{P(Q_{w(q',q)}=q'_{w(q',q)})}{\sum_{i=1}^{m}P(Q_i=q'_i)}.
\end{aligned}\tag{8}$$

Combining (7) and (8) completes the proof. □

In summary, without any prior knowledge, the user is unlikely to associate sensitive data with individuals from anonymized data set. Even with the knowledge of the distribution of the quasi-identifier, the user's ability is limited by the upper bound given in Proposition 10.

5. Quasi-sensitive knowledge preservation

Now that we have discussed the security of our anonymization method, in this section, we turn to the issue of knowledge preservation, which is important in data mining. Many *k*-anonymization methods (LeFevre et al., 2005, 2006; Meyerson & Williams, 2004) try to minimize the loss of data details while maximizing the level of privacy protection. Some recent algorithms (Fung, Wang, & Yu, 2007 ;Kifer & Gehrke, 2006) also try to retain the data utility for the purpose of data analysis and data mining. But few consider the issue of non-sensitive knowledge preservation. In Section 5.1, we first explain why we preserve the non-sensitive knowledge and also define its general form in the term "quasi-sensitive knowledge". Then in Section 5.2, we discuss the accuracy of knowledge recovery.

5.1. Quasi-sensitive knowledge

In most applications of data analysis and data mining, we are interested in finding out the data relationships among attributes, especially the associations between the Q-I and sensitive values. We call these associations the quasi-sensitive associations.

Definition 11 (*Quasi-sensitive associations*). Let Q be the set of Q-I attributes in data set D and S be the sensitive attributes. Suppose attribute sets $\hat{Q}\subseteq Q$ and $\hat{S}\subseteq S$. For any values $\hat{q}\in\hat{Q}$ and $\hat{s}\in\hat{S}$, we call the associations $\hat{q}\rightarrow\hat{s}$ the quasi-sensitive associations in D.

All the quasi-sensitive associations make up the quasi-sensitive knowledge in D.

For example, in the data set in Fig. 1, the two of its quasi-sensitive associations:

- Clerk → Hypertension (confidence: 75%)
- [30–40], Clerk → Hypertension (confidence: 67%)

which are helpful in analyzing the causes of hypertension.

We denote by $|\hat{q}|$ the number of values in $\hat{q}$ and $|Q|$ the number of attributes in Q, and call $|\hat{q}|$ the length of the association $\hat{q}\rightarrow\hat{s}$. We can see that normally the closer $|\hat{q}|$ approaches $|Q|$, the more sensitive the related association is. Especially when $|\hat{q}|=|Q|$, the related association is most sensitive because the antecedent part of the association rule can infer a specific individual. To preserve the less sensitive associations, we try to make the discovery of the "shorter" associations more accurate than the "longer" ones. In the following section, we will show how to achieve this goal.

5.2. Accuracy evaluation of knowledge recovery

We define the *relative difference*:

$$R(q)=\frac{supp_{D'}(q)-supp_D(q)}{supp_D(q)}\,100\%.$$

The frequently used measurement *relative bias* is represented by $B(q)=|E(R(q))|$. This provides a measurement of the difference between the support in D' and the support in D. Thus, from the definition of support, the smaller $B(q)$ is, the more accurate the knowledge discovery can be. In this section, we first derive the general form of the relative error in our method. Then, we analyze how the quasi-sensitive associations are treated with respect to their length.

Recall that Algorithm 1 randomly selects Q-I attribute for anonymization. Intuitively, the more Q-I attributes q involves, the more likely that q will be changed. Consequently, it is more likely that the difference $R(q)$ will be larger. The following proposition shows a relation between the relative bias $B(q)$ and the length of q. First we introduce some notations. Again, let q be a value combination in the original Q-I data. We denote $\neg q$ the set of value combinations: $\{\hat{q}$ such that $|\hat{q}|=|q|$ & $QI(\hat{q})=QI(q)$ & $\hat{q}\neq q\}$. Suppose $q=\{q_1,\ldots,q_l\}, l\leqslant m$, we denote $q\ominus q_i, i=1,\ldots,l$, the value combination $\{q_1,\ldots,q_{i-1},q_{i+1},\ldots,q_l\}$, and $q\oplus q_x$ the value combination $\{q_1,\ldots,q_l,q_x\}$. When $|q|=1$, we define $supp_D(q\ominus q)=1$. We also use $lift(q_i,q\ominus q_i)=\frac{supp(q)}{supp(q\ominus q_i)P(Q_i=q_i)}$ to measure the strength of the relationship between q_i and $q\ominus q_i$. A lift value greater than 1 indicates there is a positive association between q_i and $q\ominus q_i$, whereas a value less than 1 indicates there is a negative association (Giudici, 2003).

Proposition 12. *For a value combination q in D, the relative bias $B(q)$ is positively proportional to $\left|\sum_i\left[\frac{1}{lift(q_i,q\ominus q_i)}-1\right]\right|$.*

Proof 3. From the definition of $R(q)$, we have

$$\begin{aligned} E(R(q)) &= E\left(\frac{supp_{D'}(q) - supp_D(q)}{supp_D(q)}\right) \\ &= \frac{supp_D(q)P(q \to q) + supp_D(\neg q)P(\neg q \to q) - supp_D(q)}{supp_D(q)} \\ &= \frac{supp_D(\neg q)}{supp_D(q)} P(\neg q \to q) - P(q \to \neg q). \end{aligned} \tag{9}$$

Then we get

$$\begin{aligned} B(q) &= |E(R(q))| \\ &= \frac{1}{m}\left|\sum_{i=1}^{|q|}\left[\frac{supp_D(q \ominus q_i) - supp_D(q)}{supp_D(q)} P(Q_i = q_i) - P(Q_i \neq q_i)\right]\right| \\ &= \frac{1}{m}\left|\sum_{i=1}^{|q|}\left[\frac{1}{lift(q_i, q \ominus q_i)} - 1\right]\right|, \end{aligned} \tag{10}$$

which completes the proof. □

The proposition shows that the more attributes q has, the larger the range of relative bias $B(q)$ is. In particular, when $lift(q_i, q \ominus q_i) \geqslant 0.5$, then $B(q) \leqslant \frac{|q|}{m}$, implying that the range of the expected relative difference between $supp_{D'}$ and $supp_D$ increases linearly with the length $|q|$.

As for the variance of the relative difference, we have:

$$Var(R(q)) = \frac{1}{m^2}\sum_{i=1}^{|q|}\left(\frac{supp_D(q \ominus q_i)^2}{supp_D(q)^2}\left(P(Q_i = q_i) - P(Q_i = q_i)^2\right)\right). \tag{11}$$

Let value combination $\tilde{q} = q \oplus q_x, q_x \notin q$. If $\frac{supp_D(q \ominus q_i)}{supp_D(q)}$ varies little from $\frac{supp_D(\tilde{q} \ominus q_i)}{supp_D(\tilde{q})}$ for $i = 1, \ldots, |q|$, then $Var(R(\tilde{q})) \geqslant Var(R(q))$. That means $Var(R(q))$ grows almost linearly with the increase of $|q|$. In other words, the relative difference between $supp_{D'}(q)$ and $supp_D(q)$ becomes more uncertain as value combination gets longer. In particular, in the case of single attribute Q-I, where $m = 1$, from (10), we have zero bias:

$$B(q) = \left|\frac{1}{lift(q, q \ominus q)} - 1\right| = 0.$$

In summary, the shorter the length of value combination is, the better the quasi-sensitive knowledge is preserved, and the more accurate the knowledge recovery is. Also, since short associations are usually less sensitive than long ones, the user is more likely to discover the less sensitive associations than the sensitive ones.

6. Enhancement of privacy protection

Although the RA algorithm prevents the user from associating the individuals with their sensitive data, the prior knowledge of the associations of the Q-I data may worsen privacy leakage.

In Fig. 5, the probabilistic anonymity $Pa(D')$ of the anonymized data set is about 11 by Proposition 7. Suppose the user knows in advance the association: "Clerk, USA → [30–40]" (50%). Once the value combination {[20–30], Clerk, USA} occurs in the anonymized data, the user is able to guess with α confidence that the original Q-I value combination is {[30–40], Clerk, USA}, where

$$\alpha = P(\text{"age" is anonymized})\ 50\% \approx 17\% > \frac{1}{Pa(D')}.$$

Thus, the user's inference about the data privacy is improved. To tackle this problem, we randomly anonymize more than one Q-I attribute. This will improve privacy protection. However, knowledge recovery may become less accurate. In this section, we generalize Algorithm 1 by uniformly choosing $\lambda(\lambda \geqslant 1)$ Q-I attributes for each record and replacing their values by using their original distributions. We first present the algorithm, then analyze its privacy protection and knowledge preservation.

Algorithm 2. The Enhanced RA Algorithm

```
1: Input : the original data set D, the Q-I attributes Q, and the
   number λ(λ ⩽ m) of the attributes to be anonymized
2: Output : the original data set D is overwritten by an
   anonymized one
3: begin
4: m := |Q|;
5: n := |D|;
6: Dist := ∅;
7: for i := 1 to m do
8:    begin
9:    Dist_i := the distribution of the values of Q_i;
10:   end
11: for j := 1 to n do
12:   begin
13:   Randomly select λ attributes with equal probability in Q-I
      of the jth record;
14:   for k := 1 to λ do
15:      begin
16:      attr := kth selected Q-I attribute;
17:      Randomly generate a new value for attr based on Dist_attr;
18:      Replace the value of attr with the new value;
19:      end
20:   end
21: end
```

Apparently, the Algorithm 2 enhances privacy protection when λ gets larger. What is the impact of λ on knowledge preservation? In the following, we will show that both the bias $B(q)$ and variance $Var(R(q))$ increase as λ gets larger.

Since we choose the Q-I attributes uniformly at random, the probability that each combination of λ Q-I attributes gets selected is $1 \Big/ \binom{m}{\lambda}$. In the algorithm, a value combination cannot be anonymized to all the other forms. If two value combinations q and q' differ in η values ($\eta \leqslant \lambda$), then these two value combinations are convertible into each other, since all the η values may be anonymized. Before the discussion of $B(q)$ and $Var(E(q))$, in the following proposition, we investigate the impact of λ on the probability that a value combination is anonymized to its convertible value combination. We will show that both probabilities $P(q \to q')$ and $P(q' \to q)$ increase as λ increases.

Proposition 13. *Suppose λ is the number of the anonymized attributes in each record. Let $q \in D$ and $q' \in D'$ be partial value combinations of Q-I, $|q| = |q'| \leqslant m$, and they are convertible into each other. Then the probabilities $P(q \to q')$ and $P(q' \to q)$ increase directly as λ.*

Proof 4. Suppose that q and q' differ in η attributes, $\eta \leqslant \lambda$. Let $AttrSets^{\eta+k}$ be the set of all possible combinations of $\eta + k$ Q-I attributes where each combination $AttrSets_i^{\eta+k}, i = 1, \ldots, \binom{|q| - \eta}{k}$, contains those η attributes and other $k, k \leqslant |q| - \eta$, arbitrary Q-I attributes of q. Let $P(AttrSets_i^{\eta+k}, q')$ be the probability that the values of the $\eta + k$ attributes in $AttrSets_i^{\eta+k}$ are anonymized to the

corresponding values of q'. Then probability that q is anonymized to q' is:

$$P(q \to q') = \sum_{k=0}^{\min(\lambda-\eta,|q|-\eta)} \left[\frac{\binom{m-|q|}{\lambda-\eta-k}}{\binom{m}{\lambda}} \sum_{i=1}^{|AttrSets^{\eta+k}|} P(AttrSets_i^{\eta+k}, q') \right] = \sum_{k=0}^{\min(\lambda-\eta,|q|-\eta)} Pr_k^{\lambda,\eta}$$

where, for simplicity, $Pr_k^{\lambda,\eta}$ denotes $\frac{\binom{m-|q|}{\lambda-\eta-k}}{\binom{m}{\lambda}} \sum_{i=1}^{|AttrSets^{\eta+k}|} P(AttrSets_i^{\eta+k}, q')$. As λ is increased by 1,

$$\frac{Pr_k^{\lambda+1,\eta}}{Pr_k^{\lambda,\eta}} = \frac{\frac{\binom{m-|q|}{\lambda+1-\eta-k}}{\binom{m}{\lambda+1}}}{\frac{\binom{m-|q|}{\lambda-\eta-k}}{\binom{m}{\lambda}}} = \frac{(\lambda+1)(m-\lambda+1)}{(\lambda+1-\eta-k)(m-\lambda+1-|q|+\eta+k)}.$$

It then follows that

$$1 \leqslant \frac{Pr_k^{\lambda+1,\eta}}{Pr_k^{\lambda,\eta}} \leqslant \lambda+1,$$

since $\eta + k \leqslant \min(\lambda, |q|)$.

This shows that the probability $P(q \to q')$ increases directly as λ. The proof for $P(q' \to q)$ is similar. □

Now let us look at $B(q')$ and $Var(R(q'))$. From (9), we get

$$B(q') = \left| \sum_{\{q \in \neg q'\}} \left[\frac{supp_D(q)}{supp_D(q')} P(q \to q') \right] - P(q' \to q) \right|$$

and

$$Var(R(q')) = \sum_{\{q \in \neg q'\}} \left[\left(\frac{supp_D(q)}{supp_D(q')} \right)^2 (P(q \to q') - P(q \to q')^2) + \left(P(q' \to \neg q') - P(q' \to \neg q')^2 \right) \right].$$

As λ increases, there are more value combinations q in D convertible to q'. Also, from Proposition 13, both $P(q \to q')$ and $P(q' \to q)$ vary directly as λ. Therefore, the range of $B(q)$ increases with λ. As for the variance, $P(q \to q')$ and $P(q' \to q)$ are usually less than 1/2, then $P(q \to q') - P(q \to q')^2$ and $P(q' \to \neg q') - P(q' \to \neg q')^2$ increase with λ. Thus, the variance also varies directly as λ.

In summary, by tuning the value of λ, we can adjust the level of the average bias and the variance. Thus, we need to find a compromise between the protection of data privacy and the preservation of useful knowledge, as demonstrated in our experiments.

7. Experiments

In this section, we present our experiment results to demonstrate the effectiveness of our methods. By common aggregate queries, we first compare our RA Algorithm 1 with the l-diversity and permutation methods on the SQL queries. Then we compare our method with the permutation method on the quasi-sensitive knowledge discovery. At the same time, we demonstrate the effects of the parameter λ in our enhanced RA Algorithm 2 and the association length on knowledge discovery.

7.1. Experiment setup

The data set adopted in the experiments is the adult data set from the UCI machine learning repository (Hettich & Bay, 1999). We used the following nine original attributes to compose the quasi-identifier: "education", "race", "sex", "work-class", "marital-status", "age", "relationship", "native-country" and "salary". The attribute "occupation" was retained as the sensitive attribute. The records with missing values were removed and the resulting data set contained 30,162 records.

7.2. Accuracy of SQL queries

In the first part of the experiments, we generated three anonymized data sets by respectively applying the entropy l-diversity, permutation, and our RA Algorithm 1 to the test data set. We implemented the entropy l-diversity method based on the k-anonymization algorithm in LeFevre et al. (2006). While the permutation algorithm was implemented as in Xiao & Tao (2006). Then, we performed random queries on the three anonymized data sets and compared the average accuracy of their answers. The queries were in the forms similar to those in Xiao & Tao (2006):

```
select count (*) from tablename
where Q1 in (R1) and ... and Qm in (Rm) and S in (Rs)
```

where Q_i is the ith Q-I attribute, S the sensitive attribute, and R_i the query values for the ith attribute.

7.2.1. Record count estimation

When a data set is anonymized by the l-diversity algorithm, we cannot run the queries directly since the data is generalized. Instead, the record count of each query is obtained by the following estimate (Xiao & Tao, 2006):

$$\sum_p \left(\#_p(R_s) \prod_{i=1}^{m} \frac{\#_p(R_i \cap Values(Q_i))}{\#_p(Values(Q_i))} \right), \tag{12}$$

where $Values(Q_i)$ is the set of the distinct values of attribute Q_i and function $\#_p(x)$ counts in the pth Q-I group the occurrence of the values in the set x. By summing up the estimated record counts in all the Q-I groups, we get an estimate for the query.

When a data set is anonymized by the permutation method, the answers to the queries are estimated by

$$\sum_p \left(\#_p(R_s) \frac{\#_p(\text{records matching the Q-I conditions})}{\#_p(\text{records in the } p \text{ th group})} \right). \tag{13}$$

In this case, in the pth Q-I group, the records matching the conditions of Q-I in the `where` clause in a query can be directly retrieved as $\#_p$ (records matching the Q-I conditions). The product in the above equation assumes all the associations between the value combinations of Q-I and sensitive data are equally possible. We call this kind of associations the uniform associations between the Q-I and sensitive values.

When the data set is anonymized by the RA algorithm, we retrieve the queries directly without estimation.

7.2.2. The results

We conducted eight sets of experiments. In the jth set, $1 \leqslant j \leqslant 8$, of experiments, we performed 1000 queries on the three anonymized data sets and each query was composed of the sensitive attribute and j attributes which were selected from the quasi-identifier uniformly at random. For each selected attribute, we chose a random number of categorical values from its domain to form its query values R_i. In each set of experiments, we

calculated the average relative errors of the query results. The relative error is defined by

$$\frac{|Ans(query, D) - Ans(query, D')|}{Ans(query, D)},$$

where $Ans(query, D)$ returns the answer to *query* on data set D. Since the numerator $|Ans(query, D) - Ans(query, D')|$ compares the aggregated counts between D and D', the above definition of relative error may exceed 1.

The probabilistic anonymity of the test data set reached 34 by our RA anonymization method, meaning that the average probability that a user could correctly associate an individual with the corresponding sensitive value was 1/34. While the l-diversity and permutation methods limited that probability to $1/l$, which was further limited by the distribution of the sensitive values. In the test data set, the maximal value of l was about 10 (Entropy(occupation)$\approx \ln 10$). Thus, we set $l = 10$ for the l-diversity and permutation methods.

As shown in Fig. 6a, when the queries become more sensitive, involving more attributes, the relative error in the permuted data varies less than those in the l-diverse data and the RA anonymized data. This is because the estimate (12) for the results of the queries on the l-diverse data assumes a uniform value distribution of each Q-I attribute. Thus, the more Q-I attributes are involved, the farther away their joint distribution is from the multivariate uniform distribution. As for the RA method, from Proposition 12, the frequent and infrequent value combinations usually have higher values of $B(q)$ than other value combinations. Since the proportions of the frequent and infrequent value combinations decrease when the number of attributes involved increases, the rise of the relative error slows down as the number of attributes involved increases, as shown in Fig. 6a. For the permuted data set, the data links in the original data set are not completely uniform associations assumed by (13), causing the errors in the queries. Since the strength of those associations is not influenced by the number of the attributes involved in the queries, the relative error varies the least.

We then decreased the value of diversity l to 8 and 5 and repeated the comparison. This time, each query in the experiment consisted of the sensitive attribute and 5 random Q-I attributes. As shown in Fig. 6b, although the accuracy of the estimations in the l-diverse data and permuted data improved as l decreased, the estimation from the anonymized data set by our RA method consistently had the least amount of error. This is because the estimation from the RA anonymized data set assumes no relationships between the attributes. The RA algorithm tries to preserve the data relationships by perturbing a small portion of data set.

7.3. Quasi-sensitive knowledge preservation

In this section, we investigate knowledge preservation in several anonymization techniques, the effect of the parameter λ in the enhanced RA Algorithm 2 on knowledge discovery, and the impact of the association length on knowledge discovery.

We first ran the Apriori algorithm (Agrawal & Srikant, 1994) on the original data set D to get the quasi-sensitive associations $Asso_D$, which was used as the baseline result in the following experiments. Then we compared baseline results with the quasi-sensitive associations from each anonymized data set D'. In the comparison, we evaluated the accuracy of the knowledge discovery by calculating the "relative percentage":

$$\frac{|Asso_D \cap Asso_{D'}|}{|Asso_D|} 100\%.$$

In the test data set, the highest frequency of the values of attribute "occupation" was about 13%. Thus, in the Apriori algorithm, we set the confidence threshold to 15% and the support threshold to 8% .

Fig. 7 compares the effect of our RA Algorithm 1 with that of the permutation method with $l = 5, 8, 10$ (We did not compare with l-diversity and k-anonymization since most data associations are suppressed there). More quasi-sensitive associations were recovered from our RA anonymized data than from the permuted data. Although the previous experiments of aggregate queries showed that the accuracy of the answers from the permuted data varied the least, in the current experiment, only a small part of the associations was recovered. This is caused by the assumption of the uniform associations between the Q-I and sensitive values. In the previous experiments, most queries consisted of independent attribute values. For example, in the first set of the previous experiments, the queries consisted of the sensitive attribute and one random Q-I attribute. We ran the Chi-Square test for each contingency table consisting of a pair of attributes in the queries. Most

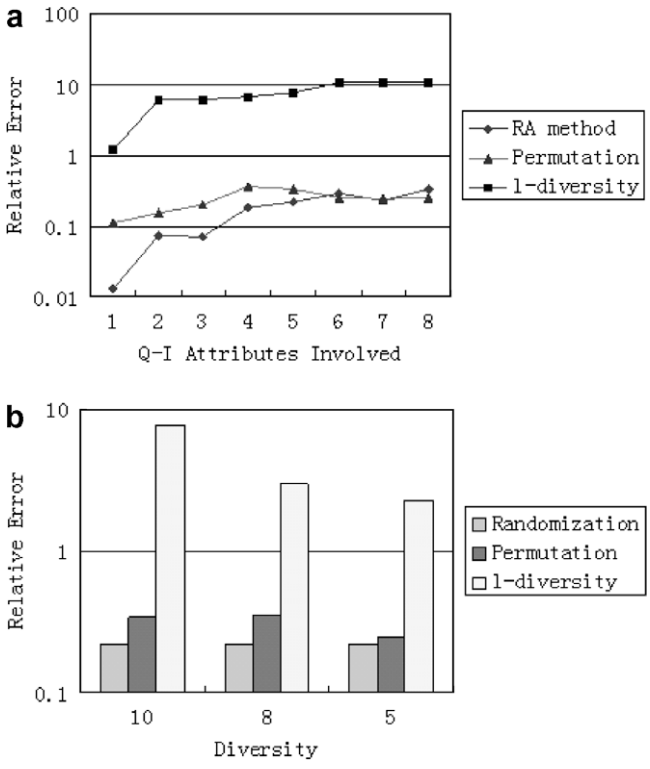

Fig. 6. (a) Comparison of the query accuracy among the data sets by different anonymization methods ($l = 10$); (b) comparison of the query accuracy with different diversity values.

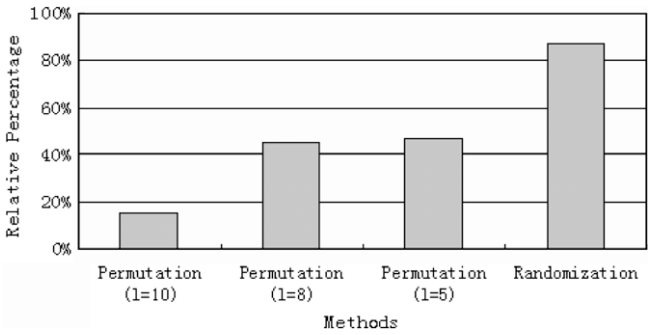

Fig. 7. Comparison of the quasi-sensitive associations among different anonymization methods.

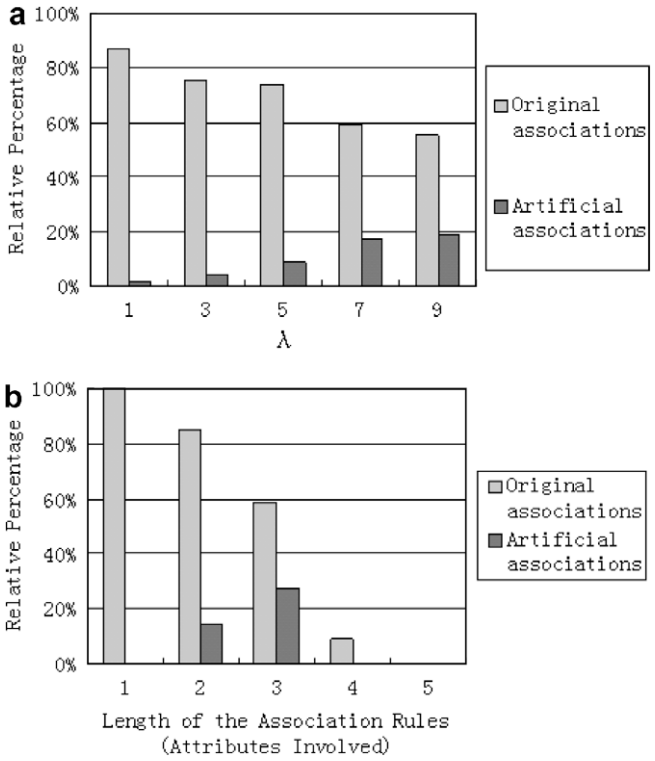

Fig. 8. (a) Discovery of the quasi-sensitive associations with different values of λ in the enhanced RA method; (b) discovery of the quasi-sensitive associations with different length ($\lambda = 5$).

results of $\frac{\chi^2}{|D|}$ was less than 0.2. Thus, the relative errors from the queries consisting of associated attribute values contributed little to the average error. However, in the association discovery, it only discovered those highly associated attribute values. The more interesting the associations are, the farther away they are from the uniform relationship. While in our RA method, the data randomization had a relatively small impact on the data relationships. Therefore, it was capable of preserving more associations.

To test the influence of λ on the discovery of the quasi-sensitive knowledge, we implemented the enhanced RA Algorithm 2 with various numbers of Q-I attributes anonymized. We set the parameter λ to 1,3,5,7,9.

As shown in Fig. 8a, with the increase of λ, more artificial associations were generated and fewer original associations were discovered. This confirms our conclusions in Section 6.

Since a quasi-sensitive association may cause more potential threats to the data privacy when it involves more attribute values, we further compared the discovery of the quasi-sensitive associations with various association lengths. In the comparison, we set $\lambda = 5$.

Fig. 8b shows that it is more accurate to recover the shorter associations than the longer ones. Most of the associations with 1, 2 or 3 attributes were discovered even when more than half of the attribute values were anonymized in each record. (Here, we denote by length 1 the frequent items in the data set.) This means that the user is more likely to discover less sensitive associations.

The above experiments demonstrate the following characteristics of our methods. The RA Algorithm 1 is capable of protecting data privacy with minor impact on the data distribution. The enhanced RA Algorithm 2 can further decrease the risk of privacy leakage when some original data associations are disclosed. Moreover, by tuning the value of λ, it can make a balance between the useful knowledge preservation and privacy protection.

8. Conclusions

In this paper, we introduce a novel data anonymization method by randomizing the data records. Different from the generalization methods such as the *k*-anonymization methods, we regard the data privacy as the links between the Q-I and sensitive values. By randomly replacing part of the values in each record while maintaining statistical relations in whole data set, the RA method not only achieves a higher level of the privacy protection but also preserves more non-sensitive knowledge than the other anonymization methods, as demonstrated by our experiments. Moreover, the useful associations which are less sensitive can be discovered more accurately than the sensitive ones. In particular, when association length is 1, our method delivers zero bias. Furthermore, the enhanced RA method provides the extra privacy protection when the user has some prior knowledge of the Q-I data.

Data anonymization is a popular direction in the research of privacy preserving data mining. Integrating our method with other privacy preserving techniques, we may accomplish more tasks. This work serves as an initial step. Additional research will include more work on other types of data knowledge and the definition of data privacy.

References

Aggarwal, C. C. (2005). On *k*-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on very large data bases* (pp. 901–909). Trondheim, Norway.

Agrawal, D., & Aggarwal, C. C. (2001). On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems* (pp. 247–255). Santa Barbara, California.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th international conference on very large data bases* (pp. 487–499). Santiago, Chile.

Agrawal, R., & Srikant R. (2000). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 439–450). Dallas, Texas.

Aggarwal, C. C., & Yu, P. S. (2008). *A general survey of privacy-preserving data mining models and algorithms. Privacy-preserving data mining* (pp. 11–52). US: Springer.

Chen, C., Jin, F., Zhu, X., & Ouyang, G. (1983). *Mathematical analysis*. Higher Education Press.

Ciriani, V., Vimercati, S., Foresti, S., & Samarati, P. (2008). *k-Anonymous data mining: A survey. Privacy-preserving data mining* (pp. 105–136). US: Springer.

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, M. Y. (2002). Tools for privacy preserving distributed data mining. *ACM SIGKDD Explorations Newsletter, 4*(2), 28–34.

Du, W., & Zhan, Z. (2003). Using randomized response techniques for privacy-preserving data mining. In *Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 505–510). Washington, DC.

Evfimievski, A., Gehrke, J., & Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the 22th ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems* (pp. 211–222). San Diego, California.

Fung, B. C. M., Wang, K., & Yu, P. S. (2007). Anonymizing classification data for privacy preservation. *IEEE Transactions on Knowledge and Data Engineering, 19*(5), 711–725.

Giudici, P. (2003). *Applied data mining: Statistical methods for business and industry*. John Wiley & Sons. Ltd.

Hettich, S., & Bay, S. D. (1999). The UCI KDD archive. URL : http://kdd.ics.uci.edu.

Kifer, D., & Gehrke, J. (2006). Injecting utility into anonymized datasets. In *Proceedings of the 2006 ACM SIGMOD international conference on management of data* (pp. 217–228). Chicago, Illinois.

Koudas, N., Srivastava, D., Yu, T., & Zhang, Q. (2007). Aggregate query answering on anonymized tables. In *Proceedings of the 23rd international conference on data engineering* (pp. 116–125). Istanbul, Turkey.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2005). Incognito: Efficient full-domain *k*-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on management of data* (pp. 49–60). Baltimore, Maryland.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006). Mondrian multidimensional *k*-anonymity. In *Proceedings of the 22nd international conference on data engineering* (p. 25). Atlanta, Georgia.

Li, N., Li, T., & Venkatasubramanian, S. (2007). *t*-Closeness: Privacy beyond *k*-anonymity and *l*-diversity. In *Proceedings of the 23rd international conference on data engineering* (pp. 106–115). Istanbul, Turkey.

Lindell, Y., & Pinkas, B. (2000). Privacy preserving data mining. In *Proceedings of the 20th annual international cryptology conference on advances in cryptology* (pp. 36–54). Santa Barbara, California.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). *l*-Diversity: Privacy beyond *k*-anonymity. In *Proceedings of the 22th international conference on data engineering* (p. 24). Los Alamitos, California.

Meyerson, A., & Williams, R. (2004). On the complexity of optimal *k*-anonymity. In *Proceedings of the 23rd ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems* (pp. 223–228). Paris, France.

Sweeney, L. (2002). Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, 10*(5), 571–588.

Xiao, X., & Tao, Y. (2006). Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd international conference on very large data bases* (pp. 139–150). Seoul, Korea.