J Intell Inf Syst (2016) 47:427–445
DOI 10.1007/s10844-015-0373-4

Permutation anonymization

Dong Li[1] · Xianmang He[2,4] · LongBin Cao[3] ·
Huahui Chen[2]

Received: 29 April 2013 / Revised: 16 May 2015 / Accepted: 14 July 2015 /
Published online: 4 August 2015

Abstract In data publishing, anonymization techniques have been designed to provide privacy protection. Anatomy is an important techniques for privacy preserving in data publication and attracts considerable attention in the literature. However, anatomy is fragile under background knowledge attack and the presence attack. In addition, anatomy can only be applied into limited applications. To overcome these drawbacks, we propose an improved version of anatomy: permutation anonymization, a new anonymization technique that is more effective than anatomy in privacy protection, and in the meanwhile is able to retain significantly more information in the microdata. We present the detail of the technique and build the underlying theory of the technique. Extensive experiments on real data are conducted, showing that our technique allows highly effective data analysis, while offering strong privacy guarantees.

Keywords Privacy preservation · Generalization · Anatomy · Permutation anonymization

✉ Xianmang He
hexianmang@nbu.edu.cn

Dong Li
lidong@nsfc.gov.cn

LongBin Cao
LongBing.Cao@uts.edu.au

Huahui Chen
chenhuahui@nbu.edu.cn

[1] Information Center, National Natural Science Foundation of China, Beijing, China

[2] School of Information Science and Engineering, Ningbo University, Ningbo, Zhejiang, China

[3] Advanced Analytics Institute, University of Technology Sydney, Ultimo, NSW, Australia

[4] School of Computer Science and Technology, Fudan University, Shanghai, China

1 Introduction

Nowadays, partly driven by many web applications, more and more data has been made publicly available and analyzed in one way or another. Privacy preserving publishing of sensitive data becomes a more and more important concern. Numerous organizations, like census bureaus and hospitals, maintain large collections of personal information (e.g., census data and medical records). Such data collections are of significant research value, and there is much benefit in making them publicly available. For example, medical records of patients may be released by a hospital to aid the medical study. Assume that a hospital wants to publish records of Table 1, which is called as microdata (T). Nevertheless, as the data is sensitive in nature, in our example, attribute *Disease* is sensitive, we need to ensure that no adversary can accurately infer the disease of any patient from the published data. For this purpose, any unique identifier of patients, such as *Name* should be anonymized or excluded from the published data. However, it is still possible for the privacy leakage if adversaries have certain background knowledge about patients. For example, if an adversary (targeting Bob's medical record) knows that Bob is of age 65 and Sex M, s/he can infer that Bob's disease is Emphysema since Age together with Sex uniquely identify each patient in Table 1. The attribute set that uniquely identify each record in a table is usually referred to as a quasi-identifier(QI for short) (Samarati and Sweeney 1998; Sweeney 2002; Samarati 2001) of the table.

The proper measures must be taken to ensure that its publication does not endanger the privacy of the individuals that contributed the data. A plethora of techniques have been proposed for privacy preserving data publishing. One way to overcome above threat is anatomy (Xiao and Tao 2006). In a typical anatomy-based solution, we first need to divide tuples into subsets (each subset is referred to as a QI-group). For example, tuples in Table 1 can be partitioned into two subsets $\{Bob, Alex, Jane, Lily, Andy\}$ (with Group-ID 1) and $\{Mary, Linda, Lucy, Sarah\}$ (with Group-ID 2), as indicated by the group ID(GID) in Table 2. Then, we release the projection of microdata on quasi-identifiers as a quasi-identifier table(QIT), and projection on sensitive attribute as a sensitive table(ST). The group ID is also added to QIT and ST. In this way, QI-attributes and sensitive attributes are distributed into two separate tables. For example, by anatomy, Table 1 is separated into two tables as shown in Table 3.

Now, given the published table of anatomy, an adversary with background knowledge of Bob can only infer that Bob belongs to group 1 from QIT and has only probability

Table 1 Microdata

Name	Age	Sex	Disease
Bob	65	M	Emphysema
Alex	50	M	Cancer
Jane	70	F	Flu
Lily	55	F	Gastritic
Andy	90	F	Dyspepsia
Mary	45	M	Flu
Linda	50	F	Pneumonia
Lucy	40	F	Gastritic
Sarah	10	M	Bronchitis

Table 2 A 4-diversity table

GID	Age	Sex	Disease
1	[50–90]	F/M	Emphysema
1	[50–90]	F/M	Cancer
1	[50–90]	F/M	Flu
1	[50–90]	F/M	Gastritic
1	[50–90]	F/M	Dyspepsia
2	[10–50]	F/M	Flu
2	[10–50]	F/M	Pneumonia
2	[10–50]	F/M	Gastritic
2	[10–50]	F/M	Bronchitis

$1/5$ to infer Bob's actual disease from ST. On the other hand, anatomy captures the exact QI-values, which retains a larger amount of data characteristics. Consequently, *anatomy is believed to be a anonymization solution of high data quality compared to generalization-based solutions that lead to heavy information loss* (Xiao and Tao 2006; Tao et al. 2009). We illustrate this in Example 1.

Example 1 Suppose, we need to estimate the following query:

$$\text{Select Count}(*) \text{ From } T \text{ Where Age } \in [40, 70] \text{ And Sex } =' \text{F}' \text{ And Disease } =' \text{Flu}'.$$

If T is the original microdata, i.e. Table 1, we get the accurate result, which is 1. However, if we evaluate the query from the result produced by a generalization based approach, large gross error will occur. Table 2 is a typical result by generalizing Table 1. In Table 2, each QI-attribute in the QI-group is replaced by a less specific form. For example, age 65 is replaced by [50, 90], sex M is replaced by F/M. When estimate the above query from Table 2, without additional knowledge, the researcher generally assumes uniform data distribution in the generalization Table 2. Consequently, we obtain an approximate answer $\frac{1}{4} + \frac{1}{8} = \frac{3}{8}$, which is much smaller than the accurate result. However, when we evaluate the same query from Table 3, we have the result result $\frac{2}{5} + \frac{2}{4} = 0.9$, which is quite close to the accurate result.

Table 3 Anatomy

	Age	Sex	Group-ID		Group-ID	Disease
QIT	65	M	1	ST	1	Emphysema
	50	M	1		1	Flu
	70	F	1		1	Cancer
	55	F	1		1	Dyspepsia
	90	F	1		1	Gastritis
	45	M	2		2	Flu
	50	F	2		2	Bronchitis
	40	F	2		2	Gastritic
	10	M	2		2	Pneumonia

1.1 Motivation

However, anatomy is still vulnerable to certain kind of background knowledge attack. For example, suppose an adversary knows that Bob is a 65-year-old male whose record is definitely involved in the microdata, the adversary can only find out that Bob is one of the first five records. With a random guess, the adversary's estimate of the probability that Bob has Emphysema is $\frac{1}{5}$. However, if adversaries have acquired the background knowledge about the correlations between Emphysema and the non-sensitive attributes Age and Sex, e.g., *'the prevalence of emphysema was appreciably higher for the 65 and older age group than the 10-64 age group for each race-sex group'* and *'the prevalence was higher in males than females and in whites than blacks'*, the adversary can infer that Bob has Emphysema rather than other diseases with high confidence.

In reality, numerous background knowledge, such as well-known facts, demographic information, public record and information about specific individuals etc., can be available to adversaries. Furthermore, in general, it is quite difficult for the data publisher to know exactly the background knowledge that will be used by an adversary. As a result of these facts, background knowledge attack arises to be one of great challenges for anatomy-based data anonymization solutions.

Besides the frangibility under background knowledge attack, limited applications of anatomy also motivate us to improve it. Note that in an anatomy solution, all QI-values are precisely disclosed, which is not permitted in those applications where presence attack is a critical concern (In a presence attack, an adversary with QI-values of an individual wants to find out whether this individual exists in the microdata). In some other real applications such as location-based services (LBS) (Kalnis et al. 2007; Mokbel et al. 2006), where all QI-attributes themselves are sensitive, the anatomy is not applicable any more.

Thus, we may wonder *whether we can improve the present anatomy techniques so that the final solution is more safe under background knowledge attack, and simultaneously can be applied into more applications such as those concerning presence attack or LBS*. To address this issue, we will propose *permutation anonymization* (PA) as an improved version of present anatomy solution. PA releases all the quasi-identifier and sensitive values after imposing random permutations in two separate tables. Combined with a grouping mechanism, this approach protects privacy, and captures a large amount of correlation in the microdata. We close this section by illustrating Example 2, which is to help us to understand the technique of permutation anonymization.

Example 2 Given Table 1 and one of its partition suggested in Table 2, one valid result of permutation anonymization is shown in Table 4. Suppose we need to evaluate the same query in Example 1, from Table 4, we have $\frac{4}{5} \times \frac{3}{5} + \frac{3}{4} \times \frac{2}{4} = 0.855$, which is quite close to the result of the anatomy approach.

1.2 Contribution and plan of this article

The work in this paper is motivated by the above-mentioned observations: the anatomy technique i) is fragile under background knowledge attack; and ii) can only be applied into limited applications. In view of the above, in this paper we present an approach, which we termed as permutation anonymization(PA), captures and quantifies their underlying what between QI-attributes. We present the details of the technique and theoretical properties of

Table 4 Permutation anonymization

	Age	Sex	Group-ID		Group-ID	Disease
PQT	50	F	1	PST	1	Emphysema
	90	M	1		1	Flu
	70	F	1		1	Cancer
	65	F	1		1	Dyspepsia
	55	M	1		1	Gastritis
	50	M	2		2	Flu
	10	F	2		2	Bronchitis
	45	M	2		2	Gastritic
	40	F	2		2	Pneumonia

the proposed approach. Based on this foundation, we also propose a generalization algorithm to implement it. The extensive experiments on real data sets show the performance and utility improvement of our approach.

In following texts, we will first formalize the major problem addressed in this paper and the main solution: PA in Section 2. In Section 3, we present an anonymization algorithm to implement PA. In Section 4, we experimentally evaluates the effectiveness of our technique. In Section 5, related works are reviewed. Finally, Section 6 concludes the paper.

2 Problem statement

In this section, we will first give the basic notations that will be used in the following texts. Then, we give the formal definition about permutation anonymization and present theoretic properties about this technique. We close this section by the definition of the major problem that will be addressed in this paper.

2.1 Basic notations

Given a microdata table $T(A_1, A_2, \cdots, A_n)$ that contains the private information of a set of individuals, and has n attributes $A_1, \cdots, A_n$, and a sensitive attribute (SA) A_s. A_s is categorical and every attribute $A_i (1 \leq i \leq n)$ can be either numerical or categorical. All attributes have finite and positive domains. For each tuple $t \in T$, $t.A_i (1 \leq i \leq n)$ denotes its value on A_i, and $t.A_s$ represents its SA value.

A *quasi-identifier* $QI = \{A_1, A_2, \cdots, A_d\} \subseteq \{A_1, A_2, \cdots, A_n\}$ is a minimal set of attributes, which can be joined with external information in order to reveal the personal identity of individual records. A *QI-group* of T is a subset of the tuples in T. A *partition* P *of* T is a set of disjoint QI-groups $QI_j (1 \leq j \leq m)$ whose union equals T (Namely, $T = \bigcup_j^m QI_j$). An *anonymization principle* (AP) is a constraint on a SA attribute. A partition P satisfies an AP if the SA attribute of every QI-group in P satisfies the constraint posed by the AP. Most notable principles include k-anonymity (Sweeney 2002; Samarati 2001), l-diversity (Machanavajjhala et al. 2006), t-Closenessc (Li et al. 2007),(ε, m)-anonymity (Li et al. 2008) etc. We are interested only in l-diverse partitions that can lead to provably good privacy guarantees.

A permutation on a set V, is an one-to-one mapping from V to itself. If $|V| = N$, there are overall $N!$ permutations on V. Let α be a permutation on a set V of tuples, we use $\alpha(t_i)$ to denote the image of t_i under α. We use S_V to denote the set of all permutations on V. If V is a QI-group of microdata T, for example, $V = QI_i$, we can independently uniformly select a permutation from S_{QI_i} at random.

2.2 The concept of permutation anonymization

Now we are ready to give the formal definition about Permutation Anonymization, which is given in the following definition.

Definition 1 (Permutation Anonymization(PA)) Let T be a table consisting of QI-attributes $A_i (1 \leq i \leq d)$ and sensitive attribute A_s. Given a partition P with m QI-groups on T, permutation anonymization is a procedure with (T, P) as input, which produces a quasi-identifier table PQT and a sensitive table PST satisfying following conditions:

(1) PQT is a table with schema $(A_1, A_2, \cdots, A_d, \text{Group-ID})$ such that for each tuple $t \in QI_i$, PQT has a tuple of the form: $\left(\alpha_{i_1}(t).A_1, \alpha_{i_2}(t).A_2, \cdots, \alpha_{i_d}(t).A_d, i\right)$, where $\{\alpha_{i_j} : (1 \leq j \leq d)\}$ is independently uniformly selected from S_{QI_i} at random.

(2) PST is a table with schema (Group-ID, A_s) such that for each tuple $t \in QI_i$, PST has a record of the form: $\left(i, \alpha_{i_s}(t).A_s\right)$, where α_{i_s} is uniformly selected from S_{QI_i} at random.

For instance, based on the 4-diverse (the l-diversity will be introduced in Defintion 2) partition suggested in Table 2, PA produces the PQT and PST in Table 4, respectively, as explained in Example 2. Similar to anatomy, PA also produce two tables. The main difference between anatomy and permutation anonymization is that: *anatomy directly releases all the QI-values without extra treatment, while PA releases attribute values after random permutation*. In this sense, anatomy can be considered as an improvement of anatomy. Random permutation of PA accounts for a variety of advantages of PA compared over anatomy. One of them is the stronger privacy preservation of PA. As shown in Theorem 1, an adversary has small probability to infer the sensitive value of a victim. Compared to naive anatomy under the same partition on microdata (where $Pr\{t.A_S = v\} = \frac{c_j(v)}{|QI_j.A_s|}$), the probability of privacy leakage of PA is significantly small(See the Theorem 1), which in the worst case equals to anatomy. Besides above advantages, PA also provides good enough data utility, which in most cases is close to data quality of the anatomy approach, and it is shown in Theorem 2. We illustrate this in Example 2.

Theorem 1 *Let T be a table with QI-attributes $A_i(1 \leq i \leq d)$ and sensitive attribute A_s, let P be a partition with m QI-groups on T. From an adversary's perspective, for any tuple $t \in QI_j$,*

$$Pr\{t.A_s = v\} = \frac{c_j(v)}{|QI_j.A_s| \cdot \prod_{i=1}^{d} |QI_j.A_i|} \tag{1}$$

where $c_j(v)$ is the number of tuples in QI_j with A_s as v, and $|QI_j.A_i|$ is the number of distinct values of QI_j on attribute A_i.

Proof Consider any tuple $t \in T$, which is contained in QI-group QI_j for some $j \in [1, m]$. The adversary, who attempts to find out $t[A_s]$, can obtain j from the PQT which, however, does not have A_s data. Hence, the adversary can only conjecture that $t.A_s$ equals one of the A_s values (pertinent to QI_j) summarized the PST. Without any other information, the adversary assumes that every tuple in QI_j has an equal chance to carry any A_s value relevant to QI_j , which leads to the above equation. □

Moreover, PA can be applied into applications such as LBS (Kalnis et al. 2007; Mokbel et al. 2006). In these applications, all attributes are sensitive. Permutation anonymization allows privacy preservation for such applications by permutating all the attributes. PA also exhibits stronger privacy preservation for presence attack, which is shown in Lemma 1 and illustrated in Example 3.

Lemma 1 *From an adversary's perspective, the probability (δ_t) to find out whether an individual t exists in the QI-group QI_j by the presence attack is at most*

$$\delta_t = \frac{\prod_i^d |n_{i_j}|}{|QI_j|^d} \tag{2}$$

where n_{i_j} denotes the number of the value $t.A_i$ on attribute A_i in QI_j.

Proof For each attribute A_i, consider that the probability of an individual $t.A_i$ exists in QI_j is $\frac{n_{i_j}}{|QI_j|}$. Thus, the lemma holds. □

Example 3 We explain Theorem 1 and Lemma 1 using Table 4. Suppose the victim is $Bob = \langle 65, M\rangle$ whose age and sex are available to adversaries, then from adversary's perspective, the probability that Bob contacted Emphysema is $\frac{10}{5\times2\times5}=\frac{1}{5}$, and the probability that Bob exists in QI_1 of Table 4 is $\frac{1}{5} \times \frac{2}{5} = \frac{2}{25}$.

2.3 Preserving correlation

In this section, we discuss the data correlation between QI-attributes and sensitive attribute. A good publication approach should preserve both privacy and data correlation between QI-attributes and sensitive attribute. Comparing a concrete query in Example 1 and Example 2, we have shown that PA allows more effective aggregation analysis than generalization, and it is not far away from that of the anatomy in practice. Next, we provide the underlying theoretical rationale.

The combination of attributes define a d-Dimension space DS. Every tuple in the table can be mapped to a point in DS. We model $t \in T$ as an approximate density function (pdf) $\eta_t(x) : DS \longrightarrow [0, 1]$:

$$\eta_t(x) = \begin{cases} 1, & \text{if}(x = t) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Now, we discuss published PA-tables. Assume QI_j as the QI-group containing the tuple t (in the underlying l-diverse partition). Let $v_1, v_2, \cdots, v_\lambda$ be all the distinct A_s values in the QI-group. Denote $c(v_h)(1 \leq h \leq \lambda)$ as the count value in the PST corresponding to v_h. The reconstructed pdf $\eta_t^{PA}(x)$ of t is

$$\eta_t^{PA}(x) = \begin{cases} c(v_1) \times \delta_t, \text{if } x = (t.A_1, t.A_2, \cdots t.A_d, v_1) \\ \cdots\cdots \\ c(v_\lambda) \times \delta_t, \text{if } x = (t.A_1, t.A_2, \cdots t.A_d, v_\lambda) \\ 0, \text{otherwise} \end{cases} \tag{4}$$

In the (4), δ_t is the probability of presence attack of tuple t. Given an approximate pdf η^{PA} (4), we quantify its error from the actual η_t(3) as follows:

$$Err(t) = \sum_{\forall t \in T} \int_{x \in DS} (\eta_t^{PA}(x) - \eta_t(x))^2 dx. \tag{5}$$

Naturally, taking into account all tuples $t \in T$, a good publication approach should minimize the following re-construction error(see Xiao and Tao 2006):

$$RCE = \sum_{\forall t \in T} Err(t) \tag{6}$$

The following Theorem 2 establishes the lower bound of the RCE (see the Equation 5) achievable by the approach PA.

Theorem 2 *RCE (5) satisfies*

$$RCE \geq \sum_{t \in T} (1 + l \times \delta_t^2 - 2\delta_t) \tag{7}$$

for any pair of PQT and PST, where $|T|$ is the cardinality of the microdata T, δ_t is the probability of presence attack of tuple t.

Proof Anatomized tables (Definition 4) are computed from an l-diverse partition. Assume that the partition contain m QI-groups $QI_1, \cdots, QI_m$. For each $j \in [1, m]$, use β_t to denote the average Err_t (5) for each tuple $t \in QI_j$. Thus, the re-construct error RCE (6) can be rewritten as

$$RCE = \sum_{j=1}^{m} (|QI_j|) \cdot \beta_t = \sum_{t \in T} \beta_t. \tag{8}$$

Without the loss of generality, assume that QI_j contains λ distinct A_s values $v_1, \cdots, v_\lambda$. In particular, there are $c(v_h)$ $(1 \leq h \leq \lambda)$ tuples in QI_j with A_s value v_h. Consider that an arbitrary tuple $t \in QI_j$ with A_s value v_h (for some $h \in [1, \lambda]$). The actual pdf η_t and approximate η_t^{PA} are given in (3) and (4), respectively. Thus, by (5), we have

$$Err_t = (1 - \delta_t \cdot c(v_h))^2 + \sum_{h'=1 \wedge h' \neq h}^{\lambda} (c(v_{h'}) \cdot \delta_t)^2. \tag{9}$$

For computing the average β_t of Err_t for each tuple $t \in QI_j$, we combine the above formula with the fact that $c(v_h)$ tuples have A_s value v_h:

$$\beta_t = \frac{\sum_{h=1}^{\lambda} c(v_h) \cdot ((1 - c(v_h) \cdot \delta_t)^2 + \sum_{h'=1, h' \neq h}^{\lambda} (\delta_t \cdot c(v_{h'}))^2}{|QI_j|}. \tag{10}$$

Therefore, it remains to solve the minimum β_t subject to the constraints

$$\sum_{h=1}^{\lambda} c(v_h) = |QI_j|, \tag{11}$$

and

$$c(v_h) \leq \frac{|QI_j|}{l} \text{, for all } h \in [1, \lambda] \tag{12}$$

Allow us to ignore the second constraint temporarily. Then, the minimization of β_t subject to the first constraint is a standard problem tackled by the Lagrange multiplier method. Application of the method result in $\beta \geq (1 - 2\delta_t + l \cdot {\delta_t}^2)$, where the equality holds only when $c(v_1) = c(v_2) = \cdots = c(v_h) = \frac{|QI_j|}{l}$.

Note that $\lambda \geq l$, consider that $\sum_{h=1}^{\lambda} c(v_h) \leq \lambda \cdot \frac{|QI_j|}{l}$. The left side of the inequality equals $|QI_j|$, therefore, the inequality indicates that $\lambda \geq l$.

Finally, we have: $\beta_t \geq \frac{|QI_j| \cdot ((1-\delta_t \cdot \frac{|QI_j|}{l})^2 + (\lambda - 1) \cdot {\delta_t}^2)}{|QI_j|} \geq (1 - \delta_t)^2 + (\lambda - 1) \cdot {\delta_t}^2 \geq (1 + {\delta_t}^2 - 2\delta_t) + (l - 1) \cdot {\delta_t}^2 = (1 + l \cdot {\delta_t}^2 - 2\delta_t)$. The theorem holds. □

The above theorem is a generalized version of Theorem 2 in the paper (Xiao and Tao 2006). In the anatomy-based approach, the probability of presence attack is bounded by $\frac{1}{l}$, that leads to the $RCE = \sum_{t \in T} (1 + l \times \delta^2 - 2\delta) = |T|(1 + l \times (\frac{1}{l})^2 - \frac{2}{l}) = |T|(1 - \frac{1}{l})$.

2.4 The optimality problem

Using PA, we can implement different security models, one of them is l-diversity (given in Definition 2), which is widely used in previous researches about privacy preservation and will be one of major objectives of this paper. Another aspect of privacy preservation is data utility. In general, high data quality or less information loss is expected. In this paper, we use normalized certainty penalty (Definition 3) to measure the information loss. Now, we are ready to give the formal definition about the problem that will be addressed in this paper.

Definition 2 (l-diversity Machanavajjhala et al. 2006) A generalized table T^* is l-diversity if each QI-group $QI_j \in T^*$ satisfies the following condition: let v be the most frequent A_s value in QI_j, and $c_j(v)$ be the number of tuples $t \in QI_j$, then $\frac{c_j(v)}{|QI_j|} \leq \frac{1}{l}$.

Definition 3 (Normalized Certainty Penalty(NCP) Xu et al. 2006) Suppose a table T is anonymized to T^*. In the domain of each attribute in T, suppose there exists a global order on all possible values in the domain. If a tuple t in T^* has range $[x_i, y_i]$ on attribute $A_i (1 \leq i \leq d)$, then the normalized certainty penalty in t on A_i is $NCP_{A_i}(t) = \frac{|y_i - x_i|}{|A_i|}$, where $|A_i|$ is the domain of the attribute A_i. For tuple t, the normalized certainty penalty in t is $NCP(t) = \sum_i^d NCP_{A_i}(t)$. The normalized certainty penalty in T is $\sum_{t \in T^*} NCP(t)$.

In general, a table T together with a partitioning P on T implicitly implies a generalization T^*, which is obtained by replacing each numeric value of $t.A_i$ by a range $[Min_i, Max_i]$ of $QI_i.A_i$ (Min_i, Max_i are the minimal and maximal value of $QI_i.A_i$), replacing each categorial value $t.A_j$ by $QI_i.A_j$. In following texts, to simplify description, we use *'NCP of a table T and its partitioning P'* to denote the NCP of their implicit generalization T^*. In some contexts with confusion, table T is also omitted.

Example 4 We calculate NCP for Table 2. Note that the domain of $\langle Age, Sex \rangle$ are $\langle [10 - 90], \{F, M\} \rangle$. The NCP of Table 2 is $\frac{90-50}{90-10} \times 5 + \frac{50-10}{90-10} \times 4 + \frac{2}{2} \times 9 = 13\frac{1}{2}$.

Definition 4 (The Optimality Problem) Given a table T and an integer l, we aim to generate PST and PQT for T by PA so that PST is l-diversity and NCP of PQT is minimized.

As shown in Xiao et al. (2010) the optimality problem to minimize the information loss is NP-hard. In the following section, we will introduce an algorithm to provide an approximate approach.

3 Generalization algorithm

In this section, we will propose an algorithm to implement PA. By the Definition 4, we can see that the key to solve the problem is to find an appropriate partition of T so that l-diversity can be achieved and information loss can be minimized. We will first present the detail of the partitioning step, which produces QI-groups $G_1, G_2, \cdots, G_n$ satisfying l-diversity. After this step, we use a populating step to implement permutation anonymization essentially.

3.1 The partitioning step

In this subsection, we will present a simple yet effective partitioning algorithm, which runs linearly and produce a partitioning satisfying l-diversity. The detailed procedure is presented in Algorithm 1.

The principle of l-diversity demands that: the number of the most frequent A_s value in each QI-group G_i can't exceed $\frac{|G_i|}{l}$. Motivated by this, we arrange the tuples in T to a list ordered by its A_s values(line 2-3), then distribute the tuples in L into $G_i (1 \leq i \leq g)$ a round-robin fashion(line 5). The resulting partitioning is guaranteed to be l-diversity, which is stated in Theorem 3. (If table T with sensitive attribute A_s satisfies $\max\{c(v) : v \in T.A_s\} > \frac{|T|}{l}$, then there exists no partition that is l-diversity.) In order to reduce the information loss, we examine each QI-group G_j, if the size of G_j is larger than $2 \cdot l - 1$, then partition the G_j further into $\frac{|G_j|}{l}$ sub-groups randomly to make each group contains no more than $2 \cdot l - 1$. It can be easily verified that the optimal partitioning of microdata does not contain QI-groups of more than $2 \cdot l - 1$ records.

Algorithm 1 A partitioning algorithm

```
Input: Microdata $T$, parameter $l$
Output: QI-groups $G_j$ that satisfy $l$-diversity;
Method:
1. If max $\{c(v) : v \in T.A_s\} \geq \frac{|T|}{l}$, Return;
2. Hash the tuples in $T$ into groups $Q_1, Q_2, \cdots, Q_\lambda$ by their $A_s$ values;
3. Insert these groups $Q_1, Q_2, \cdots, Q_\lambda$ into a list $L$ in order;
4. Let $g = \frac{|T|}{l}$, set QI-groups $G_1 = G_2 = \cdots = G_g = \emptyset$;
5. Assign tuple $t_i \in L$ $(1 \leq i \leq |L|)$ to $G_j$, where $j = (i \bmod g) + 1$
6. Examine each $G_j$:
7.    If $|G_j| \geq 2 \cdot l$
8.        Partition $G_j$ into $\frac{|G_j|}{l}$ sub-groups.
```

Theorem 3 *If table T with sensitive attribute A_s satisfies* $\max\{c(v) : v \in T.A_s\} \leq \frac{|T|}{l}$*(where $c(v)$ is the number of tuples in T with sensitive value v), the partition produced by our partitioning algorithm fulfills l-diversity.*

Proof First we observe that the number of the most frequently A_s is at most $\frac{|T|}{l}$. The partitioning algorithm produces $\frac{|T|}{l}$ QI-groups, and assign tuples having same A_s values into different QI-groups. Additionally, each QI-group contains at least l tuples, therefore, the principle of l-diversity is satisfied. □

Theorem 4 *The complexity of the partitioning algorithm is $O(|T|)$, where $|T|$ denotes the cardinality of microdata T.*

Proof Hash tuples in T cost $O(|T|)$, insert and assign tuples consume another $O(|T|)$ time, respectively, therefore, the total cost of partitioning algorithm is $O(3|T|)$. □

The above partitioning algorithm takes no account of information loss, but to obtain a l-diversity partitions. To reduce information loss, we will first preprocess the microdata following the idea: *allocate tuples sharing the same or quite similar QI-attribute values into the same sub-tables*. Then for each sub-table T_i, we call the partitioning algorithm to partition T_i into QI-groups.

The detailed preprocessing procedure is presented in Algorithm 2. Initially, S contains T itself (line 1); then, each $G \in S$ is divided into two generalizable subsets G_1 and G_2 such that $G_1 \cup G_2 = G$, $G_1 \cap G_2 = \emptyset$ (line 5-7). Then for each new subset, we check whether $G_1(G_2)$ satisfies l-diversity (line 8). If both are generalizable, we remove G from S, and add G_1, G_2 to S; otherwise G is retained in S. The attempts to partition G are tried k times and tuples of G are randomly shuffled for each time (line 3-4). Our experimental results show that most of G can be partitioned into two sub-tables by up to $k = 5$ tries. The algorithm stops when no sub-tables in S can be further partitioned.

In the above procedure, the way that we partition G into two subsets G_1 and G_2 is influential on the information loss of the resulting solution. For this purpose, we artificially construct two tuples $t_1, t_2 \in G$ with each attribute taking the maximal/minimal value of the corresponding domains, and then insert them G_1 and G_2 separately (line 6). After this step, for each tuple $w \in G$ we compute $\Delta_1 = NCP(G_1 \cup w) - NCP(G_1)$ and $\Delta_2 = NCP(G_2 \cup w) - NCP(G_2)$, and add tuple w to the group that leads to lower penalty (line 7). After successfully partitioning G, remove the artificial tuples from G_1 and G_2 (line 8).

Algorithm 2 The preprocessing algorithm

```
Input: A microdata $T$, integers $k$ and $l$
Output: A set $S$ consisting of sub-tables of $T$;
Method:
/* the parameter $k$ is number of rounds to partition $G$*/
1. $S = \{T\}$;
2. While($\exists G \in S$ that has not been partitioned)
3.    For $i = 1$ to $k$
4.        Randomly shuffle the tuples of $G$;
5.        Set $G_1 = G_2 = \emptyset$;
6.        Add tuple $t_1$ ($t_2$) of extremely maximal (minimal) value to $G_1$ ($G_2$);
7.        For each tuple $w$
              compute $\Delta_1 = NCP(G_1 \cup w) - NCP(G_1)$ and
              $\Delta_2 = NCP(G_2 \cup w) - NCP(G_2)$, respectively.
              If($\Delta_1 < \Delta_2$) then Add $w$ to $G_1$, else add $w$ to $G_2$;
8.     If both $G_1$ and $G_2$ satisfy $l$-diversity
              remove $G$ from $S$, and add $G_1 - \{t_1\}$, $G_2 - \{t_2\}$ to $S$, break;
9. Return $S$;
```

Theorem 5 *The average complexity of the partitioning step is $O(|T|log|T|)$.*

Proof In the average case, the time complexity denoted by $F(|T|)$ satisfies

$$F(|T|) = \frac{1}{|T|}\sum_{i=0}^{|T|-1}(F(i) + F(|T| - i)) + O(6|T|) = \frac{2}{|T|}\sum_{i=0}^{|T|-1}F(i) + O(6|T|) \quad (13)$$

It is a standard problem tackled by the average time of quick-sort problem. Solve this equation, the theorem holds. □

3.2 The populating step

For each QI-group $QI_j (1 \leq j \leq m)$ generated by the previous partitioning step, we independently uniformly generate $d + 1$ (d is the number of QI-attributes) permutations on QI_j and A_s, i.e. $\alpha_1, ...\alpha_m, \alpha_s$. Then, for each tuple $t \in QI_j$, insert a tuple $\langle \alpha_1(t).A_1, \cdots, \alpha_d(t).A_d, j\rangle$ into PQT, and insert $\langle j, \alpha_s(t).A_S)\rangle$ into PST. The above procedure is repeated until all QI-groups have been processed. Clearly, this step run linearly. Hence, the overall anonymization algorithm runs in $O(|T|log(|T|))$.

4 Empirical evaluation

In this section, we experimentally evaluate the effectiveness (data quality) and efficiency (computation cost) of the proposed technique. In the following experiments, we compare permutation anonymization (denoted by Permutation) against anatomy (denoted by Anatomy) from two aspects: (i) utility of the published tables for data analysis, and (ii) cost of computing these tables. For anatomy, we use the executable code that was downloaded from the author's homepage.

4.1 Data sets

For these purposes, we utilize a real data set CENSUS containing personal information of 500k American adults, which is widely used in the literature (Xu et al. 2006; Xiao and Tao 2006; Tao et al. 2009; Li et al. 2008). Each tuple describes the personal information of an American. The data set has 9 discrete attributes as summarized in Table 5 and Table 6 summarizes the parameters of our experiments.

In order to examine the influence of dimensionality and sensitive value distribution, we create two sets of microdata tables. The first sets contain five tables from CENSUS, denoted as SAL-3,$\cdots$ SAL-7, respectively. Specially, SAL-d ($3 \leq d \leq 7$), treats the first d attributes in Table 5 as the QI-attributes, and Salary-class as the sensitive attribute A_s. For example,

Table 5 Summary of attributes

Attribute	Number of distinct values	Types
Age	78	Numerical
Gender	2	Categorical
Education	17	Numerical
Marital	6	Categorical
Race	9	Categorical
Work-class	10	Categorical
Country	83	Numerical
Occupation	50	Sensitive
Salary-class	50	Sensitive

Table 6 Parameters and tested values

Parameter	Values
l	2,4,6,8,10
Cardinality n	100k,200k,300k,400k,500k
Number of QI-attributes d	3 4 5 6 7
Query dimensionality w	4, 5, ...$d+1$
Expected selectivity s	0.10, 0.15, ..., 0.3

SAL-3 is 4D, containing QI-attributes Age, Gender, and Education. The second table sets also contain 5 tables $OCC-3, \cdots, OCC-7$, where $OCC-d(3 \leq d \leq 7)$ has the same QI-attributes as $SAL-d$, but includes Occupation as the sensitive attribute.

All experiments are conducted on a PC with 1.9 GHz AMD CPU and 1 gigabytes memory. All the algorithms are implemented with Microsoft Visual C++ 2008.

4.2 Query accuracy

Anonymized data is often used for analysis and data mining. We measure the data utility by answering aggregate queries on published data, since they are the basic operations for numerous data mining tasks (e.g., decision tree learning, association rule mining, etc.). Specifically, we use queries of the following form:

Select Count $(*)$ from SAL-d(OCC-d) Where $A_1 \in b_1$ And $A_2 \in b_2$ And $\cdots$ And $A_w = b_w$.

Here, w is a parameter called the query dimensionality. A_1, ..., A_{w-1} are $w-1$ arbitrary distinct QI-attributes in SAL, but A_w is always Salary-class or Occupation, $b_i(1 \leq i \leq w-1)$ is a random interval in the domain of A_i. The generation of $b_1, \cdots, b_{w-1}$ is governed by another parameter termed volume s, which is a real number in [0, 1], and determines the length (in the number of integers) of $b_i(1 \leq i \leq w)$ as $\lfloor |A_i| \cdot s^{1/(w+1)} \rfloor$. Apparently, the query result becomes larger given a higher s. We derive the estimated answer of a query using the approach explained in (Zhang et al. 2007). The accuracy of an estimate is gauged as its relative error. Namely, let act and est be the actual and estimated results respectively, the relative error equals $|act - est|/act$. The average error rate is computed in answering a workload which contains 1000 queries.

4.2.1 *Privacy level l*

Now, we explore the influence of l on data utility. Towards this, we vary l from 2 to 10. The result is shown in Fig. 1. The error increases with the growth of l. This is expected, since a larger l demands stricter privacy preservation, which reduces data utility. Compared to anatomy that produces about 14–25 % average error on two data sets, the published data produced by our algorithm is significantly more useful. It is quite impressive to see that the error of our algorithm is consistently below 14 % despite of the growth of l. Another advantage of our method over anatomy is that the utility achieved by our model is less sensitive to domain size than anatomy. From the figures, we can see that data sets generated by permutation has a lower average query error on SAL-d than that on OCC-d $(3 \leq d \leq 7)$ due to the fact that the number of the most frequent sensitive value (Salary-Class) of SAL is smaller than that of OCC(Occupation). Such a fact implies that the information loss is positively correlated to the number of the most frequent sensitive value.

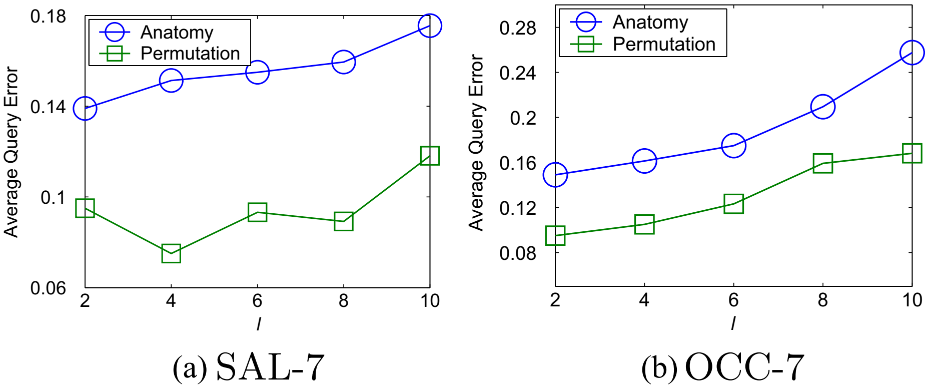


Fig. 1 Average query error vs. privacy level l

4.2.2 Query dimensionality w

To study the impact of w, Fig. 2a–f plot the error of anatomy and permutation as a function of w on two data sets, respectively. The accuracy incurs less error as w increases. To explain this, recall that all queries have the same (expected) selectivity $s = 0.1$. When w becomes larger, the values of $b_i (1 \leq i \leq w)$ queried on each attribute increases considerably, leading to a more sizable search region, which in return reduces error.

4.2.3 Volume s

Experiments of this subsection is designed to show the relation between the average query error of the two approaches and volume s. Figure 3 investigates the influence of s on data utility. Figure 3a–f plot the error of anatomy and permutation as a function of s on two

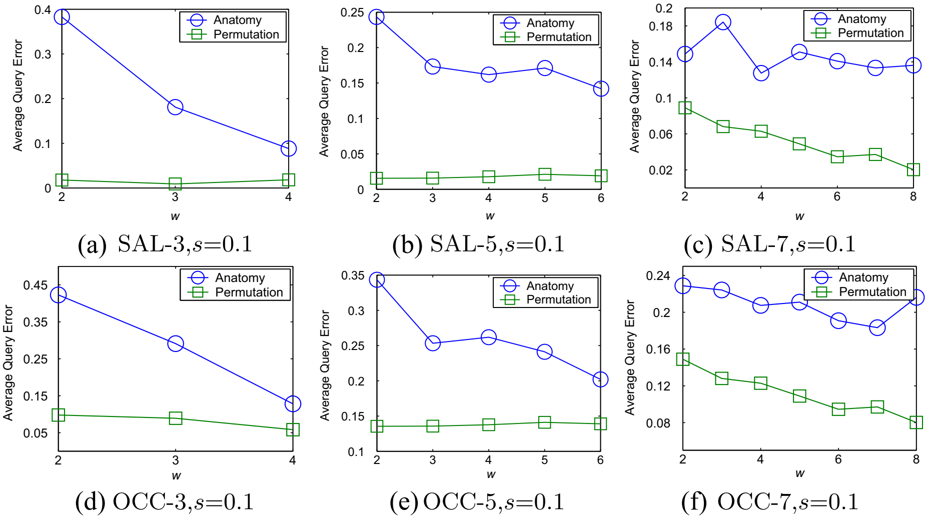


Fig. 2 Average query error vs. parameters w

data sets, respectively. Evidently, the query result becomes better if a higher *s* is given. This phenomenon is consistent with the existing understanding that both anonymization techniques provide better support for count queries when query results are larger. Apparently, our algorithm produces significantly more useful published data than anatomy.

4.2.4 Cardinality n

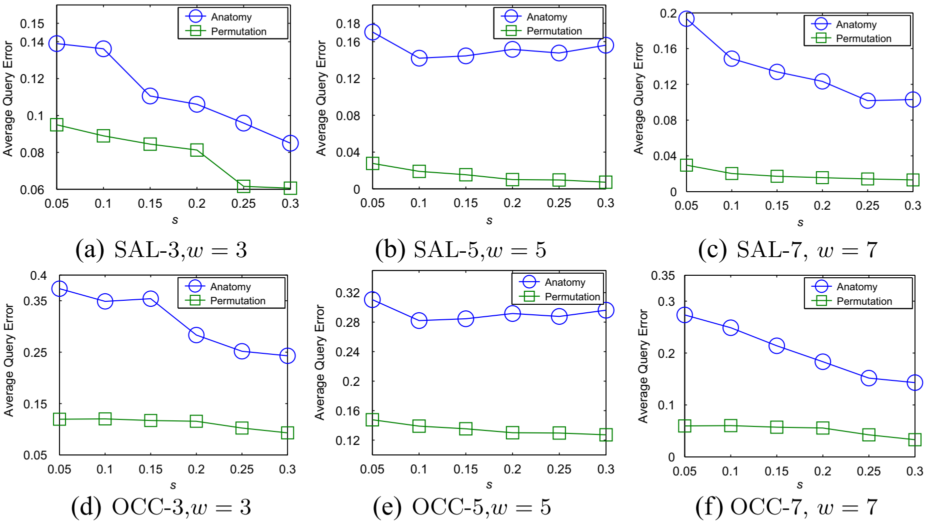
Figure 4a–b examines relationship between the accuracy of each method and the cardinality of the two data sets. As expected, the accuracy descends as *n* grows. This observation can be attributed to the fact that when the table size increases more tuples will share the same or quite similar QI-attributes. As a result, it is easier for the algorithm to find very similar tuples to generalize. Similar to previously experimental results. Again, permutation achieves significantly lower error in all cases.

4.3 Computation of overhead

Finally, we evaluate the overhead of performing anonymization. Figure 5 illustrates the cost of computing the publishable tables by two anonymization techniques, respectively, when the cardinality *n* linearly grows from 100k to 500k. The cost grows as *n* increases. This is expected, since longer time needs to be paid for larger number of tuples that participated in the anonymization.

From Fig. 5, we also can see that the advantages of our method in anonymization quality does not come for free. However, in all tested cases our algorithm can finish in less than 80 seconds, which is acceptable especially for those cases where query accuracy is the critical concern.

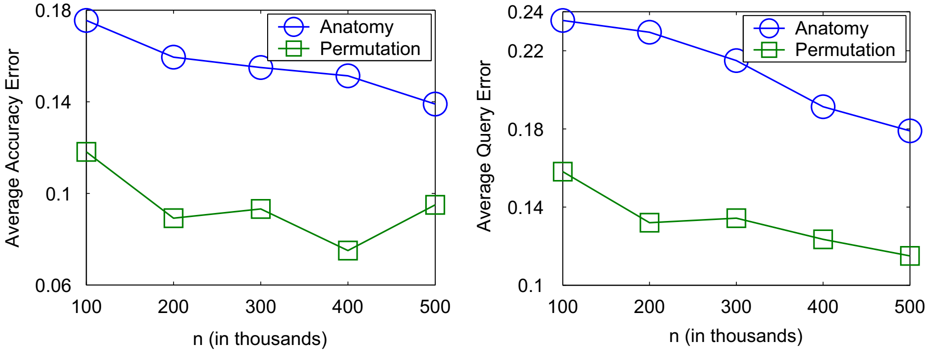
Summary The above results clearly show that permutation anonymization achieves less information loss than the anatomy in all cases. Calculations show that on the SAL data sets



(a) SAL-3,$w = 3$ (b) SAL-5,$w = 5$ (c) SAL-7, $w = 7$

(d) OCC-3,$w = 3$ (e) OCC-5,$w = 5$ (f) OCC-7, $w = 7$

Fig. 3 Average query error vs. volume *s*



(a) SAL-7 (b) OCC-7

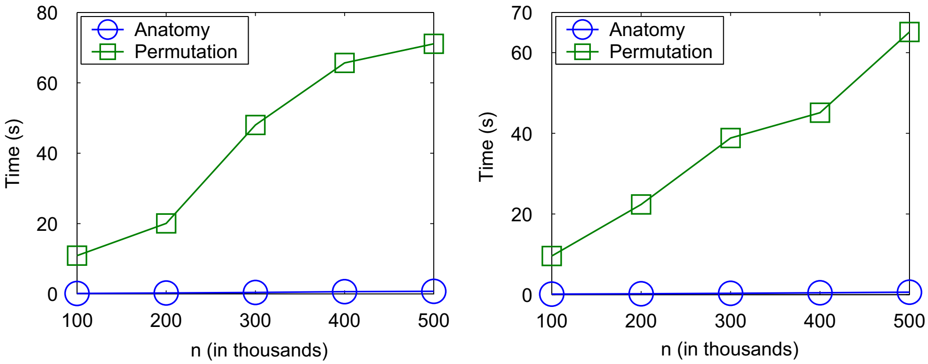
Fig. 4 Average query error vs. cardinality *n*

the PA algorithm compared with the anatomy algorithm, reduces about by 25 % probably, while on OCC data set, probably reduced by 30 %.

The trade-off of high anonymization quality is the runtime and the anatomy method is more efficient. However, the runtime of the permutation anonymization is not far away from that of the anatomy in practice. Moreover, for anonymization methods, the computation time is often a secondary consideration yielding to the quality.

5 Discussions and related work

Now let's revisit permutation anonymization with the objective to gain deep insight about this technique. The essential reasons that an attacker may recover an individual's sensitive attribute value is the existence of the following relationships: (1) the link between the identities and quasi-identifier in the public database (PD); (2) the link between quasi-identifiers and the sensitive attribute; and (3) the link among the quasi-identifier. Figure 6 illustrates



(a) SAL-7 (b) OCC-7

Fig. 5 Running time vs. cardinality *n*, *l*=10

these relationships among identities, sensitive attributes and quasi-identifier. Breaking or weakening the relationship of any of the above links will help protect privacy.

Generalization actually breaks the second and the third links. Anatomy breaks the third link and while permutation anonymization breaks the second link and weakens the third link between the QI-attributes. This is the reason why permutation anonymization can provide a stronger privacy than anatomy and higher data utility than generalization. Hence, we argue that permutation anonymization is a good tradeoff between anatomy and generalization.

Previous anonymization techniques including data perturbation Agrawal and Srikant (2000), condensation Aggarwal and Yu (2008), clustering Aggarwal et al. (2006), Swapping Fienberg and Mcintyre (2004). These techniques are employed to hide the exact values of the data. However, it may not be suitable if one wants to make inferences with 100 % confidence. Authors of paper Aggarwal and Yu (2008) proposed the condensation method, which releases only selected statistics about each QI-group. In essence, authors of paper Zhang et al. (2007) provide another version of Anatomy (Xiao and Tao 2006). Data swapping Fienberg and Mcintyre (2004) produces an alternative table by interchanging the values (of the same attribute) among the tuples in T, however, it is not designed with linking attacks in mind. Consequently, data swapping can't promise the prevention of such attacks. These technique share a same feature: to change or to replace the sensitive attribute value, which protect privacy by breaking the third link. Authors of Kifer (2009) presents an attack using data mining techniques that are based on a deep statistical theorem known as deFinetti's representation theorem. Li et al. (2012) provides a novel technique called slicing, which partitions the data both vertically and horizontally. They demonstrates that slicing preserves better data utility than generalization. Another important advantage of slicing is that it can handle high-dimensional data. More precisely, slicing is an cross approach between Anatomy and Permutation Anonymization presented in this paper.

In addition, Tao et al. (2009) proposes an interesting anonymization technique, ANGEL, to enhance the utility for privacy preserving publication. Many QI-groups of the anonymized

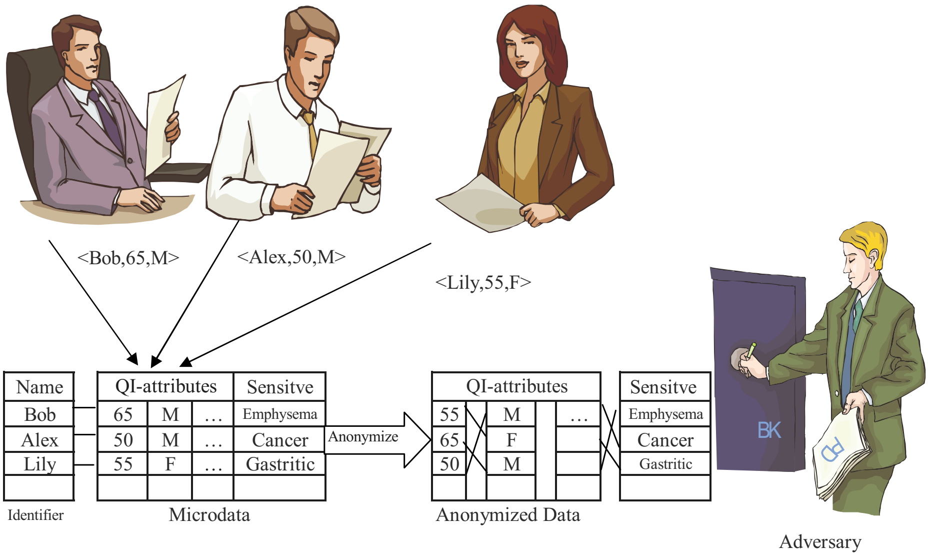


Fig. 6 the relationship among identifier, QI-attributes and sensitive attribute

table released by ANGEL may contain a large number of sensitive values, which in the worst case is quadratic to the number of tuples in the QI-group. Such facts will cause significant average error when answering aggregate queries. Differential privacy has attracted techniques for answering aggregate queries over sensitive data in a privacy preserving way by adding noise to the query answers. Their objective is typically to minimize absolute errors while satisfying differential privacy (Hardt and Talwar 2010; Dwork 2006; Dwork and Lei 2009; Dwork 2008).

To the best of our knowledge, the first link was first noted and excluded from the microdata. There is a long line of work on the third link. However, none of the previous works focus on the second link systematically, which is the major concern of this paper with the basis of (He et al. 2012).

6 Conclusion

Although anatomy is a common methodology for protecting privacy, the weakness of anatomy motivates us to develop a novel anonymization technique called permutation anonymization, which provides better privacy preservation than anatomy without sacrificing data utility. We systematically investigate the theoretic properties of this new technique and propose a corresponding algorithm to implement it. As verified by extensive experiments, our method allows significantly more effective data analysis, and simultaneously providing enough privacy preservation.

In future work, we will extend this method to the problem of personalized privacy in which allows multiple privacy levels for different records in the database. By varying the permutation in the QI-groups, it may be possible to provide personalized levels of privacy to different records in the database.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (NO.61202007), Chinese Postdoctoral Science Foundation Funded Project (No.2013M540323) and the Natural Science Foundation of NingBo (NO.2013A610110).

References

Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. In *SIGMOD '00: Proceedings of the 2009 ACM SIGMOD international conference on management of data* (pp. 439–450). New York: ACM. [Online]. Available: doi:10.1145/342009.335438.

Aggarwal, C.C., & Yu, P.S. (2008). On static and dynamic methods for condensation-based privacy-preserving data mining. *ACM Transactions on Database Systems*, *33*, 1–39.

Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., & Zhu, A. (2006). Achieving anonymity via clustering. In *PODS '06: Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 153–162). New York: ACM.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., & Wegener, I. (Eds.), *ALP06': Automata, languages and programming*, (Vol. 4052 pp. 1–12). Berlin: Springer.

Dwork, C. (2008). Differential privacy: A survey of results. *Theory and Applications of Models of Computation*, *4978*, 1–19.

Dwork, C., & Lei, J. (2009). Differential privacy and robust statistics. In *In STOC '09: Proceedings of the 41st annual ACM symposium on Theory of computing* (pp. 371–380). New York: ACM. [Online]. Available: doi:10.1145/1536414.1536466.

Fienberg, S.E., & Mcintyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. *Privacy in Statistical Databases*, 14–29.

Hardt, M., & Talwar, K. (2010). On the geometry of differential privacy. In *In STOC '10: Proceedings of the 41st annual ACM symposium on theory of computing* (pp. 705–714). New York: ACM. [Online]. Available: doi:10.1145/1806689.1806786.

He, X., Xiao, Y., Li, Y., Wang, Q., Wang, W., & Shi, B. (2012). Permutation anonymization: Improving anatomy for privacy preservation in data publication. In *New frontiers in applied data mining(Pakdd2011 workshop)*, (Vol. 7104 pp. 111–123). Berlin: Springer.

Kalnis, P., Ghinita, G., Mouratidis, K., & Papadias, D. (2007). Preventing location-based identity inference in anonymous spatial queries. *IEEE Transactions on Knowledge and Data Engineering*, *19*(12), 1719–1733.

Kifer, D. (2009). Attacks on privacy and definetti's theorem. In *In SIGMOD '09: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data* (pp. 127–138). New York: ACM. [Online]. Available: doi:10.1145/1559845.1559861.

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *ICDE '07: International conference on data engineering* (pp. 106–115).

Li, J., Tao, Y., & Xiao, X. (2008). Preservation of proximity privacy in publishing numerical sensitive data. In *SIGMOD '08: Proceedings of the 2009 ACM SIGMOD international conference on management of data* (pp. 473–486). New York: ACM.

Li, T., Li, N., Zhang, J., & Molloy, I. (2012). Slicing: A new approach for privacy preserving data publishing. *IEEE Transactions on Knowledge and Data Engineering*, *24*, 561–574.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). l-diversity: Privacy beyond k-anonymity. In *ICDE '06: International conference on data engineering* (pp. 24–35).

Mokbel, M.F., Chow, C.-Y., & Aref, G.W. (2006). The new casper: query processing for location services without compromising privacy. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases* (pp. 763–774).

Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, *13*(6), 1010–1027.

Samarati, P., & Sweeney, L. (1998). Generalizing data to provide anonymity when disclosing information (abstract). In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems* (pp. 188–195). New York: ACM.

Sweeney, L. (2002). k-anonymity: a model for protecting privacy. *International Journal for Uncertainty Fuzziness Knowledge-Based Systems*, *10*(5), 557–570.

Tao, Y., Chen, H., Xiao, X., Zhou, S., & Zhang, D. (2009). Angel: Enhancing the utility of generalization for privacy preserving publication. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1073–1087.

Xiao, X., & Tao, Y. (2006). Anatomy: simple and effective privacy preservation. In *VLDB '06: Proceedings of the 32nd international conference on Very large data bases* (pp. 139–150): VLDB Endowment.

Xiao, X., Yi, K., & Tao, Y. (2010). The hardness and approximation algorithms for l-diversity. In *EDBT '10: Proceedings of the 13th International Conference on Extending Database Technology* (pp. 135–146). New York: ACM.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu A. W.-C. (2006). Utility-based anonymization using local recoding. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 785–790). New York: ACM.

Zhang, Q., Koudas, N., Srivastava, D., & Yu, T. (2007). Aggregate query answering on anonymized tables. In *ICDE '07: International Conference on Data Engineering* (pp. 116–125).