Knowledge and Information Systems (2019) 60:1039–1080
https://doi.org/10.1007/s10115-018-1237-3

Two privacy-preserving approaches for data publishing with identity reservation

Jinyan Wang[1,2] · Kai Du[1,2] · Xudong Luo[1,2] · Xianxian Li[1,2]

Received: 22 January 2017 / Revised: 4 April 2018 / Accepted: 26 May 2018 / Published online: 29 June 2018

Abstract
Many approaches have been proposed for publishing useful information while preserving data privacy. Among them, the privacy models of identity-reserved (k, l)-anonymity and identity-reserved (α, β)-anonymity have been proposed to handle the situation where an individual could have multiple records. However, the two models fail to prevent attribute disclosure. To this end, we propose two new privacy models: enhanced identity-reserved l-diversity and enhanced identity-reserved (α, β)-anonymity. Moreover, to implement the two privacy models we design a general anonymization algorithm, called *DAnonyIR*, with clustering technique by calling different decision functions, which can decrease the information loss caused by generalization. Further, we compare *DAnonyIR* concerning our two privacy models with existing generalization method *GeneIR* concerning identity-reserved (k, l)-anonymity and identity-reserved (α, β)-anonymity, respectively. The experimental results show that our two approaches provide stronger privacy preservation, and their information loss and relative error ratio of query answering are less than those of *GeneIR*.

Keywords Privacy preservation · Data publishing · Identity reservation · Hitting set · Information loss · Anonymity

1 Introduction

Hospitals and other organizations often need to publish microdata (e.g. medical data or census data) for the purposes of scientific research and knowledge-based decision-making [1], for example, disease analysis and prediction. These data are often stored in a table in the form of D(*explicit identifier*, *quasi-identifier*, *sensitive attributes*, *other attributes*) [2], where *explicit identifier* (*ID*) can clearly identify individuals (e.g. name and social security number); *quasi-identifier* (*QI*) is a set of attributes that can potentially identify an individual, such as zip

✉ Xianxian Li
lixx@gxnu.edu.cn

1 Guangxi Key Lab of Multi-source Information Mining and Security, Guangxi Normal University, Guilin 541004, China

2 School of Computer Science and Information Engineering, Guangxi Normal University, Guilin 541004, China

code, date of birth, and gender (in general, we assume that QI is the background knowledge possessed by attackers); *sensitive attributes* consist of sensitive person-specific information (e.g. disease and salary); and *other attributes* are those attributes that are not contained in the previous three categories (e.g. visit date, if it is not contained in QI). To avoid the leakage of individual privacy, explicit identifying information has to be removed when microdata are released. However, individual privacy still could be leaked by linking other public data with QI [2,3]. Thus, the methods and tools for privacy-preserving data publishing are required. Recently, the problem has received wide attention, so many approaches have been proposed catering for different data publishing scenarios [2–9]. In particular, for relational data, main approaches are k-anonymity [3], l-diversity [10], (α, k)-anonymity [11], and t-closeness [12], all of which assume that an individual has only a record in a data table.

However, in real life often an individual could have multiple records in a data table. For example, if a patient suffers from several diseases, there are several records related to the patient in the table. In this case, privacy models above may be underprotected and are inadequate, because an equivalence class may contain less individuals. Particularly, it is possible that an equivalence class contains only an individual. An equivalence class or a group of an anonymized table is a set of records with the same values for the quasi-identifier attributes [3]. Moreover, removing the explicit identifying information will damage the relation among the values of sensitive attribute, which belong to the same individual, because it is very important in medical research (e.g. complications of a disease).

In order to solve the problem with respect to an individual with multiple records, Tong et al. [13] proposed identity-reserved (*IR*) k-anonymity, *IR* (k, l)-anonymity, and *IR* (α, β)-anonymity. *IR* k-anonymity enables to prevent *identity disclosure*, which occurs when an individual is linked to a particular record in the published table [2], because it requires that each equivalence class in the anonymous table contains at least k different individuals. Nevertheless, *IR* k-anonymity does not prevent *attribute disclosure*, which occurs when an attacker can infer individual's sensitive values from the released data [2]. Thus, *IR* (k, l)-anonymity and *IR* (α, β)-anonymity are proposed. *IR* (k, l)-anonymity needs that each equivalence class in the anonymous table satisfies *IR* k-anonymity and contains at least l different sensitive values. *IR* (α, β)-anonymity requires that the percentage of records of any individual in each equivalence class is not larger than α and the percentage of any sensitive value in each equivalence class is not larger than β. Nonetheless, they do not consider that some records in an equivalence class belong to the same individual regarding the restrictions with respect to l and β and still lead to *attribute disclosure* when a sensitive value appears in major individuals of an equivalence class, especially in every individual of an equivalence class.

Thus, in order to prevent the privacy leakage caused by *IR* (k, l)-anonymity and (α, β)-anonymity, in this paper we will propose enhanced identity-reserved (*EIR*) l-diversity and *EIR* (α, β)-anonymity. *EIR* l-diversity considers that for each equivalence class, any set of records from different individuals has at least l different sensitive values, which can ensure that there are at least l different individuals. *EIR* (α, β)-anonymity requires that the percentage of records of any individual in each equivalence class is not larger than α and especially the percentage of any sensitive value in any set of records from different individuals is not more than β. They can prevent *identity disclosure* and *attribute disclosure*. Also, to anonymize a data table and make it satisfy *EIR* l-diversity or *EIR* (α, β)-anonymity, we will design a general anonymization algorithm, called *DAnonyIR*, by using clustering technique without predefined taxonomy trees.

The main contributions of this paper are as follows. (1) We introduce the notions of reasoning set and reasoning space of an equivalence class and present two privacy models: *EIR* l-diversity and *EIR* (α, β)-anonymity. For an equivalence class, the problems whether

it satisfies the *EIR* l-diversity and *EIR* (α, β)-anonymity are changed to the problems of minimum hitting set and the highest frequency of sensitive values, respectively. (2) We define the distances between two individuals, between individual and equivalence class, and between two equivalence classes. (3) We design a general anonymization algorithm *DAnonyIR* with clustering techniques to make a dataset satisfy different identity-reserved privacy models by calling different decision functions. (4) We do lots of experiments to show the vulnerability of *IR* (k, l)-anonymity and *IR* (α, β)-anonymity, and our algorithm of *DAnonyIR* concerning *EIR* l-diversity and *EIR* (α, β)-anonymity outperforms the existing one of *GeneIR* [13] concerning *IR* (k, l)-anonymity and *IR* (α, β)-anonymity with respect to information loss and relative error ratio of query answering, respectively. At present, *GeneIR* is the only an algorithm for achieving *IR* (k, l)-anonymity and *IR* (α, β)-anonymity. Also, we show that *DAnonyIR* can achieve the two privacy models, and compare them with our enhanced approaches.

The rest of this paper is organized as follows. Section 2 recaps basic concepts and notations we will use in this paper. Section 3 introduces the concepts of reasoning set of an equivalence class and its reasoning space and proposes two privacy models for identity reservation. Section 4 defines some distance concepts about individuals and equivalence classes and designs a general anonymization algorithm based on clustering techniques for different privacy models with identity reservation. Section 5 analyses our methods experimentally. Section 6 discusses the related work. Finally, Sect. 7 concludes this paper and points out directions for further study.

2 Preliminaries

This section recaps some fundamental privacy models, generalization operations, and information metrics, which are necessary for developing our work.

2.1 Privacy models with identity reservation

Consider an original data table $D = \{ID, A_1, \ldots, A_d, A_s\}$ in which there are no duplicate records, where $A_1, \ldots, A_d$ are QI attributes.[1] For the convenience of reference, Table 1 summarizes the meanings of symbols used in the paper.

In privacy-preserving data publishing, for every privacy model π, there is a corresponding anonymization approach to transforming the original data table to an anonymous table which satisfies π. And privacy models k-anonymity, distinct l-diversity, and (α, k)-anonymity all assume that an individual has only one record. For these privacy models, their anonymous tables are in the form of $D^*(QI', A_s)$, where QI' is an anonymous version of the original QI obtained by applying anonymization approach to QI in original table D. For the problem of identity-reserved anonymity, we need to keep the information in which multiple records belong to the same individual, so the explicit identifier should not be directly deleted and we use numbers to identify different individuals. The published anonymous table of D is in the form of $D^*(Id\_num, QI', A_s)$, where the values of Id_num are numbers, which denote

[1] For the methods of privacy-preserving data publishing, the privacy model and anonymizaiton algorithm operate on QI and sensitive attribute and do not operate on the *other attributes*. That is, if some attributes in *other attributes* are published, they are the same in an original data and its anonymous version. For convenience, we omit the *other attributes*. At present, most approaches assume that the original database has a sensitive attribute. Also, it is important to solve privacy-preserving data publishing in relational data with multiple sensitive attributes. However, it is beyond the scope of this paper and so will be addressed in future.

Table 1 Summary of notations

Symbol	Description
D	Original table
D^*	The anonymous table obtained by generalizing on D
ID	Explicit identifier
QI	Quasi-identifier, consisted of $A_i (1 \leq i \leq d)$ attributes
A_s	Sensitive attribute
r_j	A record in D
$r_j[A]$	The value of record r_j in attribute A
$P = \{p_1, \ldots, p_n\}$	The set of individuals
$S = \{s_1, \ldots, s_m\}$	The set of sensitive values
$R(p_k)$	The set of records of individual p_k
$S(p_k)$	The set of sensitive values of individual p_k
$R(s_t)$	The set of records in which sensitive value is s_t
$P(s_t)$	The set of individuals whose sensitive values contain s_t
Q	An equivalence class
m_Q	The number of sensitive values in Q
n_Q	The number of individuals in Q
P_Q^{sin}	The set of single-record individuals in Q
P_Q^{mul}	The set of multi-record individuals in Q
Q_{rea}	The reasoning space of Q
Q_{rea}^i	A reasoning set in Q_{rea}

different individuals. Some notions directly related to our approaches are presented in the following.

The first one is equivalence class [3], which is the elementary unit of anonymous table. Formally, we have:

Definition 2.1 (*Equivalence class*) Let Q be a set of records in a data table. Q is called an equivalence class if $\forall r_i, r_j \in Q, r_i[QI] = r_j[QI]$.

The second is the definitions of k-anonymity [3], distinct l-diversity [10], and (α, k)-anonymity [11], which are suitable for the situation where an individual has a record only.

Definition 2.2 (*k-anonymity*, *distinct l-diversity*, *(α, k)-anonymity*) Given an original data table D, the published anonymous table D^* of D satisfies

(1) k-anonymity if any equivalence class in D^* contains at least k records;
(2) distinct l-diversity if any equivalence class in D^* contains at least l different sensitive values; and
(3) (α, k)-anonymity if for any equivalence class Q in D^*, Q satisfies k-anonymity and the percentage of any sensitive value in Q is less than or equal to α.

Distinct l-diversity and (α, k)-anonymity can prevent *identity disclosure* and *attribute disclosure*, and k-anonymity can only prevent *identity disclosure* [2].

The following is the definitions of *IR* k-anonymity, *IR* (k, l)-anonymity, and *IR* (α, β)-anonymity [13], which are obtained by extending k-anonymity, distinct l-diversity, and (α, k)-anonymity to the scenario where an individual could have multiple records, respectively.

Definition 2.3 (*IR* k-anonymity, *IR* (k, l)-anonymity, *IR* (α, β)-anonymity) Given an original data table D, for an equivalence class Q in the published anonymous table D^*, let $|Q|$ be the number of records, m_Q be the number of sensitive values, and n_Q be the number of individuals, i.e.

$$m_Q = \left| \bigcup_{p_i \in Q.Id\_num} S(p_i) \right|, \tag{1}$$

$$n_Q = \left| \bigcup_{s_j \in Q.A_s} P(s_j) \right|. \tag{2}$$

Then D^* satisfies

(1) identity-reserved (IR) k-anonymity if any equivalence class Q contains at least k different individuals, i.e. $n_Q \geq k$;
(2) identity-reserved (IR) (k, l)-anonymity if any equivalence class Q contains at least k different individuals and l different sensitive values, i.e. $m_Q \geq l$ and $n_Q \geq k$; and
(3) identity-reserved (IR) (α, β)-anonymity if for any equivalence class Q, $|R(p_i)|/|Q| \leq \alpha$, $\forall p_i \in Q.Id\_num$ and $|R(s_j)|/|Q| \leq \beta$, $\forall s_j \in Q.A_s$, where $\alpha, \beta \in (0, 1)$.

Obviously, $|R(s_j)|/|Q| \leq \beta$, the condition for *IR* (α, β)-anonymity, is equal to $|P(s_j)|/|Q| \leq \beta$ because s_j appears in any individual at most once.

2.2 Data generalization

In order to satisfy the requirement of given privacy model π, the original table usually needs to be generalized in the values of quasi-identifier. The idea is to replace a specific value by a general value. Although the generalization operation reduces the data quality of an original table, it still can retain its semantic information to some extent. Therefore, the method attracts increasing attention in the field of privacy-preserving data publishing. For an anonymization algorithm, it first needs to satisfy the given privacy model and then considers to keep data quality (when an equivalence class satisfies the given privacy model, it is unnecessary to generalize its attributes' values to higher levels) and spend time as little as possible. In general, the stronger the privacy preservation of a privacy model is, the worse its data quality is and the more runtime it needs. Many existing approaches generalize the values of attributes in quasi-identifier according to predefined taxonomy trees [10–23], which are given by the domain experts. For example, Fig. 1 shows the taxonomy trees for categorical attribute *Postcode* and numeric attribute *age*.[2]

Wang et al. [24] pointed out that the taxonomy tree restricts the choice of data generalization and causes some unnecessary information loss. If the values of $Postcode$ in two records are 10076 and 10085, respectively, we can generalize them to {10076, 10085} in order to make the two records into an equivalence class. However, they are generalized to 100* according to the taxonomy tree. If the values of Age in two records are 34 and 35, respectively, we can generalize them to [34, 35] for forming an equivalence class. However, they are generalized to [30, 39] according to the taxonomy tree. Thus, Wang et al. [24] supplied a different scheme of data generalization, focusing on both numeric and categorical attributes.

[2] For the taxonomy tree for *age*, 30, 31, 32, 33, 34 are the leaves of [30, 34] and 35, 36, 37, 38, 39 are the leaves of [35, 39]. For the sake of simplicity, we omit it.

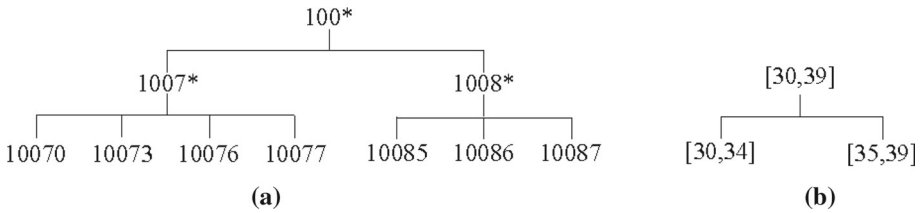


Fig. 1 Taxonomy trees for *Postcode* and *Age*. **a** Postcode. **b** Age

An interval number $[a, b]\,(a \leq b)$ is used to denote a numeric attribute's value between a and b. For a point value, we have $a = b$. Given ω records, the values in a numeric attribute $A \in QI$ are $r_i[A] = [a_i, b_i], i \in \{1, 2, \ldots, \omega\}$, respectively. We can generalize them to the interval number $[\min\{a_i\}, \max\{b_i\}]$. A set V is considered as the value of a categorical attribute. If it is point-valued, V is a set that contains only an element. Given ω records, the values in a categorical attribute $A' \in QI$ are $r_i[A'] = V_i, i \in \{1, 2, \ldots, \omega\}$, respectively, which are generalized to $\bigcup_{1 \leq i \leq \omega} V_i$. And $r_1, r_2, \ldots, r_\omega$ are generalized and have the same values of attributes in QI, which constitute an equivalence class. The identity element of the equivalence class is denoted as $\overline{r} = \delta(r_1, r_2, \ldots, r_\omega)$, where $\overline{r}[A] = [\min\{a_i\}, \max\{b_i\}]$ for each numeric attribute $A \in QI$, $\overline{r}[A'] = \bigcup_{1 \leq i \leq \omega} V_i$ for each categorical attribute $A' \in QI$, and $\overline{r}[A_s] = null$.

2.3 Information metrics

There exists information loss due to data generalization. So various metrics have been proposed for calculating how much information is lost. For the generalization with taxonomy tree, normalized certainty penalty (*NCP*) [21,25] and generalized loss metric (*GLM*) [26,27] are the two main metrics proposed. They are the same for numerical attributes but different for categorical ones. That is, for a numerical attribute A, and an interval $I = [l, u]$ from the domain $[L, U]$ of A, used to generalize A's value, the information loss associated with I is defined as follows:

$$Loss(I) = \frac{u - l}{U - L}. \tag{3}$$

For a categorical attribute A', where T is its taxonomy tree and a node p in T is used to generalize A''s value, the information loss associated with p is defined as follows (for *NCP* and *GLM*, respectively):

$$Loss_{NCP}(p) = \begin{cases} 0 & |u_p| = 1, \\ \frac{|u_p|}{|u|} & \text{otherwise}; \end{cases} \tag{4}$$

$$Loss_{GLM}(p) = \frac{|u_p| - 1}{|u| - 1}, \tag{5}$$

where u_p is the set of leaf nodes of the subtree rooted at p in T and u is the set of all the leaf nodes in T.

For example, a record has values 32 and 10073 for *Age* and *Postcode* attributes, respectively, whose taxonomy trees are shown in Fig. 1. Assume that 32 and 10073 are generalized to [30, 34] and 1007*. The information loss for *Age* by using *NCP* and *GLM* is the same: $\frac{34-30}{39-30} = \frac{4}{9}$. We can see that [30, 34] actually contains 30, 31, 32, 33, and 34, and there are five different numbers, while $34 - 30 = 4$. Similarly, [30, 39] actually contains ten different

numbers, while $39 - 30 = 9$. The information loss for categorical attribute *Postcode* by using *NCP* is $\frac{4}{7}$, while it is $\frac{4-1}{7-1} = \frac{1}{2}$ by using *GLM*. 1007* is equal to {10070, 10073, 10076, 10077} and there are four different numbers. The set of all the leaf nodes in the taxonomy tree of *Postcode* is {10070, 10073, 10076, 10077, 10085, 10086, 10087} and there are seven different numbers. We can see that the definitions of *GLM* are consistent with respect to numeric and categorical attributes, while *NCP* is not. So *GLM* is more suitable.

For *NCP* and *GLM*, they consider only the information loss between an original value and a generalized value of an attribute. When we create an equivalence class by continually adding the records of individuals until it satisfies given privacy model, it is necessary to compute the information loss from current equivalence class to the later equivalence class obtained by adding the records of an individual to current equivalence class. As a result, we need to consider the information loss caused by one generalization value to another. The definition of information loss between a generalized value (or original value) and another needs to be given.

For the generalization method without predefined taxonomy trees, Wang et al. [24] presented the information metrics for categorical and numerical attributes, as given by Eqs. (6) and (7), respectively. However, the results of Eqs. (6) and (7) are not normalized, i.e. the results (except 0) all are greater than 1. Also, it is unreasonable to denote the information loss by using the times (the number related to the later generalized value is divided by the number related to the original value or before generalized value), because the denominator is varied and so the standard is not uniform. In this paper, based on *GLM* we will improve the information metrics for numeric and categorical attributes, which are described in Sect. 4.1.

For a record r, the value is $r[A] = [a, b]$ on a numeric attribute A, which is the original value or a generalized value, and the (later) generalized value is $r^*[A] = [a^*, b^*]$. Then the information loss of r on attribute A is given by

$$Loss(r[A], r^*[A]) = \begin{cases} \frac{(b^*-a^*+1)}{b-a+1} & \text{if } r[A] \neq r^*[A], \\ 0 & \text{if } r[A] = r^*[A]. \end{cases} \tag{6}$$

For a record r, the value is the set $r[A']$ on a categorical attribute A', which is the original value or a generalized value, and the (later) generalized value is the set $r^*[A']$. Then the information loss of r on attribute A' is

$$Loss(r[A'], r^*[A']) = \begin{cases} \frac{|r^*[A']|}{|r[A']|} & \text{if } r[A'] \neq r^*[A'], \\ 0 & \text{if } r[A'] = r^*[A']. \end{cases} \tag{7}$$

3 Enhanced privacy models with identity reservation

In this section, we first give an example to show that although *IR* (k, l)-anonymity and *IR* (α, β)-anonymity can prevent *identity disclosure*, they fail to prevent *attribute disclosure*. Thus, in this section we will propose two privacy models, the *EIR* l-diversity and *EIR* (α, β)-anonymity, to prevent not only *identity disclosure* but also *attribute disclosure*.

Table 2 is a patient table, in which an individual has one or several records, where *{Gender, Age, Postcode}* is the quasi-identifier set *QI*. And Table 3 is a published table which satisfies *IR* $(3, 3)$-anonymity or *IR* $(0.4, 0.6)$-anonymity, obtained by the anonymization approach *GeneIR* [13] according to predefined taxonomy trees. If an attacker knows the *Mike*'s *QI* information (i.e. $\{M, 36, 10085\}$), obtained by some public information (i.e. voter list), and knows that $Mike$ is in the published table, then the attacker can infer that *Mike* is

Table 2 An patient table in which an individual could have multiple records

	Name	Gender	Age	Postcode	Disease
r_1	*Mike*	*M*	36	10085	*Hypertension*
r_2	*Mike*	*M*	36	10085	*Heart*
r_3	*Lily*	*F*	37	10076	*Cancer*
r_4	*Tim*	*M*	36	10086	*Hypertension*
r_5	*Jane*	*F*	33	10087	*Hypertension*
r_6	*Jane*	*F*	33	10087	*Diabetes*
r_7	*Tina*	*F*	38	10077	*HIV*
r_8	*Ella*	*F*	34	10070	*Leukaemia*
r_9	*Ella*	*F*	34	10070	*Heart*
r_{10}	*Lucy*	*F*	33	10073	*Syphilis*

Table 3 *IR* (3, 3)-anonymous or *IR* (0.4, 0.6)-anonymous table

	EC-ID	Id_num	Gender	Age	Postcode	Disease
r_3	Q_1	2	*F*	[30, 39]	1007*	*Cancer*
r_7	Q_1	5	*F*	[30, 39]	1007*	*HIV*
r_8	Q_1	6	*F*	[30, 39]	1007*	*Leukaemia*
r_9	Q_1	6	*F*	[30, 39]	1007*	*Heart*
r_{10}	Q_1	7	*F*	[30, 39]	1007*	*Syphilis*
r_1	Q_2	1	*	[30, 39]	1008*	*Hypertension*
r_2	Q_2	1	*	[30, 39]	1008*	*Heart*
r_4	Q_2	3	*	[30, 39]	1008*	*Hypertension*
r_5	Q_2	4	*	[30, 39]	1008*	*Hypertension*
r_6	Q_2	4	*	[30, 39]	1008*	*Diabetes*

in equivalence class Q_2 (see Table 3). Because Q_2 contains three different individuals, the attacker cannot know which one is corresponding to *Mike*, and thus, the *identity disclosure* is prevented. However, the attacker knows that *Mike* has *hypertension* disease because any individual in Q_2 has *hypertension* disease. Obviously, the privacy leakage is 100% in this case, so *attribute disclosure* happens.

Why does this attribute disclosure happen? This is because *IR* (k, l)-anonymity and *IR* (α, β)-anonymity do not consider that some records in an equivalence class belong to the same individual regarding the restrictions with respect to l and β, so they cannot prevent *attribute disclosure*. To address this issue, in the following we will give two enhanced privacy models, called *EIR* l-diversity and *EIR* (α, β)-anonymity, which consider that for each equivalence class, any set of records from different individuals satisfies l-diversity, and the percentage of any sensitive value in any set of records from different individuals is not more than β, respectively.

Definition 3.1 (*Single-record/multi-record individual*) Given an original data table D, for $\forall p_i \in P$, if p_i has one record, then p_i is called an single-record individual; if p_i has several records, then p_i is called an multi-record individual and these records are called related records.

For example in Table 2, r_1 and r_2 belong to the same individual, so they are related. From the example discussed in the beginning of this section, we know that the attacker cannot differentiate which individual in Q_2 is corresponding to $Mike$. However, the attacker can infer that $Mike$ has a disease with certain probability. If the probability is very big, it means that $Mike$'s privacy is leaked. The probability of a disease is the maximum percentage of the disease in any set of records from different individuals. The attacker reasons in the four sets consisting of these records from different individuals in Q_2, i.e. $\{r_1, r_4, r_5\}$, $\{r_1, r_4, r_6\}$, $\{r_2, r_4, r_5\}$, and $\{r_2, r_4, r_6\}$. The related records are not in the same set, and we need to prevent privacy leakage in any set.

Definition 3.2 (*Reasoning set*) For an equivalence class Q, let $P_Q = \{p_1, p_2, \ldots, p_{n_Q}\}$ be the set of individuals. A reasoning set of Q is a set of records which maps to different individuals from p_1 to p_{n_Q}.

All reasoning sets of Q constitute the reasoning space. The size of the reasoning space is determined by multi-record individuals, i.e.

$$|R(p_{m_1})| \times |R(p_{m_2})| \times \cdots \times |R(p_{m_{n_Q^m}})|, \tag{8}$$

where $P_Q^{sin} = \{p_i \mid |R(p_i)| = 1, p_i \in P_Q\} = \{p_{s_1}, p_{s_2}, \ldots, p_{s_{n_Q^s}}\}$ is the set of single-record individuals in Q, $P_Q^{mul} = P_Q \setminus P_Q^{sin} = \{p_{m_1}, p_{m_2}, \ldots, p_{m_{n_Q^m}}\}$ is the set of multi-record individuals, and $n_Q = n_Q^s + n_Q^m$.

For the equivalence class Q_2 in Table 2, $P_{Q_2} = \{1, 3, 4\}$, where $P_{Q_2}^{sin} = \{3\}$ and $P_{Q_2}^{mul} = \{1, 4\}$. We have that $R(3) = \{r_4\}$, $R(1) = \{r_1, r_2\}$, and $R(4) = \{r_5, r_6\}$. Then $\big\{\{r_1, r_4, r_5\}, \{r_1, r_4, r_6\}, \{r_2, r_4, r_5\}, \{r_2, r_4, r_6\}\big\}$ is the reasoning space, in which each element is a reasoning set, and the size is $|R(1)| \times |R(4)| = 2 \times 2 = 4$.

Definition 3.3 (*EIR* l-diversity) Given an original data table D, for an equivalence class Q in D^*, let $Q_{rea} = \{Q_{rea}^1, Q_{rea}^2, \ldots, Q_{rea}^q\}$ be the reasoning space of Q, where $Q_{rea}^i (1 \le i \le q)$ is a reasoning set. Then we say Q satisfies the enhanced identity-reserved (EIR) l-diversity if for $\forall Q_{rea}^i \in Q_{rea}$, Q_{rea}^i contains at least l different sensitive values (i.e. $|Q_{rea}^i.A_s| \ge l$), and D^* satisfies EIR l-diversity if all equivalence classes in published anonymous table D^* satisfy EIR l-diversity.

We do not restrict the number of individuals in an equivalence class with parameter k, because *EIR* l-diversity can ensure that there are at least l different individuals in an equivalence class.

Definition 3.4 (*EIR* (α, β)-anonymity) Given an original data table D, for an equivalence class Q in D^*, let $Q_{rea} = \{Q_{rea}^1, Q_{rea}^2, \ldots, Q_{rea}^q\}$ be the reasoning space of Q. Then Q satisfies the enhanced identity-reserved (EIR) (α, β)-anonymity if the following conditions are satisfied:

(1) $\forall p_i \in Q.Id\_num$, the percentage of p_i's records in Q is not more than α, i.e. $|R(p_i)|/|Q| \le \alpha$; and
(2) $\forall Q_{rea}^i \in Q_{rea}, \forall s_i^j \in Q_{rea}^i.A_s$, the percentage of individuals whose sensitive value is s_i^j in Q_{rea}^i is not more than β, i.e. $|P(s_i^j)|/n_Q \le \beta$.

If all equivalence classes in published anonymous table D^* satisfy EIR (α, β)-anonymity, we say that D^* satisfies EIR (α, β)-anonymity.

Definition 3.5 (*Hitting set*) [28] Let $\Psi = \{X_1, X_2, \ldots, X_t\}$ be a collection of subsets of a finite set X. If $H \subseteq X$, and $H \cap X_i \neq \emptyset, \forall X_i \in \Psi$, H is called a hitting set of Ψ. If there is no $H' \subset H$ such that H' is a hitting set of Ψ, then H is called a minimal hitting set of Ψ. If cardinality of H is smallest, then H is called a minimum hitting set of Ψ.

For example, given $X = \{x_1, x_2, x_3, x_4\}$, $\Psi = \{\{x_2, x_3\}, \{x_2, x_4\}, \{x_1, x_2\}\}$, then $\{x_1, x_3, x_4\}$ and $\{x_2\}$ are minimal hitting sets of Ψ, where $\{x_2\}$ is minimum hitting set of Ψ.

For an equivalence class Q, if the average number of records of an individual in P_Q^{mul} is r, then the size of reasoning space is $r^{|P_Q^{mul}|}$. So we need to check whether $r^{|P_Q^{mul}|}$ reasoning sets satisfy the enhanced identity-reserved privacy model. In fact, we do not check all the reasoning sets. The problems of whether Q satisfies the *EIR* l-diversity and *EIR* (α, β)-anonymity are changed to the problems of minimum hitting set and the highest frequency of sensitive values, respectively.

Theorem 3.1 *Given an equivalence class Q, $P_Q = \{p_1, p_2, \ldots, p_{n_Q}\}$ is the set of individuals in Q. Let $\Psi = \{S(p_1), S(p_2), \ldots, S(p_{n_Q})\}$ and H is a minimum hitting set of Ψ. If $|H| \geq l$, then Q satisfies the EIR l-diversity.*

Proof For each $Q_{rea}^i \in Q_{rea}$, we know that Q_{rea}^i contains a set of records which maps to different individuals from p_1 to p_{n_Q} by Definition 3.2. $Q_{rea}^i.A_s$ is the set of sensitive values of these records in Q_{rea}^i. We have that $Q_{rea}^i.A_s \cap S(p_j) \neq \emptyset, \forall j \in \{1, \ldots, n_Q\}$. So $Q_{rea}^i.A_s$ is a hitting set of Ψ. Since H is a minimum hitting set of Ψ and $|H| \geq l$, we have $|Q_{rea}^i.A_s| \geq l$. By Definition 3.3, we know that Q satisfies the EIR l-diversity. □

In this case, if $|S(P_Q^{sin})| \geq l$, where $S(P_Q^{sin})$ is the set of the values of sensitive attribute of all individuals in P_Q^{sin}, then we can obtain that $|H| \geq l$ because $S(P_Q^{sin})$ is included in any a hitting set of Ψ according to the definition of hitting set (i.e. Definition 3.5). So Q satisfies the *EIR* l-diversity.

Theorem 3.2 *Given an equivalence class Q, let*

$$p = \arg\max\{|R(p_i)| \mid p_i \in Q.Id\_num\}, \tag{9}$$

$$s = \arg\max\{|P(s_i)| \mid s_i \in Q.A_s\}. \tag{10}$$

If $|R(p)|/|Q| \leq \alpha$ and $|P(s)|/n_Q \leq \beta$, then Q satisfies the EIR (α, β)-anonymity.

Proof We know that $|R(p_i)|/|Q| \leq |R(p)|/|Q| \leq \alpha, \forall p_i \in Q.Id\_num$. For $\forall Q_{rea}^i \in Q_{rea}$, we have $\max\{|P(s_i^j)| \mid s_i^j \in Q_{rea}^i.A_s\}/n_Q \leq |P(s)|/n_Q \leq \beta$. By Definition 3.4, Q satisfies the EIR (α, β)-anonymity. □

4 Anonymization

In this section, first we will discuss how to calculate the information loss for numeric and categorical attributes, a record, and a data table, then define various concepts of distances, and finally propose an anonymization algorithm and analyse its correctness and time and space complexity.

4.1 Information metric used in our approach

We extend *GLM* to the definition of information metrics between a generalized value (or original value) and another generalized value for numeric and categorical attributes.

Definition 4.1 (*Information metric for a numeric attribute*) Let the value domain of a numeric attribute A be $[L, U]$. Let the value of a record r be $r[A] = [a, b]$ on the attribute A, which is the original value or a generalized value, and its (later) generalized value be $r^*[A] = [a^*, b^*]$ on the attribute A. Then the information loss of r on numeric attribute A from $r[A]$ to $r^*[A]$ is

$$Loss(r[A], r^*[A]) = \frac{(b^* - a^*) - (b - a)}{U - L}. \tag{11}$$

When $r[A]$ is the original value, we have $b - a = 0$, which is consistent with GLM/NCP as a result for numeric attribute. When $[a, b]$ is a generalized value, $Loss(r[A], r^*[A]) = \frac{b^*-a^*}{U-L} - \frac{b-a}{U-L}$ denotes the increment of information loss from $[a, b]$ to $[a^*, b^*]$.

Definition 4.2 (*Information metric for a categorical attribute*) Let the value domain of a categorical attribute A' be the set X. And let the value of a record r be $r[A']$ on the attribute A', which is the original value or a generalized value, and its (later) generalized value be $r^*[A']$ on the attribute A'. Then the information loss of r on categorical attribute A' from $r[A']$ to $r^*[A']$ is

$$Loss(r[A'], r^*[A']) = \frac{|r^*[A']| - |r[A']|}{|X| - 1}. \tag{12}$$

When $r[A']$ is the original value, $|r[A']| = 1$; it is consistent with *GLM* as a result for categorical attribute. When $r[A']$ is a generalized value, $Loss(r[A'], r^*[A']) = \frac{|r^*[A']|-1}{|X|-1} - \frac{|r[A']|-1}{|X|-1}$ denotes increment of information loss from $r[A']$ to $r^*[A']$.

For our anonymization approach, if there exists an individual p_i, which is added to any an equivalence class and the equivalence class does not satisfy given privacy requirement, or is added to the equivalence class, whose distance to p_i is minimum, and data quality of the equivalence class is decreased seriously, we need to suppress the records of the individual. The suppression of a value of an attribute is to replace it by a special mark (i.e. $*$, a special generalization), indicating that the replaced values are not disclosed, and its information loss is 1.

Definition 4.3 (*Information loss for a record*) The information loss of a record r generalized to r^* is

$$Loss(r, r^*) = \sum_{i=1}^{|QI|} Loss(r[A_i], r^*[A_i]), \tag{13}$$

where $Loss(r[A_i], r^*[A_i])$ is calculated by Eq. (11), if A_i is numeric attribute; otherwise, it is calculated by Eq. (12).

Definition 4.4 (*Information loss for a data table*) Given an original data table D, the information loss of D generalized to its published anonymous table D^* is

$$Loss(D, D^*) = \sum_{i=1}^{|D|} Loss(r_i, r_i^*). \tag{14}$$

The normalized information loss of D generalized to D^* is

$$NLoss(D, D^*) = \frac{Loss(D, D^*)}{|D| \cdot |QI|}. \tag{15}$$

In worst case, we suppress all records.

4.2 Definition of distances

Some approaches consider the problem of data anonymization as a clustering problem satisfying k-anonymity or l-diversity [18,19,24]. They gave some definitions about distance between two records, distance between a record and an equivalence class, and distance between two equivalence classes. In this subsection, we will also give some definitions about distances from information loss perspective, including distance between two individuals, distance between an individual and an equivalence class, and distance between two equivalence classes. Different from the concepts of distances in [18,19,24], we consider that an individual could have multiple records, and use Eqs. (11) and (12) to calculate the information loss for numeric attributes and categorical attributes, respectively.

Definition 4.5 (*Distance between two individuals*) Given the individuals p_1 and p_2, $\overline{r}_{p_1}$ and $\overline{r}_{p_2}$ are the original identity elements of p_1 and p_2, respectively. The information loss caused by generalizing the records in $R(p_1)$ and $R(p_2)$ to $\overline{r}_{p_1p_2}$ is called the distance between p_1 and p_2, defined as follows:

$$Dist(p_1, p_2) = |R(p_1)| \times Loss(\overline{r}_{p_1}, \overline{r}_{p_1p_2}) + |R(p_2)| \times Loss(\overline{r}_{p_2}, \overline{r}_{p_1p_2}), \tag{16}$$

where $\overline{r}_{p_1p_2} = \delta(\overline{r}_{p_1}, \overline{r}_{p_2})$ is the identity element of equivalence class $Q_{p_1p_2}$, which is formed by generalizing the records in $R(p_1)$ and $R(p_2)$.

To make $R(p_1) \cup R(p_2)$ become an equivalence class, we need to generalize these records to $\overline{r}_{p_1p_2}$. In Eq. (16), the first product term denotes the information loss caused by generalizing the records in $R(p_1)$ to $\overline{r}_{p_1p_2}$ and the second is the information loss caused by generalizing the records in $R(p_2)$ to $\overline{r}_{p_1p_2}$.

For example, in Table 2, r_5 and r_6 both belong to individual 4, whose original identity element is {F, 33, 10087, *null*}, and r_8 and r_9 both belong to individual 6, whose original identity element is {F, 34, 10070, *null*}. If we put the records of individuals 4 and 6 into an equivalence class, the identity element of the equivalence class is {F, [33, 34], {10070, 10087}, *null*}, i.e. the records of individuals 4 and 6 are all generalized to {F, [33, 34], {10070, 10087}, *null*}. Then the information loss caused by the generalization is the distance between individuals 4 and 6.

Definition 4.6 (*Distance between individual and equivalence class*) Let $\overline{r}_q$ be the identity element of equivalence class Q. If individual $p \notin Q.Id\_num$, then the information loss caused by generalizing the records in $R(p)$ and Q to $\overline{r}_{pq}$ is the distance between p and Q, given by:

$$Dist(p, Q) = |R(p)| \times Loss(\overline{r}_p, \overline{r}_{pq}) + |Q| \times Loss(\overline{r}_q, \overline{r}_{pq}), \tag{17}$$

where $\overline{r}_p$ is the identity element of p and $\overline{r}_{pq} = \delta(\overline{r}_p, \overline{r}_q)$ is the identity element of the equivalence class Q_{pq}, which is formed by generalizing the records in $R(p)$ and Q.

To make $R(p) \cup Q$ become an equivalence class, we need to generalize these records to $\overline{r}_{pq}$. In Eq. (17), the first product term denotes the information loss caused by generalizing

the records in $R(p_1)$ to $\overline{r}_{pq}$ and the second is the information loss caused by generalizing the records in Q to $\overline{r}_{pq}$.

Definition 4.7 (*Distance between two equivalence classes*) Let $\overline{r}_{q_1}$ and $\overline{r}_{q_2}$ be the identity elements of equivalence classes Q_1 and Q_2, respectively. The information loss caused by generalizing the records in Q_1 and Q_2 to $\overline{r}_{q_{12}}$ is the distance between Q_1 and Q_2, given by:

$$Dist(Q_1, Q_2) = |Q_1| \times Loss(\overline{r}_{q_1}, \overline{r}_{q_{12}}) + |Q_2| \times Loss(\overline{r}_{q_2}, \overline{r}_{q_{12}}), \tag{18}$$

where $\overline{r}_{q_{12}}$ is the identity element of the equivalence class $Q_{1,2}$, which is formed by generalizing the records in Q_1 and Q_2.

To make $Q_1 \cup Q_2$ become an equivalence class, we need to generalize these records to $\overline{r}_{q_{12}}$. In Eq. (18), the first product term denotes the information loss caused by generalizing the records in Q_1 to $\overline{r}_{q_{12}}$ and the second is the information loss caused by generalizing the records in Q_2 to $\overline{r}_{q_{12}}$.

4.3 Algorithm

Wang et al. [24] presented a clustering algorithm for data anonymization achieving l-diversity. In this subsection, by improving their method, we will propose the heuristic greedy clustering algorithm *DAnonyIR*, as shown in Algorithm 1. It generalizes the original data table to an anonymous table which satisfies given privacy requirement for identity reservation. To explain our algorithm, first we need the following concept:

Definition 4.8 (*Optimal clustering*) Given an original data table D and a privacy model with identity reservation π, an optimal clustering of D is a partition $P = \{Q_1, \ldots, Q_e\}$ such that $\bigcap_{i=1}^{e} Q_i = \emptyset$, $\bigcup_{i=1}^{e} Q_i \subseteq D$, and Q_i $(i = 1, \ldots, e)$ satisfies π after Q_i is generalized. The published anonymous table D^* consists of these generalized Q_i, and $Loss(D, D^*)$ is minimal.

DAnonyIR The whole clustering algorithm *DAnonyIR* is shown in Algorithm 1. Its input is the original data table, *QI* attributes, and some parameters about privacy requirement π. The output is an anonymous table. The basic idea of the algorithm is as follows: when $D \neq \emptyset$, we try to create an equivalence class Q from D; if Q satisfies π, we add it to D^*; if $D = \emptyset$ and Q still does not satisfy π, the individuals in Q are residual and we use *Handle* function to deal with them. Firstly, on line 1, we preprocess the original data. That is, recode the explicit identifier of D with numbers. As the information where several records belong to the same individual needs to be kept, the explicit identifier is replaced with a different number to denote a different individual. On lines 2–19, we try to create continually equivalence classes until $D = \emptyset$. The process of creating an equivalence class is shown on lines 3–15, First, select randomly individual p from D and initialize equivalence class Q with these records of p. $\overline{r}_q$ is the identity element of Q. Now Q only contains an individual and it does not satisfy π, so we set $SatFlag = False$, where variable $SatFlag$ denotes whether Q satisfies π. When Q does not satisfy π and $D \neq \emptyset$, we perform repeatedly lines 7–14. On lines 7 and 8, we get the individual p' and the equivalence class Q' from D and D^*, whose distances to Q are minimum, respectively. Because the current Q does not satisfy π, we need to add more individuals. We can add p' to Q, or combine Q with Q'. In order to reduce the information loss, we select the way in which less information loss is caused. If $Dist(p', Q) \leq Dist(Q', Q)$, we add $R(p')$ to Q and update the identity element of Q;

Algorithm 1 *DAnonyIR*

Input: original data table D; $quasi-identifier\ QI$; some parameters about privacy requirement π;
Output: An anonymous table D^*;

```
1: recode the explicit identifier of D with numbers;
2: while D ≠ ∅ do
3:     select randomly individual p from D; D = D − R(p);
4:     form the equivalence class Q = R(p), where r̄_q is the identity element of Q, r̄_q[QI] are the values of
       p on QI attributes, and r̄_q[A_s] = null;
5:     SatFlag = False;
6:     while !SatFlag&&D ≠ ∅ do
7:         p' = argmin_{p_i ∈ D.Id_num}{Dist(p_i, Q)};
8:         Q' = argmin_{Q_k ∈ D*}{Dist(Q_k, Q)};
9:         if Dist(p', Q) ≤ Dist(Q', Q) then
10:            D = D − R(p'); Q = Q ∪ R(p'); r̄_q = δ(r̄_q, r̄_p');
11:        else
12:            D* = D* − {Q'}; Q = Q ∪ {Q'}; r̄_q = δ(r̄_q, r̄_q');
13:        end if
14:        SatFlag = SatPriIR(Q, parameters);
15:    end while
16:    if SatFlag == True then
17:        D* = D* ∪ {Q};
18:    end if
19: end while
20: if SatFlag == False then
21:    while Q ≠ ∅ do
22:        select randomly individual p'' from Q;
23:        Q = Q − R(p'');
24:        Handle(p'');
25:    end while
26: end if
27: for ∀Q_i ∈ D* do
28:    for ∀r ∈ Q_i do
29:        substitute its values on QI attributes with Q_i's identity element;
30:    end for
31: end for
32: return D*;
```

otherwise, we merge Q with Q' and update the identity element of Q. On line 14, we call function *SaPriIR* to judge whether Q satisfies π. On lines 16 and 18, if $SatFlag = True$, i.e. Q satisfies π, we add Q to D^*. On lines 20–26, when the lines 2–19 are executed and $D = \emptyset$, if the last equivalence class Q does not satisfy π, for every individual in Q, we call function *Handle* to decide to add its records to an equivalence class or suppress it. Because these individuals are directly put in D^*, it will lead to privacy leakage. After that, we obtain the set D^* of equivalence classes which are not generalized. Then for $Q_i \in D^*$ and every record in Q_i, we substitute its values on QI attributes with its identity element, so the anonymous table satisfying π is obtained.

SatPriIR For function *SatPriIR*, the procedure is different for a different privacy model with identity reservation. For *IR* (k, l)-anonymity, *IR* (α, β)-anonymity, *EIR* l-diversity, and *EIR* (α, β)-anonymity, it is substituted with *SatPriIR_kl*, *SatPriIR_*$\alpha\beta$, *SatPriIR_El*, and *SatPriIR_E*$\alpha\beta$, respectively. The privacy models *EIR* l-diversity and *EIR* (α, β)-anonymity are proposed in the paper, so we describe the functions *SatPriIR_El* and *SatPriIR_E*$\alpha\beta$ in detail, as shown in Algorithms 2 and 4, respectively. For *SatPriIR_kl*, we scan Q once to obtain the number n_Q of individuals and the number m_Q of sensitive values appearing in Q. If $n_Q \geq k$ and $m_Q \geq l$, then Q satisfies the *IR* (k, l)-anonymity, and return $True$; otherwise,

return $False$. For *SatPriIR_*$\alpha\beta$, the algorithm is similar to *SatPriIR_E*$\alpha\beta$, and the difference is on line 12 in Algorithm 4. According to Definition 2.3 (3), if $MaxRecNum/|Q| \leq \alpha$ and $MaxSenNum/|Q| \leq \beta$, then Q satisfies *IR* (α, β)-anonymity, and return $True$; otherwise, return $False$.

Algorithm 2 *SatPriIR_El*(Q, l)

Input: the set of records Q; parameter l;
Output: $True$ or $False$;
1: get $\Psi = \{S(p_1), S(p_2), \ldots, S(p_{n_Q})\}$;
2: $\xi = \emptyset$;
3: **for** $\forall S(p_i) \in \Psi$ **do**
4: **if** $|S(p_i)| == 1$ **then**
5: $\Psi = \Psi \backslash S(p_i)$;
6: $\xi = \xi \cup S(p_i)$;
7: **end if**
8: **end for**
9: **if** $|\xi| = \emptyset$ **then**
10: $\mathcal{H} = BHS(\Psi)$;
11: find a minimum hitting set h from $\mathcal{H}$;
12: **if** $|h| \geq l$ **then**
13: **return** $True$;
14: **else**
15: **return** $False$;
16: **end if**
17: **else if** $|\xi| \geq l$ **then**
18: **return** $True$;
19: **else**
20: **for** $\forall S(p_i) \in \Psi$ **do**
21: **if** $S(p_i) \cap \xi \neq \emptyset$ **then**
22: $\Psi = \Psi \backslash S(p_i)$;
23: **end if**
24: **end for**
25: $\mathcal{H}' = BHS(\Psi)$;
26: find a minimum hitting set h' from $\mathcal{H}'$;
27: **if** $|h'| + |\xi| \geq l$ **then**
28: **return** $True$;
29: **else**
30: **return** $False$;
31: **end if**
32: **end if**

SatPriIR_El In Algorithm 2, on line 1, we get the collection of subsets Ψ, in which $S(p_i)$ is the set of sensitive values of the individual p_i. On line 2, we set a variable ξ to store single element sets in Ψ. From lines 3 to 8, we find all single element sets, delete them from Ψ and add them to set ξ. On lines 9–16, if $|\xi| = \emptyset$, we directly call function *BHS* [28] to get all minimal hitting sets of Ψ, and find a minimum hitting set h. Function *BHS* is described in Algorithm 3. According to Theorem 3.1, if $|h| \geq l$, then Q satisfies *EIR* l-diversity, and return $True$; otherwise, return $False$. On lines 17 and 18, if $|\xi| \neq \emptyset$ and $|\xi| \geq l$, then the cardinality of any minimum hitting set is not less than l, because ξ is contained in any minimal hitting set, also any minimum hitting set. So return $True$. From lines 19 to 31, we consider another case, $0 < |\xi| < l$. On lines 20 to 24, we use ξ to further simplify Ψ according to the definition of a hitting set. $\forall S(p_i) \in \Psi$, if $S(p_i) \cap \xi \neq \emptyset$, then $\Psi = \Psi \backslash S(p_i)$. On lines 25 and 26, we call *BHS* to get all minimal hitting sets of current Ψ, and then find a minimum hitting set h'. If $|h'| + |\xi| \geq l$, return $True$; otherwise, return $False$. In fact, we divide Ψ

Algorithm 3 $BHS(\Psi)$

Input: the collection of sets Ψ;
Output: all minimal hitting sets of Ψ;
1: transform Ψ to Boolean formula Π with disjunctive normal form;
2: **if** Π only contains a conjunctive item, i.e. $\Pi = \overline{s_1}\,\overline{s_2}\cdots\overline{s_\theta}$ **then**
3: **return** $s_1 + s_2 + \cdots + s_\theta$;
4: **end if**
5: simplify Π with absorption law;
6: **if** every conjunctive item in Π contain literal $\bar{s}$ **then**
7: **return** s;
8: **end if**
9: **if** there are single literal items in Π, i.e.$\overline{s'_1}, \overline{s'_2}, \cdots, \overline{s'_\vartheta}$ **then**
10: $sig = s'_1 s'_2 \cdots s'_\theta$;
11: delete $\overline{s'_1}, \overline{s'_2}, \cdots, \overline{s'_\vartheta}$ from Π;
12: **end if**
13: get the literal s' whose frequency appearing in Π is highest;
14: $sig \cdot (s' \cdot BHS(\Pi_1) + BHS(\Pi_2))$, where Π_1 and Π_2 are the results by deleting these conjunctive items which contains s', and s' from Π, respectively;

to two parts: One contains the sets whose intersection with ξ is not $\emptyset$ and the other contains the sets whose intersection with ξ is $\emptyset$. ξ is the only minimum hitting set of the first part. We need to find a minimum hitting set h' of the second part. $h' \cup \xi$ is a minimum hitting set of the whole Ψ.

BHS We combine an example to explain Algorithm 3. Let

$$\Psi = \{\{x_1, x_3\}, \{x_1, x_3, x_5\}, \{x_1, x_6\}, \{x_3, x_5\}, \{x_5, x_7\}, \{x_4\}, \{x_4, x_5\}, \{x_4, x_6\}\}.$$

On line 1, we transform Ψ to

$$\Pi = \overline{x_1}\,\overline{x_3} + \overline{x_1}\,\overline{x_3}\,\overline{x_5} + \overline{x_1}\,\overline{x_6} + \overline{x_3}\,\overline{x_5} + \overline{x_5}\,\overline{x_7} + \overline{x_4} + \overline{x_4}\,\overline{x_5} + \overline{x_4}\,\overline{x_6},$$

where xy (or $x \cdot y$) and $x + y$ denote the AND and OR results of x and y, respectively. For any hitting set of Ψ, e.g. $\{x_1, x_4, x_5\}$, we have $\Pi \cdot x_1 x_4 x_5 = 0$. On lines 2 and 3, when Π only contains a conjunctive item, and assume that $\Pi = \overline{x_1}\,\overline{x_3}$, return $x_1 + x_3$, i.e. $\{x_1\}$ and $\{x_3\}$ are minimal hitting sets, because $\Pi \cdot x_1 = 0$ and $\Pi \cdot x_3 = 0$. In this example, Π has 8 conjunctive items, the algorithm executes the fifth line. We simplify Π with the absorption law $A + AB = A$ and obtain

$$\Pi = \overline{x_1}\,\overline{x_3} + \overline{x_1}\,\overline{x_6} + \overline{x_3}\,\overline{x_5} + \overline{x_5}\,\overline{x_7} + \overline{x_4}.$$

On lines 6 and 7, if every conjunctive item in Π contains literal $\bar{s}$, then s is the only a minimal hitting sets, because $\Pi \cdot s = 0$. In this example, $\overline{x_4}$ is a single literal item, and line 9 is executed. We have $sig = x_4$, which is contained in all hitting sets. We delete $\overline{x_4}$ from Π, and

$$\Pi = \overline{x_1}\,\overline{x_3} + \overline{x_1}\,\overline{x_6} + \overline{x_3}\,\overline{x_5} + \overline{x_5}\,\overline{x_7}.$$

Select literal $\overline{x_1}$ whose frequency appearing in Π is highest for accelerating convergence. We have

$$x_4 BHS(\Pi) = x_4(x_1 BHS(\overline{x_3}\,\overline{x_5} + \overline{x_5}\,\overline{x_7}) + BHS(\overline{x_3} + \overline{x_6} + \overline{x_3}\,\overline{x_5} + \overline{x_5}\,\overline{x_7})).$$

The first part is these minimal hitting sets containing x_1 and the second part is the ones which do not contain x_1. For $BHS(\overline{x_3}\,\overline{x_5} + \overline{x_5}\,\overline{x_7})$, because the two conjunctive items both contain $\overline{x_5}$, so return x_5.

$$\begin{aligned} &BHS(\overline{x_3} + \overline{x_6} + \overline{x_3}\,\overline{x_5} + \overline{x_5}\,\overline{x_7}) \\ &= BHS(\overline{x_3} + \overline{x_6} + \overline{x_5}\,\overline{x_7}) \\ &= x_3x_6BHS(\overline{x_5}\,\overline{x_7}) \\ &= x_3x_6(x_5 + x_7) \\ &= x_3x_5x_6 + x_3x_6x_7. \end{aligned}$$

So the final return result is

$$\begin{aligned} &x_4(x_1x_5 + x_3x_5x_6 + x_3x_6x_7) \\ &= x_1x_4x_5 + x_3x_4x_5x_6 + x_3x_4x_6x_7, \end{aligned}$$

and $\{x_1, x_4, x_5\}$, $\{x_3, x_4, x_5, x_6\}$, and $\{x_3, x_4, x_6, x_7\}$ are minimal hitting sets.

SatPriIR_Eαβ For Algorithm 4, from lines 2 to 10, we obtain the $MaxRecNum$ which denotes the maximum number of records of individuals in Q and count the number of occurrences of any sensitive value appearing in Q. On line 11, we find $MaxSenNum$ which is the maximum value in $\{Num_{s_1}, \ldots, Num_{s_{m_Q}}\}$, where $s_1, \ldots, s_{m_Q}$ are sensitive values appearing in Q and Num_{s_j} is the number of occurrences of s_j. By Theorem 3.2, if $MaxRecNum/|Q| \leq \alpha$ and $MaxSenNum/n_Q \leq \beta$, Q satisfies *EIR* (α, β)-anonymity, so return $True$; otherwise, return $False$.

Algorithm 4 *SatPriIR_Eαβ*(Q, α, β)

Input: the set of records Q; parameters α and β;
Output: $True$ or $False$;
1: $MaxRecNum = 0$;
2: **for** every p_i in Q **do**
3: $RecNum = |R(p_i)|$;
4: **if** $RecNum > MaxRecNum$ **then**
5: $MaxRecNum = RecNum$;
6: **end if**
7: **for** every $s_j \in S(p_i)$ **do**
8: $Num_{s_j}++$;
9: **end for**
10: **end for**
11: $MaxSenNum = max\{Num_{s_1}, \ldots, Num_{s_{m_Q}}\}$;
12: **if** $MaxRecNum/|Q| \leq \alpha$ and $MaxSenNum/n_Q \leq \beta$ **then**
13: **return** $True$;
14: **else**
15: **return** $False$;
16: **end if**

In fact, when some records of an individual are added to Q, in order to check whether the current Q satisfies privacy requirement π, we do not call *SatPriIR_kl*, *SatPriIR_El*, *SatPriIR_αβ*, or *SatPriIR_Eαβ*. Because these individuals are added to Q one by one (when we combine an equivalence class Q' to Q, we may consider as the individuals of Q' are added to Q one by one), we can use incremental methods to check whether Q satisfies π, denoted by *SatPriIRInc_kl*, *SatPriIRInc_El*, *SatPriIRInc_αβ*, and *SatPriIRInc_Eαβ*, respectively. *SatPriIRInc_El* and *SatPriIRInc_Eαβ* are shown in Algorithms 5 and 6, respectively. For *SatPriIRInc_kl*, when an individual p is added to Q, we only need to update the number n_Q of individuals and the number m_Q of sensitive values appearing in Q according to p. For

SatPriIRInc_αβ, the updated process to $MaxRecNum$ and $MaxRecNum$ is the same as *SatPriIRInc_Eαβ*.

Algorithm 5 *SatPriIRInc_El*$(Q, \mathcal{H}_Q, p, l)$

Input: the set of records Q; all minimal hitting sets $\mathcal{H}_Q = x_1x_2 \dots x_{h_1} + y_1y_2 \dots y_{h_2} + \dots + z_1z_2 \dots z_{h_t}$ of $\Psi = \{S(p_1), S(p_2), \dots, S(p_{n_Q})\}$; individual p added to Q with $S(p) = \{s_1, s_2, .., s_r\}$; parameter l;
Output: $True$ or $False$;
1: **if** $Q = \emptyset$ **then**
2: $\quad \mathcal{H}_{Qp} = s_1 + s_2 + \dots + s_r$;
3: **else**
4: $\quad \mathcal{H}_{Qp} = (x_1x_2 \dots x_{h_1} + y_1y_2 \dots y_{h_2} + \dots + z_1z_2 \dots z_{h_t})(s_1 + s_2 + \dots + s_r)$;
5: $\quad$ simplify $\mathcal{H}_{Qp}$ with Boolean algebra;
6: $\quad$ find a minimum hitting set h from $\mathcal{H}_{Qp}$;
7: $\quad$ **if** $h \geq l$ **then**
8: $\quad\quad$ **return** $True$;
9: $\quad$ **else**
10: $\quad\quad$ **return** $False$;
11: $\quad$ **end if**
12: **end if**

SatPriIRInc_El The idea of $SatPriIRInc\_El$ is introduced by [28]. For the set of records Q,

$$\mathcal{H}_Q = x_1x_2 \dots x_{h_1} + y_1y_2 \dots y_{h_2} + \dots + z_1z_2 \dots z_{h_t}$$

contains all minimal hitting sets of $\Psi = \{S(p_1), S(p_2), \dots, S(p_{n_Q})\}$. In fact,

$$\mathcal{H}_Q = \{\{x_1, x_2, \dots, x_{h_1}\}, \{y_1, y_2, \dots, y_{h_2}\}, \dots, \{z_1, z_2, \dots, z_{h_t}\}\}.$$

For convenience, we represent $\mathcal{H}_Q$ with Boolean formula. When an individual p is added to Q, we use Algorithm 5 to get all minimal hitting sets $\mathcal{H}_{Qp}$ of $\Psi \cup S(p)$. If $Q = \emptyset$, then $\mathcal{H}_{Qp} = s_1 + s_2 + \dots + s_r$, i.e. $\mathcal{H}_{Qp} = \{s_1, s_2, \dots, s_r\}$; otherwise, we use the method shown on line 4 to get $\mathcal{H}_{Qp}$ and simplify it with Boolean algebra. Then find a minimum hitting set h from $\mathcal{H}_{Qp}$. If $h \geq l$, return *True*; otherwise, return *False*.

SatPriIRInc_Eαβ When an individual p is added to Q, Algorithm 6 is used to check whether at the moment Q satisfies *EIR* (α, β)-anonymity. We only need to check whether $|R(p_i)|$ is greater than $MaxRecNum$ and the numbers of occurrences of $s_1, s_2, .., s_r$ in Q at the moment are greater than $MaxSenNum$ to decide whether to update $MaxRecNum$ and $MaxSenNum$.

Handle Function *Handle* is shown in Algorithm 7. For every residual individual p'' in D, we need to decide to add its records to an equivalence class or suppress it. On lines 1 and 2, we set two variables $MinDis$ and Min. The initial value of $MinDis$ is a greater value. From lines 3 to 10, we find the equivalence class Q_{min} which still satisfies privacy requirement π after merging p'' (this step is ignored for *IR* (k, l)-anonymity, and *EIR* l-diversity, as the equivalence class after adding the records of an individual satisfies still them, if an equivalence class satisfies the two privacy models) and the distance to p'' is minimum. If distance $MinDis$ is less than the information loss caused by suppressing $R(p'')$, we merge p'' to Q_{min}; otherwise, we suppress $R(p'')$.

When $D = \emptyset$ and the equivalence class Q still does not satisfy privacy requirement π, we call function *Handle* to deal with these individuals in Q. In Algorithm 7, for every residual individual p'' we find Q_{min}. If p'' is added to Q_{min}, we need to generalize the records of p'' and Q_{min} to $\overline{r}_{p''q_{min}}$, which is the identity element of the equivalence class

Algorithm 6 *SatPriIRInc_Eαβ*$(Q, SenNum_Q, MaxRecNum, MaxSenNum, p, \alpha, \beta)$

Input: the set of records Q; the array $SenNum_Q$ contains the number of occurrences of every sensitive value in Q; $MaxRecNum$ is the maximum number of records of individuals in Q; $MaxSenNum$ is the maximum number of occurrences of sensitive values in Q; individual p added to Q with $S(p) = \{s_1, s_2, .., s_r\}$; parameters α and β;
Output: $True$ or $False$;
1: **if** $|R(p_i)| > MaxRecNum$ **then**
2: $MaxRecNum = |R(p_i)|$;
3: **end if**
4: **for** every $s_j \in S(p)$ **do**
5: find s_j's corresponding position k in $SenNum$;
6: $SenNum[k] + +$;
7: **if** $SenNum[k] > MaxSenNum$ **then**
8: $MaxSenNum = SenNum[k]$;
9: **end if**
10: **end for**
11: **if** $MaxRecNum/|Q| \leq \alpha$ and $MaxSenNum/n_Q \leq \beta$ **then**
12: **return** $True$;
13: **else**
14: **return** $False$;
15: **end if**

Algorithm 7 *Handle*(p'')

Input: an individual (p'');
Output: the set D^* of equivalence classes without generalization;
1: $MinDis = MaxValue$;
2: $Min = 0$;
3: **for** $\forall Q_i \in D^*$ **do**
4: **if** $SatPriIR(Q_i \cup R(p''), parameters)$ **then**
5: **if** $Dist(p'', Q_i) < MinDis$ **then**
6: $MinDis = Dist(p'', Q_i)$;
7: $Min = i$;
8: **end if**
9: **end if**
10: **end for**
11: **if** $MinDis \leq |R(p'')| \times |QI|$ **then**
12: $Q_{min} = Q_{min} \cup R(p'')$;
13: $\overline{r}_{q_{min}} = \delta(\overline{r}_{q_{min}}, \overline{r}_{p''})$;
14: **else**
15: suppress $R(p'')$;
16: **end if**

$Q_{p''q_{min}}$, formed by generalizing the records in p'' and Q_{min}, and $MinDis$ is the information loss caused by generalization. Should we generalize or suppress p''? We select the way that causes less information loss. That is, if suppression is selected, the information loss caused by suppression is less than that caused by generalization.

Running Example We utilize our *DAnonyIR* by calling *SatPriIRInc_kl*, *SatPriIRInc_El*, *SatPriIRInc_αβ* and *SatPriIRInc_Eαβ* to generalize original data table to anonymous tables, which satisfy *IR* (k, l)-anonymity, *EIR* l-diversity, *IR* (α, β)-anonymity, and *EIR* (α, β)-anonymity, denoted by *DAnonyIR_kl*, *DAnonyIR_El*, *DAnonyIR_αβ*, and *DAnonyIR_Eαβ*, respectively. Assume that we apply *DAnonyIR _El* $(l = 3)$ to anonymize data D in Table 2, in which the domains of $gender$, Age, and $Postcode$ are $\{M, F\}$, [30, 39], {10085, 10076, 10086, 10087, 10077, 10070, 10073}, respectively. First, recode the *Name* of D with num-

bers. That is, the attribute *Name* is substituted with *Id_num*, and the values of *Name* {*Mike*, *Lily*,…, *Lucy*} are substituted with {1, 2,…, 7}. So we have that $D \neq \emptyset$.

We try to create an equivalence class Q from D. Select randomly an individual p from D, and assume that individual 6 is selected. $D = D - R(6)$ and $Q = R(6)$. The identity element $\overline{r}_q$ of Q is $\{F, 34, 10070, null\}$. Because Q only contains individual 6 and it does not satisfy *EIR* 3-diversity, we set $SatFlag = False$. $!SatFlag = True$ and $D \neq \emptyset$, so we add continually an individual to Q. The individual whose distance to Q is minimum is the one with *Id_num*=7 for our example. The computation of distance between individual 7 and Q is as follows: $\overline{r}_{7q}$ is the identity element of $Q \cup R(7)$, which is $\{F, [33, 34], \{10070, 10073\}, null\}$. Thus

$$
\begin{aligned}
&Loss(7, Q)\\
&= |R(7)| \times Loss(\overline{r}_7, \overline{r}_{7q}) + |Q| \times Loss(\overline{r}_q, \overline{r}_{7q})\\
&= 1 \times \left(\frac{|\{F\}| - |\{F\}|}{2-1} + \frac{(34-33)-(33-33)}{39-30} + \frac{|\{10070, 10073\}| - |\{10073\}|}{7-1} \right)\\
&\quad + 2 \times \left(\frac{|\{F\}| - |\{F\}|}{2-1} + \frac{(34-33)-(34-34)}{39-30} + \frac{|\{10070, 10073\}| - |\{10070\}|}{7-1} \right)\\
&= 1 \times \left(0 + \frac{1}{9} + \frac{1}{6} \right) + 2 \times \left(0 + \frac{1}{9} + \frac{1}{6} \right)\\
&= 0.833,
\end{aligned}
$$

where $\overline{r}_7$ is the identity element of individual 7, which is $\{F, 33, 10073, null\}$. The distances between individuals 1, 2, 3, 4, 5 and Q are 5.556, 1.500, 4.167, 1.111, and 1.833, respectively.

We execute line 10 in Algorithm 1. $D = D - R(7)$, $Q = Q \cup R(7)$, and $\overline{r}_q$ is $\{F, [33, 34], \{10070, 10073\}, null\}$. When we check whether Q satisfies *EIR* 3-diversity, we do not call *SatPriIR_El* and use the incremental method *SatPriIRInc_El*. That is, we do not call *SatPriIR_El* to find a minimum hitting set of

$$\Psi = \Psi \cup \{\{Syphilis\}\} = \{\{Leukaemia, Heart\}, \{Syphilis\}\}.$$

By using *SatPriIRInc_El*,

$$(Leukaemia + Heart)Syphilis = Leukaemia \cdot Syphilis + Heart \cdot Syphilis.$$

The collection of all minimal hitting sets of previous Ψ is $\{\{Leukaemia\}, \{Heart\}\}$, and the collection of all minimal hitting sets of current Ψ is $\{\{Leukaemia, Syphilis\}, \{Heart, Syphilis\}\}$, whose cardinalities both are less than 3. Q does not satisfy *EIR* 3-diversity and $D \neq \emptyset$, so we add continually an individual to Q. The distance of individual 4 to Q are is minimum, which can be calculated as follows: $\overline{r}_{4q}$ is the identity element of $Q \cup R(4)$, which is $\{F, [33, 34], \{10070, 10073, 10087\}, null\}$, and $\overline{r}_4$ is the identity element of individual 4, which is $\{F, 33, 10087, null\}$, and thus

$$
\begin{aligned}
&Loss(4, Q)\\
&= |R(4)| \times Loss(\overline{r}_4, \overline{r}_{4q}) + |Q| \times Loss(\overline{r}_q, \overline{r}_{4q})\\
&= 2 \times \left(\frac{|\{F\}| - |\{F\}|}{2-1} + \frac{(34-33)-(33-33)}{39-30} + \frac{|\{10070, 10073, 10087\}| - |\{10087\}|}{7-1} \right)\\
&\quad + 3 \times \left(\frac{|\{F\}| - |\{F\}|}{2-1} + \frac{(34-33)-(34-33)}{39-30} + \frac{|\{10070, 10073, 10087\}| - |\{10070, 10073\}|}{7-1} \right)
\end{aligned}
$$

$$= 2 \times \left(0 + \frac{2}{9} + \frac{2}{6}\right) + 3 \times \left(0 + \frac{0}{9} + \frac{1}{6}\right)$$
$$= 1.389.$$

The distances between individuals 1, 2, 3, 5 and Q are 7.501, 2.278, 5.834, and 2.722, respectively.

Then $D = D - R(4)$, $Q = Q \cup R(4)$, and $\overline{r}_q$ is $\{F, [33 \sim 34], \{10070, 10073, 10087\}, null\}$. And

$$\begin{aligned}\Psi &= \Psi \cup \{\{Hypertension, Diabetes\}\} \\ &= \{\{Leukaemia, Heart\}, \{Syphilis\}, \{Hypertension, Diabetes\}\}.\end{aligned}$$

By using $SatPriIRInc\_El$, the collection of all minimal hitting sets of Ψ is $\{\{Leukaemia, Syphilis, Hypertension\}, \{Leukaemia, Syphilis, Diabetes\}, \{Leukaemia, Syphilis, Diabetes\}, \{Heart, Syphilis, Diabetes\}\}$ by

$$\begin{aligned}&(Leukaemia \cdot Syphilis + Heart \cdot Syphilis)(Hypertension + Diabetes) \\ &= Leukaemia \cdot Syphilis \cdot Hypertension + Heart \cdot Syphilis \cdot Hypertension \\ &\quad + Leukaemia \cdot Syphilis \cdot Diabetes + Heart \cdot Syphilis \cdot Diabetes,\end{aligned}$$

in which every set is a minimum hitting set, and the cardinality is equal to 3. Q satisfies *EIR* 3-diversity. Algorithm 1 executes line 17; the first equivalence class is added to D^*, i.e. $D^* = \{\{r_8, r_9, r_{10}, r_5, r_6\}\}$.

$D = \{r_1, r_2, r_3, r_4, r_7\}$ and $D \neq \emptyset$. We try to create another equivalence class from D. Select randomly individual p from D. Assume that individual 3 is selected, then $D = D - R(3)$ and $Q = R(3)$. We have $\overline{r}_q = \{M, 36, 10086, null\}$. Q only contains an individual and it does not satisfy 3-diversity. Also $D \neq \emptyset$. Thus we add continually an individual to Q. The individual is 1, whose distance to Q is 0.500 and is minimum. The distance between Q and the first equivalence class in D^* is 8.778. Therefore, $Q = Q \cup R(1)$, and $\overline{r}_q = \{M, 36, \{10085, 10086\}, null\}$. Then $R(2)$ and $R(5)$ are added consecutively to Q, and Q satisfies *EIR* 3-diversity.

$$\begin{aligned}D^* &= D^* \cup \{Q\} \\ &= \{\{r_5, r_6, r_8, r_9, r_{10}\}, \{r_1, r_2, r_3, r_4, r_7\}\}.\end{aligned}$$

Now $D \neq \emptyset$ and $SatFlag = True$, i.e. there are no residual individuals. Algorithm 1 executes line 27. For every equivalence class Q in Q^*, we substitute its values on QI attributes with Q's identity element.

Published anonymous table D^*, as shown in Table 4, satisfies *EIR* 3-diversity, and it can prevent *identity disclosure* and *attribute disclosure*. We also take *Mike* as an example. Assume that an attacker knows $Mike$'s QI information (i.e. $\{M, 36, 10085\}$) and knows that $Mike$ is in published table D^*, then the attacker can infer that *Mike* is in equivalence class Q_2. There are three different individuals in Q_2, and the attacker cannot know which one is corresponding to *Mike*. Thus, the *identity disclosure* is prevented. In Q_2, there are two reasoning sets, i.e. $\{r_1, r_3, r_4, r_7\}$ and $\{r_2, r_3, r_4, r_7\}$. Every reasoning set contains at least three different sensitive values, so the attacker cannot know which sensitive disease is corresponding to *Mike*, and so *attribute disclosure* is prevented.

Similarly, we use $DAnonyIR\_E\alpha\beta$ with $\alpha = 0.4$ and $\beta = 0.6$ to anonymize data D in Table 2. Assume the first selected individuals are 6 and 3 in creating equivalence classes Q_1 and Q_2, respectively. The anonymous table is also shown in Table 4, which satisfies *EIR* $(0.4, 0.6)$-diversity. We also take *Mike* as an example. According to the background

Table 4 *EIR* 3-diverse or *EIR* (0.4, 0.6)-anonymous table

	EC_ID	Id_num	Gender	Age	Postcode	Disease
r_5	Q_1	4	F	[33, 34]	{10070, 10073, 10087}	*Hypertension*
r_6	Q_1	4	F	[33, 34]	{10070, 10073, 10087}	*Diabetes*
r_8	Q_1	6	F	[33, 34]	{10070, 10073, 10087}	*Leukaemia*
r_9	Q_1	6	F	[33, 34]	{10070, 10073, 10087}	*Heart*
r_{10}	Q_1	7	F	[33, 34]	{10070, 10073, 10087}	*Syphilis*
r_1	Q_2	1	$\{M, F\}$	[36, 38]	{10076, 10077, 10085, 10086}	*Hypertension*
r_2	Q_2	1	$\{M, F\}$	[36, 38]	{10076, 10077, 10085, 10086}	*Heart*
r_3	Q_2	2	$\{M, F\}$	[36, 38]	{10076, 10077, 10085, 10086}	*Cancer*
r_4	Q_2	3	$\{M, F\}$	[36, 38]	{10076, 10077, 10085, 10086}	*Hypertension*
r_7	Q_2	5	$\{M, F\}$	[36, 38]	{10076, 10077, 10085, 10086}	*HIV*

knowledge of an attacker, *Mike* is inferred in equivalence class Q_2. Because the percentage of any individual's records in Q_2 is not more than 0.4, there are multiple individuals. The attacker cannot know which one is corresponding to *Mike*, and the *identity disclosure* is prevented. As shown above, there are two reasoning sets in Q_2. In every reasoning set of Q_2, the percentage of any sensitive value is not more than 0.6. So the attacker cannot know which sensitive disease is corresponding to *Mike* with certain probability (> 0.6) and *attribute disclosure* is prevented.

4.4 Analysis of algorithm

In this section, we first analyse the correctness of algorithm *DAnonyIR* and then give its time and space complexity analysis.

4.4.1 Analysis of correctness

If given original table D satisfies privacy model π, algorithm *DAnonyIR* can transform D to D^* which satisfies π. When $D \neq \emptyset$, we try to create an equivalence class Q. Firstly, we randomly select an individual and add its records to Q. If Q does not satisfy π and $D \neq \emptyset$, we add continually individuals to Q until Q satisfies π or $D = \emptyset$. If Q satisfies π, we add it to published anonymous table D^*. If Q does not satisfy π and $D = \emptyset$, then these individuals in Q are residual. We add every individual in Q to an equivalence class, which distance to the individual is smallest and still satisfies π after being combined with the individual, or we suppress the individual.

If given original table D does not satisfy π, we remove all individuals to the equivalence class Q and Q still does not satisfy π. Then these individuals in Q are residual, algorithm *DAnonyIR* calls *Handle* function to deal with them. Finally, all individuals in D are suppressed.

4.4.2 Analysis of time complexity

For our algorithm *DAnonyIR*, given an original data table D, let $|D.Id\_num| = n$ (the ID attribute of D has been substituted with Id_num), $|QI| = d$, the numbers of equivalence

classes $|D^*| = e$, $|Q.Id\_num| = q$, m be the size of domain of sensitive attribute, r be the average number of sensitive values or the average records of an individual, and h be the number of all minimal hitting sets of an equivalence class.

We check whether Q satisfies privacy requirement π, and the time is $O(nr)$ for *SatPriIR_kl*, *SatPriIR_*$\alpha\beta$, and *SatPriIR_E*$\alpha\beta$, which scan once the data with some simple comparison. For *SatPriIR_El*, we first scan the data to obtain Ψ and obtain the number of single-record individuals shown on lines 1–8 of Algorithm 2, and the time is $O(nr + n)$. In worst case, we need to call BHS to compute all minimal hitting sets and then find a minimum hitting set. The time of BHS is $O(2^m)$, because we may select every sensitive value to divide Boolean formula Π to two parts in worst case, as shown on line 14 of Algorithm 3. So the time of *SatPriIR_El* is $O(nr + n + 2^m)$, i.e. $O(nr + 2^m)$. Q contains some single-record individuals and we assume that the number is s. If $s \geq l$, algorithm *SatPriIR_El* returns $True$ and ends. When it is not true, we call BHS to calculate all minimal hitting sets.

In fact, we use incremental method *SatPriIRInc_kl*, *SatPriIRInc_El*, *SatPriIRInc_*$\alpha\beta$, or *SatPriIRInc_E*$\alpha\beta$ to check whether Q satisfies π. When an individual is added to Q, we only consider how the added individual influences the privacy. The time is $O(r)$ for *SatPriIRInc_kl*, *SatPriIRInc_*$\alpha\beta$, and *SatPriIRInc_E*$\alpha\beta$. For *SatPriIRInc_El*, we need to consider hr hitting sets when an individual is added to Q, and so the time is $O(hr)$.

For *Handle*, it scans D^* to check whether each equivalence class satisfies π after adding an individual, and computes corresponding distances in QI. For *IR* (k, l)-anonymity and *EIR* l-diversity, an equivalence class still satisfies the privacy model after adding an individual to it. So we need not to check. For *IR* (α, β)-anonymity and *EIR* (α, β)-anonymity, we call *SatPriIRInc_*$\alpha\beta$ and *SatPriIRInc_E*$\alpha\beta$ to check it. The time is $O(r)$. The main time is spent in finding the equivalence class whose distance to the individual is minimum and the time is $O(ed)$. So the time of *Handle* is $O(ed)$ for *IR* (k, l)-anonymity and *EIR* l-diversity, and $O(ed + r)$ for *IR* (α, β)-anonymity and *EIR* (α, β)-anonymity.

The time complexity analysis of our algorithm *DAnonyIR* is shown as follows. (1) On line 1, we recode explicit identifier of D with numbers, and the executed time is $O(n)$. (2) From lines 2–19, the while loop needs to be run $e+1$ times (the last obtained equivalence class may not satisfy π). Each loop creates an equivalence class and needs time is described as follows: we need to scan $q-1$ times D and D^* and compute corresponding distances in QI in order to create an equivalence class, because the first individual in the equivalence class is randomly selected; due to $|D.Id\_num| + |D^*.Id\_num| \leq n$, the time is dn for once scanning D and D^* and computing distances to obtain the individual and equivalence class, whose distances to Q are minimum (lines 7 and 8 in Algorithm 1); when an individual is added to Q or an equivalence class Q' is combined to Q, we call incremental method check whether Q satisfies privacy requirement π; in first case, the time is $O(r)$ for *SatPriIRInc_kl*, *SatPriIRInc_*$\alpha\beta$, and *SatPriIRInc_E*$\alpha\beta$; for the latter case, we can add the individuals of Q' to Q one by one, and the time is $O(qr)$. For *SatPriIRInc_El*, the time is $O(hr)$ and $O(qhr)$ for these two cases, respectively. So the time of this step is $O((e+1)(q-1)(dn+qr))$ for *DAnonyIR_kl*, *DAnonyIR_*$\alpha\beta$ and *DAnonyIR_E*$\alpha\beta$, and $O((e+1)(q-1)(dn+qhr))$ *DAnonyIR_El*. (3) When Q does not satisfy π and $D = \emptyset$, we need to run the while loop on lines 21-25. The number of residual individuals is less than q, because the residual individuals are not created to an equivalence class that satisfies π, so the loop is executed at most q times. Every loop calls function *handle*. The time of the while loop is $O(edq)$ for *IR* (k, l)-anonymity and *EIR* l-diversity, and $O(q(ed + r))$ for *IR* (α, β)-anonymity and *EIR* (α, β)-anonymity. (4) We scan every Q and substitute the values in QI with Q's identity element. The time is $O(eqd)$.

The total time is $O(n) + O((e+1)(q-1)(dn+qr)) + O(edq)$ for *DAnonyIR_kl* and $O(n)+O((e+1)(q-1)(dn+qr))+O(q(ed+r))$ for *DAnonyIR_*$\alpha\beta$ and *DAnonyIR_E*$\alpha\beta$, i.e. $O(dn^2 + qnr - end)$ for these three algorithms, because $eq \leq n$. The total time is $O(n)+O((e+1)(q-1)(dn+qhr))+O(edq)$, i.e. $O(dn^2+qnhr-end)$ for *DAnonyIR_El*.

When d increases, the time of *DAnonyIR* is increased linearly. If the parameter l increases, an equivalence class needs more individuals to make it satisfy *IR* (k, l)-anonymity or *EIR* l-diversity, and so e decreases but q increases. From the time complexity analysis, we know that the runtime is increased. As parameter α (or β) increases, q decreases and e increases, and so the runtime is decreased.

4.4.3 Analysis of space complexity

We can store the records of an individual in a node of a linked list for saving storage memory, because the QI attributes of the records of an individual are the same. A node is denoted by a *struct* of C++, and the *struct* contains $|QI| + 1$ variables and an array, which is used to store the sensitive values of an individual. We need $O((1 + d + r)n)$ units to store the data, where an individual needs $1 + d + r$ units to denote 1 explicit identifier, d values in QI, and r sensitive values. When published anonymous table D^* is outputted, we transform it in the form of relational data.

When we check whether Q satisfies privacy requirement π, $O(m)$ units are used to store the frequency of each sensitive value appearing in Q for *SatPriIR_kl*, *SatPriIR_*$\alpha\beta$, and *SatPriIR_E*$\alpha\beta$. For *SatPriIR_El*, we first scan Q to obtain Ψ, and need $O(qr)$ units to store it. In worst case, we call BHS to compute all minimal hitting sets, and need $O(2^m qr)$ units to store, because every sensitive value may be selected to divide the Boolean formula Π to two parts in worst case as shown on line 14 of Algorithm 3.

In fact, we use incremental method *SatPriIRInc_kl*, *SatPriIRInc_El*, *SatPriIRInc_*$\alpha\beta$, or *SatPriIRInc_E*$\alpha\beta$ to check whether Q satisfies π. We still need $O(m)$ units used to store the frequency of each sensitive value appearing in Q for *SatPriIRInc_kl*, *SatPriIRInc_*$\alpha\beta$, and *SatPriIRInc_E*$\alpha\beta$. For *SatPriIRInc_El*, we need $O(hm)$ units to store all the minimal hitting sets related to Q.

The space complexity analysis of our algorithm *DAnonyIR* is shown as follows. (1) We need $O((1+d+r)n)$ memory units to store the data original dataset D. (2) For lines 2–19, the while loop needs to be run at most $e+1$ times, and e equivalence classes are generated. When an equivalence class is generated, we need $O(q(1 + d + r))$ units to store it. So in this loop we need $O(m+eq(1+d+r))$ units for *DAnonyIR_kl*, *DAnonyIR_*$\alpha\beta$, and *DAnonyIR_E*$\alpha\beta$, and $O(hm + eq(1 + d + r))$ units for *DAnonyIR_El*. (3) For other steps, we do not need extra space.

The total space is $O((1 + d + r)n) + O(m + eq(1 + d + r))$, i.e. $O(m + nd + nr)$ for *DAnonyIR_kl*, *DAnonyIR_*$\alpha\beta$, and *DAnonyIR_E*$\alpha\beta$, and $O((1+d+r)n)+O(hm+eq(1+d+r))$, i.e. $O(hm + nd + nr)$ for *DAnonyIR_El*, because $eq \leq n$.

When d increases, the space that *DAnonyIR* required is increased linearly. If the parameter l increases, an equivalence class needs more individuals to make it satisfy *IR* (k, l)-anonymity or *EIR* l-diversity, and so e decreases but q increases. Because eq is invariable, it is no influence on *DAnonyIR_kl*. For *DAnonyIR_El*, when the size of an equivalence class is increased, h is increased. So does the space *DAnonyIR_El* required. As parameter α (or β) increases, q decreases and e increases. Because eq is invariable, the space is no influence on *DAnonyIR_*$\alpha\beta$ or *DAnonyIR_E*$\alpha\beta$.

5 Experimental analysis

IR (k, l)-anonymity and *IR* (α, β)-anonymity are proposed for solving the anonymous problem with multiple records with identity reservation. To the best of our knowledge, there is no further research on *IR* (k, l)-anonymity and *IR* (α, β)-anonymity, and generalization algorithm *GeneIR* [13] is only an algorithm for achieving the two privacy models. So we can only use *GeneIR* to benchmark our approaches *DAnonyIR_El* and *DAnonyIR_Eαβ*. We implement *GeneIR* for *IR* (k, l)-anonymity and *IR* (α, β)-anonymity in [13], denoted by *GeneIR_kl*, and *GeneIR_αβ*, respectively. For avoiding the influence caused by different algorithms, we also compare *DAnonyIR_kl* and *DAnonyIR_αβ* with our *DAnonyIR_El* and *DAnonyIR_Eαβ*, respectively.

In anonymous table, these equivalence classes, which satisfy *IR* (k, l)-anonymity (*IR* (α, β)-anonymity) but do not satisfy *EIR* l-diversity (*EIR* (α, β)-anonymity), are called vulnerable equivalence classes. These vulnerable equivalence classes may cause privacy leakage, because the *IR* (k, l)-anonymity does not ensure that an attacker knows the sensitive value of an individual is one of l sensitive values, and maybe it is one of less l sensitive values, but *EIR* l-diversity can ensure that. Likewise, *IR* (α, β)-anonymity does not ensure that an attacker knows the sensitive value of an individual with probability of at most β, and maybe it is with probability more than β, but *EIR* (α, β)-anonymity can ensure that.

Besides information loss, we also study the utility of the anonymized data based on the accuracy of query answering, because it is the basis of statistical analysis and many data mining applications (e.g. association rule mining and decision trees). The type of aggregation queries is defined as follows [23,29]:

SELECT COUNT(*) FROM Anonymized data D^*
WHERE $pred(A_1)$ AND …AND $pred(A_\lambda)$ AND $pred(A_s)$

where A_j is a QI attribute. The query has predicates on the λ randomly selected QI attributes and sensitive attribute A_s. Given a query, the precise result $prec$ is computed from the original table, and the estimated result est is obtained from the anonymized table, defined as follows:

$$est = \sum_{Q_i \in D^*} p_{Q_i}^{A_1} \times \ldots \times p_{Q_i}^{A_\lambda} \times num_{Q_i}^{A_s}, \tag{19}$$

where $p_{Q_i}^{A_j} = |pred(A_j) \cap \overline{r}_q.A_j| / |\overline{r}_q.A_j|$ $(j = 1, \ldots, \lambda)$ is the percentage of the intersection of $pred(A_j)$ and generalized value on attribute A_j in Q_i, and $num_{Q_i}^{A_s}$ is the number of individuals in Q_i which satisfy $pred(A_s)$. The *relative error ratio* is defined as $|est - prec|/prec$.

The purposes of our experiments are as follows. (1) We use the percentage, v, of vulnerable equivalence classes in anonymous table to show the vulnerability of *IR* (k, l)-anonymity or *IR* (α, β)-anonymity, i.e. $v = Num_{vec}/Num_{ec}$, where Num_{vec} and Num_{ec} are the numbers of vulnerable equivalence classes and all equivalence classes in an anonymized table, respectively. (2) From data quality, including information loss and accuracy of query answering, and runtime, we analyse the performance of *DAnonyIR_El*, compared with *DAnonyIR_kl* and *GeneIR_kl*; also, we analyse the performance of *DAnonyIR_Eαβ*, compared with *DAnonyIR_αβ* and *GeneIR_αβ*.

For conveniently comparing *IR* (k, l)-anonymity with *EIR* l-diversity, we set parameter $k = l$, because *EIR* l-diversity can ensure that there are at least l different individuals in an equivalence class according to Definition 3.3. The algorithms are implemented in C++ and ran on a computer with a four-core 3.2GHz CPU and 8GB RAM running Windows

Table 5 Detailed description of the dataset used in our experiment

Attribute	Distinct values
Month of birth	12
Year of birth	82
Gender	2
Race	6
EducYear	9
Marry	6
Poverty	5
HISPANX	2
Diagnosis	619

7. All experiment results are the mean values of those from 50 experiments. To compare our *DAnonyIR_El* with *DAnonyIR_kl* and *GeneIR_kl*, parameters l and $|QI|$ are varied to show variation trends with respect to the percentage of vulnerable equivalence classes, data quality, and runtime of these algorithms. To analyse the performance of *DAnonyIR_Eαβ*, *DAnonyIR_αβ*, and *GeneIR_αβ*, parameters α, β, and $|QI|$ are varied.

We use a real-world dataset, appeared in INFORMS data mining contest 2008, in which each patient has one or multiple diagnosis records.[3] This dataset includes 456,767 records of 49,384 different patients. We have used the following attributes of the dataset: *Month of Birth*, *Year of Birth*, *Gender*, *Race*, *EducYear*, *Marry*, *Poverty*, *HISPANX*, and *Diagnosis*, where *EducYear* denotes the years of education, *Marry* denotes the marital status, *Poverty* denotes the economic condition, and *HISPANX* denotes whether an individual is Hispanic. In our experiments, *Diagnosis* is a sensitive attribute. Because our approach considers a sensitive attribute, if there are multiple sensitive attributes, we handle them as follows: (1) if A'_s is in *other attributes* (we have explained how to handle *other attributes* in footnote 1), we may do not publish it, and (2) if the sensitive attribute A'_s has to be published, we consider it as the sensitive attribute which we protect, and residual attributes are considered as *QI* or *other attributes*. The detailed description of the dataset is shown in Table 5. In order to show the results with the change of $|QI|$, we set $|QI|$ from 3 to 8. When $|QI| = d$ ($d \in \{3, \ldots, 8\}$), it means that QI contains the front d attributes.

5.1 The vulnerability of *IR* anonymity

In this subsection, we discuss the vulnerability of *IR* (k, l)-anonymity and *IR* (α, β)-anonymity. The experimental results for *IR* (k, l)-anonymity and *IR* (α, β)-anonymity are as shown in Figs. 2 and 3, respectively. We can see that *GeneIR_kl* and *DAnonyIR_kl* (*GeneIR_αβ* and *DAnonyIR_αβ*) both generate many vulnerable equivalence classes, and the percentage of vulnerable equivalence classes for *DAnonyIR_kl* (*DAnonyIR_αβ*) is more than that for *GeneIR_kl* (*GeneIR_αβ*). *GeneIR* continually repeats the process: randomly selects an attribute in QI to generalize according to the predefined taxonomy tree, then checks D to gain equivalence classes, which satisfy *IR* (k, l)-anonymity or *IR* (α, β)-anonymity, and moves them to D^*. While *DAnonyIR* uses the set generalization and considers the distance among individuals in an equivalence class. If an equivalence class satisfies *IR* (k, l)-anonymity or *IR* (α, β)-anonymity, we remove the equivalence class to D^*. An equivalence class obtained

[3] https://sites.google.com/site/informsdataminingcontest.

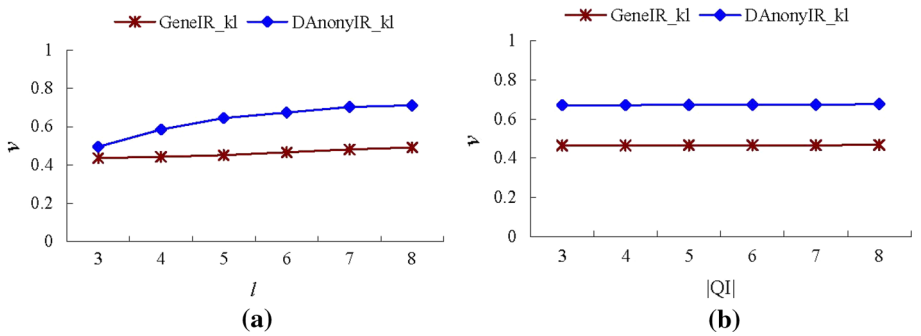


Fig. 2 The percentage of vulnerable equivalence classes in *GeneIR_kl* and *DAnonyIR_kl*. **a** $|QI| = 6$. **b** $l = 6$

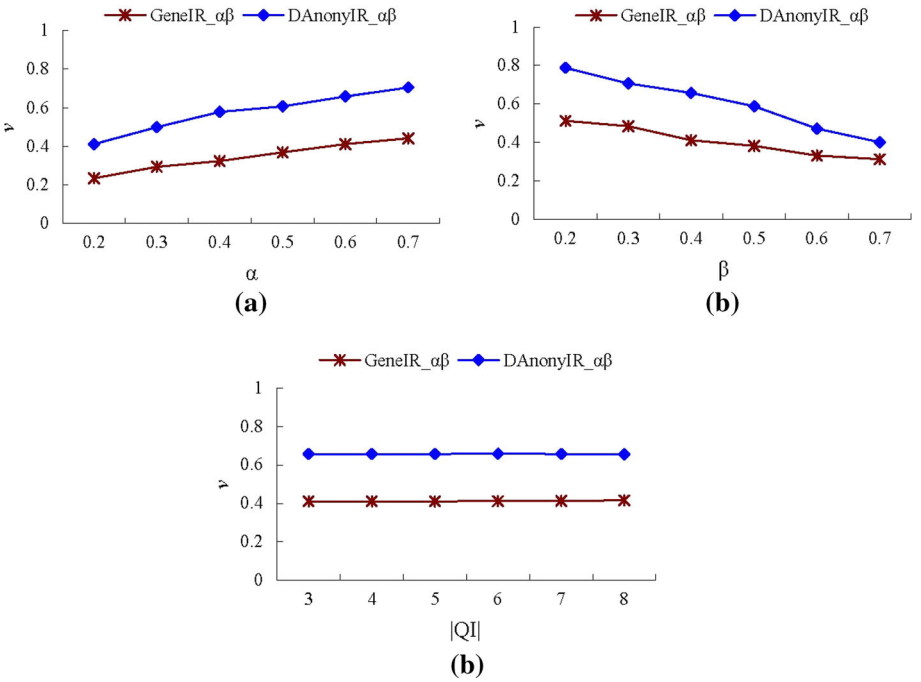


Fig. 3 The percentage of vulnerable equivalence classes in *GeneIR_αβ* and *DAnonyIR_αβ*. **a** $|QI| = 6$ and $\beta = 0.4$. **b** $|QI| = 6$ and $\alpha = 0.6$. **c** $\alpha = 0.6$ and $\beta = 0.4$

by *GeneIR* contains more individuals, so the possibility that satisfies *EIR* l-diversity or *EIR* (α, β)-anonymity is higher.

When l increases but QI is fixed (i.e. $|QI| = 6$), for an equivalence class satisfying *IR* (k, l)-anonymity, it is more difficult to satisfy *EIR* l-diversity, because *EIR* l-diversity requires that the number of different sensitive values of every reasoning set in the equivalence class is larger than or equal to l, so v increases, as shown in Fig. 2a. When α increases but parameters QI and β are fixed (i.e. $|QI| = 6$ and $\beta = 0.4$), for an equivalence class satisfying *IR* (α, β)-anonymity, the number of individuals decreases in the equivalence class. Therefore, it is more difficult to satisfy *EIR* (α, β)-anonymity, and v increases, as shown in Fig. 3a. When

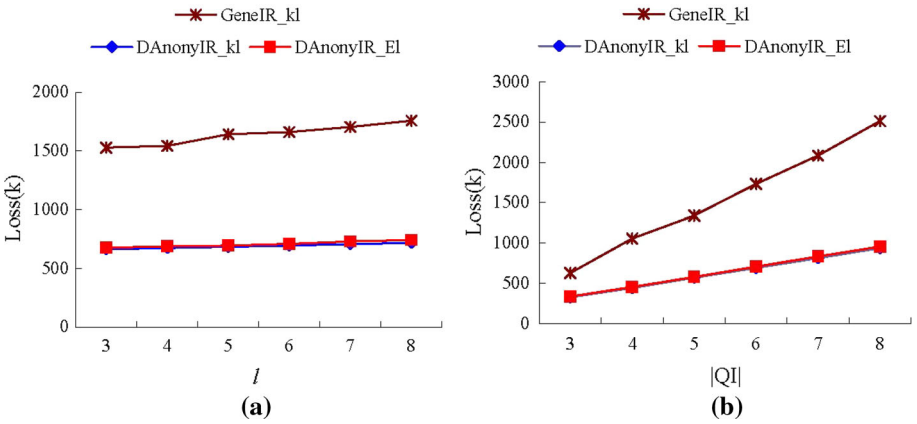


Fig. 4 The information loss in *GeneIR_kl*, *DAnonyIR_kl*, and *DAnonyIR_El*. **a** $|QI| = 6$. **b** $l = 6$

β increases but parameters QI and α are fixed (i.e. $|QI| = 6$ and $\alpha = 0.6$), the constraint is looser, for an equivalence class satisfying *IR* (α, β)-anonymity, it is easier to satisfy *EIR* (α, β)-anonymity. Consequently, v decreases, as shown in Fig. 3b.

From Figs. 2b and 3c, we observe that the change of QI size almost does not affect v for *GeneIR_kl*, *DAnonyIR_kl*, *GeneIR_*$\alpha\beta$ and *DAnonyIR_*$\alpha\beta$. When $|QI|$ increases, and parameters l, α and β are fixed (i.e. $l = 6$, $\alpha = 0.6$, and $\beta = 0.4$), the number of equivalence classes is almost not changed. That is, the size of an equivalence class is almost not affected, so does the ratio of vulnerable equivalence classes.

5.2 The analysis of data quality

In this subsection, we analyse the data quality from information loss and accuracy of query answering.

5.2.1 Information loss

Figures 4 and 5 show the information loss exhibited by *DAnonyIR* and *GeneIR* algorithms based on the setting of different values of parameters l, α, β, and $|QI|$. We can see that *GeneIR* is worse than our approach of *DAnonyIR*, because *GeneIR* makes the size of an equivalence class become very great, and causes much unnecessary information loss. In order to satisfy different privacy models with identity reservation, in *DAnonyIR* only the *SatPriIR* functions are different. If an equivalence class Q satisfies given privacy model, we do not add any individual to Q. Although an equivalence class needs more individuals for satisfying *EIR* l-diversity (*EIR* (α, β)-anonymity) than *IR* (k, l)-anonymity (*IR* (α, β)-anonymity), the increase is very small. So our *DAnonyIR* algorithms have a much closer difference than *GeneIR*.

When l or $|QI|$ increases, the information loss is increased in *GeneIR_kl*, *DAnony- IR_kl*, and *DAnonyIR_El*, as shown in Fig. 4a, b, respectively. When QI is fixed (i.e. $|QI| = 6$), as l increases, the number of individuals is increased and the number of records is also increased in each equivalence class for *GeneIR_kl*, *DAnonyIR_kl*, and *DAnonyIR_El*, and the possibility of providing more general values for the attributes per record increases.

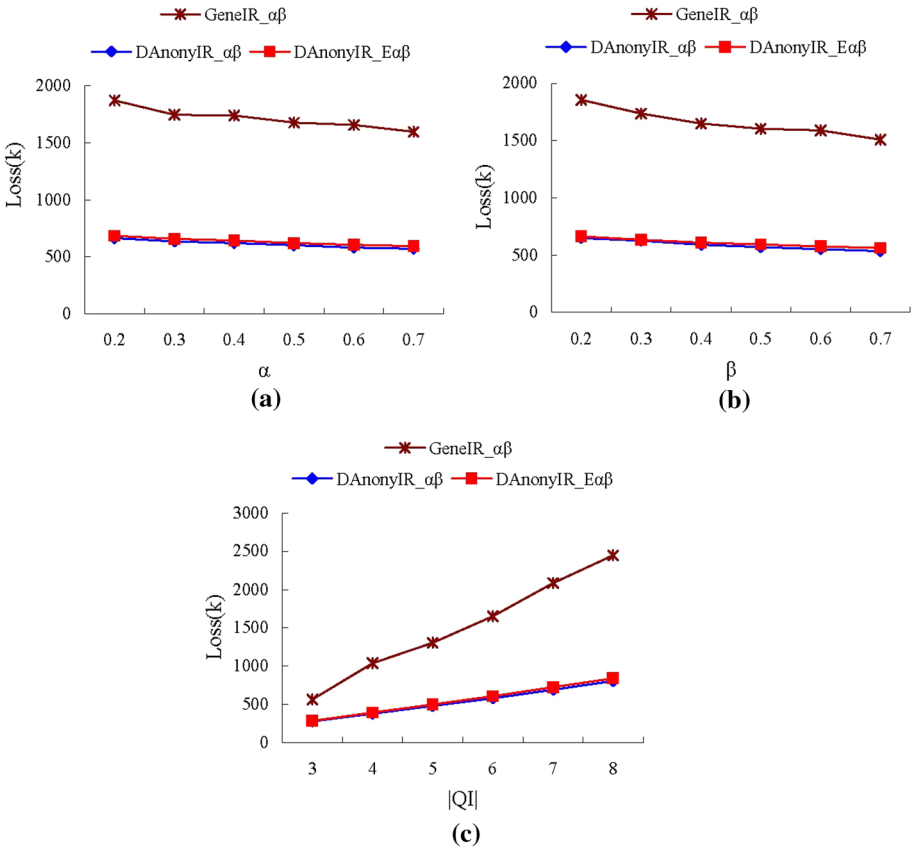


Fig. 5 The information loss in *GeneIR_αβ*, *DAnonyIR_αβ*, and *DAnonyIR_Eαβ*. **a** $|QI| = 6$ and $\beta = 0.4$. **b** $|QI| = 6$ and $\alpha = 0.6$. **c** $\alpha = 0.6$ and $\beta = 0.4$

Therefore, the information loss is increased. When $|QI|$ increases and l is fixed (i.e. $l = 6$), the number of attributes which we need to generalize is increased. That is, there are more generalized attributes for creating equivalence classes. So the information loss is increased. From Fig. 4, we can see that the increase when $|QI|$ is increased is more sharp for the algorithms with respect to the increase of l, because the increase of l only makes records do further generalization, while the increase of $|QI|$ makes records increase the generalized attributes.

When α or β increases, the information loss is decreased in *GeneIR_αβ*, *DAnony- IR_αβ* and *DAnonyIR_Eαβ*, as shown in Fig. 5a, b, respectively. When QI and β (α) are fixed (i.e. $|QI| = 6$ and $\beta = 0.4$ ($\alpha = 0.6$)) and α (β) increases, then the number of records is decreased in each equivalence class, so the information loss is decreased. From Fig. 5(c), when $|QI|$ increases and α and β are fixed ($i.e. \alpha = 0.6$ and $\beta = 0.4$), we can see that the information loss is increased, because the number of attributes that need to generalize is increased.

From Fig. 4 (Fig. 5), we can see that the information loss for *GeneIR_kl* (*GeneIR _αβ*) is much more than the information loss for *DAnonyIR_kl* (*DAnonyIR_αβ*) and *DAnonyIR_El* (*DAnonyIR_Eαβ*), and it is 1.923∼2.683 (2.018∼3.039) times, and 1.863∼2.626 (1.976∼

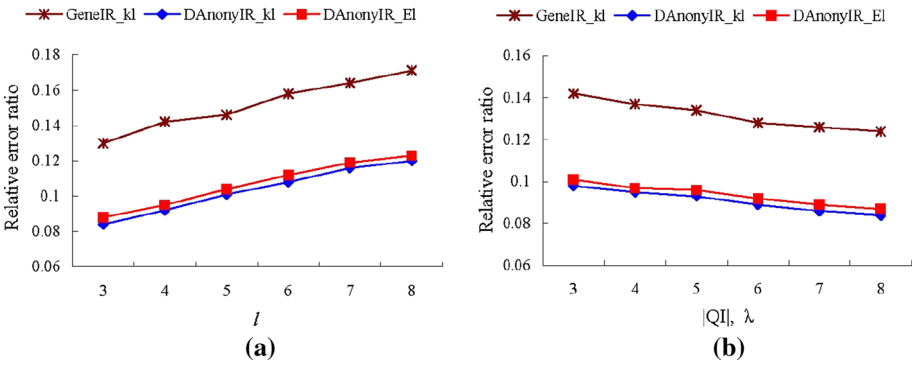


Fig. 6 The accuracy of query answering in *GeneIR_kl*, *DAnonyIR_kl*, and *DAnonyIR_El*. **a** $|QI| = 6$ and $\lambda = 1$. **b** $l = 6$

2.915) times, respectively. Compared with *DAnonyIR_kl* (*DAnonyIR_*$\alpha\beta$), the information loss for *DAnonyIR_El* (*DAnonyIR_*$\alpha\beta$) is higher, but the increment is very small and it just is 1.0148∼1.0347 (1.0127∼1.0519) times as shown in Fig. 4 (Fig. 5).

5.2.2 Aggregate query answering

The accuracy of query answering for *GeneIR_kl*, *DAnonyIR_kl* and *DAnonyIR_El* is as shown in Fig. 6. When QI and λ are fixed (i.e. $|QI| = 6$ and $\lambda = 1$), as l increases, the size of an equivalence class will increase. Thus, a more general value is needed for every attribute. So the relative error ratio is increased in these algorithms, as shown in Fig. 6a.

The accuracy of query answering for *GeneIR_*$\alpha\beta$, *DAnonyIR_*β and *DAnonyIR _E*$\alpha\beta$ is shown in Fig. 7. When QI and λ are fixed (i.e. $|QI| = 6$ and $\lambda = 1$), as α (i.e. $\beta = 0.4$) or β (i.e. $\alpha = 0.6$) increases, the size of an equivalence class will decrease. Thus a less general value is needed for every attribute. So the relative error ratio is decreased with respect to these algorithms, as shown in Fig. 7a, b.

In order to show the influence of query dimension to relative error ratio, we set $\lambda = |QI|$. From Figs. 6b and 7c, we can see that the relative error ratio is decreased, as the query dimension increases. Therefore, the anonymized data are performed better for queries with a larger query dimension. When l is fixed (i.e. $l = 6$) or α and β are fixed (i.e. $\alpha = 0.6$ and $\beta = 0.4$), as query dimension λ increases, the precise result $prec$ is decreased, and the estimated result est obtained from the anonymized table is closer to $prec$. Therefore, the relative error ratio is decreased.

From Fig. 6 (Fig. 7), we can see that the relative error ratio of *DAnonyIR_kl* and *DAnonyIR_El* (*DAnonyIR_*$\alpha\beta$ and *DAnonyIR_E*$\alpha\beta$) are less than that of *Gene- IR_kl* (*GeneIR_*$\alpha\beta$), and the relative error ratio of *DAnonyIR_El* (*DAnonyIR_E*$\alpha\beta$) is close to that of *DAnonyIR_kl* (*DAnonyIR_*$\alpha\beta$). Because *DAnonyIR* and *GeneIR* do not generalize on sensitive attribute, $\sum_{Q_i \in D^*} num_{Q_i}^{A_s}$ is invariable according to Eq. (19). When the size of every equivalence class become larger, the est is further away from $prec$. Although an equivalence class needs more individuals for satisfying *EIR* l-diversity (*EIR* (α, β)-anonymity) than *IR* (k, l)-anonymity (*IR* (α, β)-anonymity), the increase is very small.

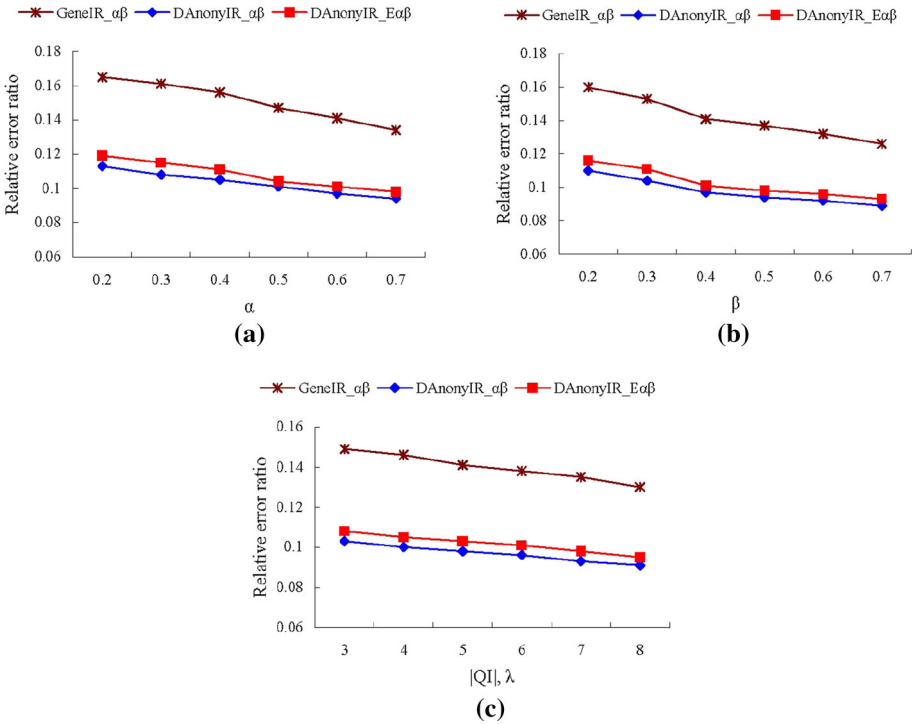


Fig. 7 The accuracy of query answering in *GeneIR_αβ*, *DAnonyIR_αβ*, and *DAnonyIR_Eαβ*. **a** $|QI| = 6$, $\beta = 0.4$, and $\lambda = 1$. **b** $|QI| = 6$, $\alpha = 0.6$, and $\lambda = 1$. **c** $\alpha = 0.6$ and $\beta = 0.4$

5.3 The analysis of efficiency

The runtimes exhibited by these algorithms based on the setting of values of different parameters l, α, β, and $|QI|$ are shown in Figs. 8 and 9. *GeneIR* can find quickly equivalence classes and check whether they satisfy *IR* (k, l)-anonymity or *IR* (α, β)-anonymity for each generalization operation. Apparently, its runtime is less than that of *DAnonyIR*. For *GeneIR* algorithm, it is possible that multiple equivalence classes are generated by scanning the dataset once or several times, while for *DAnonyIR* algorithm, an equivalence class is generated by scanning the dataset q times, where q is the number of individuals in the equivalence class. Therefore, the runtime of *GeneIR* is far less than that of *DAnonyIR*.

As shown in Fig. 8, the time of *DAnonyIR_kl* and *DAnonyIR_El* is close, because they both use the *DAnonyIR* algorithm, in which function *SatPriIR* is different, and they are *SatPriIRInc_kl* and *SatPriIRInc_El*. They both incremental methods and consider only an individual how to influence privacy, so the time of *DAnonyIR_kl* and *DAnonyIR_El* is close. It is similar to show that the time of *DAnonyIR_αβ* and *DAnonyIR_Eαβ* is close.

For *DAnonyIR_kl* and *DAnonyIR_El*, when l or $|QI|$ increases, the runtime is increased, as shown in Fig. 8, which is consistent with the time complexity analysis of *DAnonyIR*. For each equivalence class Q, the first individual is randomly selected, and we do not need to compute, so little time is spent. For every other individual in the equivalence class, we need scan D and D^* to get the individual or equivalence class with minimum distance to Q, respectively. In Fig. 8a, when l increases with fixed QI (i.e. $|QI| = 6$), the number of equivalence classes is decreased. That is, the number of individuals is increased in each equivalence class, so

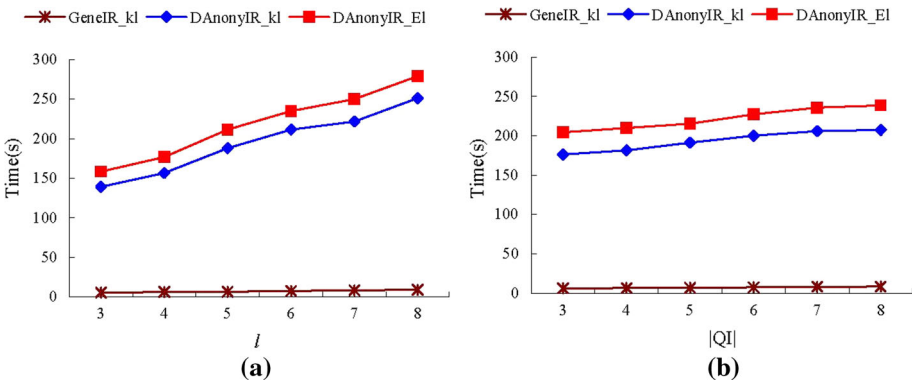


Fig. 8 The runtime of *GeneIR_kl*, *DAnonyIR_kl*, and *DAnonyIR_El*. **a** $|QI| = 6$. **b** $l = 6$

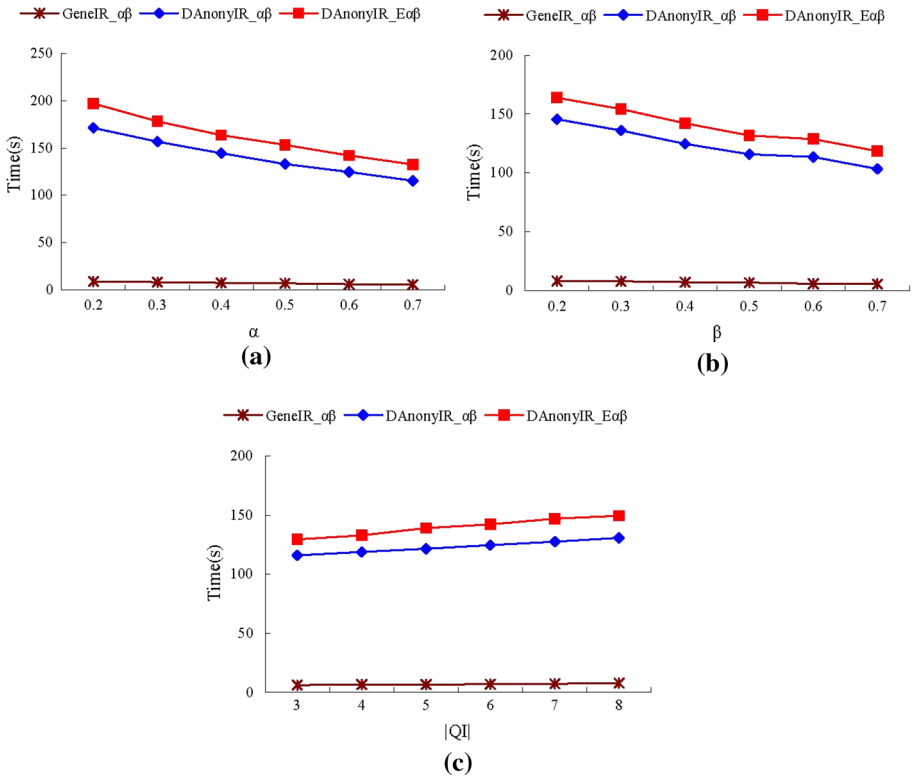


Fig. 9 The runtime of *Gene_αβ*, *DAnonyIR_αβ*, and *DAnonyIR_Eαβ*. **a** $|QI| = 6$ and $\beta = 0.4$. **b** $|QI| = 6, \alpha = 0.6$. **c** $\alpha = 0.6$ and $\beta = 0.4$

the algorithm needs more calculations, and thus the runtime increases. When $|QI|$ increases and l is fixed (i.e. $l = 6$), more attributes are considered for distance calculation, and thus the algorithm needs more time, as shown in Fig. 8b. When l or $|QI|$ increases, the runtime of *GeneIR_kl* is increased, because *GeneIR_kl* need more generalization operations.

For *DAnonyIR_αβ* and *DAnonyIR_Eαβ*, when α or β increases, the runtime is decreased, as shown in Fig. 9a, b, which is consistent with the time complexity analysis of *DAnonyIR*. When QI and β (α) are fixed (i.e. $|QI| = 6$ and $\beta = 0.4$ ($\alpha = 0.6$)), as α (β) increases, the number of records is decreased in each equivalence class. That is, the number of equivalence class is increased. So the algorithm needs less calculations, and thus the runtime is decreased. From Fig. 9c, we can see that the runtime increases with $|QI|$, which is consistent with the time complexity analysis of *DAnonyIR*. When $|QI|$ increases but α and β are fixed (i.e. $\alpha = 0.6$ and $\beta = 0.4$), more attributes are considered in calculating distance, thus the algorithm needs more time. When α (β) increases, the runtime of *GeneIR_αβ* is decreased, because the constraint condition is loosed and thus *GeneIR_αβ* needs less generalization operations. As $|QI|$ increases, *GeneIR_αβ* needs to generalize more attributes in order to obtain equivalence classes, so its runtime is increased.

Compared with *DAnonyIR_kl* (*DAnonyIR_αβ*), *DAnonyIR_El* (*DAnonyIR_Eαβ*) needs more time to judge whether an equivalence class satisfies the *EIR* l-anonymity (*EIR* (α, β)-diversity). For Fig. 8 (Fig. 9), the maximum increment is only 28.5s (25.9s).

5.4 Comprehensive analysis

From these experimental results, we can see that the percentage of vulnerable equivalence classes is fairly high. That is, if we use *IR* (k, l)-anonymity and *IR* (α, β)-diversity, there are many equivalence classes that could cause privacy leakage. Although *GeneIR_kl* and *GeneIR_αβ* can achieve quickly anonymization, in terms of the data quality and ability of privacy preservation, they are worse than our *DAnonyIR*. Comparing with *DAnonyIR_kl* and *DAnonyIR_αβ*, our *DAnonyIR_El* and *DAnonyIR _Eαβ* have higher information loss and relative error ratio for query answering, and spend more time, but the increments in these aspects are small and acceptable, because *DAnonyIR_El* and *DAnonyIR_Eαβ* supply stronger privacy preservation and the anonymized process is offline. Therefore, our enhanced privacy models and *DAnonyIR* algorithm are suitable for anonymizing just once over static datasets in an offline manner. However, when anonymization needs to take place quite frequently for data streams and execution time plays a major role, these approaches can be considered as inappropriate. The two previous privacy models do not reach a right privacy level while they suffer from great information loss by using *GeneIR*. On the other hand, the time exhibited by our *DAnonyIR* approach is not acceptable. Therefore, an appropriate approach for anonymizing data streams will be proposed in our further work.

6 Related work

In this section, we first discuss the privacy models and their anonymous approaches for static relational datasets with one-time anonymization. Then we discuss the development on privacy preservation for publishing dynamic relational datasets and data streams. Also, we discuss some privacy-preserving approaches for other data types except relational data. Finally, we show the main characteristics of our proposed approaches. An overall comparison of various anonymous approaches is shown in Table 6, where *IDis*=identity disclosure, *ADis*=attribute disclosure, *AOpe*=anonymous operation, *MSen*=multiple sensitive attributes, *MRec*=multiple records, *DUpd*=data update, *DType*=data type, *Gen(T)*=generalization with

Table 6 An summary of differences among various anonymous approaches

Privacy model	IDis	ADis	AOpe	MSen	MRec	DUpd	DType
k-Anonymity	√		Gen(T)	No	No	Sta	Rel
l-Diversity	√	√	Gen(T/S)	No	No	Sta	Rel
$(QI \rightarrow s, h)$		√	Gen(T)	No	No	Sta	Rel
(α, k)-Anonymity	√	√	Gen(T)	No	No	Sta	Rel
t-Closeness		√	Gen(T)	No	No	Sta	Rel
m-Privacy	√	√	Gen(T)	No	No	Sta	Rel
MSA l-diversity	√	√	Gen(T)	Yes	No	Sta	Rel
(ϵ^+, δ)-Dissimilarity	√	√	Gen(T)	Yes	No	Sta	Rel
MSA-diversity	√	√	Gen(T)	Yes	No	Sta	Rel
(X, Y)-Anonymity	√	√	Gen(T)	No	Yes	Sta	Rel
IR k-anonymity	√		Gen(T)	No	Yes(IR)	Sta	Rel
IR (k, l)-anonymity	√		Gen(T)	No	Yes(IR)	Sta	Rel
IR (α, β)-anonymity	√		Gen(T)	No	Yes(IR)	Sta	Rel
EIR (k, l)-anonymity	√	√	Gen(S)	No	Yes(IR)	Sta	Rel
EIR (α, β)-anonymity	√	√	Gen(S)	No	Yes(IR)	Sta	Rel
Dynamic l-diversity	√	√	Gen(T)	No	No	Dyn	Rel
m-Invariance	√	√	Gen(T)	No	No	Dyn	Rel
m-Distinct	√	√	Gen(T)	No	No	Dyn	Rel
k_s-Anonymity	√		Gen(T)	No	Yes	Str	Rel
k^m-Anonymity	√		Gen(T)	No	No	Sta	SVa
(h, k, p)-Coherence	√	√	Gen(T)	No	No	Sta	SVa
ρ-Uncertainty		√	Gen(T)	No	No	Sta	SVa
$(K, C)_L$-Anonymity	√	√	Supp	No	No	Sta	Tra
k-Degree anonymity	√		Modify	No	No	Sta	Net

predefined taxonomy tree (generalization is generally with suppression), *Gen*(*S*)=set generalization, *Supp*=suppression, *IR*=identity reservation, *Sta*=static data, *Dyn*= dynamic data, *Str*=data stream, *Rel*=relational data, *SVa*=set-valued data, *Tra*= trajectory data, *Net*=social network.

6.1 Anonymization for static datasets

Privacy preservation approaches for publishing static datasets contain these scenarios: single record and single sensitive attribute, single record and multiple sensitive attributes, and multiple records and single sensitive attribute, where single record and multiple records mean that an individual has only a record and multiple records in a data table, respectively.

6.1.1 Single record and single sensitive attribute

In recent years, the problem of privacy-preserving data publishing has been studied extensively [1–9]. Traditional privacy-preserving approaches deal with static datasets with single record and single sensitive attribute. K-anonymity is the first privacy model, proposed by

Samarati and Sweeney [3,14] in 1998. There exist many anonymization methods to implement k-anonymity, such as bottom-up generalization [15,16], top-down specialization [17] and anonymity by clustering technique [18–20]. The bottom-up generalization starts from the original data which violates k-anonymity and greedily selects a generalization operation according to a search metric, until all equivalence classes satisfy k-anonymity. In contrast to the bottom-up approach, the top-down specialization starts from the most general state in which all values are generalized to the most general values of their taxonomy trees. The specialization process terminates if no specialization can be performed without violating k-anonymity. In order to decrease the information loss caused by generalization, the anonymization approaches by clustering technique were proposed to achieve k-anonymity, which transform the problem to a clustering problem (i.e. to find a set of clusters (equivalence class), each of which contains at least k records). These records in a cluster are as similar as possible, which can ensure that less distortion is required when the records in a cluster are modified to have the same QI value.

Due to its simplicity, k-anonymity remains one of the most widely used models in the literature. It can protect against *identity disclosure*, but cannot prevent *attribute disclosure*. As a result, l-diversity has been proposed in [10]. It requires that every equivalence class contains at least l "well-represented" sensitive values, which can be defined in diverse ways, i.e. distinct l-diversity, and entropy l-diversity (the entropy of sensitive values in each equivalence class should be at least $\log l$). The *IR* (k, l)-anonymity refers to distinct l-diversity, so we also extent it to *EIR* l-diversity. There are numerous methods for achieving l-diversity [21,24]. Ghinita et al [21] proposed a fast data anonymization with low information loss for achieving l-diversity, which first maps the multi-dimensional QI attributes to 1-dimensional space, then partitions the space with considering to cover a variety of sensitive values and finally generalizes the QI attributes in each group. Wang et al. [24] argued that traditional data generalization based on the predefined taxonomy trees often causes some unnecessary information loss, so they proposed more flexible strategies for data generalization by set generalization and presented a clustering algorithm to implement l-diversity. Wang et al. [22] gave a privacy template in the form of $(QI \rightarrow s, h)$ (meaning that the confidence of inferring the sensitive value s from any group on QI is no more than h) and proposed an algorithm to minimally suppress a table to satisfy a set of privacy templates.

Furthermore, Wong et al. [11] extended k-anonymity to (α, k)-anonymity to limit the confidence of the implications from the QI to a sensitive value to within α in order to protect the sensitive information from being inferred by strong implications and proposed a bottom-up generalization algorithm to achieve (α, k)-anonymity. In order to prevent *skewness attack* and *similarity attack*, which belong to *attribute disclosure*, Li et al. [12] proposed t-closeness model. *Skewness attack* and *similarity attack* will happen when the percentages of sensitive values are skewness (some values appear with high frequency, while others appear with low frequency), and these sensitive values are similar semantically in an equivalence class, respectively. And t-closeness requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table. They also revised the Incognito algorithm [16], which is a top-down generalization method proposed for k-anonymity, to achieve t-closeness. Cao et al. [23] pointed out there is no anonymization algorithm tailored for t-closeness, and they proposed the SABRE approach for distribution-aware microdata anonymization based on the t-closeness principle. The approach first greedily partitions a table into buckets of similar sensitive values and then redistributes the tuples of each bucket into dynamically determined equivalence classes. Furthermore, Goryczka et al. [30] consider the collaborative data publishing problem for anonymizing horizontally partitioned data from multiple data providers. They introduced the

concept of m-privacy, which guarantees that the anonymized data satisfy a given privacy constraint (e.g. k-anonymity and l-diversity) against any group of up to m colluding data providers. Also, they presented a data provider-aware anonymization algorithm with adaptive m-privacy checking strategies to ensure high utility and m-privacy of anonymized data with efficiency.

6.1.2 Single record and multiple sensitive attributes

The above approaches consider a data table with only a sensitive attribute. However, they cannot be applied directly to the case of multiple sensitive attributes. Yang and Wang [31] proved that if the minimum class coverage φ_{min} in an equivalence class satisfies $\varphi_{min} \geq l$, then the equivalence class satisfies multiple sensitive attributes l-diversity (*MSA* l-diversity), which requires that the values of every sensitive attribute satisfy l-diversity in any equivalence class. Also, a anonymization approach with generalization is given based on minimum selected degree first. To preserve privacy against *proximity attack* (*similarity attack*), Zhang et al. [32] defined $(\epsilon^+, \delta)^k$-dissimilarity privacy model for scalable big data with multiple sensitive attributes, which requires that the size of any equivalence class Q is at least k, and any sensitive vector in Q must be dissimilar to at least $\delta \times (|Q| - 1)$ $(0 \leq \delta \leq 1)$ other sensitive vectors. Parameter k controls Q to prevent *identity disclosure* and δ specifies constraints on the number of ϵ^+ neighbours that each sensitive vector can own to combat *proximity attack*. Also, they proposed a clustering anonymization approach. Abdalaal et al. [33] assumed that adversaries can launch attacks by joining the quasi-identifiers with some non-membership knowledge to link individuals with the sensitive values. They proposed *MSA*-diversity, which ensures that the probability of mapping an individual to a sensitive value is bounded by $\frac{1}{l-i}$ under i bits of non-membership knowledge, but its strict grouping condition will result in excessive information loss.

6.1.3 Multiple records and single sensitive attribute

However, all the above approaches assume that each record in a data table represents a distinct owner. In fact, the case that an individual could have multiple records appears frequently in real life, if there exists $1 : N$ relationship between an individual and the sensitive attribute. For example, a student has grades in different courses, a patient suffers from different diseases, and a person has multiple hobbies. The relation among different sensitive values, which belong to the same individual, is very important for researchers and decision-makers. So we need to keep it by identifying the *ID* with numbers instead of removing the explicit identifying information. In this case, these anonymity models, in which an individual has only a record in a data table, may be underprotected, and are inadequate, and could cause privacy leakage.

At present, there exist some approaches to handle the situation that an individual could have multiple records. (X, Y)-anonymity introduced by Wang and Fung [34] specifies that each value in X is linked to at least k distinct values in Y. It provides a flexible way to specify different types of privacy requirements. If we specify k-anonymity with respect to patients by letting X be QI attributes and Y be explicit identifier of individual, in this case, several records may represent the same record owner (individual). In order to maintain such a correlation, Tong et al. [13] proposed three privacy models with identity reservation (i.e. *IR* k-anonymity, *IR* (k, l)-anonymity, and *IR* (α, β)-anonymity) and presented an anonymization method, *GeneIR*, with bottom-up generalization by predefined taxonomy trees for implementing these privacy models. They first recode the *ID* of database D with numbers and group the records

with the same QI values. If an equivalence class satisfies the given privacy model π, the group is removed from D to D^*. Then repeatedly execute the step: select an attribute in QI to generalize up a level in its taxonomy tree and check D to obtain the equivalence classes which satisfy π, until D does not satisfy π or no further generalization could to be made. If there are residual individuals in D, every residual individual is added to the closest equivalence class.

6.2 Anonymization for dynamic datasets and data streams

In this subsection, we discuss the anonymization approaches for dynamic datasets and data streams, which almost all consider the scenario with single record and single sensitive attribute.

6.2.1 Anonymization for dynamic datasets

When data are dynamically updated with record insertions and/or deletions, the re-publication is needed. Anonymizing datasets statically (i.e. each release is individually anonymous) may cause privacy leakage by comparing different releases and eliminating some possible sensitive values for a victim [2]. Byun et al. [35] were the pioneers who proposed an anonymization technique with generalization that enables privacy-preserving continuous data publishing after new records have been inserted. It guarantees that every release satisfies distinct l-diversity, and makes sure that a new anonymized table to be released does not create any inference channel with respect to the previously released tables (called dynamic l-diversity). Nevertheless, this approach supports only insertions. Xiao and Tao [36] proposed m-invariance privacy model and an anonymization method with generalization to address both record insertions and deletions. A sequence of releases $D_1^*,\ldots,D_p^*$ satisfies m-invariance if every equivalence class Q in D_i^* $(1 \leq i \leq p)$ is m-unique (i.e. Q contains at least m records and all records in Q have different sensitive values) and all equivalence classes in $D_1^*,\ldots,D_p^*$ containing record r must have the same set of sensitive vales. Li and Zhou [37] extended m-invariance to m-distinct to address external updates (the dataset is updated with record insertions and/or deletions) and internal updates (the attribute values of each record are dynamically updated).

6.2.2 Anonymization for data streams

Data streams are continuous, transient, and usually unbounded. Cao et al. [27] discussed that anonymizing data streams and anonymizing dynamic datasets are different because the inferences that may arise when anonymizing dynamic datasets and those that might happen during anonymizing of data streams are different. Anonymizing a dynamic dataset requires multiple releases of a table. The inference is happened because multiple anonymized tables contain some same records, while this inference cannot be carried out in anonymizing data stream because a record in the stream is anonymized only once. The possible inferences in anonymizing data stream are due to the fact that the attacker is able to inspect the sequence of anonymized tuples given in output. Because of the characteristics of data streams, the algorithm for data streaming can only scan the data in one pass and executes in a pipeline manner, and there is a need to offer strong guarantees on the maximum allowed delay between incoming data and the corresponding anonymized output. So the efficiency plays an important role in anonymizing data streams. Cao et al. [27] presented k_s-anonymity for privacy-preserving

data streams publishing in the case of an individual with multiple records and gave a cluster algorithm to anonymize data streams and ensure the freshness of the anonymized data by satisfying specified delay constraints. However, they put the different records of the same individual in different equivalence classes. Hence, they lose the relation among the values of sensitive attribute, which belong to the same individual. Furthermore, Guo and Zhang [38] improved algorithm of Cao et al. for data streams based on clustering by considering the time constraints on tuple publication and cluster reuse, to accelerate the anonymization process and reduce the information loss.

6.3 Anonymization for other data types except relational data

There are some studies on anonymizing set-valued data, trajectory data, and social network. Terrovitis et al. [39] presented k^m-anonymity for the set-valued or transaction data, which guarantees that an adversary with maximum knowledge of m items cannot distinguish each transaction from k transactions. They proposed two heuristic anonymization algorithms, which greedily identify itemsets that violate the anonymity requirement and choose generalization rules that fix the corresponding problems. (h, k, p)-coherence introduced by Xu et al. [40] confines an attacker with maximum knowledge of p items to identify each transaction from k transactions in which no more than $h\%$ share a common private item. They gave an algorithm for achieving (h, k, p)-coherence by suppression while preserving as much information as possible. Cao et al. [41] proposed ρ-uncertainty privacy model, which does not allow an attacker knowing any subset of a transaction t to infer a sensitive item $\alpha \in t$ with confidence higher than ρ, and presented an algorithm by combining generalization and suppression to transform a data table and make it satisfy ρ-uncertainty. Chen et al. [42] studied the privacy problem of trajectory data. They proposed $(K, C)_L$-privacy model, which requires any subsequence q of any adversary's L-knowledge to be shared by either 0 or at least K records in a trajectory database and the confidence of inferring any sensitive value from q to be at most C, and showed that the proposed suppression method can significantly improve the data utility in anonymous trajectory data. Liu and Terzi [43] proposed k-degree anonymity for anonymizing social network, which requires that all vertices have at least $k-1$ other vertices sharing the same degree, and gave an algorithm to ensure that all vertices satisfy k-degree anonymity by modifying the graph structure. Moreover, Casas-Roma et al. [8] devised an efficient algorithm for k-degree anonymity in large networks.

6.4 Characteristics of our approaches

In this paper, we deal with static relational data with multiple records and single sensitive attribute. It is significant to study identity reservation for multiple records. Two privacy models *EIR* l-diversity and *EIR* (α, β)-anonymity are proposed to solve the disadvantage of *IR* (k, l)-anonymity and *IR* (α, β)-anonymity for identity reservation (i.e. they fail to prevent *attribute disclosure*). At present, many anonymization approaches use the generalization by predefined taxonomy tree, which restricts the generalized range and causes some unnecessary information loss. Therefore, Wang et al. [24] presented the set generalization and gave a clustering algorithm *l-clustering* to implement l-diversity. Inspired by the method, we present the heuristic greedy clustering algorithm *DAnonyIR* for achieving *EIR* l-diversity and *EIR* (α, β)-anonymity. Also, we can use *DAnonyIR* to anonymize database in order to make it satisfy *IR* (k, l)-anonymity and *IR* (α, β)-anonymity by calling different decision functions.

Our approach is different from *l-clustering* in the following two aspects. (1) Our approach considers the case of an individual with multiple records, so the definitions of some distances are different. We introduce our own concepts of the distance between two individuals, the distance between individual and equivalence class, and the distance between two equivalence classes by using different information metrics for numeric and categorical attributes. (2) Our approach is used to achieve *EIR* l-diversity, *EIR* (α, β)-anonymity, *IR* (k, l)-anonymity, and *IR* (α, β)-anonymity, while *l-clustering* is used for l-diversity. We need to set different decision functions for different privacy models. Experimental results have shown our *DAnonyIR* is superior to the existing *GeneIR* for multiple records with generalization by predefined taxonomy tree in terms of the data quality. Our approaches are only used to anonymize static relational datasets with multiple records and single sensitive attribute. In the next work, it will be interesting to extend our approaches to anonymize datasets with multiple sensitive attributes, dynamic datasets, data streams, and other data types.

7 Conclusions

In this paper, we have argued that *IR* (k, l)-anonymity and *IR* (α, β)-anonymity are insufficient to prevent privacy leakage. Thus, we proposed enhanced versions of these two privacy models, called *EIR* l-diversity and *EIR* (α, β)-anonymity. Moreover, we have designed a general anonymization algorithm, called *DAnonyIR*, with clustering technique to transform the dataset to satisfy different identity-reserved privacy models by calling different decision functions. Compared with the existing approaches, i.e. *GeneIR_kl* and *GeneIR_*$\alpha\beta$ [15], respectively, our *DAnonyIR_El* and *DAnonyIR_E*$\alpha\beta$ provide stronger privacy preservation, and the information loss and relative error ratio of query answering are less than those of the *GeneIR_kl* and *GeneIR_*$\alpha\beta$, although our approaches need more runtime. To avoid the influence caused by different algorithms, we also compared our enhanced approaches with *DAnonyIR_kl* and *DAnonyIR_*$\alpha\beta$, respectively, and found that our approaches are very close to *DAnonyIR_kl* and *DAnonyIR_*$\alpha\beta$ in the aspects of information loss, relative error ratio of query answering, and runtime.

Our *EIR* l-diversity and *EIR* (α, β)-anonymity are suitable for the anonymization of relational data in which an individual could have multiple records, our *DAnonyIR* algorithm is performed just once over static datasets in an offline manner, and the clustering result is not optimal. So in future, it is worthy extending our approaches to find an optimal clustering result by analysing its average time complexity, and solve these problems considering privacy leakages caused by relation among attributes, attackers' stronger background knowledge, multiple sensitive attributes, and data publishing of dynamic datasets and data streams. Also, we will consider privacy preservation of distributed data [30] and other sorts of data, contained set-valued data [38], trajectory data [42], and social network [43].

Acknowledgements Firstly, we should thank the anonymous reviewers for their very valuable suggestions, comments and advices, which help us improve the paper significantly. Secondly, we need acknowledge that this paper was supported by the National Natural Science Foundation of China (Nos. 61502111, 61763003, 61672176, 61562007, 61662008), Guangxi Natural Science Foundation (Nos. 2016GXNSFAA380192, 2015GXNSFBA139246), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, Guangxi Special Project of Science and Technology Base and Talents (AD16380008), and Guangxi Collaborative Innovation Center of Multisource Information Integration and Intelligent Processing.

References

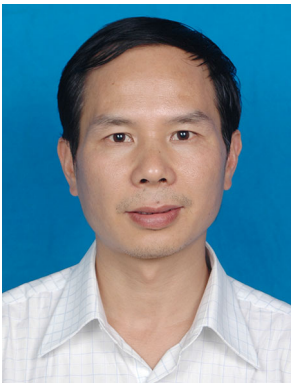
1. Li N, Li T, Venkatasubramanian S (2010) Closeness: a new privacy measure for data publishing. IEEE Trans Knowl Data Eng 22(7):943–956
2. Fung BCM, Wang K, Chen R, Yu P S (2010) Privacy-preserving data publishing: a survey of recent development. ACM Comput Surv 42(4): article 14
3. Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information. In: Proceedings of the 17th ACM symposium on principles of database systems, p 188
4. Xiao X, Tao Y (2006) Personalized privacy preservation. In: Proceedings of the 25th ACM international conference on management of data, pp 229–240
5. Terrovitis M, Liagouris J, Mamoulis N, Skiadopoulos S (2012) Privacy preservation by disassociation. In: Proceedings of the 38th international conference on very large databases, pp 944–955
6. Zakerzadeh H, Aggarwal CC, Barker K (2016) Managing dimensionality in data privacy anonymization. Knowl Inf Syst 49(1):341–373
7. Xin Y, Xie Z, Yang J (2017) The privacy preserving method for dynamic trajectory releasing based on adaptive clustering. Inf Sci 378:131–143
8. Casas-Roma J, Herrera-Joancomartí J, Torra V (2017) k-Degree anonymity and edge selection: improving data utility in large networks. Knowl Inf Syst 50(2):447–474
9. Sun Y, Yuan Y, Wang G, Cheng Y (2016) Splitting anonymization: a novel privacy-preserving approach of social network. Knowl Inf Syst 47(3):595–623
10. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M (2006) l-Diversity: privacy beyond k-anonymity. In: Proceedings of the 22nd international conference on data engineering, Article 24
11. Wong RCW, Li J, Fu AWC, Wang K (2006) (a, k)-Anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 754–759
12. 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
13. Tong Y, Tao Y, Tang S, Yang D (2010) Identity-reserved anonymity in privacy preserving data publishing. J Softw 21(4):771–781 (In Chinese)
14. Samarati P (2001) Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027
15. Wang K, Yu PS, Chakraborty S (2004) Bottom-up generalization: a data mining solution to privacy protection. In: Proceedings of the 4th international conference on data mining, pp 249–256
16. Lefevre K, Dewitt DJ, Ramakrishnan R (2005) Incognito: efficient full-domain k-anonymity. In: Proceedings of the 24th ACM international conference on management of data, pp 49–60
17. Fung BCM, Wang K, Yu ps. (2005) Top-down specialization for information and privacy preservation. In: Proceedings of the 21st international conference on data engineering, pp 205–216
18. Aggarwal G, Panigrahy R, Feder T et al (2010) Achieving anonymity via clustering. ACM Trans Algorithms 6(3): article 49
19. Li J, Wong RCW, Fu AWC, Pei J (2006) Achieving k-Anonymity by clustering in attribute hierarchical structures. In: Proceedings of the 8th international conference on data warehousing and knowledge discovery, pp 405–416
20. Byun J, Kamra A, Bertino E, Li N (2007) Efficient k-anonymization using clustering techniques. In: Proceedings of the 12th international conference on database systems for advanced applications, pp 188–200
21. Ghinita G, Karras P, Kalnis P, Mamoulis N (2009) A framework for efficient data anonymization under privacy and accuracy constraints. ACM Trans Database Syst 34(2), Article 9
22. Wang K, Fung BCM, Yu PS (2007) Handicapping attacker's confidence: an alternative to k-anonymization. Knowl Inf Syst 11(3):345–368
23. Cao J, Karras P, Kalnis P, Tan K (2011) SABRE: a sensitive sttribute bucketization and redistribution framework for t-closeness. VLDB J 20(1):59–81
24. Wang Z, Xu J, Wang W, Shi B (2010) Clustering-basd approach for data anonymization. J Softw 21(4):680–693 (In Chinese)
25. Xu J, Wang W, Pei J et al (2006) Utility-based anonymization using local recoding. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–790

26. Iyengar VS (2002) Transforming data to satisfy privacy constraints. In: Proceedings of the 8th ACM international conference on knowledge discovery and data mining, pp 279–288
27. Cao J, Carminita B, Ferrari E, Tan KL (2011) CASTLE: continuously anonymizing data streams. IEEE Trans Dependable Secure Comput 8(3):337–352
28. Jiang Y, Lin L (2003) The computation of hitting sets with Boolean formulas. J Comput 26(8):919–924 (In Chinese)
29. Xiao X, Tao Y (2008) Dynamic anonymization: accurate statistical analysis with privacy preservation. In: Proceedings of the 27th ACM SIGMOD international conference on management of data, pp 107–120
30. Goryczka S, Xiong L, Fung BCM (2014) m-Privacy for collaborative data publishing. IEEE Trans Knowl Data Eng 26(10):2520–2533
31. Yang J, Wang B (2012) Personalized *l*-diversity algorithm for multiple sensitive attributes based on minimum selected degree first. J Comput Res Dev 49(12):2603–2610 (in Chinese)
32. Zhang X, Dou W, Pei J et al (2015) Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud. IEEE Trans Comput 64(8):2293–2307
33. Abdalaal A, Nergiz ME, Saygin Y (2013) Privacy-preserving publishing of opinion polls. Comput Secur 37:143–154
34. Wang K, Fung BCM (2006) Anonymizing sequential releases. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 414–423
35. Byun JW, Sohn Y, Bertino E, Li N (2006) Secure anonymization for incremental datasets. In: Proceedings of third VLDB workshop on secure data management, pp 48–63
36. Xiao X, Tao Y (2007) *m*-Invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of the 26th ACM SIGMOD international conference on management of data, pp 689–700
37. Li F, Zhou S (2008) Challenging more updates: towards anonymous re-publication of fully dynamic datasets. Arxiv Cornell University Library
38. Guo K, Zhang Q (2013) Fast clustering-based anonymization approaches with time constraints for data streams. Knowl Based Syst 46:95–108
39. Terrovitis M, Mamoulis N, Kalnis P (2008) Privacy-preserving anonymization of set-valued data. In: Proceedings of the 34th international conference on very large data bases (VLDB), pp 610–622
40. Xu Y, Wang K, Fu AWC, Yu PS (2008) Anonymizing transaction databases for publication. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, pp 767–775
41. Cao J, Karras P, Raissi C, Tan KL (2010) ρ-Uncertainty: inference-proof transaction anonymization. In: Proceedings of the 36th international conference on very large data bases, pp 1033–1044
42. Chen R, Fung BCM, Mohammed N, Desai BC, Wang K (2013) Privacy-preserving trajectory data publishing by local suppression. Inf Sci 231:83–97
43. Liu K, Terzi E (2008) Towards identity anonymization on graphs. In: Proceedings of the 27th ACM SIGMOD international conference on management of data, pp 93–106

Jinyan Wang received the B.Sc., M.Sc., and Ph.D. degrees from the School of Computer Science and Information Technology, Northeast Normal University, Changchun, China, in 2005, 2008, and 2011, respectively. She is currently an associate professor in the School of Computer Science and Information Technology, Guangxi Normal University, Guilin, China. Her research has been supported by the National Science Foundation of China. Her research interest includes information security and automated reasoning.

Kai Du received the M.Sc. degree from the School of Computer Science and Information Technology, Guangxi Normal University, Guilin, China, in 2017. He is currently a teacher at the OneSmart Education Group Ltd., Shanghai, China. His research interest includes information security.

Xudong Luo is currently a distinguished professor at Guangxi Normal University. Before moving to this position, he worked, as a distinguished professor, at Sun Yat-sen University. Before coming back to China in 2011, he had been a senior research fellow, a research fellow, and a lecturer, respectively, at City University of Hong Kong, Nanyang Technological University, Chinese University of Hong Kong, University of Birmingham, and University of Southampton, respectively. He received a Ph.D. in Artificial Intelligence from the University of New England, Australia, an M.Sc. degree in Computer Science from the Chinese Academy of Sciences, and a B.Sc. degree in Mathematics and Computer Science from Southwest University, China. He published more than 160 papers including 2 in top journal Artificial Intelligence. He has international recognized reputation: PC chairs, members or senior members of over 100 international conferences including some major ones (such as IJCAI), and referees for first class international journals (such as Artificial Intelligence) and some major conferences (such as IJCAI). He was invited to present his work in some conferences and many universities in different countries. His research interest includes fuzzy logic, automated negotiation, game theory, and decision-making.



Xianxian Li received the Ph.D. degree from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2002. He worked as a professor at Beihang University during 2003–2010. He is currently a professor with the School of Computer Science and Information Technology, Guangxi Normal University, Guilin, China. His research interest includes information security.