Anonymization of Sensitive Quasi-Identifiers for l-Diversity and t-Closeness

Yuichi Sei, *Member, IEEE*, Hiroshi Okumura, Takao Takenouchi, Akihiko Ohsuga, *Member, IEEE*

***Abstract*—A number of studies on privacy-preserving data mining have been proposed. Most of them assume that they can separate quasi-identifiers (QIDs) from sensitive attributes. For instance, they assume that address, job, and age are QIDs but are not sensitive attributes and that a disease name is a sensitive attribute but is not a QID. However, all of these attributes can have features that are both sensitive attributes and QIDs in practice. In this paper, we refer to these attributes as sensitive QIDs and we propose novel privacy models, namely, (l1, ..., lq)-diversity and (t1, ..., tq)-closeness, and a method that can treat sensitive QIDs. Our method is composed of two algorithms: an anonymization algorithm and a reconstruction algorithm. The anonymization algorithm, which is conducted by data holders, is simple but effective, whereas the reconstruction algorithm, which is conducted by data analyzers, can be conducted according to each data analyzer's objective. Our proposed method was experimentally evaluated using real data sets.**

***Index Terms*—privacy, data mining, l-diversity, t-closeness**

I. Introduction

In recent years, numerous organizations have begun to provide services that collect large amounts of personal information. This personal information can be shared with other organizations so that they can subsequently create new services. Moreover, shared data are also very important for researchers [1]. We call an organization that has an original database a "data holder." We assume that the data holder wants to anonymize and publish the database. We call organizations that receive and use the anonymized database "data analyzers."

Many studies regarding anonymized databases of personal information have been proposed. Most existing methods consider that the data holder has a database in the form of *explicit identifiers*, *quasi-identifiers (QIDs)*, or *sensitive attributes*, where explicit identifiers are attributes that explicitly identify individuals (e.g., name), QIDs are attributes that could be potentially combined with other directories to identify individuals (e.g., zip code and age), and sensitive attributes are personal attributes of a private nature (e.g., disease and salary) [2].

TABLE I: Patient Table

Name	Age	Address	Job	Disease
Alex	41	13021	Artist	Fever
Becky	41	17025	Writer	Obesity
Carl	51	13021	Lawyer	Fever
Diana	51	14053	Lawyer	Obesity
Ewen	51	14003	Lawyer	HIV
Flora	51	16005	Lawyer	HIV
Glen	51	14003	Lawyer	Fever
Helen	51	16005	Lawyer	Obesity

Even if we remove all of the explicit identifiers from a database, disclosure may still occur. k-Anonymity [3], l-diversity [4], and t-closeness [5] are some of the major privacy models for preventing the problem. Many studies on these privacy models have been proposed, such as [6], [7], [8], [9], [10]. First, we provide a brief overview of k-anonymity, l-diversity, and t-closeness, and then we describe the problem that this paper addresses.

k-Anonymity ensures that there are k records or more that have the same QID values so that k-anonymity can protect against "identity disclosure." For example, Table I shows the original patient database that a hospital wants to publish, and Table II(a) shows one result of k-anonymity when we assume that *Name* is an explicit identifier, *Disease* is a sensitive attribute, other attributes are QIDs, and k is set to 2. Even if the data analyzer knows Ewen's QID values, he or she cannot know whether the fifth record or the sixth record is Ewen's record.

However, in some cases, k-anonymity cannot protect against "attribute disclosure." For example, in Table II(a), the data analyzer can know that Ewen surely has HIV because the fifth and the sixth records have the same disease values.

l-Diversity ensures that there are at least l "well-represented" sensitive values and protects against attribute disclosure. There are several definitions of the term "well-represented." For example, *frequency l-diversity* ensures that data analyzers cannot specify each individual's sensitive values with a confidence greater than $1/l$. Table II(b) shows one result of frequency l-diversity. The data analyzer cannot know whether HIV or fever is Ewen's disease.

However, l-diversity does not consider the rareness of each sensitive value. For example, in Table I, the probability distribution of fever, obesity, and HIV in the whole table is $\{3/8, 3/8, 2/8\}$. On the other hand, the data analyzer can know from Table II(b) that the probability distribution of

Y. Sei and A. Ohsuga are with the University of Electro-Communications, Tokyo 182-8585, Japan (e-mail: seiuny@uec.ac.jp; ohsuga@uec.ac.jp).

H. Okumura is with Enterprise Management Business Unit, Mitsubishi Research Institute, Tokyo 100-8141, Japan (email: okumurah@mri.co.jp)

T. Takenouchi is with Security Laboratories, NEC Corporation, Kanagawa 211-8666, Japan (e-mail: takenouchi@bu.jp.nec.com)

This article is an extended version of a paper presented at the 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Helsinki, Finland, 2015.

This work was supported by JSPS KAKENHI Grant Numbers 26330081, 26870201, 16K12411, 17H04705, and also supported by the Telecommunications Advancement Foundation.

The authors would like to thank the anonymous reviewers for their valuable comments and contributions to improve the paper.

TABLE II: Anonymization Results of Existing Studies

(a) 2-Anonymity

Age	Address	Job	Disease
41	13*-17*	*	Fever
41	13*-17*	*	Obesity
51	13*-14*	Lawyer	Fever
51	13*-14*	Lawyer	Obesity
51	14*-16*	Lawyer	HIV
51	14*-16*	Lawyer	HIV
51	14*-16*	Lawyer	Fever
51	14*-16*	Lawyer	Obesity

(b) 2-Diversity

Age	Address	Job	Disease
41	13*-17*	*	Fever
41	13*-17*	*	Obesity
51	13*-14*	Lawyer	Fever
51	13*-14*	Lawyer	Obesity
51	14003	Lawyer	HIV
51	16005	Lawyer	HIV
51	14003	Lawyer	Fever
51	16005	Lawyer	Obesity

(c) 0.25-Closeness

Age	Address	Job	Disease
41	13*-17*	*	Fever
41	13*-17*	*	Obesity
51	13*-16*	Lawyer	Fever
51	13*-16*	Lawyer	Obesity
51	13*-16*	Lawyer	HIV
51	13*-16*	Lawyer	HIV
51	13*-16*	Lawyer	Fever
51	13*-16*	Lawyer	Obesity

the sensitive value of Ewen is $\{1/2, 0, 1/2\}$. These two probability distributions are widely different. t-Closeness can ensure that the distance between the probability distribution of sensitive values in the records that have the same QID values and the probability distribution of sensitive values in the whole table is lower than threshold t. Table II(c) shows the result of 0.25-closeness. The data analyzer knows from the table that the probability distribution of the sensitive value of Ewen is $\{1/3, 1/3, 1/3\}$, which is near the distribution in the whole table. Moreover, t-closeness can consider the distance between sensitive values in calculating the distance between the two probability distributions.

However, in this case, we have a problem that should be addressed. Another data analyzer who knows Ewen's disease is HIV but does not know his age and job can know from Table II(b) that Ewen's age is 51 and that he works as a lawyer. This is because the values of *Disease* are protected by frequency l-diversity, but other attributes are not protected. In practice, the age, address, and job of a person might be considered as private information. In this case, we should consider that these attributes have features of both QIDs and sensitive attributes. We refer to such attributes as "sensitive QIDs." Our proposed method can protect each sensitive QID S_j by l_j-diversity for all $j = 1, \ldots, q$, where q represents the number of sensitive QIDs.

Our contributions are as follows: (1) we propose new privacy models, namely, $(l_1, \ldots, l_q)$-*diversity* and $(t_1, \ldots, t_q)$-*closeness*, which can treat databases containing several sensitive QIDs; (2) we propose a simple but effective general anonymization algorithm for $(l_1, \ldots, l_q)$-diversity and $(t_1, \ldots, t_q)$-closeness, which is conducted by data holders; and (3) we propose a novel reconstruction algorithm that can decrease the reconstructed error between the reconstructed and the original values according to each data analyzer's purpose.

The rest of this paper is organized as follows. Privacy and utility as used in this paper are defined in Section II. Section III discusses the related methods. Section IV presents the novel privacy models. Section V presents the design of our algorithm. The results of our simulations are presented in Section VI. Section VII discusses several design issues in our method. Finally, Section VIII concludes the paper.

II. BACKGROUND

A. Assumptions

A data holder has a large database that contains personal information. The personal information may contain dozens of attributes. The types of attributes are "explicit identifiers", "non-sensitive QIDs", "sensitive QIDs", "non-QID sensitive attributes", and "non-QID non-sensitive attributes". Explicit identifiers are attributes that explicitly identify individuals, non-sensitive QIDs and sensitive QIDs are attributes that could be potentially combined with other directories to identify individuals, and sensitive QIDs and non-QID sensitive attributes are personal attributes of a private nature. We refer to the attribute values of sensitive QIDs and non-QID sensitive attributes as "sensitive values."

Although the specific goals of data mining can be different, knowledge of the probability distributions of the original data is generally required [11], [12]. The probability distribution of the original data can be represented by a cross tabulation (also called a contingency table or a multi-dimensional histogram). Following existing papers (e.g., [13], [11], [12], [14]), we aim to maintain the knowledge of the probability distribution of the original data.

B. Existing Privacy Models

Let T and T^* denote the original and anonymized databases, respectively. Assume that T contains explicit identifiers, non-sensitive QIDs, and non-QID sensitive attributes only.

First we define an *equivalence class* as follows:

Definition 1 (Equivalence class) *We denote a set of records that contains all the same non-sensitive QID values as an equivalence class.*

1) l-Diversity: There are several definitions of l-diversity. We introduce the two most popular ones.

Definition 2 (Frequency l-diversity) *The anonymized database T^* satisfies the frequency l-diversity if and only if the relative frequency of each of the sensitive values does not exceed $1/l$ for each equivalence class in T^*.*

For each equivalence class in Table II(b), there are two sensitive values and the relative frequency of each of the sensitive values is $1/2$. Therefore, Table II(b) satisfies the frequency 2-diversity.

On the other hand, suppose that the attribute of *Age* is a (non-QID) sensitive attribute and that other attributes are

(non-sensitive) QIDs in Table II(b). In this case, there are eight equivalence classes that consist of one row. Therefore, Table II(b) does not satisfy the frequency 2-diversity in this case.

Definition 3 (Entropy l-diversity) *Let $D(s)$ represent the domain of a non-QID sensitive attribute s. Let $n(v)$ represent the number of occurrences of value v in a non-QID sensitive attribute of records in an equivalence class. The anonymized database T^* satisfies the entropy l-diversity if and only if the following equation holds for all equivalence classes:*

$$\begin{aligned} &-\sum_{v\in D(s)} p(v)\log(p(v)) \geq \log(l), \\ &\text{where} \quad p(v) = n(v)/\sum_{w\in D(s)} n(w). \end{aligned} \tag{1}$$

2) t-Closeness:

Definition 4 (t-Closeness) *Let d represent the number of possible values of a non-QID sensitive attribute, and let the distribution of the sensitive values in the whole database be $\boldsymbol{A} = (a_1, \ldots, a_d)$. Let the distribution of the sensitive values in an equivalence class that is inferred from the original distribution of the equivalence class and the whole distribution of the database be $\boldsymbol{B} = (b_1, \ldots, b_d)$.*

The anonymized database T^* satisfies t-closeness if and only if the following equation holds for all equivalence classes:

$$\mathfrak{D}[\boldsymbol{A}, \boldsymbol{B}] \leq t, \tag{2}$$

where $\mathfrak{D}[\boldsymbol{A}, \boldsymbol{B}]$ represents the earth mover's distance (EMD) [15] of the distributions $\boldsymbol{A}$ and $\boldsymbol{B}$.

$\mathfrak{D}(\boldsymbol{A}, \boldsymbol{B})$ measures the cost of transforming distribution $\boldsymbol{A}$ to distribution $\boldsymbol{B}$ by moving the distribution mass between them. The cost of transforming a unit mass from element i of $\boldsymbol{A}$ to element j of $\boldsymbol{B}$ is defined as the ground distance d_{ij} between i and j.

Definition 5 (Earth mover's distance (EMD)) *Let $\boldsymbol{A} = (a_1, \ldots, a_q)$ and $\boldsymbol{B} = (b_1, \ldots, b_q)$ be probability distributions, and let $d_{i,j}$ be the ground distance between element i of $\boldsymbol{A}$ and element j of $\boldsymbol{B}$. We want to find $F = [f_{i,j}]$, where $f_{i,j}$ represents the flow from element i of $\boldsymbol{A}$ to element j of $\boldsymbol{B}$ that minimizes*

$$WORK(\boldsymbol{A}, \boldsymbol{B}, F) = \sum_{i=1}^{q}\sum_{j=1}^{q} d_{i,j} f_{i,j},$$

subject to the constraints

$$\begin{aligned} &f_{i,j} \geq 0, \qquad 1 \leq i \leq q, 1 \leq j \leq q \\ &a_i - \sum_{j=1}^{q} f_{i,j} + \sum_{j=1}^{q} f_{j,i} = b_i, \quad 1 \leq i \leq q \\ &\sum_{i=1}^{q}\sum_{j=1}^{q} f_{i,j} = \sum_{i=1}^{q} a_i = \sum_{i=1}^{q} b_i = 1. \end{aligned}$$

When the optimal F is found by solving the optimization problem, the EMD is defined as

$$\mathfrak{D}[\boldsymbol{A}, \boldsymbol{B}] = WORK(\boldsymbol{A}, \boldsymbol{B}, F).$$

C. Utility Metrics

Each attribute value in the original data set is compared to each of the perturbed versions of that value in the corresponding positions in some cases. However, when we want to create a (multidimensional) histogram, we can use the difference between the distributions of the reconstructed and original histograms to measure the utility much more directly. Many existing studies have included anonymized histograms [16], [17], [18], [19], [14], [20], [21], [22], but they cannot ensure $(l_1, \ldots, l_q)$-diversity or $(t_1, \ldots, t_q)$-closeness. These studies use utility metrics to calculate the distance between the reconstructed and the original distributions.

We use L1 distance, L2 distance, and Hellinger distance as utility metrics, as these can be used to calculate the distance between two distributions. L1 distance is the first-order version of the norm of the difference; it is very common in statistics [23]. L2 distance is also very common; it has been widely used as a utility metric for privacy-preserving data mining [17], [18], [14]. The advantage of using L1 and L2 distances is that they are easy for researchers to understand.

However, because these utility metrics do not consider the *magnitude* of each value, we need another utility metric. For example, consider the distributions A_1, B_1, A_2, and B_2 with the values {10, 100}, {10, 80}, {10, 25}, and {10, 5}, respectively. The L1 distance between A_1 and B_1 is 20 ($= |10-10| + |100-80|$). The L1 distance between A_2 and B_2 is also 20 ($= |10-10| + |25-5|$). The L2 distance between A_1 and B_1 is 20 ($= \sqrt{(10-10)^2 + (100-80)^2}$). The L2 distance between A_2 and B_2 is also 20 ($= \sqrt{(10-10)^2 + (25-5)^2}$). That is, the distances between A_1 and B_1 and between A_2 and B_2 are the same for L1 and L2. This is because the difference between the first attributes is 0 for all distributions and because the difference between the second attributes of A_1 and B_1 is the same as the difference between the second attributes of A_2 and B_2 (i.e., 20). However, the value of the second attributes of A_1 is only 1.25 times that of B_1, whereas the value of the second attributes of A_2 is 5 times that of B_2.

Hellinger distance, which preserves the properties of a distance metric (non-negativity, coincidence, symmetry, and triangle inequality) [24], has often been used to quantify two distributions [25], [26]. It can be considered as the L2 distance between the square roots of two distributions. Hellinger distance can be used to consider the magnitude of each value because the difference between the square roots of two small values is larger than the difference between the square roots of two large values—even if those values have the same absolute difference. The Hellinger distance between A_1 and B_1 is 0.74, whereas that between A_2 and B_2 is 1.95. The advantages of using Hellinger distance are that it includes the magnitude of each value and that it is relatively easy to understand because it is a variation of the L2 distance, which is a very common metric.

We now introduce the notations used in this paper. Database T has N records. Let r_i denote the ith record of T, let S denote the set of all attributes of T except the

explicit identifiers, and let q denote the size of S.

Let $D(s)$ denote the domain of possible values that can appear in attribute s, and let S' be the set of attributes that a data analyzer wants to analyze; therefore, S' is a subset of S. The size of S' is q'. Let S'_j be the jth attribute of S'.

Definition 6 (Target distribution) *Let C denote all of the combinations of the elements of $D(S'_1), \ldots, D(S'_{q'})$; that is,*

$$C = D(S'_1) \times D(S'_2) \times \ldots \times D(S'_{q'}). \tag{3}$$

Let c_m denote the mth element of C, and let $c_m[i]$ denote the ith attribute value of c_m.

Example 1 *Suppose that the data analyzer wants to analyze the relationships between Job and Disease, and suppose that $D(Job)$ is {Artist, Writer, Nurse} and $D(Disease)$ is {Fever, HIV, Cancer}. In this case, C = {Artist, Writer, Nurse} × {Fever, HIV, Cancer}. Therefore, c_1 = (Artist, Fever), c_2 = (Artist, HIV), …, c_9 = (Nurse, Cancer). The value of $c_1[1]$ and $c_1[2]$ is Artist and Fever, respectively.*

A data analyzer reconstructs the distribution of values for each c_m $(m = 1, \ldots, |C|)$. Let x_m denote the number of records that are categorized to c_m according to the values of their attributes. The value x_m is unknown to data analyzers. Moreover, let $\widehat{x_m}$ denote the reconstructed number of the values of x_m.

Here, we define these utility metrics. A data analyzer reconstructs the distribution of values for each c_m $(m = 1, \ldots, |C|)$. Let x_m denote the number of records that are categorized to c_m according to the values of their attributes. The value of x_m is unknown to the data analyzers. Moreover, let $\widehat{x_m}$ denote the reconstructed number of the values of x_m.

Definition 7 (L_1 distance)

$$\text{L1 distance} = \sum_{m=1}^{|C|} |x_m - \widehat{x_m}|. \tag{4}$$

Definition 8 (L_2 distance)

$$\text{L2 distance} = \sqrt{\sum_{m=1}^{|C|} (x_m - \widehat{x_m})^2}. \tag{5}$$

Definition 9 (Hellinger distance)

$$\text{Hellinger distance} = \frac{1}{\sqrt{2}} \sqrt{\sum_{m=1}^{|C|} (\sqrt{x_m} - \sqrt{\widehat{x_m}})^2}. \tag{6}$$

III. Related Work

A. k-Anonymity, l-diversity, and t-closeness

Algorithms for k-anonymity, l-diversity, and t-closeness have been widely studied in the area of privacy-preserving data mining such as [6], [7], [8], [9], [10].

Although k-anonymity can protect individual identities, in some cases it cannot protect the sensitive attributes of these individuals by attribute-linkage attacks.

Although many algorithms for l-diversity and t-closeness have been proposed, most of these assume that they can separate sensitive attributes from QIDs. For example, Soria-Comas et al. [27] proposed a method called t-closeness aware microaggregation, which generates t-closeness data sets by refining the microaggregation algorithm usually used for k-anonymity. Their method is an efficient method; however, they [27] assume that they can separate sensitive attributes from QIDs.

Shi et al. [28] introduced a new type of attribute called "quasi-sensitive attributes," which are not sensitive by themselves, but may become sensitive when used in combination. They also assume that they can separate (quasi-) sensitive attributes from QIDs.

Terrovitis et al. [29] proposed algorithms for k-anonymity that can be applied to a situation in which several attributes have features of both sensitive attributes and QIDs. However, their method cannot be applied to l-diversity or t-closeness in the said situation as is claimed in their paper.

Soria-Comas et al. [30] proposed an algorithm called IR-SWAP for k-anonymity. It generates different anonymized databases for each data analyzer. They [30] assume that they can separate QIDs and sensitive attributes in each execution of IR-SWAP on the basis of the knowledge of each data analyzer.

Wan et al. [31] had the same motivation as our study and proposed a concept named FF-anonymity, which emphasizes that treating attributes that have a feature of both QIDs and sensitive attributes is important. However, they have not proposed their anonymization algorithm yet.

Jin et al. [11] defined versatile privacy rules that can treat attributes with a feature of both QIDs and sensitive attributes, although they did not provide the notion of $(l_1, \ldots, l_q)$-diversity and $(t_1, \ldots, t_q)$-closeness. They proposed two algorithms named Guardian Decomposition (GD) and Utility-Aware Decomposition (UAD) for satisfying the versatile privacy rules. However, UAD completely loses the relationships between attributes when the corresponding values of $l_1, \ldots, l_q$ are more than 1 or when the corresponding values of $t_1, \ldots, t_q$ are less than 1. For example, if we want to anonymize Table I so that it satisfies $(2, 2, 2, 2)$-diversity, the result is four databases that have only one attribute each: Age, Address, Job, and Disease. All the records of these databases are randomized. Therefore, the data analyzer cannot analyze any attribute correlations. The same problem can occur in GD. Moreover, the algorithm of GD may get caught in an infinite loop if two or more values of $l_1, \ldots, l_q$ are more than 1 or if the corresponding values of $t_1, \ldots, t_q$ are less than 1.

Liu et al. proposed the Rating method [32] for l-diversity, which can anonymize databases even when there are several sensitive QIDs. This method publishes the Attribute Table (AT) and the ID Table (IDT) based on each sensitive coefficient, which is similar to l_j in this paper, of different attributes. In the Rating method, each tuple of the anonymized table has several values, like in our proposed method. However, our proposed method adds sensitive QID values completely randomly, whereas the Rating method adds values based on other records' sensitive QID values. Therefore, in this method, each reconstructed value is af-

fected heavily by other reconstructed values; that is, the reconstructed values of the Rating method are much smaller than the original values when the original values are larger than the average value, and vice versa.

Ye et al. [33] proposed a method for l-diversity that can anonymize databases with multiple sensitive attributes. However, they assume that they can separate QIDs and sensitive attributes. Moreover, the method breaks the relationship between sensitive attributes. For example, if they consider that the attributes of salary and disease are sensitive, then they cannot conduct a clear analysis of the relationship between salary and disease.

As pointed out by Cao et al. [34], even if the privacy parameter t of t-closeness is set to a fixed value, the way people feel can depend on the situation. For example, assume a database with sensitive attributes HIV and fever, and we set t to 0.1. If the sensitive attribute distribution of the whole database between them is $(0.5, 0.5)$, the ratio of HIV in an equivalence class can be up to 0.6, i.e., the probability of HIV can increase by up to 1.2 times (from 0.5 to 0.6). On the other hand, if the distribution of the whole database is $(0.01, 0.99)$, the ratio of HIV in an equivalence class can be up to 0.11, i.e., the probability of HIV can increase by up to 11 times (from 0.01 to 0.11). People might feel that the privacy is less protected in the latter situation than in the former situation, even though the privacy parameter t is set to the same value.

However, this type of problem can occur in any privacy models. Although Cao et al. proposed an enhanced β-likeness to tackle the limitation of t-closeness, the way people feel could still depend on the situation. For example, assume a database with sensitive attributes HIV and fever, and we set β to 0.7. If the sensitive attribute distribution of the whole database between them is $(0.5, 0.5)$, the ratio of HIV in an equivalence class can be up to 0.846. On the other hand, if the distribution of the whole database is $(0.01, 0.99)$, the ratio of HIV in an equivalence class can be up to 0.017. People might feel that the privacy is less protected in the former situation (the probability of HIV is increased from 0.5 to 0.846) than in the latter situation (the probability of HIV is increased from 0.01 to 0.017), even though the privacy parameter β is set to the same value. Therefore, we should determine the value of the privacy parameter in considering the worst-case situation, regardless of which privacy model we use. In this paper, we target l-diversity and t-closeness because these models have been widely studied for protecting privacy.

B. Differential Privacy

Differential privacy [35] makes user data anonymous by adding noise to a data set so that an attacker cannot determine whether or not a particular point of user data is included.

Recently, several studies for the generation of differentially private data sets have been proposed. Soria-Comas et al. [14] showed that the amount of noise to be added for the generation of differentially private data sets can be reduced using a microaggregation-based k-anonymity method. In their method, they first obtain a k-anonymous database and then generate a differentially private database by adding noise to each equivalence class. Sánchez et al. [36] proposed a more efficient algorithm and proved that the algorithm can reduce information loss.

Several studies for an anonymized histogram publication with differential privacy have also been proposed [16], [21], [19]. However, if the database has many attributes, it is difficult to apply these studies to a histogram publication. Recent studies such as [22],[18],[17] proposed algorithms that prevent the high cost of the calculation. However, they do not target l-diversity or t-closeness.

C. Randomization for Association Rule Mining

Rizvi et al. [37] proposed the MASK method, which preserves privacy for frequent itemset mining. Guo et al. [12] analyzed the effectiveness of MASK. In their studies, the value of an attribute is represented by a random vector $\mathbf{X} = \{\mathbf{x_i}\}$, such that $x_i = 0$ or 1. MASK generates the randomized vector by computing $y_i = x_i \oplus r_i$, where r_i takes a value of 0 with probability p and 1 with probability $(1-p)$. In other words, MASK changes the original value to another value with some probability. The number of values is not changed.

In contrast, our proposed method does not change the original value but adds several random values to each attribute value for realizing $(l_1, \ldots, l_q)$-diversity. In regard to $(t_1, \ldots, t_q)$-closeness, our proposed method changes the original value with some probability and also adds several random values. Moreover, optimization of the parameters for $(l_1, \ldots, l_q)$-diversity and $(t_1, \ldots, t_q)$-closeness and the proofs are also our contribution.

IV. PROPOSED PRIVACY MODELS

To simplify our discussions, from here on, we assume that a database consists of explicit identifiers and sensitive QIDs only. However, our method can be used if we assume that databases contain not only explicit identifiers and sensitive QIDs but also non-sensitive QIDs, non-QID sensitive attributes, and non-QID non-sensitive attributes. We discuss this in Section VII.

Let S be the set of sensitive QIDs in a database, let q be the number of sensitive QIDs (i.e., $q = |S|$), and let $E(r, s)$ denote the value of the QID s of record r in database T.

Let S_j be the jth sensitive QID of S. Let $D(S_j)$ denote the domain of possible values that can appear in S_j.

Then, we define our novel definition of an equivalence class.

Definition 10 (Equivalence Class for S_j) *We denote a set of records that contains all the same sensitive QID values except for S_j as an equivalence class for S_j.*

Example 2 *There are four equivalence classes for S_4 in Table II(b). Equivalence classes for S_4 are the same as the original equivalence classes if we consider that S_4 is a non-QID sensitive attribute and that other attributes are non-sensitive QIDs.*

On the other hand, there are four equivalence classes each for S_1, S_2, and S_3.

A. $(l_1, \ldots, l_q)$-Diversity

We define one of our novel privacy models: $(l_1, \ldots, l_q)$-diversity.

Definition 11 (Frequency $[\![l, j]\!]$-diversity) *The anonymized database T^* satisfies frequency $[\![l, j]\!]$-diversity if and only if the relative frequency of each of the sensitive values does not exceed $1/l$ for each equivalence class for S_j in T^*.*

Definition 12 (Frequency $(l_1, \ldots, l_q)$-diversity) *The anonymized database T^* satisfies frequency $(l_1, \ldots, l_q)$-diversity if and only if frequency $[\![l_j, j]\!]$-diversity is satisfied for all $j = 1, \ldots, q$.*

Example 3 *Table II(b) satisfies frequency $[\![1, 1]\!]$-diversity, frequency $[\![1, 2]\!]$-diversity, frequency $[\![1, 3]\!]$-diversity, and frequency $[\![2, 4]\!]$-diversity; that is, it satisfies $(1, 1, 1, 2)$-diversity.*

We can also provide entropy-$(l_1, \ldots, l_q)$-diversity.

Definition 13 (Entropy $[\![l, j]\!]$-diversity) *Let $n_j(v)$ represent the number of occurrences of value v in the jth sensitive QID of records in an equivalence class for S_j. T^* satisfies entropy $[\![l, j]\!]$-diversity if and only if the following equation holds for all equivalence classes for S_j:*

$$\begin{aligned} &- \sum_{v \in D(S_j)} p_j(v) \log(p_j(v)) \geq \log(l), \\ &\text{where} \quad p_j(v) = n_j(v) / \sum_{w \in D(S_j)} n_j(w). \end{aligned} \tag{7}$$

Definition 14 (Entropy-$(l_1, \ldots, l_q)$-diversity) *The anonymized database T^* satisfies entropy $(l_1, \ldots, l_q)$-diversity if and only if entropy $[\![l_j, j]\!]$-diversity is satisfied for all $j = 1, \ldots, q$.*

B. $(t_1, \ldots, t_q)$-Closeness

We consider distance functions other than EMD to calculate the distance; however, the original paper on t-closeness [5] uses EMD as the distance function, and EMD is the most common choice for calculating t-closeness [27]. Therefore, we use EMD as the distance function for calculating $(t_1, \ldots, t_q)$-closeness.

Definition 15 ($[\![t, j]\!]$-Closeness) *Let d_j represent the size of $D(S_j)$, and let the distribution of the elements of $D(S_j)$ in the whole database be $\boldsymbol{A_j} = (a_1, \ldots, a_{d_j})$. Let the distribution of the sensitive values of S_j in an equivalence class for S_j that is inferred from the original distribution of the equivalence class for S_j and the whole distribution of the database be $\boldsymbol{B_j} = (b_1, \ldots, b_{d_j})$.*

The anonymized database T^* satisfies $[\![t, j]\!]$-closeness if and only if the following equation holds for all equivalence classes for S_j in T^*:

$$\mathfrak{D}[\boldsymbol{A_j}, \boldsymbol{B_j}] \leq t, \tag{8}$$

where $\mathfrak{D}[\boldsymbol{A_j}, \boldsymbol{B_j}]$ represents the EMD of the distributions $\boldsymbol{A_j}$ and $\boldsymbol{B_j}$.

Note that inference can be diverse, but the inference we considered in this paper is that by Bayes' theorem.

Definition 16 ($(t_1, \ldots, t_q)$-Closeness) *The anonymized database T^* satisfies $(t_1, \ldots, t_q)$-closeness if and only if $[\![t_j, j]\!]$-closeness is satisfied for all $j = 1, \ldots, q$.*

V. Proposed Algorithm

A. Outline

Our proposed method consists of two steps: a randomization conducted by the data holder and a reconstruction conducted by the data analyzer. The data holder determines the parameters of the proposed method, $p_1, \ldots, p_q$ and $\eta_1, \ldots, \eta_q$, according to the values of $(l_1, \ldots, l_q)$ or $(t_1, \ldots, t_q)$. Next, the anonymization algorithm generates an anonymized record in an "aggregated expression" from each record on the basis of the parameters and inserts the record into the anonymized database.

The data analyzer first determines which sensitive QIDs should be analyzed. Then, the number of records is estimated for each combination of the values of the selected sensitive QIDs.

Note that the anonymization algorithm and the reconstruction algorithm are common for all privacy models, even though the parameters $p_1, \ldots, p_q$ and $\eta_1, \ldots, \eta_q$ needed for these algorithms are different for each privacy model.

B. Anonymization Algorithm

First, we describe the concept of the proposed algorithm. Because the time complexity of the concept algorithm is very high, we will describe the actual algorithm later.

The proposed algorithm can be used for any of frequency $(l_1, \ldots, l_q)$-diversity, entropy $(l_1, \ldots, l_q)$-diversity, and $(t_1, \ldots, t_q)$-closeness. We first use frequency $(l_1, \ldots, l_q)$-diversity as an example because it is the easiest to understand, and then we generalize the algorithm later.

For each sensitive QID S_j of each record r_i, the data holder extracts $l_j - 1$ distinct values randomly from $D(S_j) \setminus \{E(r_i, S_j)\}$ and creates a set $R_{i,j}$ containing the extracted values and an original value $E(r_i, S_j)$. This step is conducted for every sensitive QID $S_1, \ldots, S_q$. Then, the data holder calculates the Cartesian product of $R_{i,1}, \ldots, R_{i,q}$ and inserts each element of the Cartesian product into the anonymized database. This process is conducted for every record.

Example 4 *A database has one record, which has the name, age, and disease of Alice (Table III(a)). Assume that $l_1 = l_2 = l_3 = 2$ and that the anonymization algorithm generates $R_{1,1} = \{33, 41\}$, $R_{1,2} = \{10105, 32515\}$, and $R_{1,3} = \{Cold, HIV\}$. Table III(b) represents the result of the anonymization. The sixth record represents the true record of Alice. Even if the data analyzer knows the values of any of the two sensitive QIDs for Alice, the data analyzer cannot specify Alice's value for the other sensitive QID with a confidence greater than $1/2$. For example, assume that a data analyzer knows that Alice's age is 41 and her address is 10105. The data analyzer cannot know whether the fifth*

TABLE III: Example of Anonymization by the Proposed Method for Frequency 2-Diversity

(a) Alice's original record

Name	Age	Address	Disease
Alice	41	10105	HIV

(b) Alice's anonymized records

Age	Address	Disease
33	10105	Cold
33	10105	HIV
33	32515	Cold
33	32515	HIV
41	10105	Cold
41	10105	HIV
41	32515	Cold
41	32515	HIV

(c) Aggregated expression of (b)

Age	Address	Disease
{33,41}	{10105,32515}	{Cold,HIV}

TABLE IV: Frequency (2, 2, 2, 3)-Diversity in Aggregated Expressions by Our Proposed Method in Table I

Age	Address	Job	Disease
{**41**, 23}	{12255, **13021**}	{**Artist**, Programmer}	{**Fever**, Flu, Chill}
{79, **41**}	{14000, **17025**}	{Lawyer, **Writer**}	{Cold, **Obesity**, Pus}
{42, **51**}	{**13021**, 13997}	{Writer, **Lawyer**}	{Cancer, **Fever**, Sty}
{**51**, 33}	{18002, **14053**}	{Driver, **Lawyer**}	{Cold, Flu, **Obesity**}
{15, **51**}	{**14003**, 15500}	{**Lawyer**, Teacher}	{Cold, **HIV**, Sty}
{**51**, 69}	{**16005**, 12332}	{Researcher, **Lawyer**}	{**HIV**, Fever, Sty}
{**51**, 39}	{14005, **14003**}	{**Lawyer**, Programmer}	{Flu, Cold, **Fever**}
{**51**, 60}	{**16005**, 12001}	{**Lawyer**, Writer}	{**Obesity**, HIV, Pus}

record or the sixth record is Alice's record in Table III(b) as these two records have different values of disease.

Because the size of the Cartesian product of $R_{i,1}, \ldots, R_{i,q}$ could be very large, the data holder instead generates an anonymized record in an "aggregated expression." Table III(c) shows the anonymized record in an aggregated expression of Table III(b). Note that Table III(c) is equivalent to Table III(b).

Note that each record can be grouped in several equivalence classes. For example, in Table III(b), the first and the second rows construct an equivalence class for S_3 because each row contains 33 for S_1 and 10105 for S_2. Moreover, the first and the third rows construct an equivalence class for S_2 because each row contains 33 for S_1 and Cold for S_3.

Table IV represents an example of the frequency $(2, 2, 2, 3)$-diversity of Table I in aggregated expressions. In Table IV, the bold font represents the original values. Note that it is for clarity only. Our proposed method adds $(l_j - 1)$ random values to each cell so that the jth column satisfies frequency $[\![l_j, j]\!]$-diversity.

If a data analyzer knows the values of Ewen's age, address, and job, the data analyzer can conclude that the fifth row is his, but cannot interpret from Table IV if he has a cold, HIV, or Sty. If another data analyzer knows Ewen's disease, the data analyzer cannot interpret from Table IV which of 15, 51, 60, and 69 is his age or which of 12001, 12332, 14003, 15500, and 16005 is his address, etc.

Then, we generalize the algorithm mentioned above. The data holder removes the explicit identifiers from an original database. Next, we calculate the optimized parameters $p_1, \ldots, p_q$ and $\eta_1, \ldots, \eta_q$. The method of calculating the optimized parameters is described in Section V-C. For example, in regard to frequency $(l_1, \ldots, l_q)$-diversity, $p_1, \ldots, p_q$ are set to 1 and $\eta_1, \ldots, \eta_q$ are set to $l_1, \ldots, l_q$.

For each sensitive QID S_j of each record r_i, the algorithm creates an empty set $R_{i,j}$ and tosses a coin with head probability p_j. If the coin is head, the algorithm adds an original value $E(r_i, S_j)$ and $\eta_j - 1$ distinct elements randomly extracted from $D(S_j) \backslash \{E(r_i, S_j)\}$ to set $R_{i,j}$. If the coin is tail, the algorithm adds η_j distinct elements randomly extracted from $D(S_j)$ to set $R_{i,j}$. This step is conducted for every sensitive QID $S_1, \ldots, S_q$. Then, the data holder inserts $\{R_{i,1}, \ldots, R_{i,q}\}$ into T^* as an anonymized record in an aggregated expression. This process is conducted for every record.

C. Parameter Optimization

1) Parameters for Frequency $(l_1, \ldots, l_q)$-Diversity: The expected L2 distance in regard to S_j is calculated on the basis of the study of Bayes algorithm [38] by

$$E_{L2} = \sqrt{\frac{(1 - d_j)(p_j^2 d_j + \eta_j - (p_j^2 + d_j)\eta_j)}{p_j^2 d_j N (d_j - \eta_j)}} \tag{9}$$

because L2 distance is the square root of d_j times of the MSE defined in [38].

We have the following theorem:

Theorem 1 *E_{L2} is an (or a weakly) increasing function of η_j and a (weakly) decreasing function of p_j.*

Proof. By differentiating (9) with respect to η_j, we get

$$1/\left(2p_j\sqrt{(d_j - \eta_j)^3 N \frac{-d_j p_j^2 + \eta_j(-1 + d_j + p_j^2)}{(d_j - 1)^3 d_j}}\right). \tag{10}$$

Because the value of (10) is greater than or equal to 0, the expected L2 distance decreases with decreasing η_j.

By differentiating (9) with respect to p_j, we get

$$\frac{-(d_j - 1)^2 \eta}{p^2\sqrt{(d_j - 1)d_j(d_j - \eta_j)N(-d_j p_j^2 + \eta(-1 + d_j + p_j^2))}}. \tag{11}$$

Because the value of (11) is always less than or equal to 0, the expected L2 distance decreases with increasing p_j. □

In regard to frequency $(l_1, \ldots, l_q)$-diversity, the parameters $\eta_1, \ldots, \eta_q$ should be larger than or equal to $l_1, \ldots, l_q$, respectively. Therefore, from Theorem 1, $\eta_1, \ldots, \eta_q$ are set to $l_1, \ldots, l_q$ and $p_1, \ldots, p_q$ are all set to 1 to minimize the expected L2 distance.

We have the following theorem:

Theorem 2 *The anonymization algorithm always generates database T^* of frequency $(l_1, \ldots, l_q)$-diversity by setting p_j = 1 and η_j = l_j $(j = 1, \ldots, q)$ if every d_j is larger than or equal to l_j.*

Proof. After conducting the anonymization algorithm for database T, the algorithm generates $l_1 \times \ldots \times l_q$ records from each record, in a no-aggregated expression mode. We focus on an equivalence class for S_j in anonymized database T^* and assume that its size is δ.

The equivalence class for S_j in T^* has been generated from exactly δ/l_j records in the original database T because the algorithm generates l_j records from each record that has all of the same sensitive QID values except for S_j. Hence, the possible maximum number of occurrences of each value of S_j within the equivalence class for S_j is δ/l_j because the algorithm does not generate more than one record from a record in T that has all the same sensitive QID values.

Therefore, the possible maximum relative frequency of each of the values of S_j is $1/l_j$ for each equivalence class for S_j for all $j = 1, \ldots, q$ in T^*. □

2) Parameters for Entropy $(l_1, \ldots, l_q)$-Diversity: We have the following theorem:

Theorem 3 *The anonymization algorithm always generates database T^* of entropy $(l_1, \ldots, l_q)$-diversity by setting p_j = 1 and $\eta_j = l_j$ $(j = 1, \ldots, q)$ if every d_j is larger than or equal to l_j.*

Proof. Focus on an equivalence class for S_j and assume that the size of the equivalence class for S_j is δ. The possible minimum number of occurrences of each sensitive QID value of S_j within the δ records is 1. From the proof of Theorem 2, the possible maximum number of occurrences of each value of S_j within the equivalence class for S_j is δ/l_j. Therefore, the possible value of $p_j(v)$ in (7) is from $1/\delta$ to $1/l_j$. The values of δ and l_j are larger than or equal to 2, and δ is larger than or equal to l_j. From the characteristics of the entropy, the entropy has its smallest value when $p_j(v)$ for all v is $1/l_j$. In this case, the left side of (7) is $\log(l_j)$ and this value is equal to the right side of (7). □

3) Parameters for $(t_1, \ldots, t_q)$-Closeness: Let U_j be a random variable denoting $E(r, S_j)$ of record r, and let V_j be a random variable denoting $E(r^*, S_j)$ of record r^*.

Let the distribution of the values of S_j in the whole database be $\boldsymbol{A_j} = (\boldsymbol{A_j}[1], \ldots, \boldsymbol{A_j}[d_j])$, and let the distribution of the values of S_j that is inferred from $\boldsymbol{A_j}$ and $E(r^*, S_j)$ be $\boldsymbol{B_j} = (\boldsymbol{B_j}[1], \ldots, \boldsymbol{B_j}[d_j])$.

First, we introduce the method to calculate $\boldsymbol{B_j}$.

Let κ_c denote the probability that $E(r^*, S_j)$ contains the original value and specified $\eta_j - 1$ distinct values, and let κ_d denote the probability that $E(r^*, S_j)$ does not contain the original value but contains a specified η_j distinct values. The values of κ_c and κ_d are represented by

$$\kappa_c = \frac{p_j + (1-p_j)\eta_j/d_j}{{}_{d_j-1}C_{\eta_j-1}}, \quad \kappa_d = \frac{1-(p_j+(1-p_j)\eta_j/d_j)}{{}_{d_j-1}C_{\eta_j}}.$$

From Bayes' theorem, we get

$$P(U_j = D(S_j)[m] | V_j = E(r^*, S_j)) = \frac{P(V_j = E(r^*, S_j) | U_j = D(S_j)[m]) P(U_j = D(S_j)[m])}{P(V_j = E(r^*, S_j))}.$$

Here, we have

$$P(U_j = D(S_j)[m]) = \boldsymbol{A_j}[m],$$
$$P(U_j = D(S_j)[m] | V_j = E(r^*, S_j)) = \boldsymbol{B_j}[m],$$

and

$$P(V_j = E(r^*, S_j) | U_j = D(S_j)[m]) = \begin{cases} \kappa_c & (D(S_j)[m] \in E(r^*, S_j)) \\ \kappa_d & (otherwise). \end{cases}$$

From the law of total probability,

$$P(V_j = E(r^*, S_j)) = \sum_{\beta=1}^{d_j} P(V_j = E(r^*, S_j) | U_j = D(S_j)[\beta]) P(U_j = D(S_j)[\beta]).$$

From these equations, we get

$$\boldsymbol{B_j}[m] = \frac{\Upsilon(E(r^*, S_j), D(S_j)[m])}{\sum_{\beta=1}^{d_j} \Upsilon(E(r^*, S_j), D(S_j)[\beta])}, \tag{12}$$

where

$$\Upsilon(E(r^*, S_j), D(S_j)[m]) = \begin{cases} \kappa_c \boldsymbol{A_j}[m] & (D(S_j)[m] \in E(r^*, S_j)) \\ \kappa_d \boldsymbol{A_j}[m] & (otherwise.) \end{cases}$$

If $\mathfrak{D}[\boldsymbol{A_j}, \boldsymbol{B_j}]$ is less than or equal to t_j for all records and for all j, the database satisfies $(t_1, \ldots, t_q)$-closeness from Lemma 1 described later. This is because we know from Lemma 1 that the maximum distance of any sets of records from the whole distribution can never increase when merging two sets of records.

$E[r^*, S_j]$ represents the anonymized set of values, and this set is obtained by the randomized mechanism described in Section V-B. Therefore, we calculate $\boldsymbol{B_j}$ for all possible $E[r^*, S_j]$ and then obtain the largest value of $\mathfrak{D}[\boldsymbol{A_j}, \boldsymbol{B_j}]$.

Finally, we get the optimal combination of η_j and p_j for S_j that satisfies $\mathfrak{D}[\boldsymbol{A_j}, \boldsymbol{B_j}] \leq t_j$ and minimizes the expected L2 distance. The optimized parameters might not minimize the L1 distance and Hellinger distance, but we can confirm that the parameters can reduce both the L1 distance and the Hellinger distance of the proposed method greatly by using the simulations of real data sets in Section VI.

Algorithm 1 Parameter optimization for $(t_1, \ldots, t_q)$-closeness

Input: Privacy parameters $t_1, \ldots, t_q$, Original database T
Output: Combinations of optimized parameters $(\langle \widehat{p_1}, \widehat{\eta_1} \rangle, \ldots, \langle \widehat{p_q}, \widehat{\eta_q} \rangle)$

```
for j = 1, ..., q do
    A_j ⇐ the frequency distribution of the values of D(S_j) in T
    Create associative array ParaCombs
    for η = 1, ..., η̃ do
        𝔖 ⇐ {Z | Z ⊆ D(S_j) ∧ |Z| = η}
        for each element ∈ 𝔖 do
            B_j ⇐ the result of (12)
            p ⇐ the solution of the equation t_j = 𝔇[A_j, B_j] for p
            el2 ⇐ E_L2(N, d_j, p, η)
            paraCombs.insert(⟨p, s⟩, el2)
        end for
    end for
    ⟨p̂_j, η̂_j⟩ ⇐ the key that has the minimum value in paraCombs
end for
```

Algorithm 1 shows the algorithm that calculates the optimized parameters for $(t_1, \ldots, t_q)$-closeness. First, $\boldsymbol{A_j}$ is calculated from database T (lines 2). Then, the algorithm calculates the optimal p for each η. The range of η is from 1 to $\tilde{\eta}$, which is a predefined parameter. The value of $\tilde{\eta}$ can be set to d_j, but it takes much time when we set $\tilde{\eta}$ to a larger value.

The set of subsets of $D(S_j)$ of cardinality equal to η is substituted into $\mathfrak{S}$ (line 5). For each element of $\mathfrak{S}$, the algorithm calculates the largest value of p that satisfies $\mathfrak{D}[\boldsymbol{A_j}, \boldsymbol{B_j}] \leq t_j$. We can get the largest value of p by solving the equation $\mathfrak{D}[\boldsymbol{A_j}, \boldsymbol{B_j}] = t_j$ for p. Because it is difficult to solve the equation, the algorithm performs a binary search. To narrow the search space, we can use Theorem 1.

With the parameters p and η, an expected L2 distance is calculated according to (9) (line 9). Then, the parameters and the expected L2 distance are inserted into the associative array $ParaCombs$, binding the parameters to the expected L2 distance (line 10).

Finally, the algorithm chooses the combination of the optimized parameters, which has the minimum value of the expected L2 distance (line 13).

Theorem 4 *The anonymization algorithm always generates database T^* of $(t_1, \ldots, t_q)$-closeness by setting η_j and p_j $(j = 1, \ldots, q)$ according to Algorithm 1.*

Proof. Algorithm 1 generates a combination of p_j and η_j, which satisfies $\mathfrak{D}[A_j, B_j^{(i)}] \leq t_j$, where A_j represents the distribution of the elements of $D(S_j)$ in the whole database and $B_j^{(i)}$ represents the probability distribution inferred from the distribution of $E(r_i^*, S_j)$ and the whole distribution of the database.

From Lemma 1 described below, we know that the maximum distance of any sets of records from the whole distribution of the database can never increase when merging two sets of records. Therefore, if each record r_i^* holds $\mathfrak{D}[A_j, B_j^{(i)}] \leq t_j$, the anonymized database always satisfies $[\![t_j, j]\!]$-closeness.

Because this discussion is common for all j, the anonymized database satisfies $(t_1, \ldots, t_q)$-closeness. □

Lemma 1 *Let $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{B'}$ represent distributions. We have the following equation:*

$$\mathfrak{D}[\boldsymbol{A}, (1-\lambda)\boldsymbol{B} + \lambda\boldsymbol{B'}] \leq (1-\lambda)\mathfrak{D}[\boldsymbol{A}, \boldsymbol{B}] + \lambda\mathfrak{D}[\boldsymbol{A}, \boldsymbol{B'}], \tag{13}$$

for any λ in [0,1].

Proof. Let $\boldsymbol{A}$, $\boldsymbol{A'}$, $\boldsymbol{B}$, and $\boldsymbol{B'}$ represent distributions.

First, the EMD is absolutely homogeneous, i.e.,

$$\mathfrak{D}[\lambda\boldsymbol{A}, \lambda\boldsymbol{B}] = \lambda\mathfrak{D}[\boldsymbol{A}, \boldsymbol{B}] \text{ for any } \lambda \geq 0, \tag{14}$$

because it leads to a scaled but equivalent minimization problem with solutions r^* and λr^*.

Second, the EMD is subadditive, i.e.,

$$\mathfrak{D}[\boldsymbol{A} + \boldsymbol{A'}, \boldsymbol{B} + \boldsymbol{B'}] \leq \mathfrak{D}[\boldsymbol{A}, \boldsymbol{B}] + \mathfrak{D}[\boldsymbol{A'}, \boldsymbol{B'}], \tag{15}$$

because any optimal solutions r^* and r'^* to the latter minimization problem provide a feasible (albeit not necessarily optimal) solution $r^* + r'^*$ to the additive version.

On the basis of (14) and (15), we get

$$\begin{aligned} &\mathfrak{D}[(1-\lambda)\boldsymbol{A} + \lambda\boldsymbol{A'}, (1-\lambda)\boldsymbol{B} + \lambda\boldsymbol{B'}] \\ &\leq \mathfrak{D}[(1-\lambda)\boldsymbol{A}, (1-\lambda)\boldsymbol{B}] + \mathfrak{D}[\lambda\boldsymbol{A'}, \lambda\boldsymbol{B'}] \\ &= (1-\lambda)\mathfrak{D}[\boldsymbol{A}, \boldsymbol{B}] + \lambda\mathfrak{D}[\boldsymbol{A'}, \boldsymbol{B'}] \end{aligned} \tag{16}$$

for any λ in [0,1].

By substituting $\boldsymbol{A}$ for $\boldsymbol{A'}$ in (16), we get (13). □

D. Reconstruction Algorithm

In this subsection, we consider that data analyzers receive anonymized databases in aggregated expressions. First, the data analyzers who receive the anonymized database determine their target distribution C. Then, they reconstruct the distribution of the sensitive QID values for each c_m $(m = 1, \ldots, |C|)$. We use x_m to represent the actual number of records, which are categorized to c_m. Let $\widehat{x_m}$ denote the reconstructed x_m, which is estimated by the data analyzers. We describe the ValueAdding algorithm, which is a simple algorithm, and then we extend it.

1) ValueAdding Reconstruction Algorithm: Let S' be the set of sensitive QIDs that a data analyzer wants to analyze, and let $E(r_i^*, S'_j)$ represent the values of the jth QID of S' at anonymized record r_i^* in an aggregated expression. Let the function $\mathfrak{E}(r_i^*, S')$ return the Cartesian product of $E(r_i^*, S'_j)$ for $j = 1, \ldots, q'$, i.e.,

$$\mathfrak{E}(r_i^*, S') = E(r_i^*, S'_1) \times E(r_i^*, S'_2) \times \ldots \times E(r_i^*, S'_{q'}).$$

Then, the data analyzer counts how many elements of $\mathfrak{E}(r_i^*, S')$ $(i = 1, \ldots, N)$ are categorized to c_m in the anonymized database; that is, the data analyzer calculates

$$\begin{aligned} w_m &= \sum_{i=1}^{N} H(r_i^*, c_m), \text{ where} \\ H(r_i^*, c_m) &= \begin{cases} 1 & (\mathfrak{E}(r_i^*, S') \text{ contains } c_m) \\ 0 & (otherwise) \end{cases}. \end{aligned} \tag{17}$$

Let $\eta(S'_j)$ represent the parameter η for the sensitive QID S'_j. Let $p(S'_j)$ represent the parameter p for the sensitive QID S'_j, and let $d(S'_j)$ represent the size of $D(S'_j)$.

Focus on an original record r_i. The set of the sensitive QID values of r_i can be represented by one of the elements of C. Assume that the combination of the values of sensitive QIDs selected by a data analyzer in an original record r_i is categorized to $\widehat{c_m} \in C$. The set $\mathfrak{E}(r_i^*, S')$ has $\prod_{s' \in S'} \eta(s')$ elements, and the set contains $\widehat{c_m}$ with probability $\prod_{s' \in S'}[(p(s') + (1 - p(s'))\eta/d(s')]$. On the other hand, the size of the target distribution is $|C|$, and the set $\mathfrak{E}(r_i^*, S')$ does not contain $\widehat{c_m}$ with probability $1 - \prod_{s' \in S'}[(p(s') + (1 - p(s'))\eta/d(s')]$.

On the basis of this observation, we can calculate the reconstruction results as follows:

$$\begin{aligned} \widehat{x_m} = w_m \cdot \frac{\prod_{s' \in S'}[(p(s') + (1 - p(s'))\eta/d(s')]}{\prod_{s' \in S'} \eta(s')} + \\ (N - w_m) \cdot \frac{1 - \prod_{s' \in S'}[(p(s') + (1 - p(s'))\eta/d(s')]}{|C| - \prod_{s' \in S'} \eta(s')}. \end{aligned} \tag{18}$$

This calculation is simple, but the resulting L1, L2, and Hellinger distances could be large. We describe the extended version of the reconstruction algorithm in the following subsection.

2) Proposed Reconstruction Algorithm: We extend the ValueAdding algorithm by using Bayes' technique [39], [40] for the reconstruction.

Assume that a combination of values of sensitive QIDs selected by a data analyzer in an original record r is categorized to $c_\alpha \in C$. First, the data analyzers calculate the probability $\delta_{\alpha,\beta}$ that the original record r is anonymized to a record r^* whose $\mathfrak{E}(r^*, S')$ contains $c_\beta \in C$ for each α and β.

Focus on the jth elements of c_α and c_β. The anonymized values of the jth sensitive QID of S' contain the jth element of c_α with probability $p_j + (1-p_j)\eta_j/d_j$. The anonymized values contain a specific element of $D(S'_j)$ other than the jth element of c_α with probability $p_j \cdot (\eta_j - 1)/(d_j - 1) + (1-p_j) \cdot \eta_j/d_j$.

Therefore, $\delta_{\alpha,\beta}$ is represented by

$$\delta_{\alpha,\beta} = \prod_{j=1}^{q'} F(\alpha,\beta,j), \quad \text{where } F(\alpha,\beta,j) = \begin{cases} p_j + (1-p_j)\eta_j/d_j & (c_\alpha[j] = c_\beta[j]) \\ p_j(\eta_j - 1)/(d_j - 1) + (1-p_j)\eta_j/d_j & (otherwise.) \end{cases} \tag{19}$$

Equation 19 can be calculated for all combinations of $\alpha = 1, \ldots, |C|$ and $\beta = 1, \ldots, |C|$, but the number of different results is only $2^{q'}$. Note that we usually assume that q' is less than four in creating cross tabulations [41], [42].

Here, we calculate (19) for only the $2^{q'}$ results.

Let $Z_{\alpha,\beta}$ be a function that returns bit array $b_1, \ldots, b_{q'}$, in which b_i is determined by the following equation:

$$b_j = \begin{cases} 1 & (c_\alpha[j] = c_\beta[j]) \\ 0 & (otherwise.) \end{cases} \tag{20}$$

Let $Z_{\alpha,\beta}[j]$ represent the jth bit of $Z_{\alpha,\beta}$. The number of possible values of $Z_{\alpha,\beta}$ is only $2^{q'}$. Here, we refine (19) as follows:

$$\delta_{\alpha,\beta} = \prod_{j=1}^{q'} \{Z_{\alpha,\beta}[j](p_j + (1-p_j)\eta_j/d_j) + (1 - Z_{\alpha,\beta}[j]) \cdot (p_j \cdot (\eta_j - 1)/(d_j - 1) + (1-p_j) \cdot \eta_j/d_j)\}. \tag{21}$$

Let the original state of the sensitive QID values of a record in T be given by a random variable U, and let the state of each element of $\mathfrak{E}(r^*, S')$ be given by a random variable V.

From the law of total probability, $Pr(U = \alpha)$ can be represented by

$$Pr(U = \alpha) = \frac{\sum_{\beta=1}^{|C|} Pr(V \ni \beta) Pr(U = \alpha | V \ni \beta)}{\sum_{\beta=1}^{|C|} Pr(V \ni \beta)}. \tag{22}$$

From Bayes' theorem, we have

$$Pr(U = \alpha | V \ni \beta) = \frac{Pr(V \ni \beta | U = \alpha) Pr(U = \alpha)}{\sum_{\gamma=1}^{|C|} Pr(V \ni \beta | U = \gamma) Pr(U = \gamma)} = \frac{\delta_{\alpha,\beta} \widehat{x_\alpha}}{\sum_{\gamma=1}^{|C|} \delta_{\gamma,\beta} \widehat{x_\gamma}}. \tag{23}$$

We can express $Pr(U = \alpha)$ as x_α/N, which is an unknown value. By using an estimation value $\widehat{x_\alpha}$ instead of x_α, we can introduce the following equation:

$$Pr(U = \alpha) = \widehat{x_\alpha}/N. \tag{24}$$

$Pr(V \ni \beta)$ in (23) represents the probability that $\mathfrak{E}(r^*, S')$ contains c_β. Because there are N records and c_β occurs w_β times in the anonymized database, we have

$$Pr(V \ni \beta) = w_\beta/N. \tag{25}$$

Therefore, we get from (22), (23), (24), and (25) the following:

$$\widehat{x_\alpha}^{\sharp+1} \Leftarrow \sum_{\beta=1}^{|C|} w_\beta \frac{\delta_{\alpha,\beta} \widehat{x_\alpha}^{\sharp}}{\sum_{\gamma=1}^{|C|} \delta_{\gamma,\beta} \widehat{x_\gamma}^{\sharp}}, \tag{26}$$

where an element of $\widehat{x_\alpha}^{\sharp}$ $(\alpha = 1, \ldots, |C|)$ represents the iteration at step $\sharp$. We set an initial value of $\widehat{x_\alpha}^{0}$ to w_α for all α and repeat (26) until the difference between $\widehat{x_\alpha}^{\sharp}$ and $\widehat{x_\alpha}^{\sharp+1}$ for all α is sufficiently small. We finally get

$$\widehat{x_\alpha} \Leftarrow \widehat{x_\alpha} \Big/ \prod_{s' \in S'} \eta(s'), \tag{27}$$

because we can express $\sum_{\beta=1}^{|C|} Pr(V \ni \beta)$ as $\prod_{s' \in S'} \eta(s')$.

VI. Evaluation

We evaluated the L1, L2, and Hellinger distances, and the calculation time for frequency $(l_1, \ldots, l_q)$-diversity and $(t_1, \ldots, t_q)$-closeness. The criteria for whether the values of the utility metrics are good or not can be quite different according to the purpose of the data analysis. It is our future work to propose a way to decide the criteria. The results of entropy $(l_1, \ldots, l_q)$-diversity were omitted because the results were similar to those of frequency $(l_1, \ldots, l_q)$-diversity. We used real data sets, which were the Adult data set and the US Census data set[1]. The Adult data set consists of 15 attributes (e.g., age, sex, race, relationship, etc.) and has 45,222 records after eliminating the records with unknown values. Because several attributes are continuous, we categorized each of them. The domain sizes of the categorized attributes varied from 2 to 41.

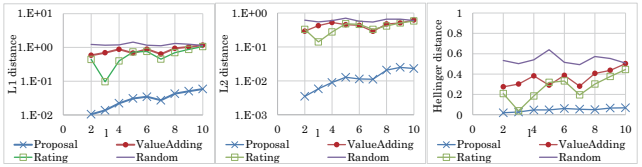
The US Census data set consists of 68 categorical attributes and has 2,458,285 records. The domain sizes of the categorical attributes varied from 2 to 18.

Many studies that use l-diversity, such as [4], [32], set l from 2 to 10. Following their setting, in this paper, we also set l from 2 to 10. In regard to t-closeness, we set t from 0.1 to 0.5, which covers the range of different privacy levels observed in existing studies [5].

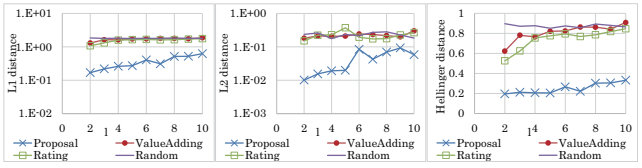
Following Xiao et al. [13], we consider that a query generated by data analyzers involves q' random sensitive QIDs. For example, when we used the Adult data set and set q'=3, $\{S'_1, S'_2, S'_3\}$ was a random three-sized subset of the 15 attributes.

We compared our proposed method with the ValueAdding method described in Section V-D1 and a baseline method called "Random," which creates a cross tabulation randomly.

[1] https://archive.ics.uci.edu/ml/datasets.html

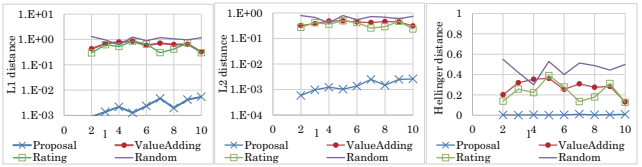


(a) L1 distance (b) L2 distance (c) Hellinger distance ($q' = 1$)

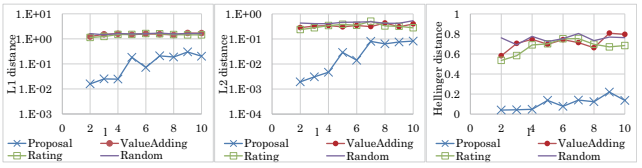


(d) L1 distance (e) L2 distance (f) Hellinger distance ($q' = 4$)

Fig. 1: Adult data set: $(l_1, \ldots, l_q)$-diversity.

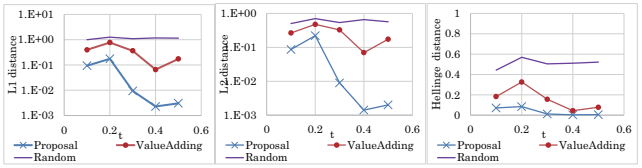


(a) L1 distance (b) L2 distance (c) Hellinger distance ($q' = 1$)

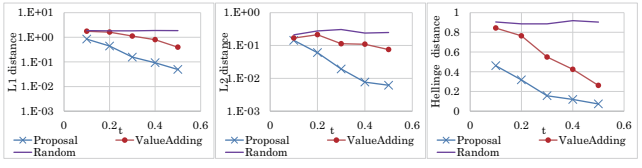


(d) L1 distance (e) L2 distance (f) Hellinger distance ($q' = 4$)

Fig. 2: US Census data set: $(l_1, \ldots, l_q)$-diversity.

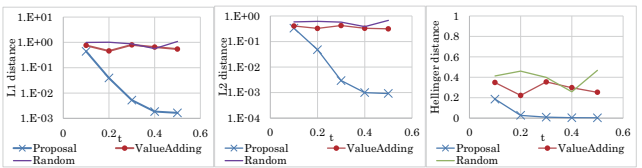


(a) L1 distance (b) L2 distance (c) Hellinger distance ($q' = 1$)

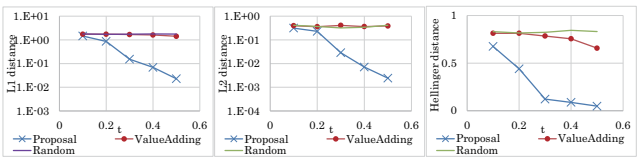


(d) L1 distance (e) L2 distance (f) Hellinger distance ($q' = 4$)

Fig. 3: Adult data set: $(t_1, \ldots, t_q)$-closeness.



(a) L1 distance (b) L2 distance (c) Hellinger distance ($q' = 1$)



(d) L1 distance (e) L2 distance (f) Hellinger distance ($q' = 4$)

Fig. 4: US Census data set: $(t_1, \ldots, t_q)$-closeness.

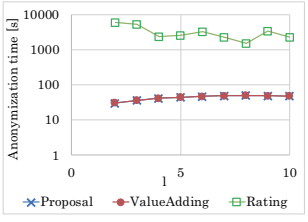


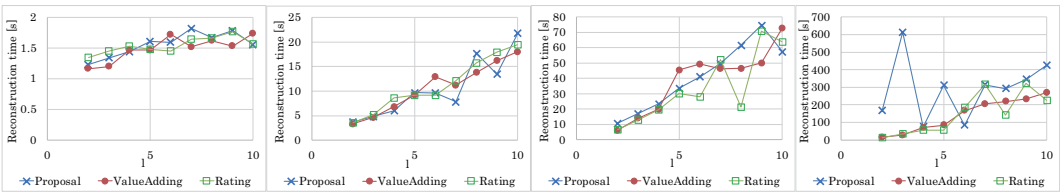
Fig. 5: Anonymization time for all records and all attributes.

Moreover, we evaluated the Rating method described in Section III in regard to $(l_1, \ldots, l_q)$-diversity. All experiments were conducted on a workstation with an Intel Xeon E5-2687W v2 CPU and 128 GB of RAM.

We set each l_j $(j = 1, \ldots, q)$ to one parameter l in regard to frequency $(l_1, \ldots, l_q)$-diversity and set each t_j $(j = 1, \ldots, q)$ to one parameter t in regard to $(t_1, \ldots, t_q)$-closeness. When d_j was less than or equal to l_j, we set l_j to $d_j - 1$. In each simulation, we generated random queries based on q' and calculated the L1 distance, L2 distance, and Hellinger distance, respectively.

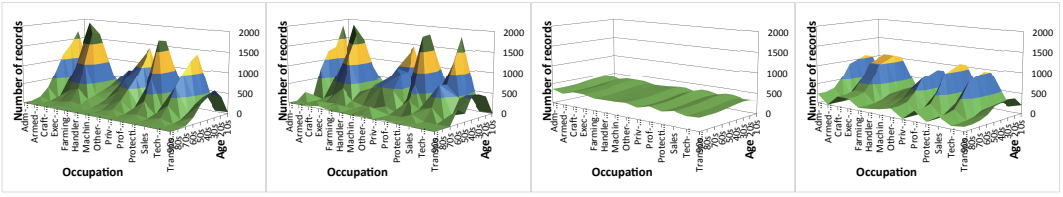
First, we conducted simulations to evaluate the L1 distance, L2 distance, and Hellinger distance for frequency $(l_1, \ldots, l_q)$-diversity. Figures 1 and 2 show the average results for frequency $(l_1, \ldots, l_q)$-diversity with the Adult data set and the US Census data set, respectively. We varied l, t, and q'. We know from the figures that our proposed method can reduce all three distances (L1, L2, and Hellinger) more so than the ValueAdding and Rating methods can. When the value of l was large, the L1, L2, and Hellinger distances increased when using our proposed method. However, the values were still lower than those of the ValueAdding and Rating methods.

Figures 3 and 4 show the average results for the L1, L2, and Hellinger distances with regard to $(t_1, \ldots, t_q)$-closeness for the Adult data set and the US Census data set, respectively. From these figures, we know that, with our proposed method, the L1, L2, and Hellinger distances decreased with increasing t. Nevertheless, we show that our proposed method can result in small distances for all three types (L1, L2, and Hellinger) even if we set t to a small value (i.e., 0.1) for both data sets.



(a) $q' = 1$ (b) $q' = 2$ (c) $q' = 3$ (d) $q' = 4$

Fig. 6: Reconstruction time with varying numbers of selected attributes.



(a) Original database (b) Result of Proposal (c) Result of ValudAdding (d) Result of Rating

Fig. 7: (a) Original Adult dataset (Occupation and Age); and (b-d) results of the anonymization and reconstruction.

The L1 and L2 distance results for ValueAdding, Rating, and Random were similar. However, note that the L1 and L2 distances for ValueAdding and Rating were lower than that for Random by about 50% when l was small or t was large.

For both the Rating method and the ValueAdding method, each tuple of the anonymized database had multiple values; however, the Rating method added values based on other records, whereas the ValueAdding method added values randomly. Moreover, the reconstruction algorithms were similar for both. We therefore consider that we have obtained similar results.

Figure 5 shows the average results of the anonymization time for the US Census data set. Our proposed method and the ValueAdding method needed less time than the Rating method. Because the anonymization algorithms of the proposed method and the ValueAdding method were the same, their anonymization times were almost the same. However, the Rating method needed more time to anonymize a database. This is because the proposed method and the ValueAdding method anonymized each record independently, whereas the Rating method anonymized each record while considering other values of the records. We know from Figure 5 that the proposed method took only several tens of seconds, even if the database has 2 million records and 68 attributes.

Figure 6 shows the average results of the reconstruction times with various l and q' for the US Census data set. Our proposed method needed more time for the reconstruction than the ValueAdding and Rating methods. If we use cross tabulation for more than three variables, the table will lose its major value [42]. In practice, the applications of cross tabulations involve only two variables at a time [41]. Therefore, we assume that the value of q' is from 1 to 2 or 3 in practice. Even if we set q' to 4, the reconstruction times of both data sets were less than 10 min on average.

Finally, we selected age and occupation as S' to show how the reconstructed histogram is similar to the original distribution. Figure 7 compares the original database of the Adult data set to the reconstructed results. We set l to 5. The sum of the reconstructed values (i.e., 45,222) was the same for all the methods. However, the reconstructed values of the ValueAdding and Rating methods were not any more precise than the original values. The reconstructed values were much smaller than the original values when the original one was larger than average, and vice versa. On the other hand, our proposed method could reconstruct the true distribution almost perfectly.

VII. DISCUSSION

A. *Analysis of the proposed anonymization algorithm*

The dominant approach to anonymize databases for l-diversity and t-closeness is based on a generalization. The generalization approach is easily understandable for data analyzers. Moreover, because the truthfulness of each record is preserved, we can obtain some information from each record.

However, the generalization approach causes high information loss when the number of attributes of the database is large because each equivalence class must have the same anonymized values for all attributes. Furthermore, if the data analyzer knows the sensitive attribute values of several records in an equivalence class, he/she has the chance to estimate the sensitive attribute values of the other records in the same equivalence class more precisely than expected.

In contrast, our approach, which adds randomized records to the original database, is relatively difficult to understand for data analyzers. Moreover, it is difficult to obtain meaningful information from each anonymized record.

However, our approach can reduce information loss, even when the number of attributes is large, because the randomization of attribute value is executed for each attribute

TABLE V: Anonymized Database in Setting $p_1 = p_2 = 1$ and $\eta_1 = d_1, \eta_2 = d_2$

Age	Disease
{1, 2, . . ., 99}	{Cancer, Cold, . . ., Chill}
{1, 2, . . ., 99}	{Cancer, Cold, . . ., Chill}

independently. Furthermore, we can ensure that, even if the data analyzer knows all the sensitive QID values of the whole database except for one record, the sensitive QID values of the record are protected.

B. Non-Sensitive QIDs, Non-QID Sensitive Attributes, and Non-QID Non-Sensitive Attributes

If the database contains non-sensitive QIDs, the data holders can set l_j to 1 for $(l_1, \ldots, l_q)$-diversity and can set t_j to ∞ for $(t_1, \ldots, t_q)$-closeness. These values are not protected, but they are considered as QIDs. In terms of non-QID sensitive attributes, the data holders can treat them as sensitive QIDs.

In our examples in this paper, we consider that age, address, job, and disease are sensitive QIDs. If the data holder decided that several attributes, such as age and job, can be treated as non-sensitive QIDs, the data analyzer can analyze the anonymized database with more precisely.

C. Discussion of $(t_1, \ldots, t_q)$-Closeness

In this paper, we used $(t_1, \ldots, t_q)$-closeness with consideration for the inference from the whole database, although we can consider $(t_1, \ldots, t_q)$-closeness without considering the inference from the whole database. This is because we consider that the former definition is more intuitive.

For example, assume that we set $p_1 = p_2 = 1$ and $\eta_1 = d_1, \eta_2 = d_2$ for anonymization. Table V represents the anonymization result. The attribute values have nothing to do with the record owners; in other words, each attribute is completely protected.

When we use $(t_1, \ldots, t_q)$-closeness with consideration for the inference from the whole database, Table V satisfies $(0, 0)$-closeness. Therefore, we can say that Table V completely protects the attribute values.

On the other hand, if we use $(t_1, \ldots, t_q)$-closeness *without* considering the inference from the whole database, we cannot say that Table V completely protects the attribute values, because in this case, Table V does not satisfy $(0, 0)$-closeness. This is counterintuitive because Table V does not leak any information about record owners in reality.

D. Extension of the Proposed Approach

We can extend our work to satisfy $(l_1, \ldots, l_q)$-diversity and $(t_1, \ldots, t_q)$-closeness *at the same time*. By letting the value of η in Algorithm 1 be changed not from 1 but from l_j, we can obtain an anonymized database that satisfies not only $(t_1, \ldots, t_q)$-closeness but also frequency $(l_1, \ldots, l_q)$-diversity and entropy $(l_1, \ldots, l_q)$-diversity. The proof is the same as the proofs of Theorem 2 and Theorem 3.

Moreover, we can define $(k_1, \ldots, k_q)$-anonymity in the same way as for $(l_1, \ldots, l_q)$-diversity.

Definition 17 ($[\![k, j]\!]$-Anonymity) *The anonymized database T^* satisfies the $[\![k, j]\!]$-anonymity if and only if every equivalence class for S_j has k or more records.*

Definition 18 ($(k_1, \ldots, k_q)$-Anonymity) *The anonymized database T^* satisfies $(k_1, \ldots, k_q)$-anonymity if and only if $[\![k, j]\!]$-anonymity is satisfied for all $1, \ldots, q$.*

It is obvious that, if an anonymized database satisfies frequency $(l_1, \ldots, l_q)$-diversity, the database also satisfies $(k_1, \ldots, k_q)$-anonymity.

There are many concepts with regard to privacy models. The application of the proposed method to other privacy models is the subject of our future work.

VIII. CONCLUSION

The models of l-diversity and t-closeness have been widely studied for protecting privacy. Most existing studies assume that they can separate QIDs from sensitive attributes, but we cannot always make such assumptions in real-world situations. Consequently, we can assume that several attributes have features of both sensitive attributes and QIDs in this paper. We proposed novel privacy models, namely, $(l_1, \ldots, l_q)$-diversity and $(t_1, \ldots, t_q)$-closeness, and algorithms of anonymization and reconstruction that can treat sensitive QIDs. Through simulations of real data sets, we have proven that our proposed method can anonymize and reconstruct databases while keeping a high quality of data within a realistic period.

REFERENCES

[1] D. Sánchez, S. Martínez, and J. Domingo-Ferrer, "Comment on "Unique in the shopping mall: On the reidentifiability of credit card metadata"," *Science*, vol. 351, no. 6279, p. 1274, 2016.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, pp. 1–53, 2010.

[3] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," in *Proc. IEEE ICDE*, 2006, p. 25.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM TKDD*, vol. 1, no. 1, p. 3, 2007.

[5] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE ICDE*, 2007, pp. 106–115.

[6] D. Riboni, L. Pareschi, and C. Bettini, "JS-Reduce: Defending Your Data from Sequential Background Knowledge Attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 3, pp. 387–400, 2012.

[7] W. M. Liu, L. Wang, P. Cheng, K. Ren, S. Zhu, and M. Debbabi, "PPTP: Privacy-Preserving Traffic Padding in Web-Based Applications," *IEEE Transactions on Dependable and Secure Computing*, vol. 11, no. 6, pp. 538–552, 2014.

[8] A. Konstantinidis, G. Chatzimilioudis, D. Zeinalipour-Yazti, P. Mpeis, N. Pelekis, and Y. Theodoridis, "Privacy-Preserving Indoor Localization on Smartphones," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3042–3055, 2015.

[9] L. Amsaleg, A. Morton, and S. Marchand-Maillet, "A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 1, pp. 152–167, 2015.

[10] H. Zakerzadeh, C. C. Aggarwal, and K. Barker, "Managing dimensionality in data privacy anonymization," *Knowledge and Information Systems*, pp. 1–33, 2015.

[11] X. Jin, M. Zhang, N. Zhang, and G. Das, "Versatile publishing for privacy preservation," in *Proc. ACM KDD*, 2010, pp. 353–362.

[12] L. Guo, S. Guo, and X. Wu, "Privacy Preserving Market Basket Data Analysis," in *Proc. PKDD*, 2007, pp. 103–114.

[13] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preservation," in *Proc. VLDB*, sep 2006, pp. 139–150.

[14] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based k-anonymity," *The VLDB Journal*, vol. 23, no. 5, pp. 771–794, 2014.

[15] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[16] G. Acs, C. Castelluccia, and R. Chen, "Differentially Private Histogram Publishing through Lossy Compression," in *Proc. IEEE ICDM*, 2012, pp. 1–10.

[17] R. Chen, Q. Xiao, Y. Zhang, and J. Xu, "Differentially Private High-Dimensional Data Publication via Sampling-Based Inference," in *Proc. ACM KDD*, 2015, pp. 129–138.

[18] W. Qardaji, W. Yang, and N. Li, "PriView: practical differentially private release of marginal contingency tables," in *Proc. ACM SIGMOD*, 2014, pp. 1435–1446.

[19] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "Improving the Utility of Differentially Private Data Releases via k-Anonymity," in *Proc. IEEE TrustCom*, 2013, pp. 372–379.

[20] X. Xiao, G. Wang, and J. Gehrke, "Differential privacy via wavelet transforms," pp. 225–236, 2010.

[21] J. Xu, Z. Zhang, X. Xiao, Y. Yang, G. Yu, and M. Winslett, "Differentially private histogram publication," *Proc. VLDB*, vol. 22, no. 6, pp. 797–822, 2013.

[22] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: private data release via bayesian networks," in *Proc. ACM SIGMOD*, 2014, pp. 1423–1434.

[23] W. K. Härdle and Z. Hlávka, "Cluster Analysis," in *Multivariate Statistics*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 225–244.

[24] N. Shlomo, "Statistical Disclosure Limitation for Health Data: A Statistical Agency Perspective," in *Medical Data Privacy Handbook*. Cham: Springer International Publishing, 2015, pp. 201–230.

[25] H.-H. Chang, M.-C. Lee, W.-J. Lee, C.-L. Chien, and N. Chen, "Feature Extraction-Based Hellinger Distance Algorithm for Non-intrusive Aging Load Identification in Residential Buildings," *IEEE Transactions on Industry Applications*, vol. 52, no. 3, pp. 2031–2039, 2016.

[26] P. Mittal, C. Papamanthou, and D. Song, "Preserving Link Privacy in Social Network Based Systems," in *Proc. NDSS*, 2013.

[27] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "t-Closeness through Microaggregation: Strict Privacy with Enhanced Utility Preservation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 11, pp. 3098–3110, 2015.

[28] P. Shi, L. Xiong, and B. Fung, "Anonymizing data with quasi-sensitive attribute values," in *Proc. ACM CIKM*, 2010, pp. 1389–1392.

[29] M. Terrovitis, N. Mamoulis, J. Liagouris, and S. Skiadopoulos, "Privacy preservation by disassociation," *Proc. VLDB*, vol. 5, no. 10, pp. 944–955, 2012.

[30] J. Soria-Comas and J. Domingo-Ferrer, "Probabilistic k-anonymity through microaggregation and data swapping," in *Proc. IEEE International Conference on Fuzzy Systems*, jun 2012, pp. 1–8.

[31] K. Wang, Y. Xu, A. W. C. Fu, and R. C. W. Wong, "FF-Anonymity: When Quasi-identifiers Are Missing," in *Proc. IEEE ICDE*, 2009, pp. 1136–1139.

[32] J. Liu, J. Luo, and J. Z. Huang, "Rating: Privacy Preservation for Multiple Attributes with Different Sensitivity Requirements," in *Proc. IEEE ICDM Workshops*, 2011, pp. 666–673.

[33] Y. Ye, Y. Liu, C. Wang, D. Lv, and J. Feng, "Decomposition: Privacy Preservation for Multiple Sensitive Attributes," in *Proc. DASFAA*, 2009, pp. 486–490.

[34] J. Cao and P. Karras, "Publishing microdata with a robust privacy guarantee," *Proc. VLDB*, vol. 5, no. 11, pp. 1388–1399, 2012.

[35] C. Dwork, "Differential Privacy," in *Proc. ICALP*, 2006, pp. 1–12.

[36] D. Sánchez, J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas, "Utility-preserving differentially private data releases via individual ranking microaggregation," *Information Fusion*, vol. 30, pp. 1–14, 2016.

[37] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proc. VLDB*, aug 2002, pp. 682–693.

[38] Y. Sei and A. Ohsuga, "Privacy Preservation for Participatory Sensing Application," in *Proc. 30th IEEE AINA*, 2016.

[39] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," in *Proc. ACM SIGMOD*, 2000, pp. 439–450.

[40] R. Agrawal, R. Srikant, and D. Thomas, "Privacy preserving OLAP," in *Proc. ACM SIGMOD*, 2005, pp. 251–262.

[41] R. Grover and M. Vriens, *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*. SAGE Publications, Inc, 2006.

[42] M. J. Wood and J. Ross-Kerr, *Basic steps in planning nursing research: From question to proposal*. Jones & Bartlett Publishers, 2010.

Yuichi Sei (M'17) received the Ph.D. degree in information science and technology from the University of Tokyo in 2009. From 2009 to 2012, he was with the Mitsubishi Research Institute. Since 2013, he has been an Assistant Professor with the University of Electro-Communications. His current research interests include pervasive computing, privacy-preserving data mining, and software engineering. He was a recipient of the IPSJ Best Paper Award in 2017.

Hiroshi Okumura received a Ph.D. degree in Economics from the Kobe University in 2012. He is working at Mitsubishi Research Institute. His research interests include statistics, econometrics, and statistical machine learning. He is a member of Japan Statistical Society and Information Processing Society of Japan (IPSJ).

Takao Takenouchi received a Ph.D. degree in engineering from the University of Electro-Communications in 2013. He is working at NEC Corporation. His research interests include privacy-preserving data mining and information security. He is a member of Information Processing Society of Japan (IPSJ) and Institute of Electronics, Information and Communication Engineers (IEICE).

Akihiko Ohsuga (M'11) received his Ph.D. degree in computer science from Waseda University in 1995. From 1981 to 2007 he was with Toshiba Corporation. He joined the University of Electro-Communications in 2007, and is currently a professor in the Graduate School of Informatics and Engineering, and is also a dean of the Graduate School of Information Systems. He is also a visiting professor at National Institute of Informatics. His research interests include agent technologies, web intelligence, and software engineering. He is a member of IEEE Computer Society (IEEE CS), Information Processing Society of Japan (IPSJ), Institute of Electronics, Information and Communication Engineers (IEICE), Japanese Society for Artificial Intelligence (JSAI), Japan Society for Software Science and Technology (JSSST), and Institute of Electrical Engineers of Japan (IEEJ). He was a chair of IEEE CS Japan Chapter. He was a member of the board of directors of JSAI and JSSST. He received the IPSJ Best Paper Awards in 1987 and 2017.