TRANSACTIONS ON DATA PRIVACY 16 (2023) 191–210

Data Sanitization for t-Closeness over Multiple Numerical Sensitive Attributes

Rajiv Bagai*, **Eric Weber****, **Vikas Thammanna Gowda***

*School of Computing, Wichita State University, Wichita, KS 67260-0083, USA.

**NetApp Inc., Wichita, KS 67226, USA.

E-mail: rajiv.bagai@wichita.edu

Received 1 June 2022; received in revised form 19 January 2023; accepted 25 February 2023

Abstract. A popular technique for preserving privacy of individuals contained in any released data is to first sanitize the data according to the t-closeness principle. This principle requires partitioning rows of the original data into equivalence classes, in a way that the distribution of sensitive values in any class is sufficiently close, within a given threshold t, to their distribution in the original data. Most existing methods for constructing t-close equivalence classes consider just one sensitive attribute in the data, which is insufficient as many real-life datasets contain multiple sensitive attributes; partitioning attempts for multiple sensitive attributes have thus far been unsatisfactory. We present a method for generating t-close equivalence classes in the presence of multiple numerical sensitive attributes, where each such attribute has its own privacy threshold. The equivalence classes are generated in a way that minimizes information loss caused later by generalizing quasi identifier values within each class. While finding an optimal solution for this problem is known to be NP-hard, we show that our approach results in an acceptable solution in polynomial time.

Keywords. Data Privacy, Anonymity, Data Publishing, t-Closeness

1 Introduction

The increasing sophistication of available data mining techniques beckons many businesses to get their data sets mined, for extracting useful hidden information, by releasing their data to third-party miners. Certain organizations, such as licensed medical centers and hospitals, also have a need for a release of data in order to simply comply with governmental regulations. The data to be released often contains uniquely identifying as well as sensitive information about individuals. In order to maintain privacy of those individuals, each individual's identifying characteristics need to be dissociated, in the released data, from their sensitive information. This dissociation is achieved by adequately sanitizing the data prior to its release.

Consider a typical relational database table, with rows and columns, that needs to be publicly released by its owner, for reasons such as third-party analysis or compliance with legal mandates. The rows in the table may correspond to individuals, such as patients in a hospital or customers of a retail store. The columns are attributes of those individuals, of which there are three relevant kinds, namely explicitly identifying, quasi identifiers, and

sensitive attributes. In order to publicly release the table, while still maintaining privacy of its individuals, it is standard practice to first remove all its explicitly identifying columns, like Name and Social Security Number. Removal of such columns is necessary to achieve individual privacy, but not sufficient.

While any quasi identifier, like Age or Race, by itself does not uniquely identify an individual, their combinations can come very close to that. In a well-known study, Sweeney [15] showed that over 87% of the U.S. population is uniquely identifiable from just their Gender, Birthdate, and Zip Code of residence. Although a more recent study by Golle [4] corrected that figure to 61-63%, that is still a very high proportion. Removal of such columns is not desirable, because that renders the data useless for mining purposes, thereby defeating the sole purpose of its release. The table, therefore, needs to be sanitized in a different way.

A widely employed sanitization approach is to first horizontally partition the table into groups of rows, called *equivalence classes*, and then generalize quasi identifiers of each row just enough to ensure that the generalized quasi identifier values of all rows within an equivalence class become identical. This way, even if the combination of an individual's quasi identifier values is in fact unique in the entire raw table, and an adversary knows those values, they can at best only place that individual in its equivalence class, without being able to further pinpoint their row in the publicly released sanitized table, with an aim to unearth their exact sensitive attribute values, like Annual Salary or Blood Pressure.

Different sanitization techniques based on this approach vary essentially in the characteristics of the constructed equivalence classes. For example, the k-anonymity technique of Sweeney [16] requires each equivalence class to be of at least a certain size, and the l-diversity technique of Machanavajjhala et al. [8] requires a sensitive attribute column of all rows in any equivalence class to have at least a certain number of "well-represented" values. Our focus in this paper is on the popular t-closeness technique of Li et al. [6], which was demonstrated by the authors to be an improvement over the aforementioned ones. This technique requires the probability distribution of all sensitive attribute values in each equivalence class to be sufficiently close, i.e. within a given privacy budget parameter t, to that in the entire raw table, thereby limiting probabilistic inference of any individual's sensitive attribute values, despite that individual's equivalence class being apparent.

1.1 Related Work

Although Li et al. [6] proposed this desirable property of equivalence classes, they did not give any method to construct such classes. Soria-Comas et al. [14] developed a method for the small class of tables in which no two distinct rows share the same sensitive attribute value. While an earlier method of Cao et al. [2] does not suffer from this severe restriction, their method still works only for tables with just one sensitive attribute.

It has long been noted that a method for achieving t-closeness over multiple sensitive attributes is needed, because many real-life applications contain multiple such attributes. As an example, even an ordinary blood test contains numerous sensitive readings, like LDL and HDL cholesterol levels, hemoglobin count, calcium, potassium, and sodium levels, etc. To date, there have been very few attempts at developing methods for creating equivalence classes that satisfy t-closeness, in the presence of multiple such attributes.

Fang et al. [3] employ one privacy budget parameter t for all sensitive attributes. Subjecting all sensitive attributes to the same t is unrealistic, because in real-life, different sensitive attributes are often sensitive to a different degree, requiring a different privacy parameter for each. As an example, people are more concerned about keeping their medical condition private, than their salary. Moreover, their method partitions the set of all sensitive

attributes into mutually disjoint subsets, and a separate t-close sub-table is published for each such subset. This loses important associations among sensitive attributes that belong to different published sub-tables.

The method of Wang et al. [17] also employs just one privacy budget parameter t, for a composite sensitive value obtained by performing Principle Component Analysis, a method developed in 1901 by Pearson [10], on all sensitive attribute values. Again, all sensitive attributes are unrealistically subjected to just one privacy parameter, and reducing multiple sensitive attribute values to a single composite one introduces inaccuracies in the t-closeness requirement over the constructed equivalence classes.

Recently, Sei et al. [12] developed a method for multiple attributes that simultaneously behave as quasi identifiers as well as sensitive attributes. However, their method involves insertion of some random rows to the original table, and modification of some existing ones, thereby adversely affecting mining utility.

1.2 Our Contribution and Paper Organization

In this paper, we develop a method for partitioning rows of a given relational table that contains multiple numerical sensitive attributes, each with its own given privacy budget parameter, into equivalence classes that satisfy t-closeness with respect to all given privacy parameters. Our method is based upon fragmenting the multi-dimensional space of all sensitive attribute values in such a way that even a random dispersion of a predetermined number of rows from each created fragment to equivalence classes results in formation of acceptable classes. We then exploit the flexibility provided by the random choices made available to us to lower the information loss incurred later due to generalizing the quasi identifier values in each class. The resulting sanitized table thereby possesses higher utility for mining.

The rest of this paper is organized as follows. Section 2 presents our mathematical framework and notations used later. Some important notions presented in this section are representing equivalence classes as matrices, the earth mover's distance, left-heavy probability distributions, the t-closeness principle, and the complete lattice of all fragmentations of the multi-dimensional sensitive attribute value space. Section 3 develops the two main phases of our method, namely of finding a desirable fragmentation of the sensitive attribute value space, and generating equivalence classes with low information loss. It also contains a complexity analysis of these phases, and shows that they can be performed in polynomial time. In most of the paper, our method is developed and explained for tables with exactly two sensitive attributes. However, our method can be generalized in a straightforward way for tables containing any arbitrary number of sensitive attributes. This section also outlines how that generalization can be carried out. Finally, Section 4 concludes our work and suggests some directions for future work.

2 Mathematical Preliminaries

In this section, we develop the necessary mathematical framework and notations used in the rest of the paper. Even though our method is for any number of sensitive attributes in a table, for ease of understanding, we develop it for tables with exactly two sensitive attributes. Generalization of our method for a larger number of sensitive attributes is straightforward, and outlined in Section 3.4.

2.1 Multiplicity Matrix and Equivalence Class Matrices

Let $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ be the *domains* of all values that may appear in the two sensitive attribute columns of the given table. We assume indices are assigned in ascending order of values in these domains, i.e. $x_1 < x_2 < \cdots < x_m$ and $y_1 < y_2 < \cdots < y_n$. Note that not every combination of these values necessarily appears in the table, while some combinations may appear in more than one row.

Let $\mathcal{M}$ be the $m \times n$ *multiplicity matrix* of values in $X \times Y$, i.e. for any i and j, $\mathcal{M}[i, j]$ is the number of rows of the given table the sensitive value pair (x_i, y_j) appears in. We will be primarily interested in the dispersion of these value pairs among the equivalence classes constructed by our method. To that end, we define a possible *equivalence class* $\mathcal{E}$ to be any $m \times n$ matrix, all whose values are within corresponding values of $\mathcal{M}$, i.e. for any i and j, $0 \leq \mathcal{E}[i, j] \leq \mathcal{M}[i, j]$. As expected, $\mathcal{M}$ itself is a possible equivalence class.

For any $p \times q$ matrix $\mathcal{A}$, let $|\mathcal{A}|$ denote the sum of all its values, i.e. $|\mathcal{A}| = \sum_{i=1}^{p} \sum_{j=1}^{q} \mathcal{A}[i, j]$. If the given table contains T rows, then clearly, $|\mathcal{M}| = T$, and for any equivalence class $\mathcal{E}$, we have that $0 \leq |\mathcal{E}| \leq T$.

In order to be able to work with probability distributions, we let $\overline{\mathcal{A}}$ denote the normalized version of any matrix $\mathcal{A}$, i.e. for any i and j, $\overline{\mathcal{A}}[i, j] = \mathcal{A}[i, j]/|\mathcal{A}|$. Clearly, $|\overline{\mathcal{A}}| = 1$, for any $\mathcal{A}$.

Row- and column-sums of matrices, as defined below, will be useful.

Definition 1 (Row- and Column-Sums)**.** Let $\mathcal{A}$ be any $p \times q$ matrix. Then $\mathcal{R}_{\mathcal{A}}$ denotes the p-vector of the *row-sums* of $\mathcal{A}$, given by:

$$\mathcal{R}_{\mathcal{A}}[i] = \sum_{j=1}^{q} \mathcal{A}[i, j], \text{ for all } 1 \leq i \leq p.$$

Also, $\mathcal{C}_{\mathcal{A}}$ denotes the q-vector of the *column-sums* of $\mathcal{A}$, given by:

$$\mathcal{C}_{\mathcal{A}}[j] = \sum_{i=1}^{p} \mathcal{A}[i, j], \text{ for all } 1 \leq j \leq q.$$

$\mathcal{R}_{\mathcal{M}}[i]$, for example, is thus the *number* of rows of the given table that contain the value x_i, and $\mathcal{R}_{\overline{\mathcal{M}}}[i]$ is the *fraction* of rows containing that value. Other vectors, like $\mathcal{C}_{\mathcal{M}}$, $\mathcal{C}_{\overline{\mathcal{M}}}$, $\mathcal{R}_{\mathcal{E}}$, $\mathcal{R}_{\overline{\mathcal{E}}}$, etc. have similar intuitive meanings.

Figure 1 shows an example multiplicity matrix $\mathcal{M}$ for a table with 300 rows, $m = 4$, and $n = 3$. Also shown in the figure is a sample equivalence class $\mathcal{E}$ with 20 rows, and the normalized versions of these matrices, namely $\overline{\mathcal{M}}$ and $\overline{\mathcal{E}}$. Vectors $\mathcal{R}_{\overline{\mathcal{M}}}$ and $\mathcal{C}_{\overline{\mathcal{M}}}$ shown in the figure are the probability distributions of sensitive values in X and Y, respectively, in the entire table. Similarly, $\mathcal{R}_{\overline{\mathcal{E}}}$ and $\mathcal{C}_{\overline{\mathcal{E}}}$ are their respective probability distributions in the equivalence class $\mathcal{E}$.

The t-closeness principle, to be defined precisely later in the paper, essentially requires $\mathcal{R}_{\overline{\mathcal{E}}}$, for any equivalence class $\mathcal{E}$, to be a close approximation of $\mathcal{R}_{\overline{\mathcal{M}}}$, and $\mathcal{C}_{\overline{\mathcal{E}}}$ to be a close approximation of $\mathcal{C}_{\overline{\mathcal{M}}}$.

2.2 The Earth Mover's Distance

A popular measure of closeness between probability distributions over some totally ordered domain, such as the numerical values of a sensitive attribute in our context, is the

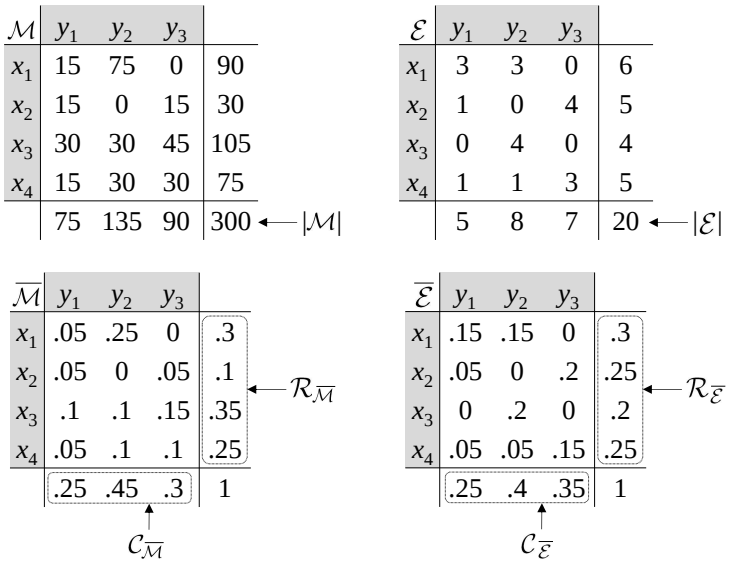

Figure 1: Examples of a multiplicity matrix and an equivalence class, along with their normalized matrices, and their row- and column-sums vectors.

earth mover's distance between them. This metric was originally proposed back in 1781 by Monge [9], but popularized recently by Rubner et al. [11].

For any integer $k > 0$ and real $\alpha > 0$, let $\mathcal{S}^{k,\alpha}$ be the set of all k-vectors of non-negative real values with sum α, i.e.:

$$\mathcal{S}^{k,\alpha} = \{\langle p_1, p_2, \ldots, p_k\rangle : \text{ each } p_i \geq 0, \text{ and } \sum_{i=1}^{k} p_i = \alpha\}.$$

Definition 2 (Earth Mover's Distance)**.** For any $P, Q \in \mathcal{S}^{k,\alpha}$, where $P = \langle p_1, p_2, \ldots, p_k\rangle$ and $Q = \langle q_1, q_2, \ldots, q_k\rangle$, the *earth mover's distance* between P and Q, denoted $\delta(P, Q)$, is given by:

$$\delta(P,Q) = \begin{cases} 0 & \text{if } k = 1, \\ \frac{1}{k-1} \sum_{i=1}^{k-1} \left| \sum_{j=1}^{i} (p_j - q_j) \right| & \text{otherwise.} \end{cases}$$

The earth mover's distance can be thought of as the sum total of the portions of the p_i values that need to be moved to other indices in P, each portion scaled by the normalized distance of its movement within the k-tuple, to turn P into Q. Its formula can be appreciated by this simple iterative algorithm:

> Start with p_1. If $p_1 \geq q_1$, move the surplus amount, $(p_1 - q_1)$, from p_1 to p_2. Otherwise, move the deficit, $(q_1 - p_1)$, from p_2 to p_1. In either case, the amount moved, normalized by the maximum distance, is $|(p_1 - q_1)|/(k-1)$, and now $p_1 = q_1$. Moving to p_2, whose value has now become $(p_1 - q_1) + p_2$, would add $|(p_1 - q_1) + (p_2 - q_2)|/(k-1)$ to the $\delta(P, Q)$ being computed. The iteration ends upon making $p_{k-1} = q_{k-1}$ when, due to the original vector sums being identical, P has been turned into Q.

As an example, consider probability distributions $P = \langle 0.2, 0.1, 0.7\rangle$, $Q = \langle 0.3, 0, 0.7\rangle$, and $R = \langle 0.1, 0, 0.9\rangle$. Then, $\delta(P,Q) = 0.1(1/2) = 0.05$, because in order to turn P into Q, 0.1 amount needs to be moved from p_2 to p_1, which is 1 index away, out of a maximum of 2 (as $k-1=2$ is the farthest movement distance in this tuple). Similarly, $\delta(Q,R) = 0.2(2/2) = 0.2$, and $\delta(P,R) = 0.1(2/2) + 0.1(1/2) = 0.15$. It is worth noting that scaling by the movement distance is necessitated by the total order that exists on the underlying set, as is the case with the numerical domains of our sensitive attribute values. The following proposition is immediate.

Proposition 3. *For any $P, Q \in \mathcal{S}^{k,\alpha}$, $\delta(P,Q) \leq \alpha$.*

Two more properties of the earth mover's distance will also be useful in the development of our method. Before presenting each, we refresh some standard notation on vectors.

Definition 4 (Vector Sum)**.** For any $P = \langle p_1, p_2, \ldots, p_k\rangle \in \mathcal{S}^{k,\alpha}$ and $Q = \langle q_1, q_2, \ldots, q_k\rangle \in \mathcal{S}^{k,\beta}$, the *sum* of P and Q, denoted $P+Q$, is the k-vector:

$$\langle p_1+q_1, p_2+q_2, \ldots, p_k+q_k\rangle \in \mathcal{S}^{k,\alpha+\beta}.$$

Proposition 5. *For any $P_1, Q_1 \in \mathcal{S}^{k,\alpha}$ and $P_2, Q_2 \in \mathcal{S}^{k,\beta}$, $\delta(P_1+P_2, Q_1+Q_2) \leq \delta(P_1,Q_1) + \delta(P_2,Q_2)$.*

Proof. One way to turn P_1+P_2 into Q_1+Q_2 is to first move the P_1 portions (of P_1+P_2) into Q_1 portions (of Q_1+Q_2), with cost $\delta(P_1,Q_1)$, and then move the remaining P_2 portions (of P_1+P_2) into the remaining Q_2 portions (of Q_1+Q_2), with additional cost $\delta(P_2,Q_2)$. While there may be other ways with lower costs, $\delta(P_1,Q_1)+\delta(P_2,Q_2)$ is clearly an upper bound of $\delta(P_1+P_2, Q_1+Q_2)$. □

Definition 6 (Vector Concatenation)**.** For any vectors $P = \langle p_1, p_2, \ldots, p_k\rangle \in \mathcal{S}^{k,\alpha}$ and $Q = \langle q_1, q_2, \ldots, q_l\rangle \in \mathcal{S}^{l,\beta}$, the *concatenation* of P and Q, denoted PQ, is the $(k+l)$-vector:

$$\langle p_1, p_2, \ldots, p_k, q_1, q_2, \ldots, q_l\rangle \in \mathcal{S}^{k+l,\alpha+\beta}.$$

Proposition 7. *For any $P_1, Q_1 \in \mathcal{S}^{k,\alpha}$ and $P_2, Q_2 \in \mathcal{S}^{l,\beta}$, $\delta(P_1P_2, Q_1Q_2) \leq \delta(P_1,Q_1) + \delta(P_2,Q_2)$.*

Proof. Let $\mathbf{0}^j$ denote the j-vector of all zero values. Clearly, $P_1P_2 = P_1\mathbf{0}^l + \mathbf{0}^kP_2$, and $Q_1Q_2 = Q_1\mathbf{0}^l + \mathbf{0}^kQ_2$. The result follows from Proposition 5, and that δ is invariant if both its arguments are extended, on the same side, by equal-length vectors with only zero values. □

2.3 The Simplex of Vectors

As illustrated in Figure 2, the space $\mathcal{S}^{k,\alpha}$ is a simplex, i.e. the polytope forming the convex hull of the k vertices of $\mathcal{S}^{k,\alpha}$, namely $V^{(1)} = \langle \alpha, 0, \ldots, 0\rangle$, $V^{(2)} = \langle 0, \alpha, \ldots, 0\rangle$, …, $V^{(k)} = \langle 0, 0, \ldots, \alpha\rangle$. Each element of $\mathcal{S}^{k,\alpha}$ is a unique convex combination of these vertices. It is also well-known that $\mathcal{S}^{k,\alpha}$ forms a metric space under δ. Rubner et al. [11] contains a proof for this, under the assumption that the ground distance between domain values at indices i and j is a metric. In our context, the ground distance is $|i-j|/(k-1)$, which is clearly a metric, as also stated in Li et al. [6].

Two of the k vertices of this simplex, namely $V^{(k)}$ and $V^{(1)}$, are especially important to our method, for the classes of vectors that are left-heavy and otherwise, respectively, as defined below.

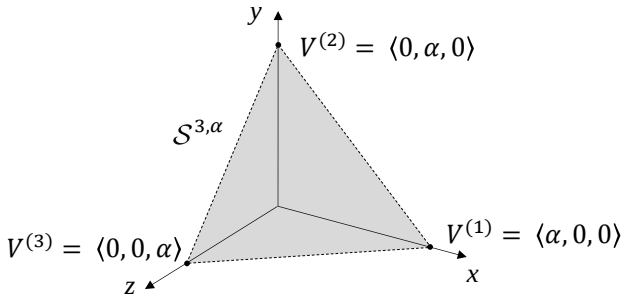

Figure 2: The simplex $\mathcal{S}^{3,\alpha}$.

Definition 8 (Center of Gravity and Left-Heavy)**.** For any $P = \langle p_1, p_2, \ldots, p_k \rangle \in \mathcal{S}^{k,\alpha}$, the *center of gravity* of P, denoted g_P, is a real value between 1 and k, such that:

$$\sum_{i<g_P} p_i(g_P - i) = \sum_{i \geq g_P} p_i(i - g_P).$$

Moreover, P is called *left-heavy* if $g_P < \frac{k+1}{2}$.

It should be noted that g_P is usually not an integer value. It is that point, somewhere between the extreme index values 1 and k, about which the sum of torques generated by all p_i values vanishes. And $\frac{k+1}{2}$, simply the mid-point between 1 and k, is also not always an integer value.

Our method is based upon a crucial observation that if P is left-heavy, then among all vectors in $\mathcal{S}^{k,\alpha}$, $V^{(k)}$ is at the farthest earth mover's distance from P. Otherwise, $V^{(1)}$ is the farthest from P. In order to prove this, we first establish the following lemma.

Lemma 9. *Let $P = \langle p_1, p_2, \ldots, p_k \rangle \in \mathcal{S}^{k,\alpha}$ be left-heavy. Suppose for some $Q \in \mathcal{S}^{k,\alpha}$, there is an index $v \geq \frac{k+1}{2}$, such that some non-zero portion λ of p_v moves backward, during the computation of $\delta(P, Q)$ to fulfill a deficit at some lower index. Then there exist indices $u_1, u_2, \ldots, u_s$, where each $u_i < g_P$, such that the combined cost contribution of λ and certain non-zero portions of all p_{u_i} to $\delta(P, Q)$ is no more than their combined contribution to $\delta(P, V^{(k)})$.*

Proof. Let v and λ be as stated. Since g_P is the center of gravity, and $g_P < v$, there must exist indices $u_1, u_2, \ldots, u_s$, where each $u_i < g_P$, as well as non-zero portions θ_i of each p_{u_i}, such that:

$$\sum_{i=1}^{s} \theta_i(g_P - u_i) = \lambda(v - g_P). \tag{1}$$

In other words, all θ_i collectively balance λ about g_P. For each i, let θ_i^- be that (possibly zero) portion of θ_i, which moves backward during the computation of $\delta(P, Q)$ to fulfill a deficit at some lower index. And let λ^- be their balancing counterpart in λ, i.e. $\sum \theta_i^-(g_P - u_i) = \lambda^-(v - g_P)$. Finally, let $\theta_i^+ = \theta_i - \theta_i^-$, and $\lambda^+ = \lambda - \lambda^-$. Clearly, we also have that $\sum \theta_i^+(g_P - u_i) = \lambda^+(v - g_P)$. Figure 3 depicts these values in a histogram view of P.

We now prove the lemma in two parts. First, consider the combined maximum cost contribution of λ^- and all θ_i^- to $\delta(P, Q)$, i.e. $\lambda^-(v-1) + \sum \theta_i^-(u_i - 1)$. Due to the relationship between λ^- and all θ_i^-, similar to Equation (1), this expression can be simplified to $\lambda^-(g_P - 1) + \sum \theta_i^-(g_P - 1)$, which is bounded above by $\lambda^-(k - v) + \lambda^-(v - g_P) + \sum \theta_i^-(g_P - 1)$,

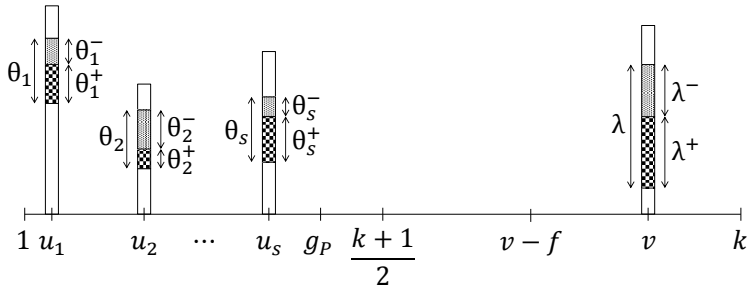

Figure 3: A histogram view of P.

since $g_P - 1 \leq k - g_P$, due to P being left-heavy. This bound simplifies to $\lambda^-(k-v) + \sum \theta_i^-(2g_P - u_i - 1)$. As $2g_P - 1 < k$, this in turn is at most $\lambda^-(k-v) + \sum \theta_i^-(k-u_i)$, which is their combined contribution to $\delta(P, V^{(k)})$.

Now consider the combined maximum contribution of λ^+ and all θ_i^+ to $\delta(P,Q)$. As shown in Figure 3, if λ moves from index v to $v-f$, for some $f \geq 0$, then $v-f$ must be the rightmost destination of any θ_i^+. The combined contribution of λ^+ and all θ_i^+ is thus at most $\lambda^+ f + \sum \theta_i^+(v - f - u_i)$. As before, due to the relationship between λ^+ and all θ_i^+, similar to Equation (1), this expression can be simplified to $\sum \theta_i^+[\frac{2g_P - u_i - v}{v - g_P} f + v - u_i]$. As $2g_P < k$, due to P being left-heavy, this value is under $\sum \theta_i^+[\frac{k-v}{v-g_P} f + v - u_i]$, which since $f \leq v - u_i$, for each i, is in turn bounded above by $\sum \theta_i^+[\frac{k-v}{v-g_P}(v-u_i) + v - u_i]$. Further algebraic simplification reduces this to $\lambda^+(k-v) + \sum \theta_i^+(k-u_i)$, which is their combined contribution to $\delta(P, V^{(k)})$. □

We now establish a key result.

Theorem 10. *For any $P, Q \in \mathcal{S}^{k,\alpha}$,*

(1) if P is left-heavy, then $\delta(P,Q) \leq \delta(P, V^{(k)})$; and

(2) otherwise, $\delta(P,Q) \leq \delta(P, V^{(1)})$.

Proof. We prove here only (1), as the proof of (2) is symmetric. Suppose P is left-heavy. If there is any index $v \geq \frac{k+1}{2}$, such that some non-zero portion λ of p_v moves backward, during the computation of $\delta(P,Q)$ to fulfill a deficit at some lower index, then by Lemma 9, there exist indices $u_1, u_2, \ldots, u_s$, such that the combined cost contribution of λ and certain non-zero portions of all p_{u_i} to $\delta(P,Q)$ is no more than their combined contribution to $\delta(P, V^{(k)})$. We first remove from P all these non-zero portions of p_v and all p_{u_i}. We then repeat this removal process for each such index v, until no more are left. Let P' denote the resulting vector.

Any backward moving portion of any particular value of P' must be at an index smaller than the midpoint $\frac{k+1}{2}$, thus contributing no more cost to $\delta(P,Q)$ than it does to $\delta(P, V^{(k)})$. Moreover, any forward moving portion of any particular value of P' could have moved no more for $\delta(P,Q)$ than it does for $\delta(P, V^{(k)})$. □

In light of the above result, we let $\Delta : \mathcal{S}^{k,\alpha} \to [0, \alpha]$ give, for any $P \in \mathcal{S}^{k,\alpha}$, the maximum earth mover's distance from P to any vertex of $\mathcal{S}^{k,\alpha}$, i.e.:

$$\Delta(P) = \begin{cases} \delta(P, V^{(k)}) & \text{if } P \text{ is left-heavy,} \\ \delta(P, V^{(1)}) & \text{otherwise.} \end{cases}$$

2.4 The t-Closeness Principle

The fundamental intent behind the t-closeness data sanitization strategy is to limit an adversary's probabilistic inference of any individual's sensitive attribute value(s) *solely* from the knowledge of the equivalence class that individual is made by the data owner to reside in. This can be achieved by the data owner by ensuring that the probability distribution of sensitive attribute value(s) in each equivalence class constructed to sanitize the data is sufficiently close, i.e. within some affordable privacy budget parameter t, to their probability distribution in the original raw table.

A naïve test of acceptability of an equivalence class $\mathcal{E}$, in the presence of sensitive attributes X and Y, is to check if $\delta(\overline{\mathcal{M}}, \overline{\mathcal{E}}) < t$, for a given privacy parameter t. This rigid approach essentially treats the attribute pair as a *single* attribute, whose values are pairs of the form (x_i, y_j), under the lexicographic ordering, thereby simultaneously subjecting both attributes to the same constraint t. Often, different sensitive attributes have different degrees of sensitivity. For example, while maintaining privacy of a patient's cancer diagnosis as well as that patient's body mass index (BMI) are both desirable, the latter goal is not *as* critical as the former. We therefore allow for independent, positive privacy budget parameters for each sensitive attribute, namely t_X and t_Y, leading to the following requirement on acceptable equivalence classes.

Definition 11 ((t_X, t_Y)-Closeness)**.** An equivalence class $\mathcal{E}$ is (t_X, t_Y)-*close* if $\delta(\mathcal{R}_{\overline{\mathcal{M}}}, \mathcal{R}_{\overline{\mathcal{E}}}) \leq t_X$ and $\delta(\mathcal{C}_{\overline{\mathcal{M}}}, \mathcal{C}_{\overline{\mathcal{E}}}) \leq t_Y$.

The overall task can now be summarized as follows:

> Given a multiplicity matrix $\mathcal{M}$ of a table, and privacy parameters t_X and t_Y, to construct a partition $\{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_e\}$ of $\mathcal{M}$ into equivalence classes, i.e. $\sum_{i=1}^{e} \mathcal{E}_i = \mathcal{M}$, such that each class $\mathcal{E}_i$ is (t_X, t_Y)-close.

Equivalence classes in a partition may be intuitively viewed as $\mathcal{M}$ sliced into "layers", which when superimposed as a stack, add up to $\mathcal{M}$, as expected.

Existence of a partition in which each equivalence class is (t_X, t_Y)-close is guaranteed. The smallest partition $\{\mathcal{M}\}$ trivially satisfies this requirement, for any given t_X and t_Y, as $\mathcal{M}$ is $(0, 0)$-close to itself. However, this solution, if adopted for sanitization of the table, would incur the maximum possible information loss, as the quasi identifier values of *all* rows in the table would need to be generalized to a common generalized value. In order for the sanitized table to be of high utility, it is thus desirable to construct a partition with the largest number of equivalence classes or, equivalently, with the smallest sizes of equivalence classes, under the constraint that each of them is (t_X, t_Y)-close.

2.5 Fragmentations of $\overline{\mathcal{M}}$

Our method for constructing the desired partition of $\mathcal{M}$ will be based upon certain contiguous sub-matrices of $\overline{\mathcal{M}}$. The entire matrix $\overline{\mathcal{M}}$ will be fragmented in such a way that equivalence classes conforming to that fragmentation, in a sense made precise later, possess the desired property.

Definition 12 (Fragments)**.** For any $m \times n$ matrix $\mathcal{A}$, and indices a, b, c, d, such that $1 \leq a \leq b \leq m$, and $1 \leq c \leq d \leq n$, the *fragment* of $\mathcal{A}$ bounded by these indices, denoted $\mathcal{A}\langle\langle a \leftrightarrow b; c \leftrightarrow d\rangle\rangle$, is the continuous sub-matrix of $\mathcal{A}$ within these index limits, given by:

$$\mathcal{A}\langle\langle a \leftrightarrow b; c \leftrightarrow d\rangle\rangle[i, j] = \mathcal{A}[i + a - 1, j + c - 1].$$

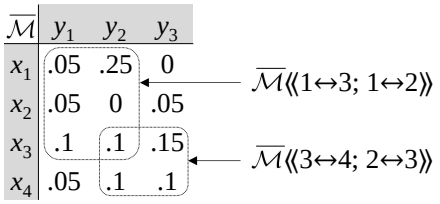

Figure 4: Examples of fragments in matrix $\overline{\mathcal{M}}$ of Figure 1.

Figure 4 shows two sample fragments of the example matrix $\overline{\mathcal{M}}$ of Figure 1.

As fragments are continuous sub-matrices, a union of two fragments into another one is possible only if the fragments are either horizontally adjacent and share the same x-indices, or vertically adjacent and share the same y-indices.

Definition 13 (Union)**.** Any fragments $F = \mathcal{A}\langle\langle a \leftrightarrow b; c \leftrightarrow d\rangle\rangle$ and $G = \mathcal{A}\langle\langle p \leftrightarrow q; r \leftrightarrow s\rangle\rangle$ may be combined into their *union* fragment, denoted $F \sqcup G$, if one of the following conditions is satisfied:

(a) *(Horizontally adjacent)* $a = p$ and $b = q$ and $\min(d, s) + 1 = \max(c, r)$. In this case, $F \sqcup G = \mathcal{A}\langle\langle a \leftrightarrow b; \min(c, r) \leftrightarrow \max(d, s)\rangle\rangle$.

(b) *(Vertically adjacent)* $c = r$ and $d = s$ and $\min(b, q) + 1 = \max(a, p)$. In this case, $F \sqcup G = \mathcal{A}\langle\langle \min(a, p) \leftrightarrow \max(b, q); c \leftrightarrow d\rangle\rangle$.

We are particulary interested in sets of fragments, on which we first define the following binary relations.

Definition 14 (Binary relations $\Rightarrow$ and $\overset{\star}{\Rightarrow}$)**.** Let $\mathcal{F}$ and $\mathcal{G}$ be any sets of fragments on some matrix. Then $\mathcal{F}$ *splits* into $\mathcal{G}$, denoted $\mathcal{F} \Rightarrow \mathcal{G}$, if $\mathcal{F} = \{H_1 \sqcup H_2, H_3, H_4, \ldots, H_k\}$ and $\mathcal{G} = \{H_1, H_2, H_3, H_4, \ldots, H_k\}$, for some fragments $H_1, H_2, \ldots, H_k$ of that matrix. We also let $\overset{\star}{\Rightarrow}$ denote the reflexive and transitive closure of $\Rightarrow$.

Intuitively, $\mathcal{F} \Rightarrow \mathcal{G}$, if $\mathcal{G}$ can be obtained by splitting exactly one fragment in $\mathcal{F}$, either horizontally or vertically. And $\mathcal{F} \overset{\star}{\Rightarrow} \mathcal{G}$, if $\mathcal{G}$ can be obtained from $\mathcal{F}$ by zero or more such steps, i.e. there exist sets of fragments $\mathcal{H}_0, \mathcal{H}_1, \ldots, \mathcal{H}_h$, for some $h \geq 0$, such that $\mathcal{F} = \mathcal{H}_0 \Rightarrow \mathcal{H}_1 \Rightarrow \ldots \Rightarrow \mathcal{H}_h = \mathcal{G}$. It is easy to see that $\overset{\star}{\Rightarrow}$ is a partial order on the collection of all sets of fragments of any matrix.

We now define the important space of fragmentations of the entire matrix $\overline{\mathcal{M}}$. From this space, our method, described later, will pick one that leads to the desired partition of $\mathcal{M}$.

Definition 15 (Fragmentations of $\overline{\mathcal{M}}$)**.** The space $\overline{\mathfrak{M}}$ of all *fragmentations* of $\overline{\mathcal{M}}$ is the smallest collection of sets of fragments of $\overline{\mathcal{M}}$ that satisfies both of the following properties:

(1) $\{\, \overline{\mathcal{M}}\langle\langle 1 \leftrightarrow m; 1 \leftrightarrow n\rangle\rangle \,\} \in \overline{\mathfrak{M}}$; and

(2) If $\mathcal{F} \in \overline{\mathfrak{M}}$ and $\mathcal{F} \Rightarrow \mathcal{G}$, then $\mathcal{G} \in \overline{\mathfrak{M}}$.

The space $\overline{\mathfrak{M}}$ is essentially the collection of all sets of pairwise non-overlapping fragments that together cover $\overline{\mathcal{M}}$. It is easily seen that $\overline{\mathfrak{M}}$ is finite and, under the partial order $\overset{\star}{\Rightarrow}$, it forms a complete lattice, as any subcollection of $\overline{\mathfrak{M}}$ has a least upper bound and a greatest lower bound. Any layer i of this complete lattice, where $1 \leq i \leq mn$, contains all fragmentations of $\overline{\mathcal{M}}$ with exactly i fragments. The unique top and bottom elements of $\overline{\mathfrak{M}}$ are, respectively:

- $\top = \{\, \overline{\mathcal{M}}\langle\langle 1 \leftrightarrow m; 1 \leftrightarrow n\rangle\rangle \,\}$, the only fragmentation of $\overline{\mathcal{M}}$ with one fragment; and
- $\bot = \{\, \overline{\mathcal{M}}\langle\langle i \leftrightarrow i; j \leftrightarrow j\rangle\rangle \mid 1 \leq i \leq m, 1 \leq j \leq n\}$, the only fragmentation of $\overline{\mathcal{M}}$ with mn fragments.

Figure 5(a) shows an example fragmentation of $\overline{\mathcal{M}}$ with three fragments, and the space $\overline{\mathfrak{M}}$ of all fragmentations of $\overline{\mathcal{M}}$ is depicted in Figure 5(b).

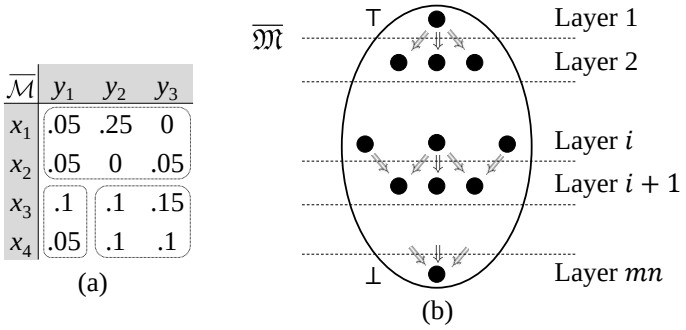

Figure 5: (a) An example fragmentation of $\overline{\mathcal{M}}$ with three fragments; (b) the complete lattice $\overline{\mathfrak{M}}$ of all fragmentations of $\overline{\mathcal{M}}$.

3 Our Method

We now present our method for constructing a partition of the given table into equivalence classes that are (t_X, t_Y)-close. The first step of the method is to find a fragmentation in $\overline{\mathfrak{M}}$, for which any equivalence class that, in a sense to be made precise soon, "conforms" to it is (t_X, t_Y)-close. Generation of equivalence classes then focuses on minimizing information loss caused by generalization, while limiting the generation to classes that conform.

Definition 16 (Conformance to a Fragment)**.** Let $\mathcal{E}$ be any equivalence class, and $F = \overline{\mathcal{M}}\langle\langle a \leftrightarrow b; c \leftrightarrow d\rangle\rangle$ be any fragment of $\overline{\mathcal{M}}$. We say that $\mathcal{E}$ *conforms* to F if

$$|F| = |G|,$$

where G is the corresponding fragment of $\overline{\mathcal{E}}$, i.e. $G = \overline{\mathcal{E}}\langle\langle a \leftrightarrow b; c \leftrightarrow d\rangle\rangle$.

Conformance of $\mathcal{E}$ to a fragment F of $\overline{\mathcal{M}}$ essentially means that, while the sizes of $\mathcal{E}$ and $\mathcal{M}$ may be different, they both contain the same *proportion* of X-Y value-pairs from the fragment F. Figure 6 shows two example fragments of $\overline{\mathcal{M}}$, and corresponding fragments of $\overline{\mathcal{E}}$, for the underlying equivalence class $\mathcal{E}$ of Figure 1. It can be seen that $\mathcal{E}$ conforms to the fragment $\overline{\mathcal{M}}\langle\langle 1 \leftrightarrow 3; 1 \leftrightarrow 2\rangle\rangle$, because the sum of all values of this fragment, namely 0.55, coincides with that of the corresponding fragment in $\overline{\mathcal{E}}$. However, $\mathcal{E}$ does not conform to the other fragment of $\overline{\mathcal{M}}$ shown in the figure, namely $\overline{\mathcal{M}}\langle\langle 3 \leftrightarrow 4; 2 \leftrightarrow 3\rangle\rangle$. Sum of values of one of the fragments in question is 0.45, while that of the other is 0.4, a different value.

We extend this concept of conformance of $\mathcal{E}$ to an entire fragmentation of $\overline{\mathcal{M}}$, which ends up placing a very useful constraint on the closeness of $\mathcal{E}$.

Definition 17 (Conformance to a Fragmentation)**.** Let $\mathcal{E}$ be any equivalence class, and $\mathcal{F}$ be any fragmentation of $\overline{\mathcal{M}}$, i.e. $\mathcal{F} \in \overline{\mathfrak{M}}$. Then $\mathcal{E}$ *conforms* to $\mathcal{F}$, if $\mathcal{E}$ conforms to each fragment in $\mathcal{F}$.

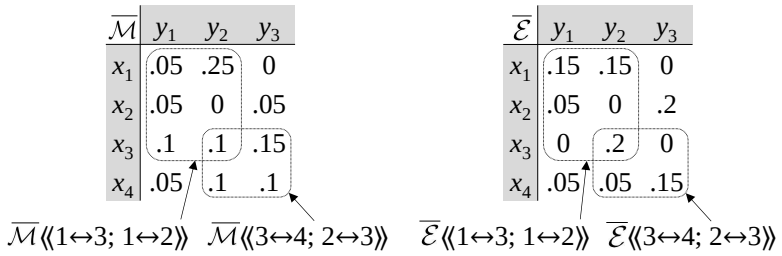

Figure 6: An equivalence class $\mathcal{E}$ conforming to one fragment of $\overline{\mathcal{M}}$, but not to the other.

Proposition 18. *Let $\mathcal{F} = \{F_1, F_2, \ldots, F_k\} \in \overline{\mathfrak{M}}$. If an equivalence class $\mathcal{E}$ conforms to $\mathcal{F}$, where $E_1, E_2, \ldots, E_k$ are the corresponding fragments of $\overline{\mathcal{E}}$, then:*

$$\begin{aligned} \delta(\mathcal{R}_{\overline{\mathcal{M}}}, \mathcal{R}_{\overline{\mathcal{E}}}) &\leq \textstyle\sum_{i=1}^{k} \delta(\mathcal{R}_{F_i}, \mathcal{R}_{E_i}), \text{ and} \\ \delta(\mathcal{C}_{\overline{\mathcal{M}}}, \mathcal{C}_{\overline{\mathcal{E}}}) &\leq \textstyle\sum_{i=1}^{k} \delta(\mathcal{C}_{F_i}, \mathcal{C}_{E_i}). \end{aligned}$$

Proof. (By induction via $\Rightarrow$) In the base case, every equivalence class $\mathcal{E}$ conforms to the top-most element $\top$ of the lattice $\overline{\mathfrak{M}}$, because in this case, $k = 1$, $F_1 = \overline{\mathcal{M}}$, $E_1 = \overline{\mathcal{E}}$, and $|\overline{\mathcal{M}}| = |\overline{\mathcal{E}}| = 1$. The above inequalities are in fact strict equalities.

For the inductive case, suppose the proposition holds for fragmentation $\mathcal{F}' = \{F_1 \sqcup F_2, F_3, F_4, \ldots, F_k\}$. Since $\mathcal{E}$ conforms to $\mathcal{F}$, we have that $|F_i| = |E_i|$, for all i. It follows that $|F_1 \sqcup F_2| = |E_1 \sqcup E_2|$, thus $\mathcal{E}$ also conforms to $\mathcal{F}'$. By the inductive hypothesis, we have:

$$\begin{aligned} \delta(\mathcal{R}_{\overline{\mathcal{M}}}, \mathcal{R}_{\overline{\mathcal{E}}}) \leq\ & \delta(\mathcal{R}_{F_1 \sqcup F_2}, \mathcal{R}_{E_1 \sqcup E_2}) + \\ & \textstyle\sum_{i=3}^{k} \delta(\mathcal{R}_{F_i}, \mathcal{R}_{E_i}). \end{aligned}$$

We now have two possible cases:

(a) *(F_1 and F_2 are horizontally adjacent)* In this case, $\mathcal{R}_{F_1 \sqcup F_2} = \mathcal{R}_{F_1} + \mathcal{R}_{F_2}$ and $\mathcal{R}_{E_1 \sqcup E_2} = \mathcal{R}_{E_1} + \mathcal{R}_{E_2}$. By Proposition 5, $\delta(\mathcal{R}_{F_1 \sqcup F_2}, \mathcal{R}_{E_1 \sqcup E_2}) \leq \delta(\mathcal{R}_{F_1}, \mathcal{R}_{E_1}) + \delta(\mathcal{R}_{F_2}, \mathcal{R}_{E_2})$.

(b) *(F_1 and F_2 are vertically adjacent)* In this case, $\mathcal{R}_{F_1 \sqcup F_2} = \mathcal{R}_{F_1} \mathcal{R}_{F_2}$ and $\mathcal{R}_{E_1 \sqcup E_2} = \mathcal{R}_{E_1} \mathcal{R}_{E_2}$. By Proposition 7, $\delta(\mathcal{R}_{F_1 \sqcup F_2}, \mathcal{R}_{E_1 \sqcup E_2}) \leq \delta(\mathcal{R}_{F_1}, \mathcal{R}_{E_1}) + \delta(\mathcal{R}_{F_2}, \mathcal{R}_{E_2})$.

Thus, in both cases, $\delta(\mathcal{R}_{\overline{\mathcal{M}}}, \mathcal{R}_{\overline{\mathcal{E}}}) \leq \sum_{i=1}^{k} \delta(\mathcal{R}_{F_i}, \mathcal{R}_{E_i})$. By a similar argument, $\delta(\mathcal{C}_{\overline{\mathcal{M}}}, \mathcal{C}_{\overline{\mathcal{E}}}) \leq \sum_{i=1}^{k} \delta(\mathcal{C}_{F_i}, \mathcal{C}_{E_i})$. □

By Theorem 10, we already know that each term on the right-hand sides of the above result, such as $\delta(\mathcal{R}_{F_i}, \mathcal{R}_{E_i})$, is in turn bounded above by $\Delta(\mathcal{R}_{F_i})$. Aggregate row and column earth mover's distance bounds for an entire fragmentation of $\overline{\mathcal{M}}$ can thus be given as follows.

Definition 19 (Aggregate Bounds)**.** For any fragmentation $\mathcal{F} = \{F_1, F_2, \ldots, F_k\} \in \overline{\mathfrak{M}}$, the aggregate row and column earth mover's distance *bounds* of $\mathcal{F}$ are given, respectively, by:

$$\widehat{\mathcal{R}}_{\mathcal{F}} = \sum_{i=1}^{k} \Delta(\mathcal{R}_{F_i}), \quad \text{and} \quad \widehat{\mathcal{C}}_{\mathcal{F}} = \sum_{i=1}^{k} \Delta(\mathcal{C}_{F_i}).$$

We now state the following key result.

Theorem 20. *If $\mathcal{E}$ conforms to a fragmentation $\mathcal{F} \in \overline{\mathfrak{M}}$, then $\mathcal{E}$ is $(\widehat{\mathcal{R}}_{\mathcal{F}}, \widehat{\mathcal{C}}_{\mathcal{F}})$-close.*

Proof. Immediate from Proposition 18, definition of aggregate bounds, and Theorem 10. □

Recall that our overall task is to construct a partition of the given $\mathcal{M}$ into equivalence classes, such that each class in the partition is (t_X, t_Y)-close, for some given t_X and t_Y values. In light of the above theorem, our method will first focus on finding an appropriate fragmentation $\mathcal{F} \in \overline{\mathfrak{M}}$, i.e. one for which $\widehat{\mathcal{R}}_{\mathcal{F}} \leq t_X$ and $\widehat{\mathcal{C}}_{\mathcal{F}} \leq t_Y$. We will then limit our search for equivalence classes to just the ones that conform to the chosen $\mathcal{F}$. Theorem 20 guarantees all such classes to satisfy the closeness requirement.

3.1 Fragmentation Search

Finding a fragmentation $\mathcal{F} \in \overline{\mathfrak{M}}$ with desired aggregate bounds, i.e. one for which $\widehat{\mathcal{R}}_{\mathcal{F}} \leq t_X$ and $\widehat{\mathcal{C}}_{\mathcal{F}} \leq t_Y$, is not difficult due to the fact that fragmentations in $\overline{\mathfrak{M}}$ are, in a somewhat weak sense, "ordered" by their aggregate bounds, as shown by the following proposition.

Proposition 21. *For any $\mathcal{F}, \mathcal{G} \in \overline{\mathfrak{M}}$, if $\mathcal{F} \Rightarrow \mathcal{G}$, then either $\widehat{\mathcal{R}}_{\mathcal{F}} \geq \widehat{\mathcal{R}}_{\mathcal{G}}$ or $\widehat{\mathcal{C}}_{\mathcal{F}} \geq \widehat{\mathcal{C}}_{\mathcal{G}}$.*

Proof. As $\mathcal{G}$ is obtained by splitting exactly one fragment in $\mathcal{F}$, it suffices to show that for any fragments P and Q of $\overline{\mathcal{M}}$ that may be combined into their union fragment $P \sqcup Q$, we have that if P and Q are vertically adjacent, then $\Delta(\mathcal{R}_{P\sqcup Q}) \geq \Delta(\mathcal{R}_P) + \Delta(\mathcal{R}_Q)$. Otherwise, $\Delta(\mathcal{C}_{P\sqcup Q}) \geq \Delta(\mathcal{C}_P) + \Delta(\mathcal{C}_Q)$. We show here the first case, as the second case can be proved in a similar manner.

Suppose P and Q are vertically adjacent, i.e. $\mathcal{R}_{P\sqcup Q} = \mathcal{R}_P\mathcal{R}_Q$. Let $\mathcal{R}_P \in \mathcal{S}^{k,\alpha}$, $\mathcal{R}_Q \in \mathcal{S}^{l,\beta}$, and $\mathcal{T}$ be the set of vectors given by:

$$\mathcal{T} = \{\langle p_1, p_2, \ldots, p_k, q_1, q_2, \ldots, q_l\rangle : \sum_{i=1}^{k} p_i = \alpha \text{ and } \sum_{i=1}^{l} q_i = \beta\}.$$

Note that $\mathcal{R}_P\mathcal{R}_Q \in \mathcal{T}$. Since $\mathcal{T} \subseteq \mathcal{S}^{k+l,\alpha+\beta}$, it follows that $\Delta(\mathcal{R}_{P\sqcup Q}) = \max\{\delta(\mathcal{R}_P\mathcal{R}_Q, S) : S \in \mathcal{S}^{k+l,\alpha+\beta}\} \geq \max\{\delta(\mathcal{R}_P\mathcal{R}_Q, T) : T \in \mathcal{T}\} = \Delta(\mathcal{R}_P) + \Delta(\mathcal{R}_Q)$. □

This ordering of aggregate bounds of fragmentations in $\overline{\mathfrak{M}}$, suggests that a simple scan along any arbitrary path starting at its topmost fragmentation $\top$ and ending at its bottommost fragmentation $\bot$, may be performed until an acceptable fragmentation is found, i.e. one whose aggregate bounds are within the given privacy budget parameters t_X and t_Y. The nondeterministic function FIND-FRAGMENTATION given below performs such a scan.

FIND-FRAGMENTATION(t_X, t_Y)
1 $\mathcal{F} \leftarrow \top$
2 **while** $\widehat{\mathcal{R}}_{\mathcal{F}} > t_X$ **or** $\widehat{\mathcal{C}}_{\mathcal{F}} > t_Y$
3 pick a fragmentation $\mathcal{G}$, such that $\mathcal{F} \Rightarrow \mathcal{G}$
4 $\mathcal{F} \leftarrow \mathcal{G}$
5 **return** $\mathcal{F}$

It should be noted that the ordering established by Proposition 21 is weak, because only one of the aggregate bounds is guaranteed to not increase at each step of the scan, and sometimes the other aggregate bound may increase. As an example of this phenomenon, consider the portion of a lattice $\overline{\mathfrak{M}}$ shown in Figure 7. It is easily verified that while $\top \Rightarrow \mathcal{F}$,

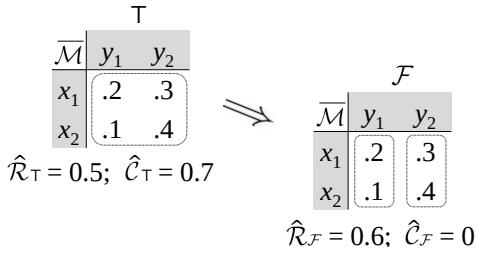

Figure 7: Example of an increase in the aggregate row earth mover's distance bound.

$\widehat{\mathcal{R}}_\top = 0.5 < \widehat{\mathcal{R}}_\mathcal{F} = 0.6$. Despite this phenomenon, however, a downward scan along *any* path of $\overline{\mathfrak{M}}$ is guaranteed to terminate with success within mn iterations because, firstly, the lattice $\overline{\mathfrak{M}}$ is finite and, secondly, the aggregate bounds of $\perp$ are zero, as shown below.

Proposition 22. *For any $\mathcal{F} \in \overline{\mathfrak{M}}$, $\widehat{\mathcal{R}}_\mathcal{F} \in [0,1]$ and $\widehat{\mathcal{C}}_\mathcal{F} \in [0,1]$. Also, $\widehat{\mathcal{R}}_\perp = \widehat{\mathcal{C}}_\perp = 0$.*

Proof. The row-sum vector, $\mathcal{R}_F$, of any fragment F of $\overline{\mathcal{M}}$, is clearly a k-vector in $\mathcal{S}^{k,|F|}$, for some k, $1 \leq k \leq m$, where k is the number of rows in F. It follows from the definition of Δ and Proposition 3 that $\Delta(\mathcal{R}_F) \in [0, |F|]$. As $\sum_{F \in \mathcal{F}} |F| = 1$, we have that $\widehat{\mathcal{R}}_\mathcal{F} \in [0,1]$. Similarly, $\widehat{\mathcal{C}}_\mathcal{F} \in [0,1]$. Moreover, each fragment $F \in \perp$ contains just one row and one column, thus $\mathcal{R}_F$ and $\mathcal{C}_F$ are 1-vectors in $\mathcal{S}^{1,|F|}$, and $\Delta(\mathcal{R}_F) = \Delta(\mathcal{C}_F) = 0$. Therefore, $\widehat{\mathcal{R}}_\perp = \widehat{\mathcal{C}}_\perp = 0$. □

While any choice of fragmentation $\mathcal{G}$ on Line 3 of the function FIND-FRAGMENTATION will lead to a somewhat usable return value of the function, upon closer inspection of $\overline{\mathfrak{M}}$, it can be seen that not all choices lead to results of equal utility. This is due to acceptable fragmentations on some downward paths being encountered earlier, i.e. in a higher layer of $\overline{\mathfrak{M}}$, than on other paths. Figure 8 shows an example of this phenomenon for privacy budgets $t_X = 0.5$ and $t_Y = 0.6$. For these budget values, from the portion of $\overline{\mathfrak{M}}$ shown in the figure, fragmentations $\top$ and $\mathcal{G}$ are not acceptable, but $\mathcal{F}$ and $\mathcal{H}$ are, because their row as well as column aggregate bounds are within their corresponding budgets. However, a downward scan from $\top$ will encounter $\mathcal{F}$ in Layer 1 of $\overline{\mathfrak{M}}$, whereas $\mathcal{H}$ will be encountered only in Layer 2. An acceptable fragmentation in a higher layer is preferable, because more equivalence classes conform to it.

Figure 9 depicts the two regions of $\overline{\mathfrak{M}}$, namely the possibly empty upper region containing fragmentations whose at least one aggregate bound exceeds its corresponding privacy budget, and the always nonempty lower region containing fragmentations both of whose aggregate bounds are within their corresponding privacy budgets. As the boundary between these regions is irregular, thereby causing some downward paths to lead to an acceptable fragmentation earlier than others, an ideal strategy for picking fragmentation $\mathcal{G}$ on Line 3 of the FIND-FRAGMENTATION function is one that terminates the iteration in the smallest number of steps. This would result in a fragmentation $\mathcal{F}$ returned by the function belonging to the highest layer of $\overline{\mathfrak{M}}$, i.e. one with the fewest fragments over all acceptable fragmentations, and one that thus has most equivalence classes conforming to it.

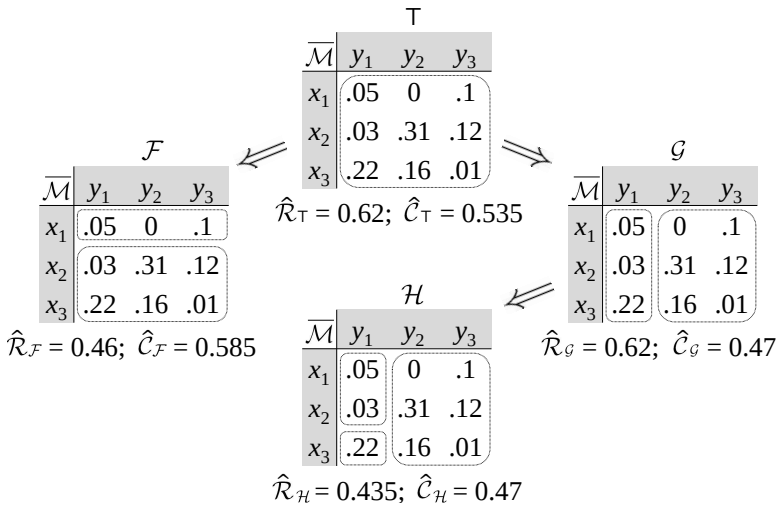

Figure 8: Fragmentations, $\mathcal{F}$ and $\mathcal{H}$ are acceptable for $t_X = 0.5$ and $t_Y = 0.6$, but lie in different layers of $\overline{\mathfrak{M}}$.

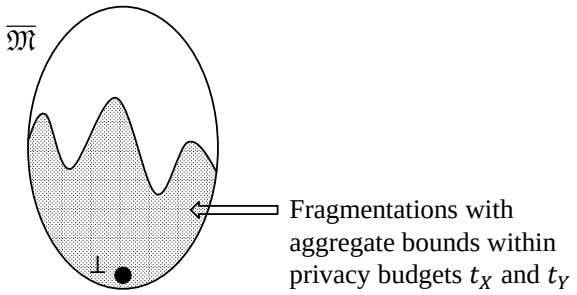

Figure 9: Fragmentations of $\overline{\mathfrak{M}}$ with acceptable aggregate bounds.

An exhaustive strategy for searching for the optimum acceptable fragmentation becomes impractical for large values of m and n, as $\overline{\mathfrak{M}}$ has mn layers and, as explained in Section 3.3, each fragmentation $\mathcal{F} \in \overline{\mathfrak{M}}$ has $\mathcal{O}(mn)$ fragmentations $\mathcal{G}$, such that $\mathcal{F} \Rightarrow \mathcal{G}$. Several different greedy strategies may be employed at Line 3 of the FIND-FRAGMENTATION function to arrive at a reasonably promising $\mathcal{G}$. One strategy is to pick the $\mathcal{G}$ whose aggregate bounds are the closest to their corresponding budgets, if not already within them. This can be achieved by selecting the $\mathcal{G}$ that minimizes $\Psi_{\mathcal{G}}$, given by:

$$\Psi_{\mathcal{G}} = \max(\widehat{\mathcal{R}}_{\mathcal{G}} - t_X, 0) + \max(\widehat{\mathcal{C}}_{\mathcal{G}} - t_Y, 0).$$

By employing the $\max$ function, $\Psi_{\mathcal{G}}$ essentially considers for comparison only those aggregate bounds that are not yet within their corresponding budgets. Picking the $\mathcal{G}$ for which $\Psi_{\mathcal{G}}$ is the smallest value, while breaking ties arbitrarily, is a greedy approach, with a hope to terminate the iteration in the FIND-FRAGMENTATION function in the smallest number of steps.

In the example $\overline{\mathfrak{M}}$ of Figure 8, where $t_X = 0.5$ and $t_Y = 0.6$, the fragmentation $\top$ is unacceptable, because $\widehat{\mathcal{R}}_{\top} > t_X$. Now, since $\Psi_{\mathcal{F}} = 0$, $\mathcal{F}$ will be picked over $\mathcal{G}$ (or any other

child of $\top$), and the iteration will terminate with $\mathcal{F}$ as the found acceptable fragmentation. The fragmentation $\mathcal{H}$, although also acceptable, is never reached, because it lies in a lower layer.

3.2 Generating Equivalence Classes

Let $\mathcal{F} = \{F_1, F_2, \ldots, F_k\} \in \overline{\mathfrak{M}}$ be the fragmentation returned by the FIND-FRAGMENTATION function of Section 3.1. We now confine ourselves to partitioning the given table into equivalence classes that conform to $\mathcal{F}$, thereby achieving (t_X, t_Y)-closeness.

With an aim to minimize information loss caused by generalization of quasi identifier values of rows contained in an equivalence class, it is desirable to generate classes with as few rows in them as possible. It is easily seen that the smallest number of rows in any equivalence class conforming to $\mathcal{F}$ is the smallest integer r, such that $r.|F_i|$ is an integer, for each i. As the given table has $|\mathcal{M}|$ rows, we now give a procedure that generates $c = |\mathcal{M}|/r$ classes, each containing r rows. The values c and r can be determined easily by observing that c is simply the greatest common divisor of all values in the set $\{|\mathcal{M}|.|F_i| : 1 \leq i \leq k\}$, and $r = |\mathcal{M}|/c$.

Of the r rows placed in any equivalence class, the conformity condition enforces placing exactly $r.|F_i|$ rows from fragment F_i, for each i. Under this constraint, within each class, information loss due to generalization can be curtailed by placing those r rows in it whose quasi identifier values are, in a sense, close to each other.

Closeness between rows of the table depend upon the underlying distance metric over them, which can vary from one application to another. As an example, if all quasi identifier attributes take numeric values, the rows may be treated as points in a Euclidean space with as many dimensions as the number of such attributes. The distance between any two rows is then just the common Euclidean distance between them.

The general procedure given below for generating equivalence classes from the chosen fragmentation is based upon the above analysis, and is independent of the underlying distance metric.

GENERATE-CLASSES

1 **for** $j \leftarrow 1$ **to** c
2 pick a row w of the table, which is from some fragment $F_i \in \mathcal{F}$, such that $|F_i| > 0$
3 $E_j \leftarrow \{w\}$
4 add to E_j, $(r.|F_i| - 1)$ other rows from F_i that are closest to w
5 from each other fragment $F_l \in \mathcal{F}$, add to E_j, $r.|F_l|$ rows that are closest to w
6 remove each row in E_j from the table
7 **return** classes $E_1, E_2, \ldots, E_c$

As each equivalence class generated by the above procedure conforms to $\mathcal{F}$, the created partition of the table is thus (t_X, t_Y)-close.

3.3 Complexity of Our Method

Liang and Yuan [7] showed that, even for one sensitive attribute, it is NP-hard to find an optimal t-close partition of a given table into equivalence classes. At the expense of

optimality, our greedy approach results in an acceptable solution in polynomial time, as shown below.

For any fragment F, computation of $\Delta(\mathcal{R}_F)$ is clearly an $\mathcal{O}(m)$ operation, and that of $\Delta(\mathcal{C}_F)$ is $\mathcal{O}(n)$. The initial computations of $\widehat{\mathcal{R}}_\top$ and $\widehat{\mathcal{C}}_\top$ in FIND-FRAGMENTATION thus take $\mathcal{O}(m+n)$ time. Any fragmentation $\mathcal{F}$ has $\mathcal{O}(mn)$ fragmentations $\mathcal{G}$, such that $\mathcal{F} \Rightarrow \mathcal{G}$, as can be seen by a worst-case of $\mathcal{F}$ containing one fragment for each of the m rows of $\overline{\mathcal{M}}$. Any such $\mathcal{G}$ is obtained by splitting exactly one fragment contained in $\mathcal{F}$ into two fragments. As all other fragments in $\mathcal{F}$ and $\mathcal{G}$ are the same, $\widehat{\mathcal{R}}_\mathcal{G}$ and $\widehat{\mathcal{C}}_\mathcal{G}$ can be obtained from $\widehat{\mathcal{R}}_\mathcal{F}$ and $\widehat{\mathcal{C}}_\mathcal{F}$, respectively, in $\mathcal{O}(m+n)$ time. Picking a $\mathcal{G}$, within any iteration, that minimizes $\Psi_\mathcal{G}$, is thus an $\mathcal{O}((m+n)mn)$ operation. Finally, as $\overline{\mathfrak{M}}$ contains exactly mn layers, which is an upper bound on the number of iterations in that function, the complexity of FIND-FRAGMENTATION is $\mathcal{O}((m+n)m^2n^2)$.

Computation of c, the number of equivalence classes to be generated, requires computation of the greatest common divisor of at most mn integers, each of which is between 1 and $|\mathcal{M}|$. While faster quasi-linear algorithms now exist for two integers, the classical Euclidean method determines that value in $\mathcal{O}(\log^2 |\mathcal{M}|)$ time (see Knuth [5] and Sorenson [13]). A straightforward iteration of this method, for mn integers, results in the complexity of determining c to be $\mathcal{O}(mn \log^2 |\mathcal{M}|)$.

Distances between rows of the table can be pre-computed for the function GENERATE-CLASSES in $|\mathcal{M}|^2$ time. With these pre-computed distances, each iteration of the function can be seen to take $\mathcal{O}(|\mathcal{M}|)$ time. As $c \leq |\mathcal{M}|$, which is the number of iterations, the complexity of GENERATE-CLASSES is thus $\mathcal{O}(|\mathcal{M}|^2)$.

3.4 Generalization to Arbitrary Number of Sensitive Attributes

We presented our method for tables with exactly two sensitive attributes only because that restriction makes understanding it significantly easier. However, our method is applicable to tables with any arbitrary number of sensitive attributes, and this section outlines how it can be generalized, in a straightforward way, from exactly two to n sensitive attributes, for any $n \geq 2$.

Instead of having domains, X and Y, for just two sensitive attributes, we now have n domains, $X_1, X_2, \ldots, X_n$, of values that may appear, respectively, in each of the n sensitive attribute columns of the given table. For any i,

$$X_i = \{x_{(i,1)}, x_{(i,2)}, \ldots, x_{(i,m_i)}\}.$$

As before, we assume that $x_{(i,1)} < x_{(i,2)} < \cdots < x_{(i,m_i)}$, for all i.

The multiplicity matrix is now the n-dimensional matrix $\mathcal{M} = \prod_{i=1}^{n} X_i$, and any cell of this matrix, $\mathcal{M}[k_1, k_2, \ldots, k_n]$, is the number of rows of the given table the sensitive value tuple $(x_{(1,k_1)}, x_{(2,k_2)}, \ldots, x_{(n,k_n)})$ appears in. Any equivalence class $\mathcal{E}$ is also now an n-dimensional matrix and, as before, any of its cells, $\mathcal{E}[k_1, k_2, \ldots, k_n]$, is a non-negative integer no larger than the corresponding cell of $\mathcal{M}$. $|\mathcal{M}|$ and $|\mathcal{E}|$ still denote the sum of all values in matrices $\mathcal{M}$ and $\mathcal{E}$, respectively, and $\overline{\mathcal{M}} = \mathcal{M}/|\mathcal{M}|$ and $\overline{\mathcal{E}} = \mathcal{E}/|\mathcal{E}|$ are still the normalized versions of these matrices.

As we no longer have only two sensitive attributes, we do away with row- and column-sum vectors, but instead now just have a slice-sum vector, within any matrix, for each sensitive attribute. For example, for any sensitive attribute i, the slice-sum vector $\mathcal{R}_{(\mathcal{M},i)}$ is an m_i-vector, where $\mathcal{R}_{(\mathcal{M},i)}[j]$ is the number of rows of the given table that contain the

value $x_{(i,j)}$ for the sensitive attribute i. Slice-sum vectors of other matrices, $\mathcal{R}_{(\overline{\mathcal{M}},i)}$, $\mathcal{R}_{(\mathcal{E},i)}$, and $\mathcal{R}_{(\overline{\mathcal{E}},i)}$, have similar intuitive meanings.

The given privacy parameters, t_X and t_Y, are now replaced by one for each sensitive attribute, $t_1, t_2, \ldots, t_n$. An equivalence class $\mathcal{E}$ is $(t_1, t_2, \ldots, t_n)$-*close* if $\delta(\mathcal{R}_{(\overline{\mathcal{M}},i)}, \mathcal{R}_{(\overline{\mathcal{E}},i)}) \leq t_i$, for each sensitive attribute i, $1 \leq i \leq n$.

A fragment of any matrix $\mathcal{A}$ now needs to have not just two, but n dimensions:

$$\mathcal{A}\langle\langle a_1 \leftrightarrow b_1; a_2 \leftrightarrow b_2; \cdots ; a_n \leftrightarrow b_n \rangle\rangle,$$

where for each i, it must be the case that $1 \leq a_i \leq b_i \leq m_i$. And a union of two fragments $\mathcal{A}\langle\langle a_1 \leftrightarrow b_1; a_2 \leftrightarrow b_2; \cdots ; a_n \leftrightarrow b_n \rangle\rangle$ and $\mathcal{A}\langle\langle p_1 \leftrightarrow q_1; p_2 \leftrightarrow q_2; \cdots ; p_n \leftrightarrow q_n \rangle\rangle$ is possible not when the fragments are either simply horizontally or vertically adjacent, but adjacent in the dimension of some sensitive attribute s, i.e. when:

- $a_i = p_i$ and $b_i = q_i$, for all $i \neq s$; and
- $\min(b_s, q_s) + 1 = \max(a_s, p_s)$.

$\overline{\mathfrak{M}}$, $\top$, and $\bot$ are now similar generalizations to n dimensions.

For any fragmentation $\mathcal{F} = \{F_1, F_2, \ldots, F_k\} \in \overline{\mathfrak{M}}$, instead of aggregate row and column earth mover's distance bounds of $\mathcal{F}$, we now have an aggregate earth mover's distance bound for each sensitive attribute i, given by:

$$\widehat{\mathcal{R}}_{(\mathcal{F},i)} = \sum_{j=1}^{k} \Delta(\mathcal{R}_{(F_j,i)}).$$

As before, if an equivalence class $\mathcal{E}$ conforms to a fragmentation $\mathcal{F} \in \overline{\mathfrak{M}}$, then:

$$\mathcal{E} \text{ is } (\widehat{\mathcal{R}}_{(\mathcal{F},1)}, \widehat{\mathcal{R}}_{(\mathcal{F},2)}, \ldots, \widehat{\mathcal{R}}_{(\mathcal{F},n)})\text{-close.}$$

An acceptable fragmentation can be found by the generalized function below.

FIND-FRAGMENTATION-GENERALIZED$(t_1, t_2, \ldots, t_n)$
1 $\mathcal{F} \leftarrow \top$
2 **while** there exists a sensitive attribute i, such that $\widehat{\mathcal{R}}_{(\mathcal{F},i)} > t_i$
3 pick a fragmentation $\mathcal{G}$, such that $\mathcal{F} \Rightarrow \mathcal{G}$
4 $\mathcal{F} \leftarrow \mathcal{G}$
5 **return** $\mathcal{F}$

Line 2 in the above function was the only place where a modification from the previous version, FIND-FRAGMENTATION, was necessary, as we now enforce t-closeness for all n sensitive attributes, instead of just two. A complexity analysis similar to the one before shows that FIND-FRAGMENTATION-GENERALIZED runs in $\mathcal{O}((\sum_{i=1}^{n} m_i)(\prod_{i=1}^{n} m_i^2))$ time, still polynomial in the given m_i values.

The above are the only generalizations to our method needed to accommodate an arbitrary number of sensitive attributes. The GENERATE-CLASSES procedure remains unchanged, as it does not depend upon the number of sensitive attributes.

4 Conclusions

A well-known technique for preserving privacy of individuals in a publicly released relational table is to sanitize the table prior to its release according to the t-closeness principle. Widely recognized, as in Aggarwal and Yu [1], for numeric attributes, t-closeness anonymization is more effective than many other privacy-preserving data mining methods. Arriving at a method for obtaining t-closeness, however, especially in the presence of multiple sensitive attributes, has thus far been challenging. In this paper, we developed a method for that task.

We presented a method for partitioning rows of a relational table containing multiple numerical sensitive attributes into equivalence classes, such that the distribution of values of each sensitive attribute X in the given table is t_X-close, for any given privacy budget parameter $0 \leq t_X \leq 1$, to the distribution of those values in any class. As different sensitive attributes are often sensitive to a different extent, our method allows each such attribute to have its own privacy budget. Although finding an optimal solution to this problem is known to be NP-hard, our method employs a greedy approach to provide an acceptable result in polynomial time.

We first showed that the multi-dimensional space of all sensitive attribute values can be fragmented in such a way that a certain predetermined number of rows can even be dispersed randomly from each fragment to create t-close equivalence classes. Our method thus proceeds by first greedily finding one such fragmentation. While generating equivalence classes, it then exploits the flexibility provided by the available random choices to lower the information loss incurred later due to generalizing the quasi identifier values in each class. The resulting sanitized table thereby possesses higher utility for mining.

For ease of understanding, we presented our method for tables with exactly two sensitive attributes, but it is straightforward to generalize our method to any number of sensitive attributes. We also explained how this generalization to an arbitrary number of sensitive attributes can be performed.

We restricted the domains of sensitive attributes to contain only numerical values. This is enough for many real-life applications, such as a standard blood test, which results in several numerical sensitive values, like white and red blood cell counts, LDL and HDL cholesterol levels, hemoglobin measure, etc. The values in the domains of these attributes are discrete and possess a total order among them. Moreover, the ground distance between any two values in a domain is simply the absolute difference between them.

An easy way to generalize our method for continuous domains, like all real values in the range $(0, 1]$, is to partition the range into discrete subranges, e.g. the 10 subranges contained in the partition $\{(0.1k, 0.1(k+1)] : 0 \leq k < 10\}$, with the absolute difference between the k values of any two subranges as the ground distance between those subranges. The granularity of the partition can be adjusted, as needed, according to the nature of the attribute values.

Many applications contain non-numerical sensitive attributes, like the disease a patient is diagnosed with. Although such domains are discrete, the ground distance between their values needs to take semantic "closeness" of values into account. For example, while arthritis and osteoporosis are different bone-related medical conditions, in a sense they are closer to each other than any of them is to AIDS, which is an immune system disease. We are currently working to generalize our method for such domains, whose values can be naturally arranged in certain categories.

References

[1] C. Aggarwal and P. Yu, editors. *Privacy-Preserving Data Mining: Models and Algorithms*. Springer, 2008.

[2] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan. Sabre: A sensitive attribute bucketization and redistribution framework for t-closeness. *The VLDB Journal*, 20(1):59–81, 2011.

[3] Y. Fang, M.Z. Ashrafi, and S.K. Ng. Privacy beyond single sensitive attribute. In *Proceedings of the 22nd International Conference on Database and Expert Systems Applications*, pages 187–201, 2011.

[4] P. Golle. Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM workshop on privacy in the electronic society*, pages 77–80, Alexandria, VA, USA, 2006.

[5] D. Knuth. *The Art of Computer Programming*, volume 2: Seminumerical Algorithms. Addison-Wesley, 1997.

[6] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd IEEE International Conference on Data Engineering (ICDE)*, pages 106–115, Istanbul, Turkey, 2007.

[7] H. Liang and H. Yuan. On the complexity of t-closeness anonymization and related problems. In W. Meng, L. Feng, S. Bressan, W. Winiwarter, and W. Song, editors, *Proceedings of the 18th International Conference on Database Systems for Advanced Applications (DASFAA)*, Lecture Notes in Computer Science, vol. 7825, Springer-Verlag, pages 331–345, Wuhan, China, 2013.

[8] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[9] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l´Académie Royale des Science de Paris, avec les Mémoires de Mathématique et de Physique*, pages 666–704, 1781.

[10] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[11] Y. Rubner, C. Tomasi, and L. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[12] Y. Sei, H. Okumura, T. Takenouchi, and A. Oshuga. Anonymization of sensitive quasi-identifiers for l-diversity and t-closeness. *IEEE Transactions on Dependable and Secure Computing*, 16(4):580–593, 2019.

[13] J. Sorenson. An analysis of Lehmer's Euclidean GCD algorithm. In *Proceedings of the ACM International Symposium on Symbolic and Algebraic Computation (ISSAC)*, pages 254–258, Montreal, Canada, 1995.

[14] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez. t-closeness through microaggregation: Strict privacy with enhanced utility preservation. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3098–3110, 2015.

[15] L. Sweeney. Uniqueness of simple demographics in the U.S. population. Technical Report LIDAPWP4, Carnegie Mellon University, Laboratory for International Data Privacy, 2000.

[16] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

[17] R. Wang, Y. Zhu, T.-S. Chen, and C.-C. Chang. Privacy-preserving algorithms for multiple sensitive attributes satisfying t-closeness. *Journal of Computer Science and Technology*, 33(6):1231–1242, 2018.