On the Complexity of t-Closeness Anonymization and Related Problems

Hongyu Liang* Hao Yuan†

Abstract

An important issue in releasing individual data is to protect the sensitive information from being leaked and maliciously utilized. Famous privacy preserving principles that aim to ensure both data privacy and data integrity, such as k-anonymity and l-diversity, have been extensively studied both theoretically and empirically. Nonetheless, these widely-adopted principles are still insufficient to prevent attribute disclosure if the attacker has partial knowledge about the overall sensitive data distribution. The t-closeness principle has been proposed to fix this, which also has the benefit of supporting numerical sensitive attributes. However, in contrast to k-anonymity and l-diversity, the theoretical aspect of t-closeness has not been well investigated.

We initiate the first systematic theoretical study on the t-closeness principle under the commonly-used attribute suppression model. We prove that for every constant t such that $0 \leq t < 1$, it is NP-hard to find an optimal t-closeness generalization of a given table. The proof consists of several reductions each of which works for different values of t, which together cover the full range. To complement this negative result, we also provide exact and fixed-parameter algorithms. Finally, we answer some open questions regarding the complexity of k-anonymity and l-diversity left in the literature.

1 Introduction

Privacy-preserving data publication is an important and active topic in the database area. Nowadays many organizations need to publish microdata that contain certain information, e.g., medical condition, salary, or census data, of a collection of individuals, which are very useful for research and other purposes. Such microdata are usually released as a table, in which each record (i.e., row) corresponds to a particular individual and each column represents an attribute of the individuals. The released data usually contain *sensitive attributes*, such as Disease and Salary, which, once leaked to unauthorized parties, could be maliciously utilized and harm the individuals. Therefore, those features that can directly identify individuals, e.g., Name and Social Security Number, should be removed from the released table. See Table 1 for example of an (imagined) microdata table that a hospital prepares to release for medical research. (Note that the IDs in the first column are only for simplicity of reference, but not part of the table.)

*Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. Email: lianghy08@mails.tsinghua.edu.cn. Supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, and the National Natural Science Foundation of China Grant 61033001, 61061130540, 61073174.

†Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China. Email: haoyuan@cityu.edu.hk. Supported by the Research Grants Council of Hong Kong under grant 9041688 (CityU 124411).

	Quasi-identifiers			Sensitive
	Zipcode	Age	Education	Disease
1	98765	38	Bachelor	Viral Infection
2	98654	39	Doctorate	Heart Disease
3	98543	32	Master	Heart Disease
4	97654	65	Bachelor	Cancer
5	96689	45	Bachelor	Viral Infection
6	97427	33	Bachelor	Viral Infection
7	96552	54	Bachelor	Heart Disease
8	97017	69	Doctorate	Cancer
9	97023	55	Master	Cancer
10	97009	62	Bachelor	Cancer

Table 1: The raw microdata table.

	Quasi-identifiers			Sensitive
	Zipcode	Age	Education	Disease
1	98⋆⋆⋆	3⋆	⋆	Viral Infection
2	98⋆⋆⋆	3⋆	⋆	Heart Disease
3	98⋆⋆⋆	3⋆	⋆	Heart Disease
4	9⋆⋆⋆⋆	⋆⋆	Bachelor	Cancer
5	9⋆⋆⋆⋆	⋆⋆	Bachelor	Viral Infection
6	9⋆⋆⋆⋆	⋆⋆	Bachelor	Viral Infection
7	9⋆⋆⋆⋆	⋆⋆	Bachelor	Heart Disease
8	970⋆⋆	⋆⋆	⋆	Cancer
9	970⋆⋆	⋆⋆	⋆	Cancer
10	970⋆⋆	⋆⋆	⋆	Cancer

Table 2: A 3-anonymous partition.

Nonetheless, even with unique identifiers removed from the table, sensitive personal information can still be disclosed due to the *linking attacks* [27, 28], which try to identify individuals from the combination of *quasi-identifiers*. The quasi-identifiers are those attributes that can reveal partial information of the individual, such as Gender, Age, and Hometown. For instance, consider an adversary who knows that one of the records in Table 1 corresponds to Bob. In addition he knows that Bob is around thirty years old and has a Master's Degree. Then he can easily identify the third record as Bob's and thus learns that Bob has a heart disease.

A widely-adopted approach for protecting privacy against such attacks is *generalization*, which partitions the records into disjoint groups and then transforms the quasi-identifier values in each group to the same form. (The sensitive attribute values are not generalized because they are usually the most important data for research.) Such generalization needs to satisfy some *anonymization principles*, which are designed to guarantee data privacy to a certain extent.

The earliest (and probably most famous) anonymization principle is the k-*anonymity* principle proposed by Samarati [27] and Sweeney [28], which requires each group in the partition to have size at least k for some pre-specified value of k; such a partition is called k-*anonymous*. Intuitively, this principle ensures that every combination of quasi-identifier values appeared in the table is

	Quasi-identifiers			Sensitive
	Zip Code	Age	Education	Disease
1	98⋆⋆⋆	3⋆	⋆	Viral Infection
2	98⋆⋆⋆	3⋆	⋆	Heart Disease
3	9⋆⋆⋆⋆	⋆⋆	⋆	Heart Disease
5	9⋆⋆⋆⋆	⋆⋆	⋆	Viral Infection
8	9⋆⋆⋆⋆	⋆⋆	⋆	Cancer
9	9⋆⋆⋆⋆	⋆⋆	⋆	Cancer
4	97⋆⋆⋆	⋆⋆	Bachelor	Cancer
6	97⋆⋆⋆	⋆⋆	Bachelor	Viral Infection
7	9⋆⋆⋆⋆	⋆⋆	Bachelor	Heart Disease
10	9⋆⋆⋆⋆	⋆⋆	Bachelor	Cancer

Table 3: A 2-diverse partition of Table 1.

indistinguishable from at least $k-1$ other records, and hence protects the individuals from being uniquely recognized by linking attacks. The k-anonymity principle has been extensively studied, partly due to the simplicity of its statement. Table 2 is an example of a 3-anonymous partition of Table 1, which applies the commonly-used *suppression method* to generalize the values in the same group, i.e., *suppresses* the conflicting values with a new symbol '⋆'.

A potential issue with the k-anonymity principle is that it is totally independent of the sensitive attribute values. This issue was formally raised by Machanavajjhala et al. [19] who showed that k-anonymity is insufficient to prevent disclosure of sensitive values against the *homogeneity attack*. For example, assume that an attacker knows that one record of Table 2 corresponds to Danny, who is an elder with a Doctorate Degree. From Table 2 he can easily conclude that Danny's record must belong to the third group, and hence knows Danny has a cancer since all people in the third group have the same disease. To forestall such attacks, Machanavajjhala et al. [19] proposed the *l-diversity* principle, which demands that at most a $1/l$ fraction of the records can have the same sensitive value in each group; such a partition is called *l-diverse*. Table 3 is an example of a 2-diverse partition of Table 1. (There are some other formulations of l-diversity, e.g., one requiring that each group comprises at least l different sensitive values.)

Li et al. [17] observed that the l-diversity principle is still insufficient to protect sensitive information disclosure against the *skewness attack*, in which the attacker has partial knowledge of the *overall* sensitive value distribution. Moreover, since l-diversity only cares whether two sensitive values are distinct or not, it fails to well support sensitive attributes with semantic similarities, such as numerical attributes (e.g., the salary).

To fix these drawbacks, Li et al. [17] introduced the *t-closeness* principle, which requires that the sensitive value distribution in any group differs from the overall sensitive value distribution by at most a threshold t. There is a metric space defined on the set of possible sensitive values, in which the maximum distance of two points (i.e., sensitive values) in the space is normalized to 1. The distance between two probability distributions of sensitive values are then measured by the *Earth-Mover Distance* (EMD) [26], which is widely used in many areas of computer science. Intuitively, the EMD measures the minimum amount of work needed to transform one probability distribution to another by means of moving distribution mass between points in the probability space. The EMD between two distributions in the (normalized) space is always between 0 and 1.

We will give an example of a t-closeness partition of Table 1 for some threshold t later in Section 2, after the related notation and definitions are formally introduced.

The t-closeness principle has been widely acknowledged as an enhanced principle that fixes the main drawbacks of previous approaches like k-anonymity and l-diversity. There are also many other principles proposed to deal with different attacks or for use of ad-hoc applications; see, e.g., [3, 20, 22, 23, 29, 30, 32, 33] and the references therein.

1.1 Theoretical Models of Anonymization

It is always assumed that the released table itself satisfies the considered principle (k-anonymity, l-diversity, or t-closeness), since otherwise there exists no feasible solution at all. Therefore, the trivial partition that puts all records in a single group always guarantees the principle to be met. However, such a solution is useless in real-world scenarios, since it will most probably produce a table full of '$\star$'s, which is undesirable in most applications. This extreme example demonstrates the importance of finding a balance between data privacy and *data integrity*.

Meyerson and Williams [21] proposed a framework for theoretically measuring the data integrity, which aims to find a partition (under certain constraints such as k-anonymity) that minimizes the number of suppressed cells (i.e., '$\star$'s) in the table. This model has been widely adopted for theoretical investigations of anonymization principles. Under this model, k-anonymity and l-diversity have been extensively studied; more detailed literature reviews will be given later.

However, in contrast to k-anonymity and l-diversity, the theoretical aspects of the t-closeness principle have not been well explored before. There are only a handful of algorithms designed for achieving t-closeness [17, 18, 8, 25]. The algorithms given by Li et al. [17, 18] incorporate t-closeness into k-anonymization frameworks (Incognito [14] and Mondrian [15]) to find a t-closeness partition. Cao et al. [8] proposed the SABRE algorithm, which is the first framework tailored for t-closeness. The information-theoretic approach in [25] works for an "average" version of t-closeness. None of these algorithms is guaranteed to have good worst-case performance. Furthermore, to the best of our knowledge, no computational complexity results of t-closeness have been reported in the literature.

1.2 Our Contributions

In this paper, we initiate the first systematic theoretical study on the t-closeness principle under the commonly-used suppression framework. First, we prove that for every constant t such that $0 \leq t < 1$, it is NP-hard to find an optimal t-closeness generalization of a given table. Notice that the problem becomes trivial when $t = 1$, since the EMD between any two sensitive value distributions is at most 1, and hence putting each record in a distinct group provides a feasible solution that does not need to suppress any value at all, which is of course optimal. Our result shows that the problem immediately becomes hard even if the threshold is relaxed to, say, 0.999. At the other extreme, a 0-closeness partition demands that the sensitive value distribution in every group must be the same with the overall distribution. This seems to restrict the sets of feasible solutions in a very strong sense, and thus one might imagine whether there exists an efficient algorithm for dealing with this special case. Our result dashes the hope for this idea. The proof of our hardness result actually consists of several different reductions. Interestingly, each of these reductions only work for a set of special values of t, but altogether they cover the full range $[0, 1)$. We note that

the hardness of t_1-closeness does not directly imply that of t_2-closeness for $t_1 \neq t_2$, since they may have very different optimal objective values.

As a by-product of our proof, we establish the NP-hardness of k-anonymity when $k = cn$, where n is the number of records and c is any constant in $(0, 1/2]$. To the best of our knowledge, this is the first hardness result for k-anonymity that works for $k = \Omega(n)$. The existing approaches for proving hardness of k-anonymity all fail to generalize to this range of k due to inherent limits of the reductions. We note that $k = n/2$ is the largest possible value for which k-anonymity can be hard, because when $k > n/2$, any k-anonymous partition can only contain one group, namely the table itself.

To complement our negative results, we also provide exact and fixed-parameter algorithms for obtaining the optimal t-closeness partition. Our exact algorithm for t-closeness runs in time $2^{O(n)} \cdot O(m)$, where n and m are respectively the number of rows and columns in the input table. Together with a reduction that we derive (Lemma 1), this gives a $2^{O(n)} \cdot O(m)$ time algorithm for k-anonymity for *all* values of k, thus generalizing the result in [4] which only works for constant k. We then prove that the problem is fixed-parameter tractable when parameterized by m and the alphabet size of the input table. This implies that an optimal t-closeness partition can be found in polynomial time if the number of quasi-identifiers and that of distinct attribute values are both small (say, constants), which is true in many real-world applications. (We say a problem is *fixed-parameter tractable* with respect to some parameters $k_1, \ldots, k_r$, if there is an algorithm solving the problem that runs in time $h(k_1, \ldots, k_r)n^{O(1)}$, where n is the size of the input and h is an arbitrary computable function depending only on the parameters. Parameterized complexity has become a very active research area. For standard notation and definitions in parameterized complexity, we refer the reader to [10].) We obtain our fixed-parameter algorithm by reducing t-closeness to a special *mixed integer linear program* in which some variables are required to take integer values while others are not. The integer linear program we derived for characterizing t-closeness may have its own interest in future applications. We note that both of our algorithms work for all values of t.

Last but not least, we review the problems of finding optimal k-anonymous and l-diverse partitions, and answer two open questions left in the literature.

- We prove that the 2-diversity problem can be solved in polynomial time, which complements the NP-hardness results for $l \geq 3$ given in [31]. (We notice that the authors of [9] claimed that 2-diversity was proved to be polynomial by [31]. However what [31] actually proved is that the special 2-diversity instances, in which there are only two distinct sensitive values, can be reduced to the matching problem and hence solved in polynomial time. They do not have results for general 2-diversity. To the best of our knowledge, ours is the first work to demonstrate the tractability of 2-diversity.)

- We then present an m-approximation algorithm for k-anonymity that runs in polynomial time for all values of k. (Recall that m is the number of quasi-identifiers.) This improves the $O(k)$ and $O(\log k)$ ratios in [1, 24] when k is relatively large compared to m. We note that the performance guarantee of their algorithms cannot be reduced even for small values of m, due to some intrinsic limitations (for example, [24] uses the tight $\Theta(\log k)$ approximation for k-set cover).

1.3 Related Work

It is known that finding an optimal k-anonymous partition of a given table is NP-hard for every fixed integer $k \geq 3$ [21], while it can be solved optimally in polynomial time when $k \leq 2$ [4]. The NP-hardness result holds even for very restricted cases, e.g., when $k = 3$ and there are only three quasi-identifiers [5, 6]. On the other hand, Blocki and Williams [4] gave a $2^{O(n)} \cdot O(m)$ time algorithm that finds an optimal k-anonymous partition when $k = O(1)$, where n and m are the number of records and attributes (i.e., rows and columns) of the input table respectively. They also showed this problem to be fixed-parameter tractable when m and $|\Sigma|$ (the alphabet size of the table) are considered as parameters. The parameterized complexity of k-anonymity has also been studied in [6, 7, 11] with respect to different parameters.

Meyerson and Williams [21] gave an $O(k \log k)$ approximation algorithm for k-anonymity, i.e., it finds a k-anonymous partition in which the number of suppressed cells is at most $O(k \log k)$ times the optimum. The ratio was later improved to $O(k)$ by Aggarwal et al. [1] and to $O(\log k)$ by Park and Shim [24]. We note that the algorithms in [21, 24] run in time $n^{O(k)}$, and hence are guaranteed to be polynomial only if $k = O(1)$, while the algorithm in [1] has a truly polynomial running time for all k. There are also a number of heuristic algorithms for k-anonymity (e.g., Incognito [14]), which work well in many real datasets but have poor worst-case performance guarantee.

Xiao et al. [31] are the first to establish a systematic theoretical study on l-diversity. They showed that finding an optimal l-diverse partition is NP-hard for every fixed integer $l \geq 3$ even if m, the number of quasi-identifiers, is any fixed integer not smaller than l. They also provided an $(l \cdot m)$-approximation algorithm. Dondi et al. [9] proved an inapproximability factor of $c \ln(l)$ for l-diversity where $c > 0$ is some constant, and showed that the problem remains APX-hard even if $l = 4$ and the table consists of only three columns. They also presented an m-approximation algorithm when the number of distinct sensitive values is constant, and gave some parameterized hardness results and algorithms.

1.4 Paper Organization

The rest of this paper is organized as follows. Section 2 introduces notation and definitions used throughout the paper, and then formally defines the problems. Section 3 is devoted to proving the hardness of finding the optimal t-closeness partition, while Section 4 provides exact and parameterized algorithms. Sections 5 and 6 present our results for k-anonymity and 2-diversity, respectively. Finally, the paper is concluded in Section 7 with some discussions and future research directions.

2 Preliminaries

We consider a raw database that contains m quasi-identifiers (QIs) and a sensitive attribute (SA).* Each record t in the database is an $(m+1)$-dimensional vector drawn from Σ^{m+1}, where Σ is the alphabet of possible values of the attributes. For $1 \leq i \leq m$, $t[i]$ is the value of the i-th QI of t, and $t[m+1]$ is the value of the SA of t. Let $\Sigma_s \subseteq \Sigma$ be the alphabet of possible SA values. A microdata table (or table, for short) $\mathcal{T}$ is a multiset of vectors (or rows) chosen from Σ^{m+1}, and we denote by $|\mathcal{T}|$ the size of $\mathcal{T}$, i.e., the number of vectors contained in $\mathcal{T}$. We will let $n = |\mathcal{T}|$ when the table $\mathcal{T}$

*Following previous approaches, we only consider instances with one sensitive attribute. Our hardness result indicates that one sensitive attribute already makes the problem NP-hard. Meanwhile, it is easy to verify that our algorithms also work for the case where multiple sensitive attributes exist.

	Quasi-identifiers			Sensitive
	Zip Code	Age	Education	Disease
1	98765	38	Bachelor	Viral Infection
2	98654	39	Doctorate	Heart Disease
3	98543	32	Master	Heart Disease
After generalization:				
1	98⋆⋆⋆	3⋆	⋆	Viral Infection
2	98⋆⋆⋆	3⋆	⋆	Heart Disease
3	98⋆⋆⋆	3⋆	⋆	Heart Disease

Table 4: The first three records in Table 1.

is clear in the context. Note that $\mathcal{T}$ may contain identical vectors since it can be a multiset. We also use $\mathcal{T}[j]$ to denote the j-th vector in $\mathcal{T}$ under some ordering, e.g., $\mathcal{T}[3][m+1]$ is the SA value of the third vector of $\mathcal{T}$.

Let $\star$ be a fresh character not in Σ. For each vector $t \in \mathcal{T}$, let t^* be the *suppressor* of t (inside $\mathcal{T}$) defined as follows:

- $t^*[m+1] = t[m+1]$;
- for $1 \le i \le m$, $t^*[i] = t[i]$ if $t[i] = t'[i]$ for all $t' \in \mathcal{T}$, and $t^*[i] = \star$ otherwise.

The *cost* of a suppressor t^* is $cost(t^*) = |\{1 \le i \le m \mid t^*[i] = \star\}|$, i.e., the number of '$\star$'s in t^*. It is easy to see that all vectors in $\mathcal{T}$ have the same suppressor if we only consider the quasi-identifiers. The *generalization* of $\mathcal{T}$ is defined as $Gen(\mathcal{T}) = \{t^* \mid t \in \mathcal{T}\}$. (Note that $Gen(\mathcal{T})$ is also a multiset.) The *cost* of the generalization of $\mathcal{T}$ is $cost(\mathcal{T}) = \sum_{t^* \in Gen(\mathcal{T})} cost(t^*)$, i.e., the sum of costs of all the suppressors. Since all suppressors in $\mathcal{T}$ have the same cost, we can equivalently write $cost(\mathcal{T}) = |\mathcal{T}| \cdot cost(t^*)$ for any $t^* \in Gen(\mathcal{T})$.

As an illustrative example, Table 4 consists of the first three record of Table 1, which contains eight QIs (we regard each digit of Zip-code and Age as a separate QI) and one SA. The generalization of Table 4 is also shown. In this case all suppressors have cost 5, and the cost of this generalization is $5 \cdot 3 = 15$.

A *partition* $\mathcal{P}$ of table $\mathcal{T}$ is a collection of pairwise disjoint non-empty subsets of $\mathcal{T}$ whose union equals $\mathcal{T}$. Each subset in the partition is called a *group* or a *sub-table*. The cost of the partition $\mathcal{P}$, denoted by $cost(\mathcal{P})$, is the sum of costs of all its groups. For example, the partition of Table 1 given by Table 2 has cost $5 \cdot 3 + 6 \cdot 4 + 5 \cdot 3 = 54$.

2.1 t-Closeness Principle

We formally define the t-closeness principle introduced in [17] for protecting data privacy. Let $\mathcal{T}$ be a table, and assume without loss of generality that $\Sigma_s = \{1, 2, \ldots, |\Sigma_s|\}$. The *sensitive attribute value space* (SA space) is a normalized metric space (Σ_s, d), where $d(\cdot, \cdot)$ is a distance function defined on $\Sigma_s \times \Sigma_s$ satisfying that (1)$d(i,i) = 0$ for any $i \in \Sigma_s$; (2)$d(i,j) = d(j,i)$ for all $i, j \in \Sigma_s$; (3)$d(i,j) + d(j,k) \ge d(i,k)$ for $i, j, k \in \Sigma_s$ (this is called the triangle inequality); and (4)$\max_{i,j \in \Sigma_s} d(i,j) = 1$ (this is called the normalized condition).

For a sub-table $M \subseteq \mathcal{T}$ and $i \in \Sigma_s$, denote by $n(M,i)$ the number of vectors whose SA value equals i. Clearly $|M| = \sum_{i \in \Sigma_s} n(M,i)$. The *sensitive attribute value distribution* (SA distribution) of M, denoted by $\mathbf{P}(M)$, is a $|\Sigma_s|$-dimensional vector whose i-th coordinate is $\mathbf{P}(M)[i] =$

$n(M,i)/|M|$ for $1 \leq i \leq |\Sigma_s|$. Thus $\mathbf{P}(M)$ can be seen as the probability distribution of the SA values in M, assuming that each vector in M appears with the same probability. For a threshold $0 \leq t \leq 1$, we say M have *t-closeness* (with $\mathcal{T}$) if $\text{EMD}(\mathbf{P}(M), \mathbf{P}(\mathcal{T})) \leq t$, where $\text{EMD}(\mathbf{X}, \mathbf{Y})$ is the *Earth-Mover Distance* (EMD) between distributions $\mathbf{X}$ and $\mathbf{Y}$ [26]. A *t-closeness partition* of $\mathcal{T}$ is one in which every group has t-closeness with $\mathcal{T}$.

Intuitively, the EMD measures the minimum amount of work needed to transform one probability distribution to another by means of moving distribution mass between points in the probability space; here a unit of work corresponds to moving a unit amount of probability mass by a unit of ground distance. The EMD between two SA distributions $\mathbf{X}$ and $\mathbf{Y}$ can be formally defined as the optimal objective value of the following linear program [26, 17]:

$$\begin{aligned}
\text{Minimize} \sum_{i=1}^{|\Sigma_s|}\sum_{j=1}^{|\Sigma_s|} d(i,j)f(i,j) \qquad & \text{subject to:} \\
\sum_{j=1}^{|\Sigma_s|} f(i,j) = \mathbf{X}[i], \qquad & \forall 1 \leq i \leq |\Sigma_s| \\
\sum_{i=1}^{|\Sigma_s|} f(i,j) = \mathbf{Y}[j], \qquad & \forall 1 \leq j \leq |\Sigma_s| \\
f(i,j) \geq 0, \qquad & \forall 1 \leq i,j \leq |\Sigma_s|.
\end{aligned}$$

The above constraints are a little different from those in [17]; however they can be proved equivalent using the triangle inequality condition of the SA space. It is also easy to see that $\text{EMD}(\mathbf{X}, \mathbf{Y}) = \text{EMD}(\mathbf{Y}, \mathbf{X})$. By the normalized condition of the SA space, we have $0 \leq \text{EMD}(\mathbf{X}, \mathbf{Y}) \leq 1$ for any SA distributions $\mathbf{X}$ and $\mathbf{Y}$.

The *equal-distance space* refers to a special SA space in which each pair of distinct sensitive values have distance exactly 1. There is a concise formula for computing the EMD between two SA distributions in this space.

Fact 1 ([17])**.** *For any two SA distributions* $\mathbf{X}$ *and* $\mathbf{Y}$ *in the equal-distance space, we have*

$$\text{EMD}(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\sum_{i=1}^{|\Sigma_s|} |\mathbf{X}[i] - \mathbf{Y}[i]| = \sum_{1 \leq i \leq |\Sigma_s| : \mathbf{X}[i] \geq \mathbf{Y}[i]} (\mathbf{X}[i] - \mathbf{Y}[i]).$$

Therefore, in the equal-distance space, the EMD coincides with the *total variation distance* between two distributions.

Let us go back to Table 1 for an example. We let 1,2,and 3 denote the sensitive values "Viral Inspection", "Heart Disease", and "Cancer", respectively. Let the SA space be the equal-distance space. The SA distribution of the whole table is then $(0.3, 0.3, 0.4)$. Suppose we set the threshold $t = 0.3$. It can be verified that Table 3, although being a 2-diverse partition, is not a t-closeness partition of Table 1. In fact, the SA distribution of the first group is $(0.5, 0.5, 0)$, and hence the EMD between it and the overall distribution is 0.4. (This example also reflects some property of the skewness attack that l-diversity suffers from. If an attacker can locate the record of Alice in the first group of Table 3, then he knows that Alice does not have a cancer. If he in addition knows that Alice comes from some district where people have a very low chance to have heart disease, then he

	Quasi-identifiers			Sensitive
	Zipcode	Age	Education	Disease
1	9⋆⋆⋆⋆	⋆⋆	⋆	Viral Infection
2	9⋆⋆⋆⋆	⋆⋆	⋆	Heart Disease
4	9⋆⋆⋆⋆	⋆⋆	⋆	Cancer
3	9⋆⋆⋆⋆	⋆⋆	⋆	Heart Disease
5	9⋆⋆⋆⋆	⋆⋆	⋆	Viral Infection
8	9⋆⋆⋆⋆	⋆⋆	⋆	Cancer
9	9⋆⋆⋆⋆	⋆⋆	⋆	Cancer
6	9⋆⋆⋆⋆	⋆⋆	Bachelor	Viral Infection
7	9⋆⋆⋆⋆	⋆⋆	Bachelor	Heart Disease
10	9⋆⋆⋆⋆	⋆⋆	Bachelor	Cancer

Table 5: A 0.3-closeness partition

would be confident that Alice has a viral infection.) We instead give a 0.3-closeness partition in Table 5. We can actually verify that it is even a 0.1-closeness partition.

Now we are ready to define the main problem studied in this paper.

Problem 1. *Given an input table $\mathcal{T}$, an SA space (Σ_s, d), and a threshold $t \in [0, 1]$, the* t-Closeness *problem requires to find a t-closeness partition of $\mathcal{T}$ with minimum cost.*

Finally we review another two widely-used principles for privacy preserving, namely k-anonymity and l-diversity, and the combinatorial problems associated with them. A partition is called *k-anonymous* if all its groups have size at least k. A (sub-)table $\mathcal{M}$ is called *l-diverse* if at most $|\mathcal{M}|/l$ of the vectors in $\mathcal{M}$ have an identical SA value. A partition is called l-diverse if all its groups are l-diverse.

Problem 2. *Let $\mathcal{T}$ be a table given as input. The* k-Anonymity *(*l-Diversity*) problem requires to find a k-anonymous (l-diverse) partition of $\mathcal{T}$ with minimum cost.*

3 NP-hardness Results

In this section we study the complexity of the t-Closeness problem. The problem is trivial if the given threshold is $t = 1$, since putting each vector in a distinct group produces a 1-closeness partition with cost 0, which is obviously optimal. Our main theorem stated below indicates that this is in fact the only easy case.

Theorem 1. *For any constant t such that $0 \leq t < 1$,* t-Closeness *is NP-hard.*

We will prove Theorem 1 via several reductions, each covering a particular range of t, which altogether prove the theorem. We first present a result that relates t-Closeness to k-Anonymity.

Lemma 1. *There is a polynomial-time reduction from* k-Anonymity *to* t-Closeness *with equal-distance space and $t = 1 - k/n$.*

Proof. Let $\mathcal{T}$ be an input table of k-Anonymity. We properly change the SA values of vectors in T to ensure that all their SA values are distinct; this can be done because the SA values are irrelevant

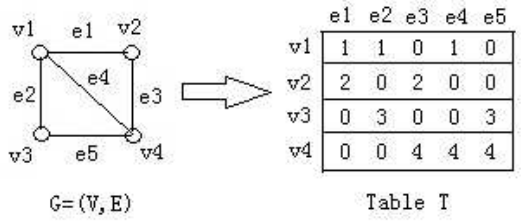

Figure 1: Reduction from MinBisection to $(n/2)$-Anonymity.

to the objective of the k-Anonymity problem. Assume w.l.o.g. that the SA values are $\{1, 2, \ldots, n\}$. Consider an instance of t-Closeness with the same input table $\mathcal{T}$, in which $t = 1 - k/n$ and the SA space is the equal-distance space. The SA distribution of $\mathcal{T}$ is $(1/n, 1/n, \ldots, 1/n)$. In the SA distribution of each size-r group $\mathcal{T}_r$, there are exactly r coordinates equal to $1/r$ and $n - r$ coordinates equal to 0. It is easy to see that $\text{EMD}(\mathbf{P}(\mathcal{T}), \mathbf{P}(\mathcal{T}_r)) = (n - r)(1/n) = 1 - r/n$. Hence, a group has t-closeness if and only if it is of size at least k. Therefore, each k-anonymous partition of $\mathcal{T}$ is also a t-closeness partition, and vice versa. The lemma follows. □

By Lemma 1 we can directly deduce the NP-hardness of t-Closeness when the threshold t is given as input, using e.g. the NP-hardness of 3-Anonymity [21]. However, to show hardness for constant t that is bounded away from 1, we need $k/n = \Omega(1)$ and thus $k = \Omega(n)$. Unfortunately, the existing hardness results for k-Anonymity only work for $k = O(1)$ and cannot be generalized to large values of k. For example, most hardness proofs use reductions from the k-dimensional matching problem, but this problem can be solved in polynomial time when $k = \Omega(n)$. Below we show the NP-hardness of k-Anonymity for $k = \Omega(n)$ via reductions different from all previous approaches in the literature.

Theorem 2. *For any constant c such that $0 < c \leq 1/2$, (cn)-Anonymity is NP-hard.*

To the best of our knowledge, Theorem 2 is the first hardness result for k-Anonymity when $k = \Omega(n)$. We note that the constant $1/2$ is the best possible, since for any $k > n/2$, a k-anonymous partition can only contain one group, namely the table itself. We first prove the following result, which will be used as a starting point in further reductions.

Theorem 3. $(n/2)$-Anonymity *is NP-hard.*

Proof. We will present a polynomial-time reduction from the minimum graph bisection (MinBisection) problem to $(n/2)$-Anonymity. MinBisection is a well-known NP-hard problem [12, 13] defined as follows: given an undirected graph, find a partition of its vertices into two equal-sized halves so as to minimize the number of edges with exactly one endpoint in each half.

Let $G = (V, E)$ be an input graph of MinBisection, where $|V| \geq 4$ is even. Suppose $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$. In what follows we construct a table $\mathcal{T}$ of size $n = |V|$ that contains $m = |E|$ quasi-identifiers. (The sensitive attributes are useless in k-Anonymity so they will not appear.) This table will serve as the input to the k-Anonymity problem with $k = n/2$. Intuitively each row (or vector) of $\mathcal{T}$ corresponds to a vertex in V, while each column (or QI) of $\mathcal{T}$ corresponds to an edge in E. The alphabet Σ is $\{1, 2, \ldots, n\}$. For each $i \in [n]$ and $j \in [m]$,[†] let $\mathcal{T}[i][j] = i$ if $v_i \in e_j$, and $\mathcal{T}[i][j] = 0$ if $v_i \notin e_j$. Thus each column contains exactly

[†]We use $[q]$ to interchangeably denote $\{1, 2, \ldots, q\}$.

two non-zero elements corresponding to the two endpoints of the associated edge. See Figure 1 for a toy example. It is easy to see that $\mathcal{T}$ can be constructed in polynomial time.

Before delving into the reduction, we first prove a result concerning the partition cost of $\mathcal{T}$. Any $(n/2)$-anonymous partition of $\mathcal{T}$ contains at most two groups. For the trivial partition that only contains $\mathcal{T}$ itself, the cost is $n \cdot m$ because all elements in $\mathcal{T}$ should be suppressed. Thus an $(n/2)$-anonymous partition with minimum cost should consist of exactly two groups. Suppose $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2\}$ is an $(n/2)$-anonymous partition of $\mathcal{T}$ where $|\mathcal{T}_1| = |\mathcal{T}_2| = n/2$. Let $\{V_1, V_2\}$ be the corresponding partition of V (recall that each vector in $\mathcal{T}$ corresponds to a vertex in V). Consider $Gen(\mathcal{T}_1)$, the generalization of $\mathcal{T}_1$. For any column $j \in [m]$, if some endpoint of e_j, say v_i, belongs to V_1, then $\mathcal{T}[i][j] = i$. By our construction of $\mathcal{T}$, any other element in the j-th column does not equal to i. Since $|\mathcal{T}_1| \geq n/2 \geq 2$, column j of $\mathcal{T}_1$ must be suppressed to $\star$. On the other hand, if none of e_j's endpoints belongs to V_1, then column j of $\mathcal{T}_1$ contains only zeros and thus can stay unsuppressed. Therefore, we obtain

$$cost(\mathcal{T}_1) = |\mathcal{T}_1| \cdot (|E_{11}| + |E_{12}|) = n(|E_{11}| + |E_{12}|)/2, \tag{1}$$

where E_{pq} denotes the set of edges with one endpoint in V_p and another in V_q, for $p, q \in \{1, 2\}$. Similarly we have $cost(\mathcal{T}_2) = n(|E_{22}| + |E_{12}|)/2$. Hence the cost of the partition $\mathcal{P}$ is

$$cost(\mathcal{P}) = \sum_{p=1}^{2} cost(\mathcal{T}_p) = n(|E| + |E_{12}|)/2, \tag{2}$$

noting that $|E| = |E_{11}| + |E_{12}| + |E_{22}|$.

We now prove the correctness of the reduction. Let OPT be the minimum size of any cut $\{V_1, V_2\}$ of G with $|V_1| = |V_2|$, and OPT' be the minimum cost of any $(n/2)$-anonymous partition of $\mathcal{T}$. We prove that $OPT' = n(|E| + OPT)/2$, which will complete the reduction from MinBisection to $(n/2)$-Anonymity. Let $\{V_1, V_2\}$ be the cut of G achieving the optimal cut size OPT, where $|V_1| = |V_2| = n/2$. Using notation introduced before, we have $OPT = |E_{12}|$. Let $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2\}$ where $\mathcal{T}_p = \{\mathcal{T}[i] \mid v_i \in V_p\}$ for $p \in \{1, 2\}$. Clearly $\mathcal{P}$ is an $(n/2)$-anonymous partition of $\mathcal{T}$. By Equation (2) we have $OPT' \leq cost(\mathcal{P}) = n(|E| + OPT)/2$.

On the other hand, let $\mathcal{P}' = \{\mathcal{T}_1', \mathcal{T}_2'\}$ be an $(n/2)$-anonymous partition with $cost(\mathcal{P}') = OPT'$. We have $|\mathcal{T}_1'| = |\mathcal{T}_2'| = n/2$. Consider the partition $\{V_1', V_2'\}$ of V with $V_p' = \{v_i \mid \mathcal{T}[i] \in \mathcal{T}_p'\}$ for $p \in \{1, 2\}$. Since $|V_1'| = |V_2'| = n/2$, we have $OPT \leq |E_{12}'|$ where E_{12}' denotes the set of edges with one endpoint in V_1' and another in V_2'. By Equation (2) we have $OPT' = n(|E| + |E_{12}'|)/2 \geq n(|E| + OPT)/2$. Combined with the previously obtained inequality $OPT' \leq n(|E| + OPT)/2$, we have shown that $OPT' = n(|E| + OPT)/2$. By the analyses we also know that an optimal $(n/2)$-anonymous partition of $\mathcal{T}$ can easily be transformed to an optimal equal-sized cut of G. This finishes the reduction from MinBisection to $(n/2)$-Anonymity, and completes the proof of Theorem 3. □

Theorem 4. *For any constant c such that $0 < c \leq 1/3$, (cn)-Anonymity is NP-hard.*

Proof. Fix $0 < c \leq 1/3$. We reduce $(n/2)$-Anonymity to (cn)-Anonymity, which will prove the NP-hardness of the latter due to Theorem 3. Let $\mathcal{T}$ be an instance of $(n/2)$-Anonymity with n rows and m QI columns. Choose two fresh symbols λ_1, λ_2 not appearing in $\mathcal{T}$. We construct a

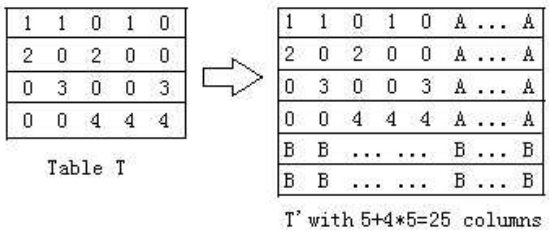

Figure 2: Reduction from $(n/2)$-ANONYMITY to $(n/3)$-ANONYMITY.

new table $\mathcal{T}'$ with $n' = n/2c$ rows and $m' = m + nm$ QI columns as follows[‡]. For all $1 \le i \le n$, let $\mathcal{T}'[i][j] = \mathcal{T}[i][j]$ for $1 \le j \le m$, and $\mathcal{T}'[i][j] = \lambda_1$ for $m+1 \le j \le m'$. For all $n+1 \le i \le n'$ and $1 \le j \le m'$, let $\mathcal{T}'[i][j] = \lambda_2$. This finishes the description of $\mathcal{T}'$. See Figure 2 for an example where $c = 1/3$, $\lambda_1 =$ 'A', and $\lambda_2 =$ 'B'. Clearly $\mathcal{T}'$ can be constructed in polynomial time.

Let OPT denote the minimum cost of an $(n/2)$-anonymous partition of $\mathcal{T}$ and OPT' be the minimum cost of a (cn')-anonymous partition of $\mathcal{T}'$. We next prove $OPT = OPT'$, which will complete the reduction from $(n/2)$-ANONYMITY to (cn)-ANONYMITY.

On one hand, let $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2\}$ be an $(n/2)$-anonymous partition of $\mathcal{T}$ with $cost(\mathcal{P}) = OPT$. We have $|\mathcal{T}_1| = |\mathcal{T}_2| = n/2$. Define a partition $\mathcal{P}'$ of $\mathcal{T}'$ as $\{\mathcal{T}'_1, \mathcal{T}'_2, \mathcal{T}'_3\}$, where $\mathcal{T}'_p = \{\mathcal{T}'[i] \mid \mathcal{T}[i] \in \mathcal{T}_p, 1 \le i \le n\}$ for $p \in \{1, 2\}$, and $\mathcal{T}'_3 = \{\mathcal{T}'[i] \mid n+1 \le i \le n'\}$. We have $|\mathcal{T}'_1| = |\mathcal{T}'_2| = n/2 \ge c(n/2c) = cn'$, and $|\mathcal{T}'_3|/n' = (n/2c - n)/(n/2c) = 1 - 2c \ge c$ as $c \le 1/3$. Hence $\mathcal{P}'$ is a (cn)-anonymous partition of $\mathcal{T}'$. It is easy to verify that $cost(\mathcal{P}') = cost(\mathcal{P}) = OPT$, implying that $OPT' \le OPT$.

On the other hand, let $\mathcal{P}' = \{\mathcal{T}'_1, \ldots, \mathcal{T}'_r\}$ be a (cn)-anonymous partition of $\mathcal{T}'$ with $cost(\mathcal{P}') = OPT'$. For the simplicity of expression, we call $\mathcal{T}'[i]$ an *old* row if $1 \le i' \le n$, and call it a *new* row if $n+1 \le i \le n'$. First assume that there exists $\mathcal{T}'_p \in \mathcal{P}'$ that contains both an old row and a new row. By our construction of $\mathcal{T}'$, an old row and a new row differ in all the last nm coordinates, and thus the cost for generalizing $\mathcal{T}'_p$ is at least $2nm$. Since $OPT \le nm$, we have $OPT' > OPT$ in this case, which cannot happen since we already proved $OPT' \le OPT$. Therefore, for any $\mathcal{T}'_p \in \mathcal{P}'$, it either contains only old rows or contains only new rows. Assume w.l.o.g. that $\mathcal{T}'_1, \ldots, \mathcal{T}'_{r'}$ are the sub-tables in $\mathcal{P}'$ that contain only old rows. Since all new rows are identical by our construction, we have

$$OPT' = cost(\mathcal{P}') = \sum_{p=1}^{r'} cost(\mathcal{T}'_p). \tag{3}$$

We now define a partition of $\mathcal{T}$ as $\mathcal{P} = \{\mathcal{T}_1, \ldots, \mathcal{T}_{r'}\}$, where $\mathcal{T}_p = \{\mathcal{T}[i] \mid \mathcal{T}'[i] \in \mathcal{T}'_p\}$ for all $1 \le p \le r'$. Because $|\mathcal{T}_p| = |\mathcal{T}'_p| \ge cn' = c(n/2c) = n/2$, $\mathcal{P}$ is an $(n/2)$-anonymous partition of $\mathcal{T}$. As the last $nm+1$ columns are identical for all old rows of $\mathcal{T}'$, we have $OPT \le cost(\mathcal{P}) = \sum_{p=1}^{r'} cost(\mathcal{T}_p) = \sum_{p=1}^{r'} cost(\mathcal{T}'_p) = OPT'$ by Equation (3). Combined with that $OPT' \le OPT$ obtained previously, we obtain that $OPT = OPT'$, and that an optimal (cn)-anonymous partition of $\mathcal{T}'$ can be easily transferred to an optimal $(n/2)$-anonymous partition of $\mathcal{T}$. This finishes the reduction from $(n/2)$-ANONYMITY to (cn)-ANONYMITY, and completes the proof of Theorem 4. □

So far we have shown the hardness of (cn)-ANONYMITY for $c \in (0, 1/3] \cup \{1/2\}$. For the remaining case $c \in (1/3, 1/2)$ we need a different reduction.

[‡]Here we assume $n/2c$ is an integer, otherwise we can use $\lfloor n/2c \rfloor$ instead and get the same result, with more tedious analyses. Similar issues appear also in other proofs, which we will not mention again.

Theorem 5. *For any constant c such that $1/3 < c < 1/2$, (cn)-*Anonymity *is NP-hard.*

Proof. Fix $1/3 < c < 1/2$. We will present a polynomial reduction from the following problem to (cn)-Anonymity: given an undirected graph $G = (V, E)$, decide whether G contains a clique (i.e., a subgraph in which every pair of vertices have an edge between them) that contains exactly $|V|/2$ vertices. Call this problem HalfClique. The NP-hardness of HalfClique easily follows from that of the well-known maximum clique problem, as can be seen as follows. We reduce the classical Clique problem to HalfClique. Given a graph $G = (V, E)$ and an integer $k \leq |V|$, the Clique problem asks whether G contains a clique with exactly k vertices. This is a well-known NP-hard problem [12]. Now construct another graph G' based on G as follows: if $k \geq |V|/2$, then add $2k - |V|$ new isolated vertices to V; if $k < |V|/2$, then add $|V| - 2k$ new vertices to V and connecting them with each other as well as all original vertices in V. Let V' be the new vertex set. It is easy to verify that G has a clique of size k if and only if G' has a clique of size $|V'|/2$, which completes the reduction.

Let $G = (V, E)$ be an input graph of HalfClique with $|V| = n \geq 4$ and $|E| = m$. Assume $V = \{v_1, \ldots, v_n\}$ and $E = \{e_1, \ldots, e_m\}$. We construct a table $\mathcal{T}$ with $n' = n/2c$ rows and m QI columns as follows. For $1 \leq i \leq n$ and $1 \leq j \leq m$, let $\mathcal{T}[i][j] = i$ if $v_i \in e_j$, and $\mathcal{T}[i][j] = 0$ otherwise. For $n + 1 \leq i \leq n'$ and $1 \leq j \leq m$, let $\mathcal{T}[i][j] = i$. (Note that, in some sense, this construction can be seen as a combination of those used in the proof of Theorems 3 and 4; however the analysis will be different and more intriguing.)

We first prove a result regarding the structure of an optimal (cn)-partition of $\mathcal{T}$. Call $\mathcal{T}[i]$ an old row if $1 \leq i \leq n$, and a new row if $n + 1 \leq i \leq n'$. We assume that $\frac{n}{2c} - n \geq 2$, i.e., $\mathcal{T}$ contains at least two new rows; this is without loss of generality because c is a constant smaller than $1/2$. Since $c > 1/3$, any (cn)-anonymous partition contains at most two groups. The trivial partition that consists of $\mathcal{T}$ itself need to suppress every coordinate in the table, because a new row and an old row do not share common values. Therefore, the minimum cost (cn)-partition of $\mathcal{T}$ contains exactly two groups.

Denote by OPT the minimum cost of a (cn)-anonymous partition of $\mathcal{T}$. We claim that G contains a clique of size $n/2$ if and only if $OPT \leq n'm - (n/2)\binom{n/2}{2}$. First consider the "only if" part. Assume $V_2 \subseteq V$ is a clique of size $n/2$, and let $V_1 = V \setminus V_2$. Then $|V_2| = |V_1| = n/2$. For $p, q \in \{1, 2\}$, denote by E_{pq} the set of edges with one endpoint in V_p and another in V_q. We define a partition $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2\}$ of $\mathcal{T}$ by letting $\mathcal{T}_1 = \{\mathcal{T}[i] \mid v_i \in V_1\}$ and $\mathcal{T}_2 = \mathcal{T} \setminus \mathcal{T}_1'$. Since $|\mathcal{T}_1| = n/2 = c(n/2c) = cn'$ and $|\mathcal{T}_2|/n' = (n/2c - n/2)/(n/2c) = 1 - c \geq c$, $\mathcal{P}$ is a (cn)-anonymous partition. Similar to the proof of Theorem 3, we have $cost(\mathcal{T}_1) = n(|E_{11}|+|E_{12}|)/2$ (see Equation (1) and its proof). Since $\mathcal{T}_2$ contains both old and new rows, we have $cost(\mathcal{T}_2) = |\mathcal{T}_2| \cdot m = (n' - n/2)m$. Therefore,

$$\begin{aligned} OPT &\leq cost(\mathcal{P}) = cost(\mathcal{T}_1) + cost(\mathcal{T}_2) \\ &= n(|E_{11}| + |E_{12}|)/2 + (n' - n/2)m \\ &= n(m - |E_{22}|)/2 + (n' - n/2)m \\ &= n'm - (n/2)|E_{22}| \\ &= n'm - (n/2)\binom{n/2}{2}, \end{aligned}$$

where the last equality holds because V_2 is a clique of size $n/2$. This proves the "only if" part of the claim.

Next we consider the "if" direction. Let $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2\}$ be a (cn)-partition with $cost(\mathcal{P}) = OPT \leq n'm - (n/2)\binom{n/2}{2}$. As argued before, every sub-table that contains both old and new rows need to be suppressed totally. Thus, if both $\mathcal{T}_1$ and $\mathcal{T}_2$ contain both old and new rows, then $cost(\mathcal{P}) = n'm$, which is worst possible. In this case we can change $\mathcal{T}_1$ to be any set of $n/2$ old rows and let $\mathcal{T}_2 = \mathcal{T} \setminus \mathcal{T}_1$ to obtain a (cn)-partition with no worse cost. Therefore, in what follows we assume w.l.o.g. that $\mathcal{T}_1$ consists of only old rows.

Let $V_1 = \{v_i \mid \mathcal{T}[i] \in \mathcal{T}_1\}$ and $V_2 = V \setminus V_1$. Define E_{pq} analogously as before for $p, q \in \{1, 2\}$. Similar to the proof of Theorem 3, we have $cost(\mathcal{T}_1) = |V_1|(|E_{11}| + |E_{12}|)$ (just replace $n/2$ with $|V_1|$ in Equation (1)). Also $cost(\mathcal{T}_2) = |\mathcal{T}_2| \cdot m = (n' - |V_1|)m$ since $\mathcal{T}_2$ contains both old and new rows. Hence, $cost(\mathcal{P}) = |V_1|(|E_{11}|+|E_{22}|)+(n'-|V_1|)m = |V_1|(m-|E_{22}|)+(n'-|V_1|)m = n'm-|V_1|\cdot|E_{22}|$. On the other hand, $cost(\mathcal{P}) = OPT \leq n'm - (n/2)\binom{n/2}{2}$. Thus we have $|V_1| \cdot |E_{22}| \geq (n/2)\binom{n/2}{2}$. As $|V_1| + |V_2| = n$ and $|E_{22}| \leq \binom{|V_2|}{2}$, we obtain that

$$(n - |V_2|)\binom{|V_2|}{2} \geq |V_1| \cdot |E_{22}| \geq \frac{n}{2}\binom{n/2}{2}. \tag{4}$$

Because $|V_1| = |\mathcal{T}_1| \geq cn' = n/2$, we have $|V_2| \leq n/2$. Define a fucntion $f : [0, n/2] \to \mathbb{R}$ as $f(x) = (n - x)\binom{x}{2} = (n - x)x(x - 1)/2$ for all $0 \leq x \leq n/2$. Then Equation (4) indicates that $f(|V_2|) \geq f(n/2)$. Since $f(0) = 0$, $|V_2| \geq 1$ holds. Let f' be the derivative of f with respect to x. It is easy to verify that $f'(x) = \frac{1}{2}(-3x^2 + 2(n + 1)x - n)$. The minimum value of $f'(x)$ when $1 \leq x \leq n/2$ can only be obtained at $x \in \{1, n/2, (n + 1)/3\}$. Simple calculations show that $f'(1), f'(n/2)$, and $f'((n+1)/3)$ are all positive. Hence $f'(x) > 0$ for all $1 \leq x \leq n/2$, which means that $f(x)$ is strictly monotone increasing on $[1, n/2]$. Since we know that $f(|V_2|) \geq f(n/2)$ and that $1 \leq |V_2| \leq n/2$, it must hold that $|V_2| = n/2$. Therefore (4) holds with two equalities. We thus have $|E_{22}| = \binom{n/2}{2}$, implying that V_2 is a clique of size $n/2$.

We have shown that G has a clique of size $n/2$ if and only if $\mathcal{T}$ has a (cn)-anonymous partition of cost at most $n'm-(n/2)\binom{n}{2}$. This completes the reduction from HalfClique to (cn)-Anonymity, from which Theorem 5 follows. □

Now Theorem 2 follows straightforward from Theorems 3, 4 and 5. Interestingly, the three reductions work for disjoint ranges of c, which altogether give the desired result. By Lemma 1 we obtain:

Corollary 1. *For any constant t such that $1/2 \leq t < 1$, t-Closeness is NP-hard even with equal-distance space.*

We next show the hardness of t-Closeness for $0 \leq t < 1/2$ by two different reductions from the 3-dimensional matching problem, each of which covers a different range of t.

Theorem 6. *For any constant t such that $0 \leq t < 1/3$, t-Closeness is NP-hard even if $|\Sigma_s| = 3$.*

Proof. Fix $0 \leq t < 1/2$. We perform a polynomial-time reduction from the 3-dimensional matching problem (3D-Matching) to t-Closeness. The input of 3D-Matching consists of three equal-sized pairwise-disjoint sets X, Y, and Z, together with a collection S of 3-tuples from $X \times Y \times Z$. The goal is to decide whether there exists a set of $|X|$ tuples from S that covers each element of $X \cup Y \cup Z$ exactly once. This problem is well known to be NP-hard [12].

Consider an instance of 3D-Matching. Assume $|X| = |Y| = |Z| = n$, $U = X \cup Y \cup Z = \{v_1, v_2, \ldots, v_{3n}$, and the set of tuples is $S = \{e_1, e_2, \ldots, e_m\}$. Each tuple in S is regarded as a

subset of U of size 3. The reduction that we will use is similar to that in [31]. We construct an instance of t-CLOSENESS as follows. The table $\mathcal{T}$ has $3n$ rows and m QI columns as well as an SA column. For every $1 \leq i \leq 3n$ and $1 \leq j \leq m$, let $\mathcal{T}[i][j] = i$ if $v_i \notin e_j$ and $\mathcal{T}[i][j] = 0$ if $v_i \in e_j$. Let $\mathcal{T}[i][m+1]$ be 1, 2, or 3, if v_i belongs to X, Y, or Z, respectively. The SA space is the equal-distance space. Notice that each QI column of $\mathcal{T}$ contains exactly three zeros, corresponding to the three elements in the tuple associated with this column. Also note that the SA distribution of $\mathcal{T}$ is $\mathbf{P}(\mathcal{T}) = (1/3, 1/3, 1/3)$ since $|X| = |Y| = |Z|$.

We will prove that, there exists n tuples of S whose union equals $U = X \cup Y \cup Z$ if and only if $\mathcal{T}$ has a t-closeness partition of cost at most $3n(m-1)$. This will complete the reduction from 3D-MATCHING to t-CLOSENESS.

First consider the "only of" direction. Assume that there exists $S' \subseteq S$, $|S'| = n$, such that $\bigcup_{e \in S'} e = U$. We assume w.l.o.g. that $S' = \{e_1, e_2, \ldots, e_n\}$. Define a partition $\mathcal{P}$ of $\mathcal{T}$ as follows: $\mathcal{P} = \{\mathcal{T}_1, \ldots, \mathcal{T}_n\}$, where $\mathcal{T}_p = \{\mathcal{T}[i] \mid v_i \in e_p\}$ for all $1 \leq p \leq n$. Clearly $|\mathcal{T}_1| = \ldots = |\mathcal{T}_n| = 3$. Since each e_p contains exactly one element from each of X, Y and Z, we have $\mathbf{P}(\mathcal{T}_p) = (1/3, 1/3, 1/3) = \mathbf{P}(\mathcal{T})$. Hence $\mathcal{P}$ is a t-closeness (and in fact 0-closeness) partition of $\mathcal{T}$. By our construction, for each $p \in [n]$, the p-th column of $\mathcal{T}_p$ consists of three zeros, and every other column contains at least two different QI values. Thus $cost(\mathcal{T}_p) = 3(m-1)$, and $cost(\mathcal{P}) = \sum_{p=1}^{n} cost(\mathcal{T}_p) = 3n(m-1)$. The "only if" direction is proved.

We next consider the "if" direction. Let $\mathcal{P} = \{\mathcal{T}_1, \ldots, \mathcal{T}_r\}$ be a t-closeness partition of $\mathcal{T}$ with cost at most $3n(m-1)$. We claim that $|\mathcal{T}_p| \geq 3$ for all $p \in [r]$. Assume to the contrary that $|\mathcal{T}_p| \leq 2$ for some p. Then $\mathbf{P}(\mathcal{T}_p)$ is either $(0, 1/2, 1/2)$ or $(0, 0, 1)$ up to permutations of the coordinates. It is easy to verify that $\text{EMD}(\mathbf{P}(\mathcal{T}_p), \mathbf{P}(\mathcal{T})) \geq 1/3 > c$ in both cases, which contradicts the fact that $\mathcal{P}$ is a t-closeness partition. Hence, $|\mathcal{T}_p| \geq 3$. If $|\mathcal{T}_p| \geq 4$, then $cost(\mathcal{T}_p) = |\mathcal{T}_p| \cdot m$, because each column of $\mathcal{T}$ consists of three zeros and $3n-3$ distinct non-zero values and thus needs to be suppressed entirely in $\mathcal{T}_p$. If $|\mathcal{T}_p| = 3$, then $cost(\mathcal{T}_p) = 3(m-1)$ if there is a tuple in S that contains the three elements associated with the vectors in $\mathcal{T}_p$ (in which case the column corresponding to this tuple needs not be suppressed), and $cost(\mathcal{T}_p) = 3m$ otherwise. Since $cost(\mathcal{P}) = 3n(m-1)$, every group $\mathcal{T}_p$ is of size 3 and induces a tuple, say $e_p \in S$. Then $\{e_1, \ldots, e_p\}$ is a set of n tuples whose union equals U, proving the "if" direction. This completes the reduction from 3D-MATCHING to t-CLOSENESS, and Theorem 6 follows. □

Finally we come to the last part $t \in [1/3, 1/2)$.

Theorem 7. *For any constant t such that $1/3 \leq t < 1/2$, t-CLOSENESS is NP-hard even if $|\Sigma_s| = 4$.*

Proof. Fix $1/3 \leq t < 1/2$. We give a reduction from 3D-MATCHING to t-CLOSENESS similar to that used in the the proof of Theorem 6, with some more ingredients. Consider an instance of 3D-MATCHING. The element set is $U = X \cup Y \cup Z = \{v_1, v_2, \ldots, v_{3n}\}$ where $|X| = |Y| = |Z| = n$. The tuple set is $S = \{e_1, \ldots, e_m\}$ where each e_i, $1 \leq i \leq m$, is a subset of U of size 3 that consists of exactly one element from each of X, Y, and Z. The goal is to decide whether there exists $S' \subseteq S$, $|S'| = n$, such that $\bigcup_{e \in S'} e = U$.

We set up an instance of t-CLOSENESS as follows. The table $\mathcal{T}$ consists of $n' = 3n/(1-2t)$ rows, m QI columns, and an SA column. For all $1 \leq i \leq 3n$ and $1 \leq j \leq m$, $\mathcal{T}[i][j] = i$ if $v_i \notin e_j$ and $\mathcal{T}[i][j] = 0$ if $v_i \in e_j$. For $1 \leq i \leq 3n$, $\mathcal{T}[i][m+1] = 1$ if $v_i \in X$, $\mathcal{T}[i][m+1] = 2$ if $v_i \in Y$, and $\mathcal{T}[i][m+1] = 3$ if $v_i \in Z$. For $3n+1 \leq i \leq n'$, $\mathcal{T}[i][j] = i$ for $1 \leq j \leq m$, and $\mathcal{T}[i][m+1] = 4$. Note that $\Sigma_s = \{1, 2, 3, 4\}$. Define the distance function of the SA space as $d(1,2) = d(1,3) = d(2,3) = 1$ and $d(4,1) = d(4,2) = d(4,3) = 1/2$; this clearly forms a metric on Σ_s. It is easy to verify that

$\mathbf{P}(\mathcal{T}) = (\frac{1-2t}{3}, \frac{1-2t}{3}, \frac{1-2t}{3}, 2t)$. The goal is to decide whether $\mathcal{T}$ has a t-closeness partition. Before showing the correctness of the reduction, we present a formula for computing the EMD between two distributions under this metric. Let $\mathbf{A} = (a_1, a_2, a_3, a_4)$ and $\mathbf{B} = (b_1, b_2, b_3, b_4)$ be two SA distributions with $a_4 \geq b_4$. Then,

$$\mathrm{EMD}(\mathbf{A}, \mathbf{B}) = \frac{1}{2}(a_4 - b_4) + \sum_{i \in \{1,2,3\}: a_i \geq b_i} (a_i - b_i). \tag{5}$$

This can be seen as follows. Let $S_{\geq} = \{1 \leq i \leq 4 \mid a_i \geq b_i\}$ and $S_{<} = \{1, 2, 3, 4\} \setminus S_{\geq}$. We have $4 \in S_{\geq}$ and $\sum_{i \in S_{\geq}} (a_i - b_i) = \sum_{j \in S_{<}} (b_j - a_j)$. To transform $\mathbf{A}$ to $\mathbf{B}$, we need to move $M = \sum_{i \in S_{\geq}} (a_i - b_i)$ amount of mass from $S_{\geq}$ to $S_{<}$. $a_4 - b_4$ amount of mass at point 4 can be moved out by distance $1/2$, while the remaining amount must be moved by distance 1. Therefore $\mathrm{EMD}(\mathbf{A}, \mathbf{B}) = \frac{1}{2}(a_4 - b_4) + \sum_{i \in S_{\geq} \setminus \{4\}} (a_i - b_i) = \frac{1}{2}(a_4 - b_4) + \sum_{i \in \{1,2,3\}: a_i \geq b_i} (a_i - b_i)$.

We prove that the answer to the matching instance is yes if and only if $\mathcal{T}$ has a t-closeness partition of cost at most $3n(m-1)$. First consider the "only if" direction. Assume w.l.o.g. that $S' = \{e_1, \ldots, e_n\}$ satisfies $\bigcup_{e \in S'} e = U$. Define a partition $\mathcal{P} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n\} \cup \{\mathcal{T}'_{3n+1}, \mathcal{T}'_{3n+2}, \ldots, \mathcal{T}'_{n'}\}$, where $\mathcal{T}_p = \{\mathcal{T}[i] \mid i \in e_p\}$ for $1 \leq p \leq n$ and $\mathcal{T}'_p = \{\mathcal{T}[p]\}$ for $3n + 1 \leq p \leq n'$, i.e., each $\mathcal{T}'_p$ consists of a single row. By similar arguments as in the proof of Theorem 6, $cost(\mathcal{T}_p) = 3(m-1)$ for $1 \leq p \leq n$, and obviously $cost(\mathcal{T}'_p) = 0$ for $3n+1 \leq p \leq n'$. Hence $cost(\mathcal{P}) = 3n(m-1)$. It remains to show that $\mathcal{P}$ is a t-closeness partition. For $1 \leq p \leq n$, $\mathbf{P}(\mathcal{T}_p) = (1/3, 1/3, 1/3, 0)$, by Equation (5) we have $\mathrm{EMD}(\mathbf{P}(\mathcal{T}_p), \mathbf{P}(\mathcal{T})) = \frac{1}{2} \cdot 2t = t$. Since $\mathbf{P}(\mathcal{T}'_q) = (0, 0, 0, 1)$ for all $3n + 1 \leq q \leq n'$, $\mathrm{EMD}(\mathbf{P}(\mathcal{T}'_q), \mathbf{P}(\mathcal{T})) = \frac{1}{2}(1 - 2t) \leq t$ as $t \geq 1/3$ (actually this holds for all $t \geq 1/4$). This proves that $\mathcal{P}$ is a t-closeness partition, and hence the "only if" direction.

Now consider the "if" direction. Let $\mathcal{P} = \{\mathcal{T}_1, \ldots, \mathcal{T}_r\}$ be a t-closeness partition of $\mathcal{T}$ with cost at most $3n(m-1)$. Call $\mathcal{T}[i]$ an old row if $1 \leq i \leq 3n$, and a new row if $i > 3n$. By our construction of $\mathcal{T}$, it is clear that $cost(\mathcal{T}_p) = |\mathcal{T}_p| \cdot m$ if $|\mathcal{T}_p| \geq 2$ and $\mathcal{T}_p$ contains at least one new row. Now let $\mathcal{T}_p$ be a group containing only old rows. If $|\mathcal{T}_p| \leq 2$, then $\mathbf{P}(\mathcal{T}_p))$ is equivalent to $(1, 0, 0, 0)$ or $(1/2, 1/2, 0, 0)$ up to permutations of the first three coordinates. By (5) and the fact that $t < 1/2$, we can verify that $\mathrm{EMD}(\mathbf{P}(\mathcal{T}_p), \mathbf{P}(\mathcal{T})) > t$ in both cases. Therefore $|\mathcal{T}_p| \geq 3$. Analogous to the proof of Theorem 6, we know that $cost(\mathcal{T}_p) = 3(m-1)$ if $\mathcal{T}_p$ consists of three old rows corresponding to three elements in the same tuple, and $cost(\mathcal{T}_p) = |\mathcal{T}_p| \cdot m$ otherwise. Thus for $cost(\mathcal{P}) = 3n(m-1)$ it must be the case that there exist n groups each of which consists of three old rows, and each of the remaining groups consists of exact one new row. As groups are disjoint, they together cover all the $3n$ old rows, which naturally induces n tuples of S whose union equals U. The "if" direction is thus proved. This completes the reduction from 3D-Matching to t-Closeness, and Theorem 7 follows. □

4 Exact and Fixed-Parameter Algorithms

In this section we design exact algorithms for solving t-Closeness. Notice that the size of an instance of t-Closeness is polynomial in n and $m + 1$. The brute-force approach that examines each possible partition of the table to find the optimal solution takes $n^{O(n)}m^{O(1)} = 2^{O(n \log n)}m^{O(1)}$ time. We first improve this bound to single exponential in n. (Note that it cannot be improved to polynomial unless P = NP.)

Theorem 8. *The t-Closeness problem can be solved in $2^{O(n)} \cdot O(m)$ time.*

Proof. Consider an input table $\mathcal{T}$ of the t-Closeness problem. Assume that $\mathcal{P} = \{\mathcal{T}_1, \ldots, \mathcal{T}_r\}$ is an optimal t-closeness partition of $\mathcal{T}$ (note that we do not know $\mathcal{P}$; it is only used for analysis). Obviously there is at most one group $\mathcal{T}_p$ with $|\mathcal{T}_p| > n/2$. We claim that, if $|\mathcal{T}_p| \leq n/2$ for all $p \in [r]$, then there is a disjoint partition (A_1, A_2) of $\{1, 2, \ldots, r\}$ such that $n/4 \leq |\bigcup_{p \in A_i} \mathcal{T}_p| \leq 3n/4$ for any $i \in \{1, 2\}$. This can be seen as follows. Denote by $n(A) = |\bigcup_{p \in A} \mathcal{T}_p|$ for any $A \subseteq [r]$. Let (A_1, A_2) be the partition of $[r]$ that minimizes $|n(A_1) - n(A_2)|$. Assume w.l.o.g. that $n(A_1) \leq n(A_2)$. If $n(A_2) \leq 3n/4$, the claim is proved. Otherwise, A_2 contains at least two groups, and we move an arbitrary group from A_2 to A_1 resulting in a new partition (A_1', A_2'). If $n(A_1') \leq n(A_2')$, then $|n(A_1') - n(A_2')| = n(A_2') - n(A_1') < n(A_2) - n(A_1) = |n(A_1) - n(A_2)|$, which contradicts the way in which (A_1, A_2) is chosen. We thus have $n(A_1') > n(A_2')$, and so $n(A_1') \geq n/2$. Since each group has size at most $n/2$, we have $n/2 \leq n(A_1') \leq n(A_1) + n/2 < 3n/4$, and hence $n/2 \geq n(A_2') = n - n(A_1') > n/4$. This proves the claim.

For any $M \subseteq \mathcal{T}$, let $OPT(M)$ denote the minimum cost of any partition of M in which each group is t-close to $\mathcal{T}$; thus the optimal cost of the problem is $OPT(\mathcal{T})$. We now have a natural recursive algorithm for computing $OPT(\mathcal{T})$: Enumerate all $\mathcal{T}_1 \subseteq \mathcal{T}$ with $n/4 \leq |\mathcal{T}_1| \leq 3n/4$ and find the one minimizing $OPT(\mathcal{T}_1) + OPT(\mathcal{T} \setminus \mathcal{T}_1)$; denote this minimum cost by OPT_1. We also exhaustively find $\mathcal{T}_1' \subseteq \mathcal{T}$ with $|\mathcal{T}_1'| > n/2$ that minimizes $OPT(\mathcal{T}_1') + OPT(\mathcal{T} \setminus \mathcal{T}_1')$, which is denoted by OPT_2. By our previous analysis, $OPT(\mathcal{T}) = \min\{OPT_1, OPT_2\}$ and thus we can solve t-Closeness by taking the better solution. Two notes on the recursive steps: (1) If we have a table of constant size (say, less than 10) then we can directly solve it in $O(m)$ time by the brute-force approach. (2) If we have a table $\mathcal{T}'$ such that $\mathrm{EMD}(\mathbf{P}(\mathcal{T}'), \mathbf{P}(\mathcal{T})) > t$ then we return with cost $+\infty$.

We now analyze the running time of the algorithm. Let $f(s)$ denote the running time on a sub-table of $\mathcal{T}$ of size s. When $s \leq 10$ we have $f(s) = O(m)$, and when $s > 10$,

$$\begin{aligned} f(s) &\leq \sum_{i=s/4}^{3s/4} \binom{s}{i} \cdot 2f(3s/4) + \sum_{i=s/2}^{s} \binom{s}{i} f(s/2) + O(2^s) \\ &\leq 2^{s+2} f(3s/4) + O(2^s). \end{aligned}$$

In the first inequality, the first term stands for the time of enumerating $\mathcal{T}_1$ with $n/4 \leq |\mathcal{T}_1| \leq 3n/4$, the second term is for the enumeration of $\mathcal{T}_1'$ with $|\mathcal{T}_1'| > n/2$, and the third term is responsible for other works such as recording the subsets. It is easy to verify that this recursion gives $f(n) \leq 2^{O(n)} \cdot O(m)$. □

In many real applications, there are usually only a small number of attributes and distinct attribute values. Thus it is interesting to see whether t-Closeness can be solved more efficiently when m and $|\Sigma|$ is small. We answer this question affirmatively in terms of fixed-parameter tractability.

Theorem 9. *t-Closeness is fixed-parameter tractable when parameterized by m and $|\Sigma|$. Thus we can solve t-Closeness optimally in polynomial time when m and $|\Sigma|$ are constants.*

Proof. Consider an input table $\mathcal{T}$ with n rows and $m + 1$ columns (of which m are QIs and one is SA). For $v \in \Sigma^m$ and $s \in \Sigma_s$, denote by $R_{v,s}$ the set of vectors in $\mathcal{T}$ that is identical to (v, s), and let $r_{v,s} = |R_{v,s}|$. We thus have $\sum_{v \in \Sigma^m, s \in \Sigma_s} r_{v,s} = n$. We write a integer linear program to characterize the minimum cost of a t-closeness partition of $\mathcal{T}$. For every $v \in \Sigma^m$ and $s \in \Sigma_s$ such that $R_{v,s} \neq \emptyset$, and every $v^* \in (\Sigma \cup \{\star\})^m$ that generalizes v, there is a nonnegative integer variable

$x(v^*, v, s)$ which means the number of vectors in $R_{v,s}$ that is generalized to (v^*, s) in the partition. We clearly have

$$\sum_{v^*: v^* \text{ generalizes } v} x(v^*, v, s) = r_{v,s}, \quad \forall (v, s) \text{ s.t. } R_{v,s} \neq \emptyset. \tag{6}$$

Each $v^* \in (\Sigma \cup \{\star\})^m$ induces a group, denoted G_{v^*}, which consists of all vectors whose QI values are generalized to v^*. Those groups together form a partition (note that some group may be empty). Denoting by C_{v^*} the number of '$\star$'s in v^*, the cost of the partition is precisely $\sum_{v^*,v,s} C_{v^*} \cdot x(v^*, v, s)$. Thus the objective function is

$$\text{Minimize} \sum_{v^*,v,s} C_{v^*} \cdot x(v^*, v, s) \ . \tag{7}$$

We still need other constraints to ensure that each group G_{v^*} either is empty or has t-closeness. We do this by adding a set of constraints, for every v^*, that characterizes the transportation between the SA distributions of G_{v^*} and $\mathcal{T}$ as in the definition of EMD. First assume that G_{v^*} is non-empty. We have $|G_{v^*}| = \sum_{v \in \Sigma^m, s \in \Sigma_s} x(v^*, v, s)$. The probability mass of $i \in \Sigma_s$ in $\mathbf{P}(G_{v^*})$ is $\sum_{v \in \Sigma^m} x(v^*, v, i)/|G_{v^*}|$, and that in $\mathbf{P}(\mathcal{T})$ is $\sum_{v \in \Sigma^m} r_{v,i}/n$. For $i, j \in \Sigma_s$, let $f(v^*, i, j)$ denote the amount of mass moved from i to j in order to transform $\mathbf{P}(G_{v^*})$ to $\mathbf{P}(\mathcal{T})$. Let $d_{i,j}$ be the distance between i and j in the SA space. To guarantee the t-closeness of G_{v^*} we can write the following constraints:

$$\begin{aligned}
\sum_{j \in \Sigma_s} f(v^*, i, j) &= \sum_{v \in \Sigma^m} x(v^*, v, i)/|G_{v^*}|, \ \forall i \in \Sigma_s \\
\sum_{i \in \Sigma_s} f(v^*, i, j) &= \sum_{v \in \Sigma^m} r_{v,j}/n, \ \forall j \in \Sigma_s \\
\sum_{i,j \in \Sigma_s} d_{i,j} \cdot f(v^*, i, j) &\leq t \\
f(v^*, i, j) &\geq 0, \ \forall i, j \in \Sigma_s \ .
\end{aligned}$$

The first constraint above is not linear. To overcome this, we define $g(v^*, i, j) = f(v^*, i, j) \cdot |G_{v^*}|$, substitute $g(v^*, i, j)$ for $f(v^*, i, j)$ in the above constraints, and expand $|G_{v^*}|$. This produces the following equivalent constraints:

$$\begin{aligned}
\sum_{j \in \Sigma_s} g(v^*, i, j) &= \sum_{v \in \Sigma^m} x(v^*, v, i), \ \forall i \in \Sigma_s \\
n \sum_{i \in \Sigma_s} g(v^*, i, j) &= \sum_{v \in \Sigma^m} r_{v,j} \sum_{v \in \Sigma^m, s \in \Sigma_s} x(v^*, v, s), \forall j \in \Sigma_s \\
\sum_{i,j \in \Sigma_s} d_{i,j} \cdot g(v^*, i, j) &\leq t \cdot \sum_{v \in \Sigma^m, s \in \Sigma_s} x(v^*, v, s) \\
g(v^*, i, j) &\geq 0, \ \forall i, j \in \Sigma_s \ .
\end{aligned}$$

Note that these constraints hold even if G_{v^*} is empty. Thus they force group G_{v^*} to be t-closeness or empty. The set of such constraints for all v^*, together with (6) and (7), compose a *mixed* integer linear program (i.e., only some of the variables are required to take integer values) that precisely characterizes the t-Closeness problem on $\mathcal{T}$.[§] The number of variables in the program

[§] A technical issue here is that, in order to apply results for mixed integer linear program, t needs to be a rational number. Nevertheless, for irrational t we can use rationals to approximate the value of t to an arbitrary precision.

is $N \leq |\Sigma|^m(|\Sigma|+1)^m|\Sigma_s| + (|\Sigma|+1)^m|\Sigma_s|^2 \leq 2(|\Sigma|+1)^{2m+1}$. The time spent on constructing and writing down this linear program is polynomial in n, m, and N. By the result in [16] (Section 5 of it deals with mixed ILP), a mixed linear integer program with N variables can be solved in $N^{O(N)}L$ time, where L is the number of bits used to encode the program. In our case L is polynomial in n and m. Therefore, we can solve this program, and hence solve t-Closeness, in $h(m, |\Sigma|)n^{O(1)}$ time for some function h. This shows that t-Closeness is fixed-parameter tractable when parameterized by m and $|\Sigma|$. □

5 Approximation Algorithm for k-Anonymity

In this section we give a polynomial-time m-approximation algorithm for k-Anonymity, which improves the previous best ratio $O(k)$ [1] and $O(\log k)$ [24] when k is relatively large compared with m. (We note that the $O(\log k)$-approximation algorithm given in [24] is not guaranteed to run in polynomial time for super-constant k, while our result holds for all k.)

Theorem 10. *k-Anonymity can be approximated within factor m in polynomial time.*

Proof. Consider a table $\mathcal{T}$ with n rows and m QI columns. Denote by OPT the minimum cost of any k-anonymous partition of $\mathcal{T}$. Partition $\mathcal{T}$ into "equivalence classes" $C_1, \ldots, C_R$ in the following sense: any two vectors in the same class are identical, i.e., they have the same value on each attribute, while any two vectors from different classes differ on at least one attribute. Assume $|C_1| \leq |C_2| \leq \ldots \leq |C_R|$. If $|C_1| \geq k$, then these classes form a k-anonymous partition with cost 0, which is surely optimal. Thus we assume $|C_1| < k$, and let $L \in [R]$ be the maximum integer for which $|C_L| < k$. Then $|C_{L'}| \geq k$ for all $L < L' \leq R$. It is clear that each vector in $C_1 \cup \ldots \cup C_L$ contributes at least one to the cost of any partition of $\mathcal{T}$. Thus $OPT \geq \sum_{i=1}^{L} |C_i|$.

Case 1: $\sum_{i=1}^{L} |C_i| \geq k$. In this case we partition $\mathcal{T}$ into $R - L + 1$ groups: $\{C_1 \cup \ldots \cup C_L, C_{L+1}, C_{L+2}, \ldots, C_R\}$. This is a k-anonymous partition of cost at most $m \cdot \sum_{i=1}^{L} |C_i| \leq m \cdot OPT$.

Case 2: $\sum_{i=1}^{L} |C_i| < k$ and $\sum_{i=1}^{L} |C_i| + \sum_{i=L+1}^{R}(|C_i| - k) \geq k$. We choose $C_i' \subseteq C_i$ for $L+1 \leq i \leq R$ satisfying that $|C_i \setminus C_i'| \geq k$ and $\sum_{i=1}^{L} |C_i| + \sum_{i=L+1}^{R} |C_i'| = k$. This can be done because of the second condition of this case. We partition $\mathcal{T}$ into $R - L + 1$ groups: $\{\bigcup_{i=1}^{L} C_i \cup \bigcup_{i=L+1}^{R} C_i', C_{L+1} \setminus C_{L+1}', \ldots, C_R \setminus C_R'\}$. This is a k-anonymous partition of cost at most $m \cdot k \leq m \cdot OPT$, since $OPT \geq k$.

Case 3: $\sum_{i=1}^{L} |C_i| + \sum_{i=L+1}^{R}(|C_i| - k) < k$. We claim that there exists $i \in \{L+1, \ldots, R\}$ such that any vector in C_i contributes at least one to the cost of any k-anonymous partition. Assume the contrary. Then there exists a k-anonymous partition such that, for every $L+1 \leq i \leq R$, there is a vector $v \in C_i$ whose suppression cost is 0, which means that v belongs to a group that only contains vectors in C_i; denote this group by C_i'. We also know that there is at least one group in the partition that has positive cost. However, by removing all C_i', $L+1 \leq i \leq R$, from $\mathcal{T}$, the number of vectors left is at most $n - k(R-L) = \sum_{i=1}^{R} |C_i| - k(R-L) = \sum_{i=1}^{L} |C_i| + \sum_{i=L+1}^{R}(|C_i| - k) < k$, due to the condition of this case. This contradicts with the property of k-anonymous partitions. Therefore the claim holds, i.e., there exists $j \in \{L+1, \ldots, R\}$ such that any vector in C_j contributes at least one to the partition cost. Thus we have $OPT \geq \sum_{i=1}^{L} |C_i| + |C_j| \geq \sum_{i=1}^{L+1} |C_i|$. We partition $\mathcal{T}$ into $R - L$ groups: $\{\bigcup_{i=1}^{L+1} C_i, C_{L+2}, \ldots, C_R\}$. This is a k-anonymous partition with cost at most $m \cdot \sum_{i=1}^{L+1} |C_i| \leq m \cdot OPT$.

By the above case analyses, we can always find in polynomial time a k-anonymous partition of $\mathcal{T}$ with cost at most $m \cdot OPT$. This completes the proof of Theorem 10. □

We note that Theorem 10 implies that k-ANONYMITY can be solved optimally in polynomial time when $m = 1$. This is in contrast to l-DIVERSITY, which remains NP-hard when $m = 1$ (with unbounded l) [9].

6 Algorithm for 2-Diversity

In this part we give the first polynomial time algorithm for solving 2-DIVERSITY. Let $\mathcal{T}$ be an input table of 2-DIVERSITY. The following lemma is crucial to our algorithm.

Lemma 2. *There is an optimal 2-diverse partition of $\mathcal{T}$ in which every group consists of 2 or 3 vectors with distinct SA values.*

Proof. It suffices to show that any 2-diverse sub-table $M \subseteq \mathcal{T}$ can be further partitioned into groups each of which consists of 2 or 3 vectors with distinct SA values (note that partitioning a group does not increase the generalization cost). We use induction on the size of M. When $|M| = 2$ or 3 it can be verified directly. Now consider $M \subseteq \mathcal{T}$ of size $t \geq 4$. Suppose M contains k SA values $\{1, 2, \ldots, k\}$ where $k \geq 2$. Let a_i be the number of vectors in M with SA value i, for $i \in [k]$. Assume w.l.o.g. that $a_1 \geq a_2 \geq \ldots \geq a_k$. Let A_1 and A_2 be two vectors with SA value 1 and 2, respectively. Partition M into $\{A_1, A_2\}$ and $M' = M \setminus \{A_1, A_2\}$. We only need to show that M' is 2-diverse, so that we can use induction on it. We perform a case analysis as follows.

- $a_1 = 1$. Then M' consists of at least two vectors with distinct SA values, and thus is 2-diverse.
- $k = 2$. Since M is 2-diverse, we have $a_1 = a_2$. Then M' still contains the same number of SA values 1 and 2, so it remains 2-diverse.
- $a_1 \geq 2, k \geq 3, a_1 > a_3$. The highest frequency of any SA value in M' is $a_1 - 1 \leq |M|/2 - 1 = |M'|/2$, and thus M' is 2-diverse.
- $a_1 = a_3 \geq 2, k \geq 3$. In this case $|M| \geq 3a_3$. The highest frequency of an SA value in M' is a_3. We have $|M'| - 2a_3 = |M| - 2 - 2a_3 \geq a_3 - 2 \geq 0$, so M' is 2-diverse.

All the cases are covered above and hence Lemma 2 is proved. □

Giving Lemma 2, the rest of the proof is basically the same with that of the polynomial-time tractability of 2-ANONYMITY given in [4]. We restate the proof for completeness. We reduce 2-DIVERSITY to a combinatorial problem called SIMPLEX MATCHING introduced in [2], which admits a polynomial algorithm [2]. The input of SIMPLEX MATCHING is a hypergraph $H = (V, E)$ containing edges of sizes 2 and 3 with nonnegative edge costs $c(e)$ for all edges $e \in E$. In addition H is guaranteed to satisfy the following *simplex condition*: if $\{v_1, v_2, v_3\} \in E$, then $\{v_1, v_2\}, \{v_2, v_3\}, \{v_3, v_1\}$ are also in E, and $c(\{v_1, v_2\}) + c(\{v_2, v_3\}) + c(\{v_1, v_3\}) \leq 2 \cdot c(\{v_1, v_2, v_3\})$. The goal is to find a perfect matching of H (i.e., a set of edges that cover every vertex $v \in V$ exactly once) with minimum cost (which is the sum of costs of all chosen edges).

Let $\mathcal{T}$ be an input table of 2-DIVERSITY. We construct a hypergraph $H = (V, E)$ as follows. Let $V = \{v_1, v_2, \ldots, v_n\}$ where v_i corresponds to the vector $\mathcal{T}[i]$. For every two vectors $\mathcal{T}[i], \mathcal{T}[j]$ (or three vectors $\mathcal{T}[i], \mathcal{T}[j], \mathcal{T}[k]$) with distinct SA values, there is an edge $e = \{v_i, v_j\}$ (or $e = \{v_i, v_j, v_k\}$) with cost equal to $cost(\{\mathcal{T}[i], \mathcal{T}[j]\})$ (or $cost(\{\mathcal{T}[i], \mathcal{T}[j], \mathcal{T}[k]\})$). Consider any 3D edge $e = \{v_i, v_j, v_k\}$. Since each column that needs to be suppressed in $\{\mathcal{T}[i], \mathcal{T}[j]\}$ must also

be suppressed in $\{\mathcal{T}[i], \mathcal{T}[j], \mathcal{T}[k]\}$, we have $c(e)/3 \geq c(\{v_i, v_j\})/2$. Similarly, $c(e)/3 \geq c(\{v_i, v_k\})/2$ and $c(e)/3 \geq c(\{v_j, v_k\})/2$. Summing the inequalities up gives $2c(e) \geq c(\{v_i, v_j\}) + c(\{v_i, v_k\}) + c(\{v_j, v_k\})$. Therefore H satisfies the simplex condition, and it clearly can be constructed in polynomial time. Call a 2-diverse partition of $\mathcal{T}$ *good* if every group in it consists of 2 or 3 vectors with distinct SA values. Lemma 2 shows that there is an optimal 2-diverse partition that is good. By the construction of H, each good 2-diverse partition of $\mathcal{T}$ can be easily transformed to a perfect matching of H with the same cost, and vice versa. Hence, we can find an optimal 2-diverse partition of $\mathcal{T}$ by using the polynomial time algorithm for Simplex Matching [2]. We thus have:

Theorem 11. *2-*Diversity *is solvable in polynomial time.*

7 Conclusions

This paper presents the first theoretical study on the t-closeness principle for privacy preserving. We prove the NP-hardness of the t-Closeness problem for every constant $t \in [0, 1)$, and give exact and fixed-parameter algorithms for the problem. We also provide conditionally improved approximation algorithm for k-Anonymity, and give the first polynomial time exact algorithm for 2-Diversity.

There are still many related problems that deserve further explorations, amongst which the most interesting one to the authors is designing polynomial time approximation algorithms for t-Closeness with provable performance guarantees. We conjecture that the best approximation ratio may be dependent on n (e.g., $O(\log n)$). The parameterized complexity of t-Closeness with respect to other sets of parameters are also of interest. Some interesting parameters that have been studied for k-anonymity can be found in [11, 6, 7].

References

[1] G. Aggarwal, T. Feder, K. K. R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu. Anonymizing tables. In *ICDT*, pages 246–258, 2005.

[2] E. Anshelevich and A. Karagiozova. Terminal backup, 3D matching, and covering cubic graphs. *SIAM Journal on Computing*, 40(3):678–708, 2011.

[3] M. M. Baig, J. Li, J. Liu, and H. Wang. Cloning for privacy protection in multiple independent data publications. In *CIKM*, pages 885–894, 2011.

[4] J. Blocki and R. Williams. Resolving the complexity of some data privacy problems. In *ICALP*, pages 393–404, 2010.

[5] P. Bonizzoni, G. D. Vedova, and R. Dondi. Anonymizing binary and small tables is hard to approximate. *Journal of Combinatorial Optimization*, 22(1):97–119, 2011.

[6] P. Bonizzoni, G. D. Vedova, R. Dondi, and Y. Pirola. Parameterized complexity of k-anonymity: hardness and tractability. *Journal of Combinatorial Optimization*, in press.

[7] R. Bredereck, A. Nichterlein, R. Niedermeier, and G. Philip. The effect of homogeneity on the complexity of k-anonymity. In *FCT*, pages 53–64, 2011.

[8] 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.

[9] R. Dondi, G. Mauri, and I. Zoppis. The l-diversity problem: Tractability and approximability. *Theoretical Computer Science*, 2012, in press. DOI: 10.1016/j.tcs.2012.05.024.

[10] R. G. Downey and M. R. Fellows. *Parameterized Complexity*. Springer, 1999.

[11] P. A. Evans, T. Wareham, and R. Chaytor. Fixed-parameter tractability of anonymizing data by suppressing entries. *Journal of Combinatorial Optimization*, 18(4):362–375, 2009.

[12] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

[13] M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified NP-complete problems. In *STOC*, pages 47–63, 1974.

[14] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD*, pages 49–60, 2005.

[15] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE*, 2006.

[16] H. W. Lenstra, Jr. Integer programming with a fixed number of variables. *Mathematics of Operations Research*, 8(4):538–548, 1983.

[17] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy beyond k-Anonymity and l-Diversity. In *ICDE*, pages 106–115, 2007.

[18] N. Li, T. Li, and S. Venkatasubramanian. Closeness: A new privacy measure for data publishing. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):943–956, 2010.

[19] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data (TKDD)*, 1(1), 2007.

[20] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y. Halpern. Worst-case background knowledge for privacy preserving data publishing. In *ICDE*, pages 126–135, 2007.

[21] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, 2004.

[22] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu. Differentially private data release for data mining. In *KDD*, 2011.

[23] M. E. Nergiz, M. Atzori, and C. Clifton. Hiding the presence of individuals from shared databases. In *SIGMOD*, pages 665–676, 2007.

[24] H. Park and K. Shim. Approximate algorithms for k-anonymity. In *SIGMOD*, 2007.

[25] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *IEEE Transactions on Knowledge and Data Engineering*, 22(11):1623–1636, 2010.

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

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

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

[29] X. Xiao and Y. Tao. Personalized privacy preservation. In *SIGMOD*, pages 229–240, 2006.

[30] X. Xiao and Y. Tao. m-invariance: Towards privacy preserving re-publication of dynamic datasets. In *SIGMOD*, pages 689–700, 2007.

[31] X. Xiao, K. Yi, and Y. Tao. The hardness and approximation algorithms for l-diversity. In *EDBT*, pages 135–146, 2010.

[32] M. Xue, P. Karras, C. Raissi, and H. K. Pung. Utility-driven anonymization in data publishing. In *CIKM*, pages 2277–2280, 2011.

[33] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *ICDE*, pages 116–125, 2007.