CERIAS Tech Report 2007-78
t-Closeness: Privacy Beyond k-Anonymity and
by Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity

Ninghui Li Tiancheng Li Suresh Venkatasubramanian
Department of Computer Science, Purdue University AT&T Labs – Research
{ninghui, li83}@cs.purdue.edu suresh@research.att.com

Abstract

The k-anonymity privacy requirement for publishing microdata requires that each equivalence class (i.e., a set of records that are indistinguishable from each other with respect to certain "identifying" attributes) contains at least k records. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure. The notion of ℓ-diversity has been proposed to address this; ℓ-diversity requires that each equivalence class has at least ℓ well-represented values for each sensitive attribute.

In this paper we show that ℓ-diversity has a number of limitations. In particular, it is neither necessary nor sufficient to prevent attribute disclosure. We propose a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). We choose to use the Earth Mover Distance measure for our t-closeness requirement. We discuss the rationale for t-closeness and illustrate its advantages through examples and experiments.

1. Introduction

Agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. Typically, such data is stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories. (1) Attributes that clearly identify individuals. These are known as *explicit identifiers* and include *Social Security Number*, *Address*, and *Name*, and so on. (2) Attributes whose values when taken together can potentially identify an individual. These are known as *quasi-identifiers*, and may include, e.g., *Zip-code*, *Birthdate*, and *Gender*. (3) Attributes that are considered sensitive, such as *Disease* and *Salary*.

When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in the literature [4, 9]: *identity disclosure* and *attribute disclosure*. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure. It has been recognized that even disclosure of false attribute information may cause harm [9]. An observer of a released table may incorrectly perceive that an individual's sensitive attribute takes a particular value, and behave accordingly based on the perception. This can harm the individual, even if the perception is incorrect.

While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly-available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. A common anonymization approach is generalization, which replaces quasi-identifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. We define an *equivalence class* of an anonymized table to be a set of records that have the same values for the quasi-identifiers.

To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table. To this end, Samarati and Sweeney [15, 16, 18] introduced *k-anonymity* as the property that each record is indistinguishable with at

least *k-1* other records with respect to the quasi-identifier. In other words, k-anonymity requires that each equivalence class contains at least k records.

While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. To address this limitation of k-anonymity, Machanavajjhala et al. [12] recently introduced a new notion of privacy, called ℓ-diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least ℓ "well-represented" values.

One problem with l-diversity is that it is limited in its assumption of adversarial knowledge. As we shall explain below, it is possible for an adversary to gain information about a sensitive attribute as long as she has information about the global distribution of this attribute. This assumption generalizes the specific background and homogeneity attacks used to motivate ℓ-diversity. Another problem with privacy-preserving methods in general is that they effectively assume all attributes to be categorical; the adversary either does or does not learn something sensitive. Of course, especially with numerical attributes, being close to the value is often good enough.

We propose a novel privacy notion called t-closeness that formalizes the idea of global background knowledge by requiring that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). This effectively limits the amount of individual-specific information an observer can learn. Further, in order to incorporate distances between values of sensitive attributes, we use the Earth Mover Distance metric [14] to measure the distance between the two distributions. We discuss the rationale for t-closeness and illustrate its advantages through examples and experiments.

The rest of this paper is organized as follows. We give an overview of ℓ-diversity in Section 2 and discuss its limitations in Section 3. We present the rationale and definition of t-closeness in Section 4, and discuss how to calculate the Earth Mover Distance in Section 5. Experimental results are presented in Section 6. Related work is discussed in Section 7. In Section 8, we discuss limitations of our approach and avenues for future research.

2. From k-Anonymity to ℓ-Diversity

The protection k-anonymity provides is simple and easy to understand. If a table satisfies k-anonymity for some value k, then anyone who knows only the quasi-identifier values of one individual cannot identify the record corresponding to that individual with confidence grater than $1/k$.

While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. This has been recognized by several authors, e.g., [12, 19, 21]. Two attacks were identified in [12]: the homogeneity attack and the background knowledge attack.

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

Table 1. Original Patients Table

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	≥ 40	Flu
5	4790*	≥ 40	Heart Disease
6	4790*	≥ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

Table 2. A 3-Anonymous Version of Table 1

Example 1 Table 1 is the original data table, and Table 2 is an anonymized version of it satisfying 3-anonymity. The *Disease* attribute is sensitive. Suppose Alice knows that Bob is a 27-year old man living in ZIP 47678 and Bob's record is in the table. From Table 2, Alice can conclude that Bob corresponds to one of the first three records, and thus must have heart disease. This is the homogeneity attack. For an example of the background knowledge attack, suppose that, by knowing Carl's age and zip code, Alice can conclude that Carl corresponds to a record in the last equivalence class in Table 2. Furthermore, suppose that Alice knows that Carl has very low risk for heart disease. This background knowledge enables Alice to conclude that Carl most likely has cancer.

To address these limitations of k-anonymity, Machanavajjhala et al. [12] introduced ℓ-diversity as a stronger notion of privacy.

Definition 1 (The ℓ-diversity Principle) *An equivalence class is said to have ℓ-diversity if there are at least ℓ "well-represented" values for the sensitive attribute. A table is said to have ℓ-diversity if every equivalence class of the table has ℓ-diversity.*

Machanavajjhala et al. [12] gave a number of interpretations of the term "well-represented" in this principle:

1. **Distinct ℓ-diversity**. The simplest understanding of "well represented" would be to ensure there are at least ℓ *distinct* values for the sensitive attribute in each equivalence class. Distinct ℓ-diversity does not prevent probabilistic inference attacks. An equivalence class may have one value appear much more frequently than other values, enabling an adversary to conclude that an entity in the equivalence class is very likely to have that value. This motivated the development of the following two stronger notions of ℓ-diversity.

2. **Entropy ℓ-diversity**. The entropy of an equivalence class E is defined to be

$$Entropy(E) = -\sum_{s \in S} p(E, s) \log p(E, s)$$

in which S is the domain of the sensitive attribute, and $p(E, s)$ is the fraction of records in E that have sensitive value s.

A table is said to have entropy ℓ-diversity if for every equivalence class E, $Entropy(E) \geq \log \ell$. Entropy ℓ-diversity is strong than distinct ℓ-diversity. As pointed out in [12], in order to have entropy ℓ-diversity for each equivalence class, the entropy of the entire table must be at least $\log(\ell)$. Sometimes this may too restrictive, as the entropy of the entire table may be low if a few values are very common. This leads to the following less conservative notion of ℓ-diversity.

3. **Recursive (c, ℓ)-diversity**. Recursive (c, ℓ)-diversity makes sure that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely. Let m be the number of values in an equivalence class, and $r_i, 1 \leq i \leq m$ be the number of times that the i^{th} most frequent sensitive value appears in an equivalence class E. Then E is said to have recursive (c, ℓ)-diversity if $r_1 < c(r_l + r_{l+1} + ... + r_m)$. A table is said to have recursive (c, ℓ)-diversity if all of its equivalence classes have recursive (c, ℓ)-diversity.

3. Limitations of ℓ-Diversity

While the ℓ-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it has several shortcomings that we now discuss.

ℓ-diversity may be difficult and unnecessary to achieve.

Example 2 Suppose that the original data has only one sensitive attribute: the test result for a particular virus. It takes two values: positive and negative. Further suppose that there are 10000 records, with 99% of them being negative, and only 1% being positive. Then the two values have very different degrees of sensitivity. One would not mind being known to be tested negative, because then one is the same as 99% of the population, but one would not want to be known/considered to be tested positive. In this case, 2-diversity is unnecessary for an equivalence class that contains only records that are negative. In order to have a distinct 2-diverse table, there can be at most $10000 \times 1\% = 100$ equivalence classes and the information loss would be large. Also observe that because the entropy of the sensitive attribute in the overall table is very small, if one uses entropy ℓ-diversity, ℓ must be set to a small value.

ℓ-diversity is insufficient to prevent attribute disclosure. Below we present two attacks on ℓ-diversity.

Skewness Attack: When the overall distribution is skewed, satisfying ℓ-diversity does not prevent attribute disclosure. Consider again Example 2. Suppose that one equivalence class has an equal number of positive records and negative records. It satisfies distinct 2-diversity, entropy 2-diversity, and any recursive $(c, 2)$-diversity requirement that can be imposed. However, this presents a serious privacy risk, because anyone in the class would be considered to have 50% possibility of being positive, as compared with the 1% of the overall population.

Now consider an equivalence class that has 49 positive records and only 1 negative record. It would be distinct 2-diverse and has higher entropy than the overall table (and thus satisfies any Entropy ℓ-diversity that one can impose), even though anyone in the equivalence class would be considered 98% positive, rather than 1% percent. In fact, this equivalence class has exactly the same diversity as a class that has 1 positive and 49 negative records, even though the two classes present very differen levels of privacy risks.

Similarity Attack: When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information. Consider the following example.

Example 3 Table 3 is the original table, and Table 4 shows an anonymized version satisfying distinct and entropy 3-diversity. There are two sensitive attributes: *Salary* and *Disease*. Suppose one knows that Bob's record corresponds to one of the first three records, then one knows that Bob's salary is in the range [3K–5K] and can infer that Bob's salary is relatively low. This attack applies not only to numeric attributes like "Salary", but also to categorical attributes like "Disease". Knowing that Bob's record belongs to the first equivalence class enables one to conclude that Bob has some stomach-related problems, because all three diseases in the class are stomach-related.

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

Table 3. Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

Table 4. A 3-diverse version of Table 3

This leakage of sensitive information occurs because while ℓ-diversity requirement ensures "diversity" of sensitive values in each group, it does not take into account the semantical closeness of these values.

Summary In short, distributions that have the same level of diversity may provide very different levels of privacy, because there are semantic relationships among the attribute values, because different values have very different levels of sensitivity, and because privacy is also affected by the relationship with the overall distribution.

4. t-Closeness: A New Privacy Measure

Intuitively, privacy is measured by the information gain of an observer. Before seeing the released table, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief. The novelty of our approach is that we separate the information gain into two parts: that about the whole population in the released data and that about specific individuals.

To motivate our approach, let us perform the following thought experiment: First an observer has some prior belief B_0 about an individual's sensitive attribute. Then, in a hypothetical step, the observer is given a completely generalized version of the data table where all attributes in a quasi-identifier are removed (or, equivalently, generalized to the most general values). The observer's belief is influenced by $\mathbf{Q}$, the distribution of the sensitive attribute value in the whole table, and changes to B_1. Finally, the observer is given the released table. By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in, and learn the distribution $\mathbf{P}$ of sensitive attribute values in this class. The observer's belief changes to B_2.

The ℓ-diversity requirement is motivated by limiting the difference between B_0 and B_2 (although it does so only indirectly, by requiring that $\mathbf{P}$ has a level of diversity). We choose to limit the difference between B_1 and B_2. In other words, we assume that $\mathbf{Q}$, the distribution of the sensitive attribute in the overall population in the table, is public information. We do not limit the observer's information gain about the population as a whole, but limit the extent to which the observer can learn additional information about specific individuals.

To justify our assumption that $\mathbf{Q}$ should be treated as public information, we observe that with generalizations, the most one can do is to generalize all quasi-identifier attributes to the most general value. Thus as long as a version of the data is to be released, a distribution $\mathbf{Q}$ will be released.[1] We also argue that if one wants to release the table at all, one intends to release the distribution $\mathbf{Q}$ and this distribution is what makes data in this table useful. In other words, one wants $\mathbf{Q}$ to be public information. A large change from B_0 to B_1 means that the data table contains a lot of new information, e.g., the new data table corrects some widely held belief that was wrong. In some sense, the larger the difference between B_0 and B_1 is, the more valuable the data is. Since the knowledge gain between B_0 and B_1 is about the whole population, we do not limit this gain.

We limit the gain from B_1 to B_2 by limiting the distance between $\mathbf{P}$ and $\mathbf{Q}$. Intuitively, if $\mathbf{P} = \mathbf{Q}$, then B_1 and B_2 should be the same. If $\mathbf{P}$ and $\mathbf{Q}$ are close, then B_1 and B_2 should be close as well, even if B_0 may be very different from both B_1 and B_2.

Definition 2 (The t-closeness Principle:) *An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.*

[1] Note that even with suppression, a distribution will still be released. This distribution may be slightly different from the distribution with no record suppressed; however, from our point of view, we only need to consider the released distribution and the distance of it from the ones in the equivalence classes.

Of course, requiring that $\mathbf{P}$ and $\mathbf{Q}$ to be close would also limit the amount of useful information that is released, as it limits information about the correlation between quasi-identifier attributes and sensitive attributes. However, this is precisely what one needs to limit. If an observer gets too clear a picture of this correlation, then attribute disclosure occurs. The t parameter in t-closeness enables one to trade off between utility and privacy.

Now the problem is to measure the distance between two probabilistic distributions. There are a number of ways to define the distance between them. Given two distributions $\mathbf{P} = (p_1, p_2, ..., p_m), \mathbf{Q} = (q_1, q_2, ..., q_m)$, two well-known distance measures are as follows. The *variational distance* is defined as:

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2}|p_i - q_i|.$$

And the Kullback-Leibler (KL) distance [8] is defined as:

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i} = H(\mathbf{P}) - H(\mathbf{P}, \mathbf{Q})$$

where $H(\mathbf{P}) = \sum_{i=1}^{m} p_i \log p_i$ is the entropy of $\mathbf{P}$ and $H(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{m} p_i \log q_i$ is the cross-entropy of $\mathbf{P}$ and $\mathbf{Q}$.

These distance measures do not reflect the semantic distance among values. Recall Example 3 (Tables 3 and 4), where the overall distribution of the Income attribute is $\mathbf{Q} = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$.[2] The first equivalence class in Table 4 has distribution $\mathbf{P_1} = \{3k, 4k, 5k\}$ and the second equivalence class has distribution $\mathbf{P_2} = \{6k, 8k, 11k\}$. Our intuition is that $\mathbf{P_1}$ results in more information leakage than $\mathbf{P_2}$, because the values in $\mathbf{P_1}$ are all in the lower end; thus we would like to have $\mathsf{D}[\mathbf{P_1}, \mathbf{Q}] > \mathsf{D}[\mathbf{P_2}, \mathbf{Q}]$. The distance measures mentioned above would not be able to do so, because from their point of view values such as $3k$ and $6k$ are just different points and have no other semantic meaning.

In short, we have a metric space for the attribute values so that a ground distance is defined between any pair of values. We then have two probability distributions over these values and we want the distance between the two probability distributions to be dependent upon the ground distances among these values. This requirement leads us to the the Earth Mover's distance (EMD) [14], which is actually a Monge-Kantorovich transportation distance [5] in disguise.

The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. Intuitively, one distribution is seen as a mass of earth spread in the space and the other as a collection of holes in the same space. EMD measures the least amount of work needed to fill the holes with earth. A unit of work corresponds to moving a unit of earth by a unit of ground distance.

EMD can be formally defined using the well-studied transportation problem. Let $\mathbf{P} = (p_1, p_2, ...p_m), \mathbf{Q} = (q_1, q_2, ...q_m)$, and d_{ij} be the ground distance between element i of $\mathbf{P}$ and element j of $\mathbf{Q}$. We want to find a flow $F = [f_{ij}]$ where f_{ij} is the flow of mass from element i of $\mathbf{P}$ to element j of $\mathbf{Q}$ that minimizes the overall work:

$$WORK(\mathbf{P}, \mathbf{Q}, F) = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij}$$

subject to the following constraints:

$$f_{ij} \geq 0 \qquad 1 \leq i \leq m, 1 \leq j \leq m \qquad (c1)$$

$$p_i - \sum_{j=1}^{m} f_{ij} + \sum_{j=1}^{m} f_{ji} = q_i \qquad 1 \leq i \leq m \quad (c2)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} q_i = 1 \qquad (c3)$$

These three constraints guarantee that $\mathbf{P}$ is transformed to $\mathbf{Q}$ by the mass flow F. Once the transportation problem is solved, the EMD is defined to be the total work,[3] i.e.,

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = WORK(\mathbf{P}, \mathbf{Q}, F) = \sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij} f_{ij}$$

We will discuss how to calculate the EMD between two distributions in Section 5. We now observe two useful facts about EMD.

Fact 1 *If $0 \leq d_{ij} \leq 1$ for all i, j, then $0 \leq \mathsf{D}[\mathbf{P}, \mathbf{Q}] \leq 1$.*

The above fact follows directly from constraint $(c1)$ and $(c3)$. It says that if the ground distances are normalized, i.e., all distances are between 0 and 1, then the EMD between any two distributions is between 0 and 1. This gives a range from which one can choose the t value for t-closeness.

Fact 2 *Given two equivalence classes E_1 and E_2, let $\mathbf{P_1}$, $\mathbf{P_2}$, and $\mathbf{P}$ be the distribution of a sensitive attribute in E_1, E_2, and $E_1 \cup E_2$, respectively. Then*

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] \leq \frac{|E_1|}{|E_1| + |E_2|}\mathsf{D}[\mathbf{P_1}, \mathbf{Q}] + \frac{|E_2|}{|E_1| + |E_2|}\mathsf{D}[\mathbf{P_2}, \mathbf{Q}]$$

[2] We use the notation $\{v_1, v_2, \cdots, v_m\}$ to denote the uniform distribution where each value in $\{v_1, v_2, \cdots, v_m\}$ is equally likely.

[3] More generally, the EMD is the total work divided by the total flow. However, since we are calculating distance between two probability distributions, the total flow is always 1, as shown in formula $(c3)$.

It follows that $\mathsf{D}[\mathbf{P}, \mathbf{Q}] \leq \max(\mathsf{D}[\mathbf{P_1}, \mathbf{Q}], \mathsf{D}[\mathbf{P_2}, \mathbf{Q}])$. This means that when merging two equivalence classes, the maximum distance of any equivalence class from the overall distribution can never increase. Thus t-closeness is achievable for any $t \geq 0$.

The above fact entails that t-closeness with EMD satisfies the following two properties.

Generalization Property *Let $\mathcal{T}$ be a table, and let A and B be two generalizations on $\mathcal{T}$ such that A is more general than B If $\mathcal{T}$ satisfies t-closeness using B, then T also satisfies t-closeness using A.*

Proof Since each equivalence class in A is the union of a set of equivalence classes in B and each equivalence class in B satisfies t-closeness, we conclude that each equivalence class in A also satisfies t-closeness. Thus $\mathcal{T}$ satisfies t-closeness using A.

Subset Property *Let $\mathcal{T}$ be a table and let C be a set of attributes in $\mathcal{T}$. If $\mathcal{T}$ satisfies t-closeness with respect to C, then T also satisfies t-closeness with respect to any set of attributes D such that $D \subset C$.*

Proof Similarly, each equivalence class with respect to D is the union of a set of equivalence classes with respect to C and each equivalence class with respect to C satisfies t-closeness, we conclude that each equivalence class with respect to D also satisfies t-closeness. Thus $\mathcal{T}$ satisfies t-closeness with respect to D.

The two properties guarantee that the t-closeness using EMD measurement can be incorporated into the general framework of the Incognito algorithm [10].

5. How to Calculate the EMD

To use t-closeness with EMD, we need to be able to calculate the EMD between two distributions. One can calculate EMD using solutions to the transportation problem, such as a min-cost flow[1]; however, these algorithms do not provide an explicit formula. In the rest of this section, we derive formulas for calculating EMD for the special cases that we need to consider.

5.1. EMD for Numerical Attributes

Numerical attribute values are ordered. Let the attribute domain be $\{v_1, v_2 ... v_m\}$, where v_i is the i^{th} smallest value.

Ordered Distance: The distance between two values of is based on the number of values between them in the total order, i.e., $ordered\_dist(v_i, v_j) = \frac{|i-j|}{m-1}$.

It is straightforward to verify that the ordered-distance measure is a metric. It is non-negative and satisfies the symmetry property and the triangle inequality. To calculate EMD under ordered distance, we only need to consider flows that transport distribution mass between adjacent elements, because any transportation between two more distant elements can be equivalently decomposed into several transportations between adjacent elements. Based on this observation, minimal work can be achieved by satisfying all elements of $\mathbf{Q}$ sequentially. We first consider element 1, which has an extra amount of $p_1 - q_1$. Assume, without loss of generality, that $p_1 - q_1 < 0$, an amount of $q_1 - p_1$ should be transported from other elements to element 1. We can transport this from element 2. After this transportation, element 1 is satisfied and element 2 has an extra amount of $(p_1 - q_1) + (p_2 - q_2)$. Similarly, we can satisfy element 2 by transporting an amount of $|(p_1 - q_1) + (p_2 - q_2)|$ between element 2 and element 3. This process continues until element m is satisfied and $\mathbf{Q}$ is reached.

Formally, let $r_i = p_i - q_i$,(i=1,2,...,m), then the distance between $\mathbf{P}$ and $\mathbf{Q}$ can be calculated as:

$$\begin{aligned}\mathsf{D}[\mathbf{P}, \mathbf{Q}] &= \frac{1}{m-1}(|r_1| + |r_1 + r_2| + ... + |r_1 + r_2 + ... r_{m-1}|) \\ &= \frac{1}{m-1} \sum_{i=1}^{i=m} \left| \sum_{j=1}^{j=i} r_j \right|\end{aligned}$$

5.2. EMD for Categorical Attributes

For categorical attributes, a total order often does not exist. We consider two distance measures.

Equal Distance: The ground distance between any two value of a categorical attribute is defined to be 1. It is easy to verify that this is a metric. As the distance between any two values is 1, for each point that $p_i - q_i > 0$, one just needs to move the extra to some other points. Thus we have the following formula:

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = \frac{1}{2} \sum_{i=1}^{m} |p_i - q_i| = \sum_{p_i \geq q_i} (p_i - q_i) = - \sum_{p_i < q_i} (p_i - q_i)$$

Hierarchical Distance: The distance between two values of a categorical attribute is based on the minimum level to which these two values are generalized to the same value according to the domain hierarchy. Mathematically, let H be the height of the domain hierarchy, the distance between two values v_1 and v_2(which are leaves of the hierarchy) is defined to be $level(v_1, v_2)/H$, where $level(v_1, v_2)$ is the height of the lowest common ancestor node of v_1 and v_2. It is straightforward to verify that this hierarchical-distance measure is also a metric.

Given a domain hierarchy and two distributions $\mathbf{P}$ and $\mathbf{Q}$, we define the *extra* of a leaf node that corresponds to element i, to be $p_i - q_i$, and the *extra* of an internal node N to be the sum of *extras* of leaf nodes below N. This *extra* function can be defined recursively as:

$$extra(N) = \begin{cases} p_i - q_i & \text{if N is a leaf} \\ \sum_{C \in Child(N)} extra(C) & \text{otherwise} \end{cases}$$

where $Child(N)$ is the set of all leaf nodes below node N. The $extra$ function has the property that the sum of $extra$ values for nodes at the same level is 0.

We further define two other functions for *internal nodes*:

$$pos\_extra(N) = \sum_{C \in Child(N) \wedge extra(C) > 0} |extra(C)|$$

$$neg\_extra(N) = \sum_{C \in Child(N) \wedge extra(C) < 0} |extra(C)|$$

We use $cost(N)$ to denote the cost of movings between N's children branches. An optimal flow moves exactly $extra(N)$ in/out of the subtree rooted at N. Suppose that $pos\_extra(N) > neg\_extra$, then $extra(N) = pos\_extra(N) - neg\_extra(N)$ and $extra(N)$ needs to move out. (This cost is counted in the cost of N's parent node.) In addition, one has to move neg_extra among the children nodes to even out all children branches; thus,

$$cost(N) = \frac{height(N)}{H} \min(pos\_extra(N), neg\_extra(N))$$

Then the earth mover's distance can be written as:

$$\mathsf{D}[\mathbf{P}, \mathbf{Q}] = \sum_N cost(N)$$

where N is a non-leaf node.

5.3 Analysis of t-Closeness with EMD

We now revisit Example 3 in Section 3, to show how t-closeness with EMD handles the difficulties of ℓ-diversity. Recall that $\mathbf{Q} = \{3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k\}$, $\mathbf{P_1} = \{3k, 4k, 5k\}$, and $\mathbf{P_2} = \{6k, 8k, 11k\}$. We calculate $\mathsf{D}[\mathbf{P_1}, \mathbf{Q}]$ and $\mathsf{D}[\mathbf{P_2}, \mathbf{Q}]$ using EMD. Let $v_1 = 3k, v_2 = 4k, ...v_9 = 11k$, we define the distance between v_i and v_j to be $|i - j|/8$, thus the maximal distance is 1. We have $\mathsf{D}[\mathbf{P_1}, \mathbf{Q}] = 0.375$,[4] and $\mathsf{D}[\mathbf{P_2}, \mathbf{Q}] = 0.167$.

For the disease attribute, we use the hierarchy in Figure 1 to define the ground distances. For example, the distance between "Flu" and "Bronchitis" is $1/3$, the distance between "Flu"and "Pulmonary embolism" is $2/3$, and the distance between "Flu" and "Stomach cancer" is $3/3 = 1$. Then the distance between the distribution {gastric ulcer, gastritis, stomach cancer} and the overall distribution is 0.5 while the distance between the distribution {gastric ulcer, stomach cancer, pneumonia} is 0.278.

Table 5 shows another anonymized version of Table 3. It has 0.167-closeness w.r.t Salary and 0.278-closeness w.r.t. Disease. The *Similarity Attack* is prevented in Table 5. For example, Alice cannot infer that Bob has a low salary or Bob has stomach-related diseases based on Table 5.

[4]One optimal mass flow that transforms $\mathbf{P_1}$ to $\mathbf{Q}$ is to move $1/9$ probability mass across the following pairs: (5k→11k), (5k→10k), (5k→9k), (4k→8k), (4k→7k), (4k→6k), (3k→5k), (3k→4k). The cost of this is $1/9 \times (6 + 5 + 4 + 4 + 3 + 2 + 2 + 1)/8 = 27/72 = 3/8 = 0.375$.

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

Table 5. Table that has 0.167-closeness w.r.t. Salary and 0.278-closeness w.r.t. Disease

We note that t-closeness protects against attribute disclosure, but does not deal with identity disclosure. Thus, it may be desirable to use both t-closeness and k-anonymity at the same time. Further, it should be noted that t-closeness deals with the homogeneity and background knowledge attacks on k-anonymity not by guaranteeing that they can never occur, but by guaranteeing that if such attacks can occur, then similar attacks can occur even with a fully-generalized table. As we argued earlier, this is the best one can achieve if one is to release the data at all.

6. Experiments

The main goals of the experiments are to study the effect of *Similarity Attack* on real data and to investigate the performance implications of the t-closeness approach in terms of efficiency and data quality.

The dataset used in the experiments is the adult dataset from the UC Irvine machine learning repository, which is comprised of data collected from the US census. We used nine attributes of the dataset, as shown in Figure 2. Records with missing values are eliminated and there are 30162 valid records in total. We use our Java implementation of the Incognito [10] algorithm. The experiments are run on a 3.4GHZ Pentium 4 machine with 2GB memory.

Similarity Attack We use the first 7 attributes as the quasi-identifier and treat *Occupation* as the sensitive attribute. We divide the 14 values of the *Occupation* attribute into three roughly equal-size groups, based on the semantic closeness of the values. Any equivalence class that has all values falling in one group is viewed as vulnerable to the similarity attack. We use Incognito to generate all entropy 2-diversity tables. In total, there are 21 minimal tables and 13 of them suffers from the *Similarity* attack. In one table, a total of *916* records can be inferred about their sensitive value class. We

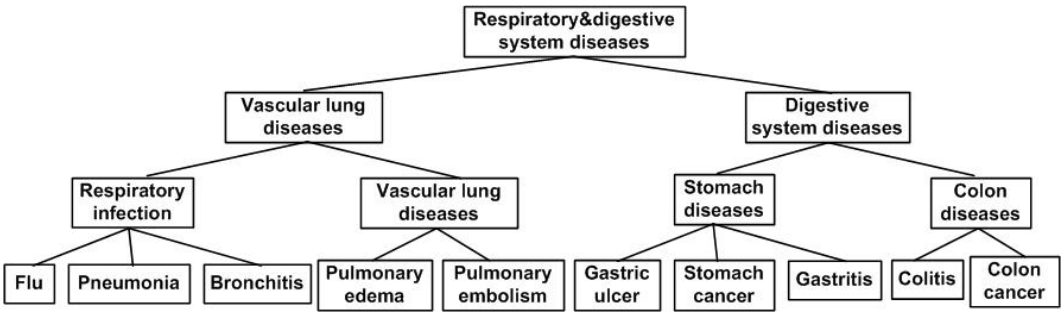


Figure 1. Hierarchy for categorical attributes *Disease*.

	Attribute	Type	# of values	Height
1	Age	Numeric	74	5
2	Workclass	Categorical	8	3
3	Education	Categorical	16	4
4	Country	Categorical	41	3
5	Marital_Status	Categorical	7	3
6	Race	Categorical	5	3
7	Gender	Categorical	2	2
8	Occupation	Sensitive	14	3
9	Salary	Sensitive	2	2

Figure 2. Description of the *Adult* dataset used in the experiment

also generate all 26 minimal recursive $(4, 4)$-diversity tables, and found that 17 of which are vulnerable to the similarity attack.

Efficiency We compare the efficiency and data quality of five privacy measures: (1) k-anonymity; (2) entropy ℓ-diversity; (3) recursive (c, ℓ) diversity; (4) k-anonymity with t-closeness($t = 0.2$); and (5) k-anonymity with t-closeness($t = 0.15$).

Results of efficiency experiments are shown in Figure 3. Again we use the *Occupation* attribute as the sensitive attribute. Figure 3(a) shows the running times with fixed $k = 5, \ell = 5$ and varied quasi-identifier size s, where $2 \leq s \leq 7$. A quasi-identifier of size s consists of the first s attributes listed in Table 2. Figure 3(b) shows the running times of the five privacy measures with the same quasi-identifier but with different parameters for k and ℓ. As shown in the figures, entropy ℓ-diversity run faster than the other four measures; the difference gets larger when ℓ increases. This is because with large ℓ, entropy ℓ-diversity prunes the search lattice earlier.

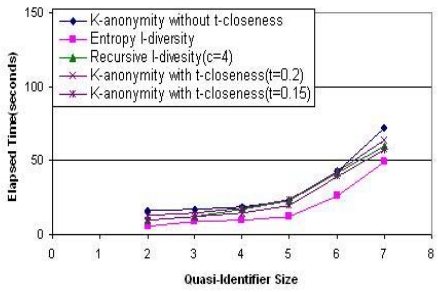
Data Quality Our third set of experiments compare the data quality of the five privacy measures using the discernibility metric [2] and Minimal Average Group Size [10, 15]. The first metric measures the number of records that are indistinguishable from each other. Each record in an equivalence class of size t gets a penalty of t while each suppressed tuple gets a penalty equal to the total number of records. The second metric is the average size of the equivalence classes generated by the anonymization algorithm.

We use the 7 regular attributes as the quasi-identifier and *Occupation* as the sensitive attribute. We set different parameters for k, ℓ, and compare the resulted dataset produced by different measurements. Figure 4 summarizes the results. We found that entropy ℓ-diversity tables has worse data quality than the other measurements. We also found that the data quality of k-anonymous tables without t-closeness is slightly better than k-anonymous tables with t-closeness. This is because t-closeness requirement provides extra protection to sensitive values and the cost is decreased data quality. When choosing $t = 0.2$, the degradation in data quality is minimal.

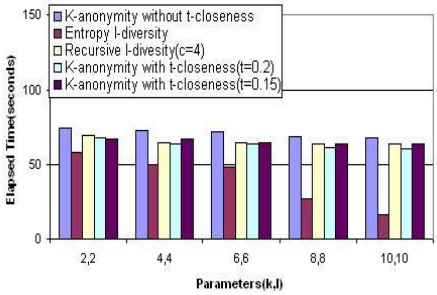
7. Related Work

The problem of information disclosure has been studied extensively in the framework of statistical databases. A number of information disclosure limitation techniques [3] have been designed for data publishing, including *Sampling*, *Cell Suppression*, *Rounding*, and *Data Swapping and Perturbation*. These techniques, however, compromised data integrity of the tables. Samarati and Sweeney [15, 16, 18] introduced the k-anonymity approach and used generalization and suppression techniques to preserve information truthfulness. Numerous algorithms [2, 6, 11, 10, 16, 17] have been proposed to achieve k-anonymity requirement. Optimal k-anonymity has been proved to be NP-hard for $k \geq 3$ [13].

Recently, a number of authors have recognized that k-anonymity does not prevent attribute disclosure, e.g., [12, 19, 21]. Machanavajjhala et al. [12] proposed ℓ-diversity. As we discuss in detail in Section 3, while ℓ-diversity is an important step beyond k-anonymity, it has a number of limitations. Xiao and Tao [21] observe that ℓ-diversity can-

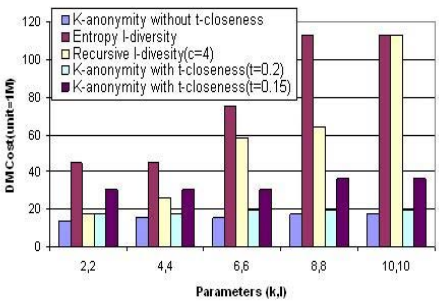


(a) Varied QI size for $k = 5, l = 5$

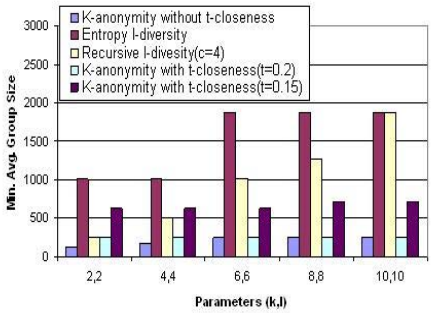


(b) Varied parameters k and l

Figure 3. Efficiency of the Five Privacy Measures.



(a) Discernibility metric cost



(b) Minimal average group size

Figure 4. Data Quality of the Five Measures.

not prevent attribute disclosure, when multiple records in the table corresponds to one individual. They proposed to have each individual specify privacy policies about his or her own attributes. We identify limitations of ℓ-diversity even when each record corresponds to one individual and proposes t-closeness, an alternative privacy measure without the need for individual policies. Xiao and Tao [20] proposed Anatomy, an data anonymization approach that divides one table into two for release; one table includes original quasi-identifier and a group id, and the other include the association between the group id and the sensitive attribute values. Anatomy uses ℓ-diversity as the privacy measure; we believe that t-closeness can be used to provide more meaningful privacy.

In the current proceedings, Koudas et al. [7] examine the anonymization problem from the perspective of answering downstream aggregate queries. They develop a new privacy-preserving framework based not on generalization, but on permutations. Their work, like ours, addresses the problem of dealing with attributes defined on a metric space; their approach is to lower bound the *range* of values of a sensitive attribute in a group.

8. Conclusions and Future Work

While k-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The notion of ℓ-diversity attempts to solve this problem by requiring that each equivalence class has at least ℓ well-represented values for each sensitive attribute. We have shown that ℓ-diversity has a number of limitations and have proposed a novel privacy notion called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). One key novelty of our approach is that we separate the information gain an observer can get from a released data table into two parts: that about all population in the released data and that about specific individuals. This enables us to limit only the second kind of information gain. We use the Earth Mover Distance measure for our t-closeness requirement; this has the advantage of taking into consideration the semantic closeness of attribute values. Below we discuss some interesting open research issues.

Multiple Sensitive Attributes Multiple sensitive attributes

present additional challenges. Suppose we have two sensitive attributes U and V. One can consider the two attributes separately, i.e., an equivalence class E has t-closeness if E has t-closeness with respect to both U and V. Another approach is to consider the joint distribution of the two attributes. To use this approach, one has to choose the ground distance between pairs of sensitive attribute values. A simple formula for calculating EMD may be difficult to derive, and the relationship between t and the level of privacy becomes more complicated.

Other Anonymization Techniques t-closeness allows us to take advantage of anonymization techniques other than generalization of quasi-identifier and suppression of records. For example, instead of suppressing a whole record, one can hide some sensitive attributes of the record; one advantage is that the number of records in the anonymized table is accurate, which may be useful in some applications. Because this technique does not affect quasi-identifiers, it does not help achieve k-anonymity and hence has not been considered before. Removing a value only decreases diversity; therefore, it does not help to achieve ℓ-diversity. However, in t-closeness, removing an outlier may smooth a distribution and bring it closer to the overall distribution. Another possible technique is to generalize a sensitive attribute value, rather than hiding it completely. An interesting question is how to effectively combine these techniques with generalization and suppression to achieve better data quality.

Limitations of using EMD in t-closeness The t-closeness principle can be applied using other distance measures. While EMD is the best measure we have found so far, it is certainly not perfect. In particular, the relationship between the value t and information gain is unclear. For example, the EMD between the two distributions $(0.01, 0.99)$ and $(0.11, 0.89)$ is 0.1, and the EMD between $(0.4, 0.6)$ and $(0.5, 0.5)$ is also 0.1. However, one may argue that the change between the first pair is much more significant than that between the second pair. In the first pair, the probability of taking the first value increases from 0.01 to 0.11, a 1000% increase. While in the second pair, the probability increase is only 25%. In general, what we need is a measure that combines the distance-estimation properties of the EMD with the probability scaling nature of the KL distance.

References

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network flows: theory, algorithms, and applications*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[2] R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *Proc. 21st Intnl. Conf. Data Engg. (ICDE)*, pages 217–228, Washington, DC, USA, 2005. IEEE Computer Society.

[3] G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig. Disclosure limitation methods and information loss for tabular data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 135–166. Elsevier, 2001.

[4] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *J. Am. Stat. Assoc.*, pages 10–28, 1986.

[5] C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *Michigan Math J.*, 31:231–240, 1984.

[6] V. S. Iyengar. Transforming data to satisfy privacy constraints. In *Proc. 8th ACM KDD*, pages 279–288, 2002.

[7] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang. Aggregate query answering on anonymized tables. In *Proc. 23rd Intnl. Conf. Data Engg. ICDE*, 2007.

[8] S. L. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:79–86, 1951.

[9] D. Lambert. Measures of disclosure risk and harm. *J. Official Stat.*, 9:313, 1993.

[10] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proc. ACM SIGMOD International Conference on Management of Data (SIGMOD'05)*, pages 49–60, 2005.

[11] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proc. 22nd Intnl. Conf. Data Engg (ICDE)*, 2006.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. In *Proc. 22nd Intnl. Conf. Data Engg. (ICDE)*, page 24, 2006.

[13] A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In *PODS*, pages 223–228. ACM Press, 2004.

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

[15] P. Samarati. Protecting respondent's privacy in microdata release. *IEEE T. Knowl. Data En.*, 13(6):1010–1027, 2001.

[16] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[17] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzz.*, 10(6):571–588, 2002.

[18] L. Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzz.*, 10(5):557–570, 2002.

[19] T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In *Proceedings of the 22nd International Conference on Data Engineering Workshops, the Second Intenational Workshop on Privacy Data Management (PDM'06)*, page 94, 2006.

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

[21] X. Xiao and Y. Tao. Personalized privacy preservation. In *Proceedings of ACM Conference on Management of Data (SIGMOD'06)*, pages 229–240, June 2006.