Differential Privacy via t-Closeness in Data Publishing

Jordi Soria-Comas* and Josep Domingo-Ferrer†
Universitat Rovira i Virgili
Dept. of Computer Engineering and Maths
UNESCO Chair in Data Privacy
Av. Països Catalans 26
43007 Tarragona, Catalonia
* Email: jordi.soria@urv.cat
† Email: josep.domingo@urv.cat

***Abstract*—k-Anonymity and ε-differential privacy are two main privacy models proposed within the computer science community. Whereas the former was proposed for privacy-preserving data publishing, *i.e.* data set anonymization, the latter initially arose in the context of interactive databases and was later extended to data publishing. We show here that t-closeness, one of the extensions of k-anonymity, can actually yield ε-differential privacy in data publishing when $t = \exp(\varepsilon)$. We detail a construction based on bucketization that realizes the previous implication; hence, as an ancillary result, we provide a new computational procedure to achieve t-closeness and ε-differential privacy in data publishing.**

I. Introduction

k-Anonymity and ε-differential privacy are two mainstream privacy models originated within the computer science community. Their approaches towards disclosure limitation are quite different: k-anonymity is a model for releases of microdata (*i.e.* individual records) that seeks to prevent record re-identification by hiding each original record within a group of k indistinguishable anonymized records, while ε-differential privacy originated as a model for interactive databases and seeks to limit the knowledge that users obtain from query responses. Both models are often presented as antagonistic: ε-differential privacy supporters view k-anonymity as an old-fashioned privacy notion that offers only poor disclosure limitation guarantees, while ε-differential privacy detractors criticize the limited utility of ε-differentially private outputs and the cumbersomeness of not having access to the data set.

A. Contribution and plan of this paper

This paper shows that, for data set anonymization, the t-closeness extension of k-anonymity is closely related to ε-differential privacy. This relation is demonstrated both versus uninformed intruders (having access only to the released data set) and informed intruders (having also background knowledge). For uninformed intruders we prove that $\exp(\varepsilon)$-closeness is equivalent to ε-differential privacy. For informed intruders, the strict equivalence we obtain for uninformed intruders does not hold; however, we show that $\exp(\varepsilon)$-closeness can be seen as a good approximation to ε-differential privacy. Our approach is a constructive one: we specify a computational procedure based on bucketization that, given an original data set, builds a t-close version of it. In the case of uninformed intruders, this version turns out to be differentially private as well; in the case of informed intruders, it is approximately differentially private.

Section II reviews background on k-anonymity, t-closeness and ε-differential privacy. Section III reviews partitioning strategies used to achieve k-anonymity and its extensions, including t-closeness. Specifically, we first examine the shortcomings of achieving k-anonymity and t-closeness in the classical sense, that is, by modifying the quasi-identifier attributes to create groups of at least k indistinguishable records. We then review two approaches which leave the quasi-identifier attributes unaltered and which will be used as building blocks of our computational procedure to reach t-closeness and ε-differential privacy. Section IV develops in detail our proposed construction to reach t-closeness. Section V shows that $\exp(\varepsilon)$-closeness reached with the previous construction implies: i) ε-differential privacy in the case of uninformed intruders; ii) approximate ε-differential privacy in the case of informed intruders. Conclusions and future research issues are summarized in Section VI.

II. Background

A. k-Anonymity

k-Anonymity [12], [11] protects against identity disclosure: an intruder cannot determine the anonymized record corresponding to a specific target individual T for whom the intruder knows the values of the quasi-identifier attributes. The reason is that, in a k-anonymous data set, each combination of values of the quasi-identifier attributes is shared by at least k records. Hence, at most the intruder can determine the group G_T of k anonymized records which contains the anonymized record corresponding to the target individual.

The most usual computational procedure to attain k-anonymity is generalization of the quasi-identifier attributes [11], but an alternative approach is based on microaggregation of the quasi-identifier attributes [4].

B. t-Closeness

While k-anonymity protects against identity disclosure, as mentioned above, it does not protect in general against attribute disclosure: if the values for one (or several) confidential

attribute(s) in all records of G_T are the same, then the intruder learns the values of that (those) attribute(s) for the target individual T.

The property of l-diversity [10] has been proposed as an extension of k-anonymity which tries to address the attribute disclosure problem. A data set is said to satisfy l-diversity if, for each group of records sharing a combination of quasi-identifier attribute values, there are at least l "well-represented" values for each confidential attribute. Achieving l-diversity in general implies more distortion than just achieving k-anonymity. Yet, l-diversity may fail to protect against attribute disclosure if the l values of a confidential attribute are very similar or are strongly skewed. p-Sensitive k-anonymity [14] is a property similar to l-diversity, which shares similar shortcomings. See [5] for a summary of criticisms to l-diversity and p-sensitive k-anonymity.

t-Closeness [9] is another extension of k-anonymity which also tries to solve the attribute disclosure problem. A data set is said to satisfy t-closeness if, for each group of records sharing a combination of quasi-identifier attribute values, the distance between the distribution of each confidential attribute in the group and the distribution of the same confidential attribute in the whole data set is no more than a threshold t. This property clearly solves the attribute disclosure vulnerability, although the original t-closeness paper did not propose a computational procedure to achieve this property and did not mention the large utility loss that this property is likely to inflict on the original data.

C. ε-Differential privacy

Differential privacy was originally proposed by [6] as a privacy model in the interactive setting, that is, to protect the outcomes of queries to a database. The assumption is that an anonymization mechanism sits between the user submitting queries and the database answering them.

Definition 1. (ε-Differential privacy) A randomized function κ gives ε-differential privacy if, for all data sets D_1, D_2 such that one can be obtained from the other by modifying a single record, and all $S \subset Range(\kappa)$

$$\Pr(\kappa(D_1) \in S) \leq \exp(\varepsilon) \times \Pr(\kappa(D_2) \in S) \qquad (1)$$

The computational mechanism to attain ε-differential privacy is often called ε-differentially private *sanitizer*. A usual sanitization approach is noise addition: first, the real value $f(D)$ of the response to a certain user query f is computed, and then a random noise, say $Y(D)$, is added to mask $f(D)$, that is, a randomized response $\kappa(D) = f(D) + Y(D)$ is returned. To generate $Y(D)$, a common choice is to use a Laplace distribution with zero mean and $\Delta(X)/\varepsilon$ scale parameter, where:

- ε is the differential privacy parameter;
- $\Delta(f)$ is the L_1-sensitivity of f, that is, the maximum variation of the query function between neighbor data sets, *i.e.*, sets differing in at most one record.

Differential privacy was also proposed for the non-interactive setting in [2], [7], [8], [3]. Even though a non-interactive data release can be used to answer an arbitrarily large number of queries, in all these proposals, this is obtained at the cost of preserving utility only for restricted classes of queries (typically count queries). This contrasts with the general-purpose utility-preserving data release offered by the k-anonymity model.

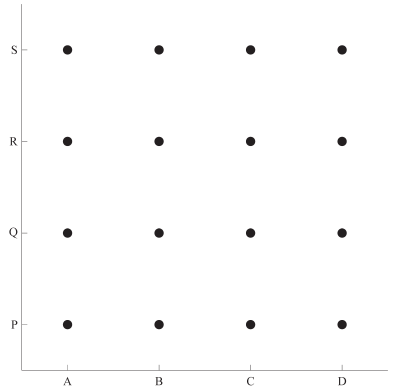


Fig. 1. Projection of the records in the data set on the quasi-identifier attributes Q_1 and Q_2

III. Partitioning strategies for k-anonymity

In k-anonymity and its extensions (including l-diversity and t-closeness), the partitioning strategy to create groups of indistinguishable records is a key point for data utility. Assume a data user who wants to analyze a group of individuals that has been selected based on the value of the quasi-identifier attributes. The utility that this user derives from the k-anonymous data depends on how well the target group of individuals can be approximated by the groups of indistinguishable records. The best utility is achieved when the target group of individuals can be approximated as the union of groups of indistinguishable records.

As an example, consider a data set with 16 records, two quasi-identifier attributes Q_1 and Q_2, and one confidential attribute C. Assume that Q_1 and Q_2 take values in the sets $\{A, B, C, D\}$ and $\{P, Q, R, S\}$, respectively. Figure 1 represents the projection of this data set on the quasi-identifier attributes; each point is the projection of one record on the quasi-identifiers. In this case, we assume that each possible combination of quasi-identifiers occurs in exactly one record. The dominant approach towards k-anonymity uses generalization and suppression to partition the data set into groups of k indistinguishable records. Assume that the generalization hierarchies are those in Figure 2, and that we want to obtain a 4-anonymous data set. Figure 3 depicts the 4-anonymous data sets produced by minimal generalizations. A data user interested in the group of individuals with $Q_1 = A$ would prefer the 4-anonymous data set on the left, which still allows distinguishing that group. On the contrary, the data set on the right of Figure 3 is the worst option for a user interested in the individuals with $Q_1 = A$, because all values of Q_1 are lumped together. However, a user interested in the group of individuals with $Q_2 = P$ would prefer the 4-anonymous data set on the right of Figure 3.

Therefore, the selected partitioning of the records is essential for the protected data set to deliver high utility. If the data collector is aware of the kind of analyses that data users are

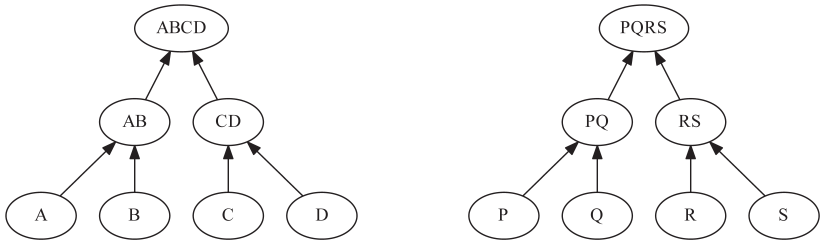


Fig. 2. Generalization hierarchies for attributes Q_1 (left) and Q_2 (right)

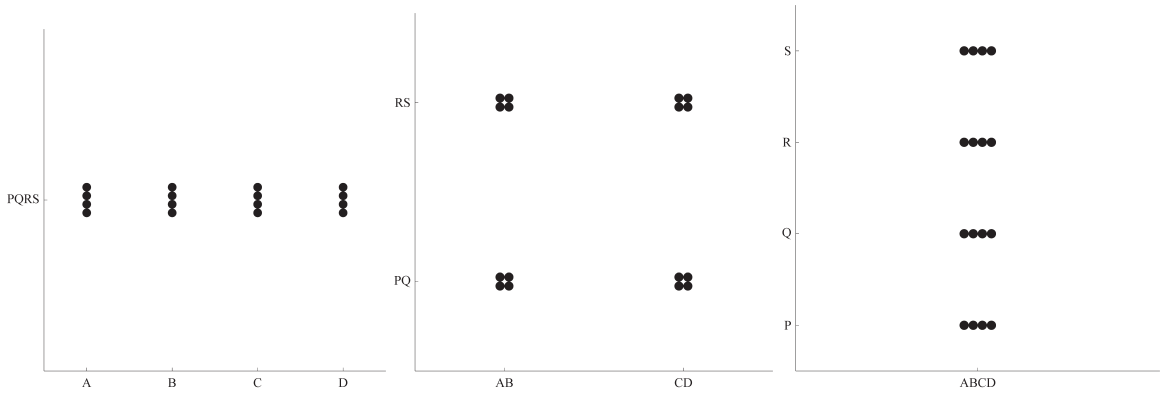


Fig. 3. Minimal 4-anonymous generalizations

interested in, then the collector can tailor the partitioning to those analyses. However, most of the time the data collector is unaware of the intended use of the data; thus, a customized partitioning is not feasible. Even if the data collector knew the relevant analyses, different analyses may require different partitions, but releasing several versions of the same data set using a different partition each is not advisable, as it would endanger whatever anonymity is gained by partitioning.

Another problem of the generalization approach is related to the number of quasi-identifiers. When there is a large number of quasi-identifiers, all of them need to be generalized to satisfy k-anonymity, which results in a large information loss. This is known as "the curse of dimensionality" [1].

This section aims at a method to generate k-anonymous data sets that mitigates the issues described above:

- We generate a partition of the records that preserves as much information as possible. To construct such a partition, instead of partitioning based on the quasi-identifiers, we will do it based on the confidential attributes.
- To avoid losing information on the quasi-identifier attributes, we replace generalization of the quasi-identifiers by an approach that preserves both the quasi-identifiers and the confidential attributes. In particular, we propose to use either the Anatomy [15] or the probabilistic k-anonymity [13] methods.

If our goal is to construct a k-anonymous data set, the Anatomy method is better, as it preserves more information, namely the distribution of the confidential attributes within each set in the partition. However, if our goal is to achieve t-closeness, we will show that the probabilistic k-anonymity approach is preferable.

A. Partitioning based on the confidential attribute

We have argued above that customizing the k-anonymous partition to specific data analysis requirements is not an option. Hence, the utility of the data depends on the amount of variability of the confidential attributes. For instance, if we target a specific individual, the quasi-identifiers allow us to determine a group of k records that must contain that individual; thus, we know that each of the values of a confidential attribute within the group has probability $1/k$ of corresponding to the target individual. The amount of knowledge we get (and thus the utility) depends on the variability of the confidential attribute within the group: the more similar the confidential attribute values, the more knowledge for the user, but also the higher the risk of attribute disclosure.

For the sake of clarity, we consider a single confidential attribute in what follows. If there are several confidential attributes, we can treat them as a single compound confidential attribute and partition the data set according to a proximity criterion that takes into account all the components (*e.g.* microaggregation over confidential attributes [4]).

To limit the variability of the confidential attribute within groups of indistinguishable records, we propose to partition the records in the data set based on the value of the confidential attribute. We focus on a numerical or ordinal confidential

TABLE I. DATA SET WITH QUASI-IDENTIFIERS QI AND A CONFIDENTIAL ATTRIBUTE C

	QI	C
individual 1	q_1	c_1
$\vdots$	$\vdots$	$\vdots$
individual N	q_N	c_N

Algorithm 1 Optimal partitioning based on the confidential attribute

let $D = \{(q_i, c_i) | i = 1, \ldots, N\}$ be the original data set
let $P = \emptyset$ the partition of D to be returned
let $O = ((oq_1, oc_1), \ldots, (oq_N, oc_N))$ be the list of records of D ordered by ascending values c_i
while $|O| \geq 3k$ **do**
  let P_{min} be the set containing the first k records of O
  insert P_{min} into P
  remove the first k records from O
  let P_{max} be the set containing the last k records of O
  insert P_{max} into P
  remove the last k records from O
end while
if $|O| \geq 2k$ **then**
  let P_{min} be the set containing the first k records of O
  insert P_{min} into P
  remove the first k records from O
end if
let P_{rest} be the set with the records remaining in O
insert P_{rest} into P
return P

attribute. Let D be a data set with quasi-identifiers collectively denoted as QI, and a confidential attribute C, as represented in Table I. To minimize the variability of C, we sort the records by C, and generate the partition by taking the k minimal and maximal records, iteratively (see Algorithm 1).

B. Anatomy: reducing information loss in quasi-identifiers

We have mentioned above the "curse of dimensionality" information loss problem inherent to generalizations affecting many quasi-identifier attributes. The problem may get even worse if we construct the partition based on the confidential attribute, as proposed in the previous section. The values of the quasi-identifier attributes in each group of the partition may span the whole domains of those attributes (or substantial fractions of them). Therefore, replacing all values of each quasi-identifier attribute within a group by a single generalized value would lead to a great utility loss. Moreover, note that the generalized values for the quasi-identifiers might coincide for different groups.

To overcome this difficulty, we propose to use the Anatomy approach to k-anonymity, which preserves the original values of the quasi-identifiers. To dissociate (break the relation between) quasi-identifiers and confidential attributes, two tables are generated: the first one assigns a group identifier to the quasi-identifiers, and the second one relates each group identifier to the confidential attributes. We illustrate this in Tables II, III and IV. Table II shows the original de-identified data. Table III presents a 3-anonymous version of the data obtained by generalization of the attributes *Date of Birth* and *Sex*. Note

TABLE II. ORIGINAL DE-IDENTIFIED MEDICAL DATA

Ethnicity	Date of Birth	Sex	Problem
asian	09/27/64	female	hypertension
asian	05/08/61	female	obesity
asian	04/18/64	male	chest pain
black	03/13/63	male	hypertension
black	03/18/63	male	shortness of breath
black	09/07/64	female	obesity
white	05/14/61	male	chest pain
white	05/08/63	male	obesity
white	09/15/61	female	shortness of breath

TABLE III. 3-ANONYMOUS DATA SET

Ethnicity	Date of Birth	Sex	Problem
asian	[61,64]	-	hypertension
asian	[61,64]	-	obesity
asian	[61,64]	-	chest pain
black	[61,64]	-	hypertension
black	[61,64]	-	shortness of breath
black	[61,64]	-	obesity
white	[61,64]	-	chest pain
white	[61,64]	-	obesity
white	[61,64]	-	shortness of breath

that we have used the greatest level of generalization for those attributes, and thus the information loss is large. In contrast, we observe in Table IV that, by using a group identifier to relate quasi-identifier attributes and confidential attributes, we achieve exactly what we wanted: we k-anonymize *the relation* between quasi-identifiers and confidential attributes, while preserving the values of quasi-identifier attributes and the confidential attribute. In particular, the distribution of the confidential attribute within each group (records sharing the same group identifier) is preserved.

C. Comparison of partitioning strategies

When generating the partition based on the quasi-identifiers, small values of the parameter k are typically used. Usually, the variability of the confidential attribute thus obtained is large enough not to lead to attribute disclosure. However, when basing the partition on the confidential attribute, small values of k will almost certainly lead to attribute disclosure, because in this case the within-group variability of the confidential attribute is small.

However, constructing the partition based on the confidential attribute has one important advantage: it allows fixing the desired level of variability for the confidential attribute. Indeed, parameter k can be increased to a value that provides effective disclosure limitation guarantees. For instance, by setting k to $0.1 \times N$, we guarantee that the confidential attribute for any individual is hidden inside a group of individuals that amount to a 10% of the actual sample.

TABLE IV. LEFT, RELATION BETWEEN QUASI-IDENTIFIERS AND GROUP IDENTIFIER. RIGHT, RELATION BETWEEN GROUP IDENTIFIER AND CONFIDENTIAL ATTRIBUTE.

Ethnicity	Date of Birth	Sex	ID	ID	Problem
asian	09/27/64	female	1	1	hypertension
asian	05/08/61	female	1	1	obesity
asian	04/18/64	male	1	1	chest pain
black	03/13/63	male	2	2	hypertension
black	03/18/63	male	2	2	shortness of breath
black	09/07/64	female	2	2	obesity
white	05/14/61	male	3	3	chest pain
white	05/08/63	male	3	3	obesity
white	09/15/61	female	3	3	shortness of breath

Note that if partitioning is based on the quasi-identifiers, we cannot control the level of variability of the confidential attribute inside each of the k-anonymous groups: some of them may exhibit a large variability (which offers protection against attribute disclosure, but poor data utility) and others may not (which offers good data utility, but high risk of attribute disclosure). The underlying problem is the impossibility of enforcing a *predetermined* amount of variability: variability increases with k, but the relationship between k and the amount of variability of the confidential attribute is not clear. Usually, k must be small (if any utility is to be provided), which results in poor disclosure limitation guarantees.

IV. A BUCKETIZATION CONSTRUCTION TO ACHIEVE t-CLOSENESS

It has been argued above that when partitioning is based on the confidential attribute, the value of k must be increased to provide effective disclosure limitation. In this section we seek to enforce a stronger disclosure limitation criterion: t-closeness. t-Closeness limits the knowledge gain that an intruder can derive from the k-anonymous groups. The distribution of the confidential attribute within each of the k-anonymous groups is required to be similar to the distribution of the confidential attribute on the whole data set.

As recalled in Section II-B above, for t-closeness to be satisfied the distance between the data set-level and the group-level distribution of the confidential attribute must be less than t for any group. When t-closeness was introduced, the Earth Mover's distance (EMD) was proposed [9]. The EMD measures the minimal amount of work required to transform one distribution to another by moving probability mass between each other.

The kind of guarantee that t-closeness offers depends on the distance function used. We aim at achieving an ε-differentially privacy-like guarantee, and this requires us to use a different distance. ε-Differential privacy guarantees that, for any two data sets that differ in one individual, the probability for a query response computed on either data set to belong to an arbitrary set S differs at most by a factor $\exp(\varepsilon)$. The distance function we propose mimics the ε-differential privacy criterion.

Definition 2. Given two random distributions $\mathcal{D}_1$ and $\mathcal{D}_2$, we define the distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ as:

$$d(\mathcal{D}_1, \mathcal{D}_2) = \max_S \{\frac{\Pr_{\mathcal{D}_1}(S)}{\Pr_{\mathcal{D}_2}(S)}, \frac{\Pr_{\mathcal{D}_2}(S)}{\Pr_{\mathcal{D}_1}(S)}\}$$

where S is an arbitrary (measurable) set, and we take the quotients of probabilities to be zero, if both $\Pr_{\mathcal{D}_1}(S)$ and $\Pr_{\mathcal{D}_2}(S)$ are zero, and to be infinity if only the denominator is zero.

In terms of t-closeness, this distance implies that the distribution of the confidential attribute in the k-anonymous groups must differ from the global distribution of the confidential attribute by a factor not greater than t (and not smaller than $1/t$).

If the distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ are discrete (as it is the case for the sampling distribution of the confidential attribute in a microdata set), computing the distance between them is simpler: taking the maximum over the possible individual values suffices.

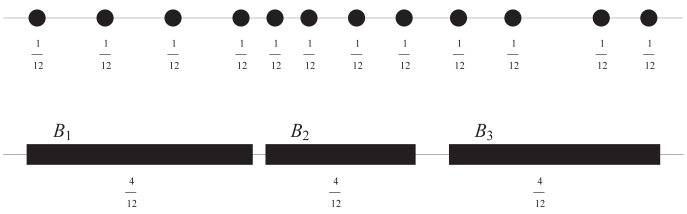


Fig. 4. Top, original confidential attribute values. Bottom, bucketized confidential attribute values.

Proposition 1. *If distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ take values in a discrete set $\{x_1, \ldots, x_N\}$, then the distance $d(\mathcal{D}_1, \mathcal{D}_2)$ can be computed as*

$$d(\mathcal{D}_1, \mathcal{D}_2) = \max_{i=1,\ldots,N} \{\frac{\Pr_{\mathcal{D}_1}(x_i)}{\Pr_{\mathcal{D}_2}(x_i)}, \frac{\Pr_{\mathcal{D}_2}(x_i)}{\Pr_{\mathcal{D}_1}(x_i)}\} \qquad (2)$$

To satisfy t-closeness, the groups in the partition must be selected such that the distance of the distribution of the confidential attribute on the whole data set and the distribution on each of the groups is less than t. When using the previously defined distance, if we work with the sampling distribution of the confidential attribute (assuming that at least one of the values of the confidential attribute has multiplicity less than the cardinality of the partition), the distance is always infinity. The reason is that the distance due to values of the confidential attribute that do not appear within the group is infinity (according to Definition 2). To avoid this issue, instead of working with the sampling distribution of the confidential attribute, we work with a bucketized version of it, where several points are clustered into a set of buckets $B_1, \ldots, B_n$. In Figure 4 the values of the confidential attribute in the original data have been clustered in buckets B_1, B_2 and B_3 that contain four points each. From this step we get a distribution for the confidential attribute with diminished granularity.

By using the proposed bucketization it is feasible to attain t-closeness for a finite t. For instance, Figure 5 shows a 4-anonymous partition of the data set that satisfies 1.5-closeness, according to the previously defined distance. The sampling distribution of the original data assigns probability $1/3$ to each of the buckets B_1, B_2 and B_3; hence, the bucket-level distribution $\mathcal{D}$ of the confidential attribute in the original data set is $\Pr(B_1) = \Pr(B_2) = \Pr(B_3) = 1/3$. Each of the groups in the partition (P_1, P_2, P_3) takes either one or two points from each bucket. Hence, the bucket-level distribution $\mathcal{D}(P_1)$ of the confidential attribute for group P_1 is $\Pr(B_1) = 1/2$ and $\Pr(B_2) = \Pr(B_3) = 1/4$; for group P_2 the distribution, denoted by $\mathcal{D}(P_2)$, is $\Pr(B_1) = \Pr(B_3) = 1/4$ and $\Pr(B_2) = 1/2$; for group P_3 the distribution, denoted by $\mathcal{D}(P_3)$, is $\Pr(B_1) = \Pr(B_2) = 1/4$ and $\Pr(B_3) = 1/2$. By using Equation (2) to measure the distance between $\mathcal{D}$ and $\mathcal{D}(P_i)$, for all i, we conclude that the generated partition satisfies 1.5-closeness. In Table 5 we have depicted both the original set of values of the confidential attribute, and the generated buckets.

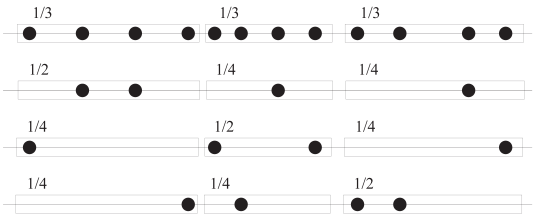


Fig. 5. Sample partition that satisfies 1.5-closeness

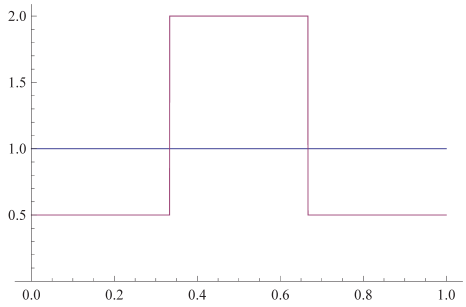


Fig. 6. Probability distributions satisfying 2-closeness with the distance of Definition 2

Let the points in the original data set depicted in Figure 5 be of the form (qi_i, c_i), where c_i is the value of the confidential attribute and $c_i \geq c_j$ for $i \geq j$. According to the Anatomy approach to k-anonymity, the 4-anonymous 1.5-close resultant data, associated to partition $\{P_1, P_2, P_3\}$ in Figure 5 and bucketization $\{B_1, B_2, B_3\}$ in Figure 4, consists of the two linked tables displayed in Table V.

A. Bucketization of the original data

The selected bucketization of the confidential attribute has a large impact on data utility: if the bucketization is too coarse, the information loss in the confidential attribute is large; if the bucketization is too fine, it may not be possible to attain t-closeness. In this section we seek to determine the optimal size (in terms of probability mass) of the buckets.

Figure 6 illustrates two probability distributions: the uniform distribution represents the global distribution of the confidential attribute (over the whole data set), and the other distribution corresponds to the confidential attribute restricted to a group P_i. Assuming a uniform $\mathcal{U}[0, 1]$ distribution can be done without loss of generality: the bucketization construction we propose depends on the relative ordering of the attribute values, but not on the actual values of the attribute. These two distributions satisfy 2-closeness with the distance of Definition 2: the density of the restriction to P_i equals $1/2$ for all the range of values of the confidential attribute, except for a range of values that has density 2.

When bucketizing the distributions in Figure 6, the range of values with density 2 should exactly correspond to a bucket or a union of buckets, in order to maximize the utility of the data. This is illustrated in Figure 7, whose top row shows bucketized versions of the distributions of Figure 6 using *three* buckets: top left graph, bucketized version of the global distribution; top right graph, bucketized version of the restriction to P_i. Note that, for each of the buckets, the global probability and the probability restricted to P_i differ by a multiplicative factor of two; that is, we attain 2-closeness with equality for each of the buckets. The bottom row of Figure 7 shows the bucketized versions of the distributions in Figure 6 using *two* buckets. It can be seen that, with the two proposed buckets, both bucketized distributions are identical; that is, we get 1-closeness, which is stronger than the intended 2-closeness, but comes at the cost of data utility loss. Therefore, the number and hence the probability mass of the optimal buckets is dependent on the level of t-closeness that we want.

In general, the smaller the number b of buckets, the easier it is to achieve t-closeness, for any given t. In the extreme case $b = 1$, all bucketized distributions are 1-close (*e.g.* there is a single bucketized distribution). In the other extreme case $b = N$ (no bucketization) it has been argued above (right after Proposition 1) that the distance between the distributions of the confidential attribute on the global data set and on a particular group is infinity; hence, one can only achieve ∞-closeness. Hence, at most b can be k, the number of values in each group, and buckets should be large enough so that, when restricted to any group, any bucket contains at least one value.

On the other hand, if the privacy requirement is t-closeness, for a certain t, it seems reasonable to use up the allowed distance t between the global distribution of the confidential attribute and the restriction of that distribution within each group. Using up the allowed distance between the confidential attribute distributions enables forming groups that are more homogeneous in terms of the values of the confidential attribute they contain, and hence decreases information loss. We want each of the k-anonymous groups to emphasize a specific bucket; that is, the probability distribution of the restriction to the partition must differ from the global distribution by a factor of t for a specific bucket, and by a factor of $1/t$ for the rest of buckets. Now, in the distribution of the confidential attribute for the original data set each bucket accumulates probability mass $1/b$, and the total probability mass of the distribution restricted to a group must add to 1. Hence, we have

$$t \times 1/b + (1/t) \times (1 - 1/b) = 1$$

which yields a number of buckets $b = t + 1$.

Ideally, each bucket should contain a probability mass $1/b$; however, as we are dealing with a finite data set, a probability mass exactly equal to $1/b$ may be unreachable, and we may need to approximate it (by taking either $\lfloor N/b \rfloor$ or $\lceil N/b \rceil$ records in each bucket). For instance, assuming that the records can be sorted by the confidential attribute c_i (this is possible if c_i is numerical or ordinal) we can list the contents of the approximating buckets as follows:

$$B_1 = \{c_1, \ldots, c_{\left[\frac{N}{t+1}+0.5\right]}\}$$
$$B_2 = \{c_{\left[\frac{N}{t+1}+0.5\right]+1}, \ldots, c_{\left[2\times\frac{N}{t+1}+0.5\right]}\}$$
$$\vdots$$
$$B_{t+1} = \{c_{\left[t\times\frac{N}{t+1}+0.5\right]+1}, \ldots, c_N\}$$

This is a balanced partition of the confidential attribute in $b = t + 1$ buckets; that is, the cardinality of each of the buckets differs at most in one record. Therefore this is the best approximation to a partition in b buckets with probability mass $1/b$ that we can get.

TABLE V. 4-ANONYMOUS 1.5-CLOSE DATA SET ASSOCIATED TO THE PARTITION $\{P_1, P_2, P_3\}$ IN FIGURE 5 AND BUCKETIZATION IN FIGURE 4

QI	qi_1	qi_2	qi_3	qi_4	qi_5	qi_6	qi_7	qi_8	qi_9	qi_{10}	qi_{11}	qi_{12}
Group Id	P_2	P_1	P_1	P_3	P_2	P_3	P_1	P_2	P_3	P_3	P_1	P_2
Group Id	P_1	P_1	P_1	P_1	P_2	P_2	P_2	P_2	P_3	P_3	P_3	P_3
Bucket	B_1	B_1	B_2	B_3	B_1	B_2	B_2	B_3	B_1	B_2	B_3	B_3

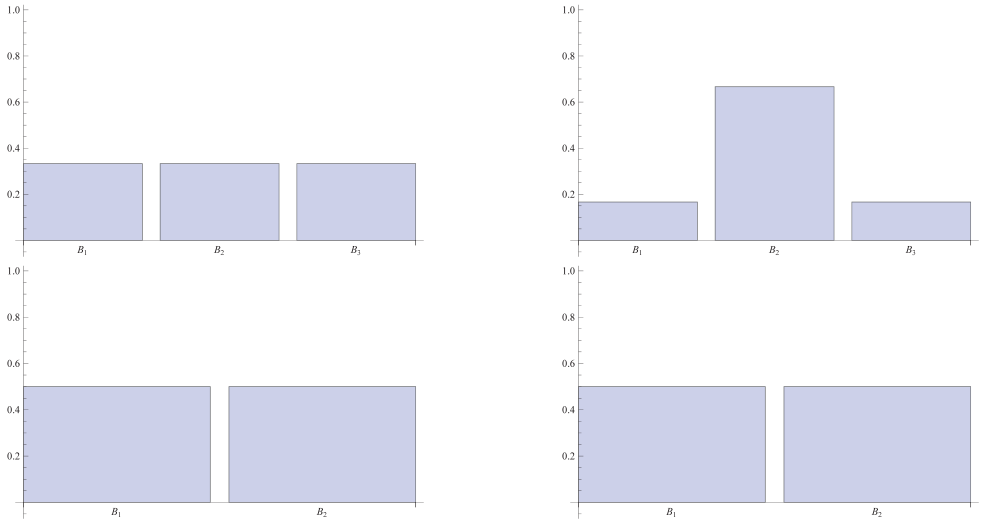


Fig. 7. Bucketized distributions of the confidential attribute for the whole data (left) and for a group P_i (right). Three buckets are considered in the top distributions, and two in the bottom ones.

See Algorithm 2 for a description of the process of generation of a t-close data set that includes the bucketization just described.

B. t-Closeness construction

Consider the original data set $D = \{(qi_i, c_i)|i = 1, \ldots, N\}$, where qi_i refers to the quasi-identifier attributes, and c_i to the confidential attribute. We want to generate a k-anonymous t-close data set D'. In this section we assume that D requires no additional bucketization; that is, either D has already been bucketized $D = \{(qi_i, B_{j(i)}), i = 1, \ldots, N\}$, or no bucketization was needed in the first place.

To generate the k-anonymous t-close data set, we need to partition D in groups of k (or more) records so that t-closeness is satisfied. Not all values of k are equally suitable. For instance, it must be $k \geq t + 1$, because we showed in Section IV-A that $b \leq k$ and $b = t + 1$. It was seen in that same section that, to improve the utility of the t-close data set, we must select a number of buckets b that allows attaining distance t for each of them. The selection of k also has an impact on the utility of the t-close data set. To improve utility, each of the k-anonymous groups should emphasize a single bucket; that is, it should increase the probability by a factor t in the selected bucket and decrease the probability by a factor $1/t$ in the rest of buckets. To be able to do so, the total number of k-anonymous groups must be a multiple of the number of buckets b. We can write:

$$k = \lfloor \frac{N}{bl} \rfloor = \lfloor \frac{N}{(t+1)l} \rfloor$$

where $l \geq 1$ is a natural number that counts the number of groups that emphasize each of the buckets.

TABLE VI. THEORETICAL PROBABILITY MASS OF THE DISTRIBUTION OF THE CONFIDENTIAL ATTRIBUTE IN EACH OF THE BUCKETS CORRESPONDING TO THE DISCRETIZATION OF THE CONFIDENTIAL ATTRIBUTE.

	B_1	B_2	...	B_{t+1}
Original data	$1/t+1$	$1/t+1$	...	$1/t+1$
P_1	$t/t+1$	$1/t(t+1)$	...	$1/t(t+1)$
P_2	$1/t(t+1)$	$t/t+1$	...	$1/t(t+1)$
⋮	⋮	⋮		⋮
P_{t+1}	$1/t(t+1)$	$1/t(t+1)$	...	$t/t+1$

In fact, if we take into account the previous inequality $k \geq t+1$, we conclude that l belongs to the set $\{1, \ldots, \left\lfloor \frac{N}{(t+1)^2} \right\rfloor\}$. Similarly to the discretization of the confidential attribute, the value of k produced by the previous formula may not be exact. In that case we need to adjust the size k_i of each group P_i to either $\lfloor N/(bl) \rfloor$ or $\lceil N/(bl) \rceil$.

Table VI gives the theoretical probability mass of each bucket of the confidential attribute for each of the groups. We assume that $l = 1$ and that group P_1 emphasizes bucket B_1, P_2 emphasizes bucket B_2, and so on. The exact theoretical probability masses may not be achievable due to the discrete nature of the data. First of all, it may not be possible to obtain a discretization of the confidential attribute in buckets with probability mass $1/(t + 1)$. Also, when generating the k-anonymous partition $P_1, \ldots, P_{t+1}$, it may not be possible for each of the groups to contain exactly k records. Let k_i be the number of records in P_i and let p_j be the probability that a record in the original data set belongs to bucket B_j. For t-closeness to be achieved, the following must hold for every group P_i: (i) at most $\lfloor k_i p_i t \rfloor$ records in P_i must have B_i as the value for the confidential attribute; and (ii) at least

Algorithm 2 $\lfloor N/(t+1) \rfloor$-anonymous t-close data set, by generating $b = t+1$ buckets and b $\lfloor N/(t+1) \rfloor$-anonymous groups

let $D = \{(q_i, c_i) | i = 1, \ldots, N\}$ be the original data set, sorted by the confidential attribute
let t be the desired level of t-closeness
let $b := t+1$ be the desired number of buckets

/ Generate a bucketized data set* $\bar{D} = \{(qi_i, B_j) | i = 1, \ldots, N \wedge c_i \in B_j\}$ *by clustering the* N *values of the confidential attribute in* D *into* b *buckets in such a way that: a) all buckets accumulate the same probability mass* $1/b$ *(or approximates it as closely as possible). b) values within each bucket are as similar as possible (*e.g. *for a numerical confidential attribute, each bucket would contain consecutive values). */*
$\bar{D} =$ Bucketize(D, b)

/ Partition the records in* $\bar{D}$ *into* b *groups* $P_1, \ldots, P_b$*, in such a way that every group satisfies that: a)* k_i*, the cardinality of* P_i*, is greater than* $\lfloor N/(t+1) \rfloor$*, in view of achieving* $\lfloor N/(t+1) \rfloor$*-anonymity; b) no bucket* B_j *contains a proportion of the confidential attribute values of the group* P_i *higher than* tp_j*, where* $p_j = |B_j|/N$ *(that is, more than* $\lfloor tk_ip_i \rfloor$ *records) or lower than* p_j/t *(that is, less than* $\lceil k_ip_j/t \rceil$ *records)*/*
$\{P_1, \ldots, P_b\} =$ Partition($\bar{D}$, b)
return $P_1, \ldots, P_b$

$\lceil k_ip_j/t \rceil$ records in P_i must have B_j as confidential attribute. For these conditions to hold, we can start by populating each P_i with $\lceil k_ip_j/t \rceil$ records with confidential attribute value B_j, for each j, and then evenly distributing the remaining records with confidential attribute value B_j among the groups P_i that emphasize bucket B_j. Note that due to the required approximations in the cardinality of B_j and P_i, we may not attain exact t-closeness, but t'-closeness for some $t' < t$.

The whole process of bucketization and generation of the t-close data set that we have described is summarized, for the optimal values $b = t+1$ and $k = \lfloor N/b \rfloor$, in Algorithm 2.

V. From t-closeness to ε-differential privacy

t-Closeness and ε-differential privacy take approaches towards disclosure limitation that are essentially different. However, for microdata releases a link between them can be found if we make some assumptions on the prior knowledge of intruders:

1) The marginal distribution of the confidential attribute is known to the intruder; actually, this assumption is a requirement, because a t-close data release preserves this marginal distribution.
2) The intruder knows whether an individual's record is in the data set; this is also a requirement, as either of the approaches proposed to generate the t-close data set, Anatomy and probabilistic k-anonymity, preserves the quasi-identifiers.
3) When regular k-anonymity is used, another assumption on the intruder's knowledge is required: the intruder's knowledge about the confidential attribute is limited to its marginal distribution.

We aim at showing that, in the case of a microdata release, $\exp(\varepsilon)$-closeness implies ε-differential privacy. In other words, we want to show that the information that an intruder obtains from accessing the released $\exp(\varepsilon)$-close microdata set (generated as per Section IV) satisfies the ε-differential privacy condition. It was seen in Section IV that attaining t-closeness for our particular distance was not possible for all data sets, and a bucketization construction was proposed to turn non-suitable data sets into suitable ones. The implication between $\exp(\varepsilon)$-closeness and ε-differential privacy is completely general, in the sense that it does not depend on whether any bucketization was used to construct the t-close data set or not.

Let I be a specific individual in the data set. Before accessing the data set, the intruder views the value of the confidential attribute of individual I as being distributed according to the distribution of the confidential attribute over the whole data set. Given the assumption that limits the prior knowledge to the marginal distribution of the confidential attribute, that is the most precise information that the intruder has about I. ε-Differential privacy guarantees that the knowledge gain obtained from the response to a query that asks for I's confidential attribute is at most $\exp(\varepsilon)$; that is, the distribution of the response must differ at most by a factor of $\exp(\varepsilon)$ from the assumed prior knowledge. Note that if we did not take into account the intruder's prior knowledge (usual ε-differentially private mechanisms do not assume any prior knowledge), the $\exp(\varepsilon)$-differentially private distribution for the confidential attribute of individual I's would be different. However, the possibility of using the available prior knowledge exists, and thus any distribution that differs from it by a factor of $\exp(\varepsilon)$ satisfies $\exp(\varepsilon)$-differential privacy.

t-Closeness is an improvement of k-anonymity. As such, it seeks to thwart record re-identification by making each record indistinguishable from $k-1$ other records as far as the quasi-identifiers are concerned. Apart from that, t-closeness requires that the sampling distribution of the confidential attribute within each of the k-anonymous groups be similar to the sampling distribution over the whole data set. Hence, t-closeness effectively limits the knowledge gain that the intruder obtains, that is, it achieves differential privacy.

A. Uninformed intruders

Consider an uninformed intruder. By inspecting the released $\exp(\varepsilon)$-close k-anonymous data set, the intruder associates a k-anonymous group of records to individual I. In this way, the intruder learns the distribution of the confidential attribute within the k-anonymous group P that contains I. As the intruder's prior knowledge is limited to the marginal distribution of the confidential attribute, after accessing the data, the best the intruder can do is to associate the distribution of the confidential attribute in P to individual I. As the released data set satisfies $\exp(\varepsilon)$-closeness (generated as per Section IV), the distribution of P differs at most in a factor $\exp(\varepsilon)$ from the distribution of the whole data set; that is, it satisfies ε-differential privacy.

B. Informed intruders

For an informed intruder (whose knowledge goes beyond the distribution of the confidential attribute over the whole data

set), in general t-closeness does not imply differential privacy. To see this, consider an intruder who knows the value of the confidential attribute for $k-1$ of the k individuals in one of the k-anonymous groups. Such an intruder can determine (with certainty) the confidential attribute value for the remaining individual in the group by simple inspection of the released data; in differential privacy terms, access to the data set has produced a infinite knowledge gain on the confidential attribute of that specific individual.

The above situation is unavoidable if, as k-anonymity does, we intend to preserve the thruthfulness of the confidential attribute inside the k-anonymous groups. However, as we showed in Section III, by increasing k, the problem is mitigated. The greatest mitigation is achieved when k equals the number of records in the data set. In such case, for the intruder to determine the confidential attribute value of any individual with certainty, he should know the confidential attribute for all the other individuals in the data set (strictly speaking, it would be enough to know that none of the other individuals take one of the values in the released data set). The problem with such a large k is that it is likely to severely damage utility.

The destruction of data utility can be mitigated if we hide the k-anonymous groups. In this way, a smaller k can be used, thereby preserving the utility of the data, and a protection equivalent to taking k equal to N is attained. Hiding the k-anonymous groups is feasible if instead of regular k-anonymity, we enforce probabilistic k-anonymity [13]. Probabilistic k-anonymity can be seen as an instance of the Anatomy method for k-anonymity: instead of associating a sampling distribution to each of the groups, the values of the confidential attribute are permuted and assigned to individual records. This process can be viewed as taking a sample of the sampling distribution for each record. When using probabilistic k-anonymity, the intruder cannot determine which records form each of the groups; thus, she cannot use the information about a specific individual to increase her knowledge on the other individuals of the group.

VI. Conclusions and future research

We have shown that the k-anonymity family of models is powerful enough to achieve ε-differential privacy in the context of data publishing. Specifically, using a suitable construction, we have shown that $\exp(\varepsilon)$-closeness implies ε-differential privacy for uninformed intruders and approximate ε-differential privacy for informed intruders. Our t-closeness construction based on bucketization is also a contribution in its own right.

Future research will include extending the proposed approach for nominal confidential attributes, which cannot be ordered. We will also provide a generalization to multiple confidential attributes. Experimental work will be reported to compare the utility of the ε-differentially private data sets obtained via bucketized $\exp(\varepsilon)$-closeness. The very nature of our construction, based on k-anonymity, gives reasonable hopes that more utility may be preserved than the one offered by the Laplace noise addition typically used to achieve ε-differential privacy: for one thing, by design, our approach does not yield any off-range values, which may however appear in noise addition procedures. We will also explore the exact privacy guarantees offered by the approximate ε-differential privacy obtained with our construction for the case of informed intruders.

Acknowledgments and disclaimer

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TIN2011-27076-C03-01 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission under FP7 projects "DwB" and "Inter-Trust". The second author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In: *Proc. of the 31st International Conference on Very Large Data Bases-VLDB 2005*, pp. 901-909, 2005.

[2] A. Blum, K. Ligett and A. Roth. A learning theory approach to non-interactive database privacy. In: *Proc. of the 40th Annual Symposium on the Theory of Computing-STOC 2008*, pp. 609-618, 2008.

[3] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai and L. Xiong. Publishing set-valued data via differential privacy. In: *37th Intl. Conference on Very Large Data Bases-VLDB 2011/Proc. of the VLDB Endowment*, 4(11):1087-1098, 2011.

[4] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195-212, 2005.

[5] J. Domingo-Ferrer. A critique of k-anonymity and some of its enhancements. In: *Proc. of ARES/PSAI 2008*, IEEE Computer Society, pp. 990-993, 2008.

[6] C. Dwork. Differential privacy. In: *Proc. of 33rd International Colloquium on Automata, Languages and Programming-ICALP 2006*, LNCS 4052, Springer, pp. 1-12, 2006.

[7] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In: *Proc. of the 41st Annual Symposium on the Theory of Computing-STOC 2009*, pp. 381-390, 2009.

[8] M. Hardt, K. Ligett and F. McSherry. A simple and practical algorithm for differentially private data release. Preprint arXiv:1012.4763v1, 21 Dec 2010.

[9] N. Li, T. Li and S. Venkatasubramanian. t-Closeness: privacy beyond k-anonymity and l-diversity. In: *Proc. of the 23rd IEEE Intl. Conference on Data Engineering-ICDE 2007*,IEEE, pp. 106-115, 2007.

[10] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian. l-Diversity: privacy beyond k-anonymity. In: *Proc. of the 22nd IEEE Intl. Conference on Data Engineering-ICDE 2006*,IEEE, 2006.

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

[12] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI International Report, 1998.

[13] J. Soria-Comas and J. Domingo-Ferrer. Probabilistic k-anonymity through microaggregation and data swapping. In: *Proc. of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2012*, IEEE, pp. 1-8, 2012.

[14] T. M. Truta and B. Vinay. Privacy protection: p-sensitive k-anonymity property. In: *Proc. of the 22nd Intl. Conference of Data Engineering Workshops*, p. 9, 2006.

[15] X. Xiao and Y. Tao. Anatomy: simple and effective privacy preservation. In: *Proc. of the 32nd Intl. Conference on Very Large Data Bases-VLDB 2006*,pp. 139-150, 2006.