A Non-Parametric Model for Accurate and Provably Private Synthetic Data Sets

Jordi Soria-Comas
Universitat Rovira i Virgili
Dept. of Computer Eng. and Math
UNESCO Chair in Data Privacy
Av. Països Catalans 26
Tarragona, Catalonia E-43007
jordi.soria@urv.cat

Josep Domingo-Ferrer
Universitat Rovira i Virgili
Dept. of Computer Eng. and Math
UNESCO Chair in Data Privacy
Av. Països Catalans 26
Tarragona, Catalonia E-43007
josep.domingo@urv.cat

ABSTRACT

Generating synthetic data is a well-known option to limit disclosure risk in sensitive data releases. The usual approach is to build a model for the population and then generate a synthetic data set solely based on the model. We argue that building an accurate population model is difficult and we propose instead to approximate the original data as closely as privacy constraints permit. To enforce an *ex ante* privacy level when generating synthetic data, we introduce a new privacy model called ε-synthetic privacy. Then, we describe a synthetic data generation method that satisfies ε-synthetic privacy. Finally, we evaluate the utility of the synthetic data generated with our method.

KEYWORDS

Synthetic data, non-parametric methods, formal privacy, ε-synthetic privacy

ACM Reference format:
Jordi Soria-Comas and Josep Domingo-Ferrer. 2017. A Non-Parametric Model for Accurate and Provably Private Synthetic Data Sets. In *Proceedings of ARES Conference, Reggio Calabria, Italy, Aug 29–Sep 1, 2017 (ARES'17),* 10 pages.
DOI: 10.475/123_4

1 INTRODUCTION

The development of data science has boosted the demand for access to microdata sets, that is, data sets whose records correspond to individual subjects. Microdata sets can be generated in-house or by external parties that release them for secondary use. Microdata release faces two competing objectives that must be reconciled. On the one side, the privacy of the subjects to whom the records correspond must be preserved, and this is usually attained by modifying the original data. On the other side, the released data must preserve the utility of the original data as much as possible; that is, the released data must retain enough detail and accuracy. Statistical disclosure control (SDC) [6, 7, 12], also known as anonymization, comprises a variety of techniques whose aim is to find an appropriate trade-off between utility and privacy. These techniques are classified in three broad categories according to the kind of data modifications they perform: perturbative masking (methods that modify the original data without preserving their truthfulness), non-perturbative masking (methods that reduce the level of detail of the original data while preserving their truthfulness) and synthetic data (methods that build a new data set from scratch based on a model of the original data). The focus of this paper is on synthetic data sets.

When building a synthetic data set, there are two possible views of the original data. We may think of the original data set as containing a sample from an underlying population or as containing the whole population of interest. When the original data set is regarded as a sample, the confidential attribute values are only known for individuals in the sample; then a model relating the confidential attributes and the non-confidential attributes is built based on the sample; after that, the population is resampled and the confidential attributes for the resampled individuals (of which only the non-confidential attributes are known) are computed from their non-confidential attributes using the model, which yields the synthetic data set. When the original data set is viewed as containing the whole population, disclosure risk limitation is attained by replacing by simulated values the original attribute values of these individuals.

Generating a fully synthetic data set [18] takes three steps: (i) a model for the population is proposed, (ii) the proposed model is adjusted to the original data set, and (iii) the synthetic data set is generated by drawing from the model (without any further dependency on the original data). The utility of fully synthetic data sets is highly dependent on the accuracy of the adjusted model. If the adjusted model fits well the population, the synthetic data set should be as good as the original data set to make inferences on the population. In this sense, synthetic data are superior in terms of data utility to masking techniques (which always lead to some utility loss). This advantage of synthetic data is, however, mostly theoretical, as (except for any relations between attributes that are known beforehand) the model must be built from the analysis of the original data. Thus, proposing a model that appropriately captures all the properties of the population is, in general, not feasible: there may be dependencies between attributes that are difficult to model or, even, to observe in the original data. Given that only the properties that are included in the model will be present in the synthetic data, it is important to include all the properties of

ARES'17, Reggio Calabria, Italy

DOI: 10.475/123_4

the data that we want to preserve. To reduce the dependency on the models, alternatives to fully synthetic data have been proposed: partially synthetic data [16] and hybrid data [4, 15]. Yet, these alternatives normally result in more disclosure risk than full data synthesis.

Fully synthetic data generation is regarded as quite safe in terms of disclosure. Because the synthetic data are generated based solely on the adjusted model, analyzing the risk of disclosure of the synthetic data can be reduced to analyzing how disclosive is the information that the model incorporates about the original data. Since this information usually merely consists of some statistical properties of the original data, disclosure risk can be thought of being intrinsically limited. However, an overfitted model can still produce synthetic data that disclose too much about the original data. A privacy model such that satisfying it provides an *ex ante* privacy guarantee would make a lot of sense also for synthetic data.

Contribution and plan of this paper

We argue in this paper that current proposals to generating fully synthetic data sets capture the relations between attributes only to a limited extent. Instead of relying on partially synthetic or hybrid data to overcome these limitations, we propose to maximize the amount of information about the original data that is incorporated by the model. Of course, doing so increases the risk of disclosure, so the maximization is necessarily constrained by the tolerable risk. This motivates us to propose a new privacy model, ε-synthetic privacy, which gives a guarantee similar to the one of differential privacy [10]: the generated synthetic data should not reveal the presence or absence of any individual in the original data set.

In Section 2, background on synthetic data generation and differential privacy is given. In Section 3, we first show that differential privacy is not suitable to account for privacy in synthetic data; then, we propose ε-synthetic privacy, that is more suited to this purpose. In Section 4, we describe a method to attain ε-synthetic privacy for absolutely continuous attributes. Section 5 reports an assessment of the utility of the generated synthetic data. Finally, Section 6 contains conclusions and future research directions.

2 BACKGROUND

This section introduces some basic concepts about synthetic data and also about differential privacy.

2.1 Types of synthetic data sets

By synthetic data we mean simulated data, *i.e.* data obtained with a random process. Thus, synthetic data bear no direct relation to any specific record in the original data set. This is in contrast with perturbative and non-perturbative masking, where each record in the anonymized data set is obtained from one or more original records. Three types of synthetic data sets are usually considered: (i) fully synthetic data sets, where every data item has been synthesized, (ii) partially synthetic data sets, where only some attributes of some records are synthesized (usually the ones that present a greater risk of disclosure), and (iii) hybrid data sets, where the original data are mixed with the synthesized data.

2.2 Disclosure risk in synthetic data sets

The generation of a synthesized data item depends only on the adjusted model. Hence, if the information on the original data incorporated by the model is not disclosive, the ensuing synthetic data will be safe. This is the case of existing methods such as multiple imputation [18], which seeks to preserve the distribution of the population; IPSO [2], which seeks to preserve the estimates that can be obtained with the original data; and synthetic data generated by Latin hypercube sampling [3], which seeks to preserve the marginal distributions and rank correlations between attributes. However, limiting the information included in the model also limits the accuracy of the synthetic data, especially when the original data have complex interdependencies. Using more flexible models has been proposed [17], but doing so requires evaluating how much is the disclosure risk thereby increased.

Although the above holds for synthesized data items in both fully and partially synthetic data sets, there are important differences in the disclosure risk limitation provided by these two approaches. If we view the original data set as a sample of some underlying population, a fully synthetic data set can be viewed as a resampling of the population; the population units (individuals) included in the synthetic data set are randomly selected from the original sample. Resampling breaks the relation between the population units in the original and the synthetic data sets and, thus, the risk of re-identification is minimized. This is not the case in partial synthesis, where the population units in the original data set are the same population units present in the partially synthetic data set. The disclosure risk protection in a partially synthetic data set comes from replacing the values of the original data set with higher risk of disclosure by simulated values. The simulated values that are assigned to an individual should be representative but are not directly related to her. Regarding hybrid data sets, we cannot even say that they are independent from the original data; in fact, the specific mixture of original and synthetic data determines the amount of dependence (the more weight given to the original data, the more dependence between original and hybrid data).

Disclosure risk in synthetic data sets is usually evaluated *ex post*, once the synthetic data set has been generated. For example, record linkage can be used for this assessment [5]. This contrasts with the approach based on privacy models proposed by the computer science community [10, 13, 14, 19, 22], that seeks to provide *ex ante* privacy guarantees: a privacy model is actually just a specific target privacy guarantee, which is attained with surety if certain masking methods with certain parameterizations are used to generate the protected data from the original data.

2.3 Differential privacy

Differential privacy [8, 10, 11] is a privacy model that has gained a lot of attention because of the strong privacy protection it offers. It is defined as follows:

Definition 2.1. (ε-differential privacy) A randomized function $\mathcal{K}$ gives ε-differential privacy if for all data sets D_1 and D_2 differing at most in one record, and all $S \subseteq Range(\mathcal{K})$,

$$\Pr(\mathcal{K}(D_1) \in S) \leq \exp(\varepsilon) \times \Pr(\mathcal{K}(D_2) \in S).$$

The definition is illustrated in Figure 1: two data sets D_1 and D_2 that differ in one record are queried with function f, and the distributions of the responses $\mathcal{K}_f(D_1)$ and $\mathcal{K}_f(D_2)$ differ by a factor of at most $\exp(\varepsilon)$. That means that the output does not allow discriminating between D_1 and D_2. In other words, if r is the record that differs between D_1 and D_2, it is not possible to determine the presence or absence of r in the original data set. Since D_1 and D_2 are arbitrary data sets that differ in one record, record r is also arbitrary. That is, the disclosure risk incurred by *any* individual due to her participation in the data sets is limited. Differential privacy is a relative measure: it guarantees that the variation in the disclosure risk between neighbor data sets is similar.

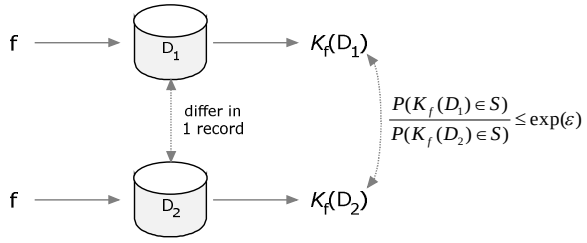

Figure 1: Illustration of ε-differential privacy

3 ACCURATE AND PROVABLY PRIVATE SYNTHETIC DATA

The specification of the model used for synthetic data generation is a critical point. A simple choice may be to use a predictive model based on a sufficient statistic. Such a model lacks the complexity needed to approximate the original data closely in a consistent manner. Hence, the synthetic data set obtained with such a model may substantially differ from the original data set, so its utility is likely to be low (the good side is that its disclosure risk will also be low). By increasing the complexity of the predictive model (e.g. by adding several polynomial terms to a linear regression), we improve its adjustment to the original data, and hence we obtain synthetic data sets that are closer to the original data set. However, an overly complex model has two main shortcomings: on the one hand, it may result in too high a disclosure risk, and on the other hand, it yields worse out-of-sample predictions. In technical terms this is known as overfitting: rather than capturing the population distribution of the attributes of the original data set, the model mimics the values of the original data set. In other words, the bias (the error due to the model) is reduced because the model is complex enough to be adjusted to a broader range of scenarios, but the variance (the error that results from adjusting the model also to the random noise in the original data values) is increased. There is a trade-off between bias and variance. In general, in predictive models, the goal is to minimize the sum of the errors introduced by both bias and variance.

When generating synthetic data for disclosure risk limitation, we may take a different approach. Rather than trying to build a predictive model out of the original data, we may seek a model that approximates the original data as closely as possible. Of course, this approach will lead to overfitting but, if we are interested in the original data rather than in obtaining predictions out of the model, the only shortcoming of overfitting we should take care of is the disclosure risk. Thus, we want protected data which stay as close to the original data as it is possible with tolerable disclosure risk. This is precisely the usual goal of statistical disclosure control.

To implement our approach, we want to generate a joint distribution that follows as closely as possible the empirical distribution of the original data. Unlike the usual parametric modeling in multiple imputation, we take a non-parametric approach. That is, we do not assume the joint distribution to follow any specific family of probability distributions; we only assume that the entries in the original data set are independent from each other.

3.1 Data model

To analyze the risk of disclosure, we focus on the simplified view of the synthetic data generation process depicted in Figure 2: a probability distribution (*i.e.*, a data generation model, not to be confused with the privacy model mentioned above) is fitted to the input data and is then repeatedly sampled to generate the synthetic data set.

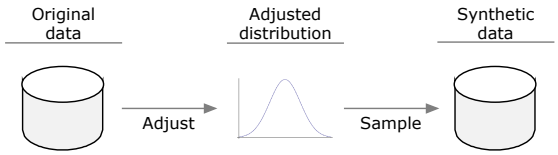

Figure 2: Simplified view of the synthetic data generation process

3.2 Assessing disclosure risk via differential privacy

As mentioned above, fitting a distribution too closely to the original data would increase the disclosure risk beyond what is acceptable. We plan to use a privacy model to assess disclosure in synthetic data, in order to determine how accurate a data generation model we can afford. We first examine whether we can use ε-differential privacy. *A priori*, this privacy model has at least two attractive features:

- It provides strong privacy guarantees.
- Generating synthetic data by repeatedly and independently drawing from the data model (as described in Section 3.1) is comparable to generating an ε-differentially private output in the sense that in both cases the output is a random draw from a given distribution. Thus, differential privacy is readily applicable as a privacy model for synthetic data.

Unfortunately, the outcome of applying differential privacy to the proposed synthetic data generation process turns out to be disappointing. The problem is that, if ε-differential privacy is to be satisfied, the data model cannot mimic the original data set as closely as it should to obtain useful synthetic data. This problem

becomes especially serious for large original data sets, as stated in the following proposition.

Proposition 3.1. *Let D_1 and D_2 be two arbitrary data sets that differ in a single record. Let P_1 and P_2 be the probability distributions used for synthetic data generation that have been fitted to D_1 and D_2, respectively. For any generated synthetic data set of size n to satisfy ε-differential privacy, the probability distributions P_1 and P_2 must differ, at any point, at most by a factor* $\exp(\varepsilon/n)$.

Proof. For differential privacy to hold, the probability of getting a specific synthetic data set must differ at most by a factor of $\exp(\varepsilon)$ between original data sets that differ in one record.

For the case $n = 1$, the proposition immediately results from the definition of differential privacy. Let $S = \{r_1\}$ be a synthetic data set of size 1. For differential privacy to hold it must be $P_1(S) \leq \exp(\varepsilon)P_2(S)$; that is, P_1 and P_2 must differ, at any point, by at most a factor $\exp(\varepsilon)$.

Extending the proof to an arbitrary n is simple. We show it for P_1 and P_2 being discrete probability distributions. Let $S = \{r_1, \ldots, r_n\}$ be a synthetic data set that contains n records obtained by independently drawing from the adjusted distribution. Since the records in S are not ordered, we must take into account that several ordered combinations of records lead to the same S. If S contains m different records $r_1, \ldots, r_m$ each of them appearing n_i times (with $\sum_{i=1}^{m} n_i = n$), then the probability of S is $P(S) = \frac{n!}{n_1! \ldots n_m!} \prod_{i=1}^{m} P(r_i)^{n_i}$. For differential privacy to hold, it must be $P_1(S) \leq \exp(\varepsilon)P_2(S)$. This inequality can be rewritten as

$$\frac{n!}{n_1! \cdots n_m!} \prod_{i=1}^{m} P_1(r_i)^{n_i} \leq \exp(\varepsilon) \frac{n!}{n_1! \cdots n_m!} \prod_{i=1}^{m} P_2(r_i)^{n_i}.$$

Since r_i are arbitrary points in the data domain and the number of repetitions of each r_i is also arbitrary (provided that $\sum_i n_i = n$), we have that P_1 and P_2 must differ, at any point, by a factor at most $\exp(\varepsilon/n)$. To check this, assume there is a point r in the data domain such that the probabilities $P_1(r)$ and $P_2(r)$ differ by a factor greater than $\exp(\varepsilon/n)$. If we consider the data set S that contains n repetitions of r then we have $P_1(S) = P_1(r)^n$ and $P_2(S) = P_2(r)^n$, and the probabilities $P_1(S)$ and $P_2(S)$ differ by a factor greater than $\exp(\varepsilon)$. □

When n grows, the factor $\exp(\varepsilon/n)$ in Proposition 3.1 quickly converges to 1, which implies $P_1 = P_2$ and, thus, P_1 is not representative of D_1 or P_2 is not representative of D_2 anymore (because $D_1 \neq D_2$). In the remainder of the section, we propose a privacy model with strong privacy guarantees (in the line of differential privacy) that is suitable for the synthetic data generation process described in Section 3.1. First of all, we introduce some additional notation.

3.3 Notation

We can view the generation of a synthetic data set as the realization of a random variable X that outputs the synthetic data set. Let $\mathcal{D}_{ds}$ be set of synthetic data sets that can be generated by the adjusted model. Then the synthetic data set is generated by drawing from the random variable $X : \Omega \rightarrow \mathcal{D}_{ds}$. We use P_X to refer to the probability distribution over $\mathcal{D}_{ds}$ induced by X.

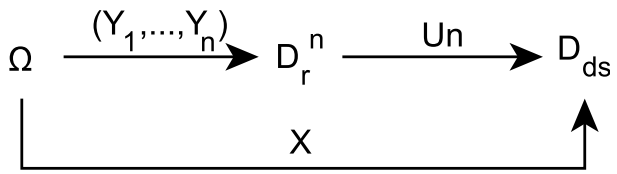

Figure 3: Relation between the random variable X and the random vector $(Y_1, \ldots, Y_n)$

Using the random variable X, the synthetic data set is generated in one shot. However, in Section 3.1 we restricted to a more specific scenario where the synthetic data set is generated by drawing from a set of identically distributed random variables, each of them returning one of the synthetic records. Let $\mathcal{D}_r$ be the domain of the synthetic records and let the identically distributed random variables used to synthesize the records be $Y_i : \Omega \rightarrow \mathcal{D}_r$ for $i = 1, \ldots, n$. We will use P_Y to refer to the probability distribution over $\mathcal{D}_r$ induced by any of the Y_i. Notice that, since a data set is an unordered n-tuple (rather than an ordered one), the output of the random vector $(Y_1, \ldots, Y_n)$ is not exactly a data set. To obtain a data set, we need to compose $(Y_1, \ldots, Y_n)$ with a function that assigns to every ordered n-tuple its corresponding unordered n-tuple. If we use Un to refer to such function, we have $X = Un(Y_1, \ldots, Y_n)$ (see Figure 3).

3.4 A privacy model for synthetic data sets

As shown in Section 3.2, differential privacy is not suitable to assess the disclosure risk in data sets synthesized according to Section 3.1. This is not a general criticism against differential privacy (which has resulted in a wealth of contributions also in the context of data publishing [20, 24, 26, 27]); we only mean that it is not useful in our context. Relaxations of differential privacy (such as (ε, δ)-differential privacy [9] and ε-individual differential privacy [25]) are equally unsuitable. Differentially private data sets are usually generated by approximating a histogram [26] but, except for simple data domains, this approach is problematic both in terms of computational complexity and in terms of utility. Alternative approaches based on records masking with a previous sensitivity reduction phase have been proposed [20, 24]. The latter approaches are more computationally efficient and scale better to complex data domains.

What causes the problem described by Proposition 3.1 is that the condition stated by differential privacy must be enforced even for synthetic data sets that are not representative of the original data set. However, when evaluating disclosure risk, we should be primarily concerned with data sets that are representative of the original data. By chance, accurate inferences about the individuals in the original data set could be made from a non-representative data set, but it unlikely that they qualify as privacy breaches.

In this section, we propose another privacy model with strong privacy guarantees (similar to the ones of differential privacy) that is more adapted to the synthetic data generation process described in Section 3.1.

As previously stated, ε-differential privacy seeks to make the probability of the response to a query similar between data sets that differ in one record. In other words, given a response, we want to

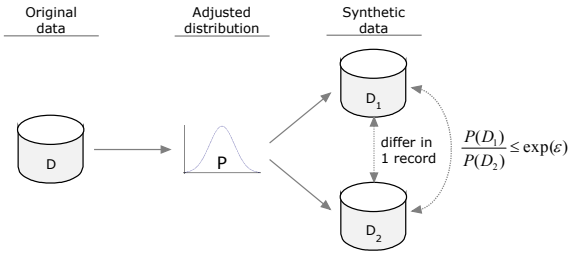

Figure 4: Use of ε-synthetic privacy to evaluate the disclosure risk in synthetic data sets

prevent it from being associated to any specific individual. The fact that we deal with synthetic data allows a different approach to reach the same objective, based on bounding the ratio of probabilities between output synthetic data sets that differ in one record. This yields the following new privacy model:

Definition 3.2. (ε-synthetic privacy) Let $X : \Omega \to \mathcal{D}_{ds}$ be a random variable that is sampled to generate the synthetic data set in a single shot. We say that X is ε-synthetically private if, for all U_1 and U_2 subsets of $\mathcal{D}_{ds}$ whose data sets differ in one record, we have

$$P_X(U_1) \leq \exp(\varepsilon) P_X(U_2).$$

When we say in Definition 3.2 that U_1 and U_2 are subsets of $\mathcal{D}_{ds}$ whose data sets differ in one record, we mean that there are two records r_1 and r_2 (with $r_1 \neq r_2$) in $\mathcal{D}_r$ such that U_2 can be obtained from U_1 by replacing one instance of r_1 by r_2 in each of the data sets contained in U_1. For technical reasons the above definition is expressed in terms of subsets of $\mathcal{D}_{ds}$, but the intent of the definition becomes apparent when dealing with a single data set: $U_1 = \{D_1\}$ and $U_2 = \{D_2\}$, where D_2 can be obtained from D_1 by replacing one instance of r_1 by r_2 (see Figure 4).

If X is ε-synthetically private, it is not possible to determine with enough certainty if any of the records in the generated output data set was in the original data set. A comparison with ε-differential privacy may help understand our proposal. Figures 1 and 4 illustrate differential privacy and ε-synthetic privacy, respectively. In differential privacy, the probability of obtaining an output data set D' is similar between input data sets D_1 and D_2 that differ in one record r. Therefore, given the output dataset D', it is not possible to determine whether it came from D_1 or from D_2. In other words, it is not possible to determine whether r was in the original data set D. In ε-synthetic privacy, given the input data set D and an arbitrary record r, the probability of obtaining as output a data set D_1 or a data set D_2 that differ in record r is similar. Therefore, it is not possible to determine if the record r was in the input data set.

4 SYNTHETIC PRIVACY FOR CONTINUOUS DATA

This section shows how to adjust the data model distribution to the original data when the random variable X returning the synthetic data set is absolutely continuous (a.c.).

4.1 The privacy model in the continuous case

First, we show that Definition 3.2 can be simplified when X is a.c. by expressing it in terms of the density function of X evaluated on data sets D_1 and D_2 that differ in one record.

Definition 4.1. We say that an a.c. random variable $X : \Omega \to \mathcal{D}_{ds}$ is ε-synthetically private if, for all D_1 and D_2 in $\mathcal{D}_{ds}$ that differ in one record, we have

$$f_X(D_1) \leq \exp(\varepsilon) f_X(D_2),$$

where f_X is the density function associated to X.

We next show that Definition 4.1 implies Definition 3.2.

Proposition 4.2. *Let $X : \Omega \to \mathcal{D}_{ds}$ be an a.c. random variable. If X is ε-synthetically private according to Definition 4.1, then it is ε-synthetically private according to Definition 3.2.*

Proof. Let S_1 and S_2 be subsets of $\mathcal{D}_{ds}$ whose data sets differ in one record in the sense explained right after Definition 3.2 above. We want to show that $P_X(S_1) \leq \exp(\varepsilon) P_X(S_2)$.

Let r_1 and r_2 be the records that differ between the data sets in S_1 and S_2. Let $g : \mathcal{D}_{ds} \to \mathcal{D}_{ds}$ be the function that when applied to a data set replaces one instance of r_1 by r_2. S_2 is obtained by applying g to each of the data sets in S_1.

The probability $P_X(S_1)$ can be computed as the integral of the density f_X over S_1:

$$P_X(S_1) = \int_{D_1 \in S_1} f_X(D_1).$$

Since Definition 4.1 holds for X, we have $f_X(D_1) \leq \exp(\varepsilon) f_X(g(D_1))$. By substituting in the previous integral we have

$$P_X(S_1) \leq \exp(\varepsilon) \int_{D_1 \in S_1} f_X(g(D_1)).$$

As $f_X \geq 0$ and $S \subseteq g^{-1}(g(S))$ we have

$$P_X(S_1) \leq \exp(\varepsilon) \int_{D_1 \in g^{-1}(g(S_1))} f_X(g(D_1)).$$

Since $D_2 = g(D_1)$ and $D_1 \in g^{-1}(g(S_1)) \Leftrightarrow D_2 \in S_2$, we have

$$P_X(S_1) \leq \exp(\varepsilon) \int_{D_2 \in S_2} f_X(D_2) = \exp(\varepsilon) P_X(S_2).$$

□

In both Definitions 3.2 and 4.1, the generated synthetic data set is viewed as the realization of a random variable X, but in Section 3.3 we mentioned that the synthetic data set was generated by drawing from a list of random variables $Y_1, \ldots, Y_n$. Let us express the ε-synthetic privacy condition in terms of $Y_1, \ldots, Y_n$. Given that the ε-synthetic privacy condition for an a.c. X can be expressed in terms of the density f_X, we simply need to express f_X in terms of $f_{(Y_1,\ldots,Y_n)}$. For illustration purposes, we consider first the case $\mathcal{D}_r \subset \mathbb{R}$ and $n = 2$ (see Figure 5). We can split $\mathcal{D}_r \times \mathcal{D}_r$ (except by a set of measure 0) into open sets U_i such that the function Un restricted to each of these sets is bijective and of class C^1 (the open sets are those above and below the dotted line in the left graph of Figure 5). Then the density function f_X can be expressed in terms of $f_{(Y_1,Y_2)}$ as

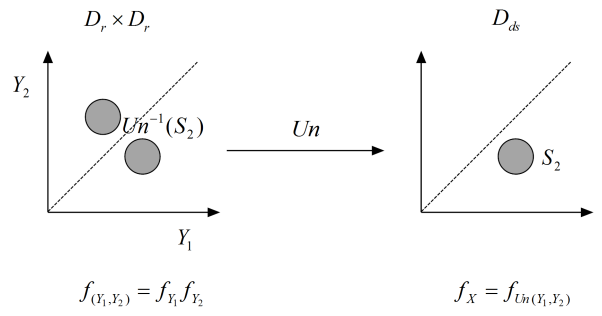

Figure 5: Relation between ordered 2-tuples (to the left) and unordered 2-tuples (to the right)

$$f_X(\{r_1, r_2\}) = \sum_{i:\{r_1,r_2\}\in Un(U_i)} f_{(Y_1,Y_2)}(Un^{-1}_{|U_i}(\{r_1, r_2\}))|J_{Un^{-1}_{|U_i}}(\{r_1, r_2\})|.$$

The previous formula may seem rather complex at a first sight, but it is easily simplified. First of all, for each i, the function $Un^{-1}_{|U_i}$ is a Euclidean plane isometry; thus, the Jacobian $J_{Un^{-1}_{|U_i}}$ is either 1 or -1. In any case, by taking the absolute value we get 1. Secondly, the goal of the summation is to consider all the possible ordered tuples that lead to the unordered tuple $\{r_1, r_2\}$. This is more clearly expressed in terms of permutations. As a result we have:

$$f_X(\{r_1, r_2\}) = \sum_{\sigma\in\mathcal{S}_2} f_{(Y_1,Y_2)}(r_{\sigma(1)}, r_{\sigma(2)}),$$

where $\mathcal{S}_2$ is the set of all possible permutations of two elements.

It is straightforward to generalize the previous formula to data sets of size n. In this case, the ordered tuples corresponding to the unordered tuple $\{r_1, \ldots, r_n\}$ are given by the permutations in S_n

$$f_X(\{r_1, \ldots, r_n\}) = \sum_{\sigma\in\mathcal{S}_n} f_{(Y_1,\ldots,Y_n)}(r_{\sigma(1)}, \ldots, r_{\sigma(n)}). \quad (1)$$

To avoid losing generality, we have not required the random variables $Y_1, \ldots, Y_n$ to be independent from one another. However, when they are independent, the ε-synthetic privacy condition can be further simplified and expressed in terms of f_Y. This is shown in the following proposition.

Proposition 4.3. *Let $Y_1, \ldots, Y_n : \Omega \to \mathcal{D}_r$ be independent identically distributed (i.i.d.) random variables with density f_Y. Let $Un : \mathcal{D}_r^n \to \mathcal{D}_{ds}$ be the function that assigns the corresponding unordered n-tuple to an ordered n-tuple. Then $X = Un(Y_1, \ldots, Y_n)$ is ε-synthetically private if, and only if, $\forall r, s \in \mathcal{D}_r \setminus Z$, with Z a set of measure 0, $f_Y(r) \leq \exp(\varepsilon) f_Y(s)$.*

Proof. According to Proposition 4.2, $X = Un(Y_1, \ldots, Y_n)$ is ε-synthetically private if, and only if, $f_X(D_1) \leq \exp(\varepsilon) f_X(D_2)$ for any two data sets D_1 and D_2 differing in one record. Let $D_1 = \{r_1, r_2, \ldots, r_n\}$ and $D_2 = \{s_1, s_2, \ldots, s_n\}$ be data sets, where $r_i = s_i$ for all $i \geq 2$ and $r_1 = r \neq s_1 = s$ are the records that differ. Given the previous discussion expressing f_X in terms of $f_{(Y_1,\ldots,Y_n)}$, the ε-synthetic privacy condition can be stated as

$$\sum_{\sigma\in\mathcal{S}_n} f_{(Y_1,\ldots,Y_n)}(r_{\sigma(1)}, \ldots, r_{\sigma(n)}) \leq \exp(\varepsilon) \sum_{\sigma\in\mathcal{S}_n} f_{(Y_1,\ldots,Y_n)}(s_{\sigma(1)}, \ldots, s_{\sigma(n)}).$$

As the random variables $Y_1, \ldots, Y_n$ are i.i.d., $f_{(Y_1,\ldots,Y_n)}(y_1, \ldots, y_n)$ is equal to $f_Y(y_1) \cdots f_Y(y_n)$. Thus, the ε-synthetic privacy condition can be rewritten as

$$\sum_{\sigma\in\mathcal{S}_n} f_Y(r_{\sigma(1)}) \cdots f_Y(r_{\sigma(n)}) \leq \exp(\varepsilon) \sum_{\sigma\in\mathcal{S}_n} f_Y(s_{\sigma(1)}) \cdots f_Y(s_{\sigma(n)}).$$

Now, all terms in the left-hand side summation are equal; thus, the summation $\sum_{\sigma\in\mathcal{S}_n} f_Y(r_{\sigma(1)}) \cdots f_Y(r_{\sigma(n)})$ equals $n! f_Y(r_1) \cdots f_Y(r_n)$. Similarly, $\sum_{\sigma\in\mathcal{S}_n} f_Y(s_{\sigma(1)}) \cdots f_Y(s_{\sigma(n)})$ equals $n! f_Y(s_1) \cdots f_Y(s_n)$. Hence, the ε-synthetic privacy condition can be written as

$$n! f_Y(r_1) \cdots f_Y(r_n) \leq \exp(\varepsilon) n! f_Y(s_1) \cdots f_Y(s_n),$$

and by simplifying (recall that $r_i = s_i$ for all $i \geq 2$) we get $f_Y(r) \leq \exp(\varepsilon) f_Y(s)$. □

4.2 Adjusting the density to the original data set

We have seen that for i.i.d. absolutely continuous random variables, ε-synthetic privacy requires the density function f_Y to satisfy $f_Y(r_1) \leq \exp(\varepsilon) f_Y(r_2)$ for any r_1 and r_2 in $\mathcal{D}_r$. In this section we propose a method to adjust such a probability distribution to the original data.

Essentially, we want the probability mass of Y to be concentrated around the points in the original data set as much as the previous constraint on the density function allows (similarly as done in [21, 23]). For this reason, we consider a density function that takes two different values: α_u in the regions that are close to the records in the original data set, and $\alpha_d = \exp(-\varepsilon)\alpha_u$ all over the rest.

Let U_u be the region with density α_u and U_d be the region with density α_d. Since the total probability mass must be 1, we have

$$\alpha_u \times m(U_u) + \alpha_d \times m(U_d) = 1, \quad (2)$$

where $m(U)$ is the size of set U.

Notice that this equality is only possible when both $m(U_u)$ and $m(U_d)$ are finite. Note also that the values of α_u and the size of the regions that satisfy the previous equality are not unique. In general, if m is the measure of $\mathcal{D}_r$,

$$\alpha_u \in [1/m, \exp(\varepsilon)/m],$$

where the above bounds are obtained from Expression (2) using that $m(U_u) + m(U_d) = m$, $\alpha_d = \exp(-\varepsilon)\alpha_u$ and $0 \leq m(U_u) \leq m$.

Also, we can obtain the following expression of the measure of U_u as a function of α_d, which will be used in the next section:

$$m(U_u) = \frac{1 - \alpha_d m}{\alpha_d(\exp(\varepsilon) - 1)}. \quad (3)$$

Finally, since we want to give each of the records in the original data set equal weight in the synthetic data set, the higher-probability region U_u is evenly distributed around all original

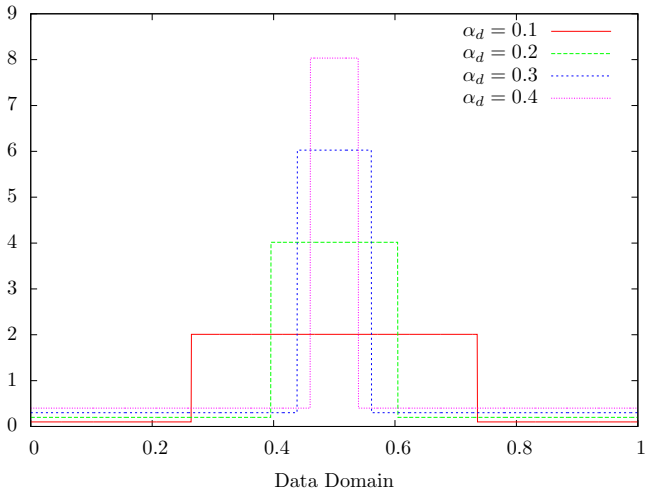

Figure 6: Effect of α_d on the concentration of the probability mass around the original records

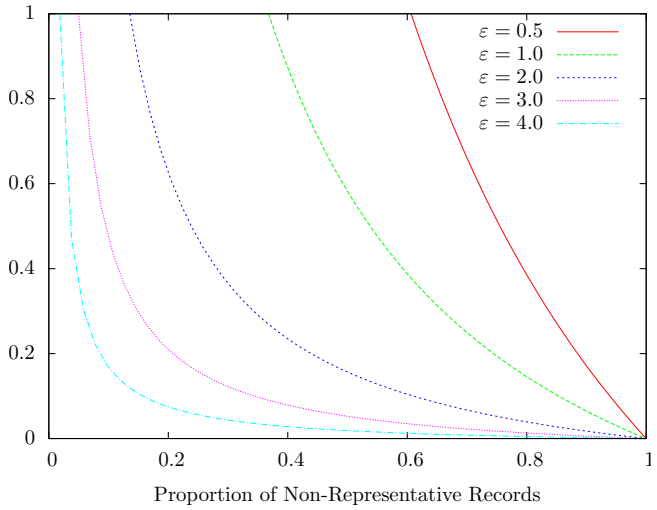

Figure 7: Tradeoff between the proportion of non-representative records (abscissae) and the dispersion around the original value of the records that are representative (ordinates), for different values of ε

records: the measure of the subset of U_u around each original record is the same.

5 EVALUATION

We examine the resemblance of the adjusted distribution to the original data. In particular, we evaluate: i) the trade-off between the proportion of records in the synthetic data set that are representative of the original data and the level of concentration of the probability mass around the original records; ii) the effect of the cardinality of the data set on the concentration of the probability mass around the original records; and iii) the absolute error in range count queries.

5.1 Non-representative records and dispersion

The distribution proposed in Section 4.2 tries to concentrate the probability mass around the records in the original data set as much as possible. Due to the constraints introduced by ε-synthetic privacy, this is only possible to a limited extent that is determined by the values α_d and $\alpha_u = \exp(\varepsilon)\alpha_d$ selected. On one side, we want α_d to be as small as possible, because the greater α_d, the more records will appear in the synthetic data set that bear no relation to the original data. On the other side, reducing α_d also reduces α_u; hence, the extent to which we can decrease the dispersion of the probability mass and concentrate it around the records in the original data set is limited. Figure 6 shows, for $\varepsilon = 3$ and a domain of measure 1, ε-synthetically private density functions for different values of α_d. For the sake of clarity we assume that the original data set contains a single attribute that takes values in $[0, 1]$ and a single record with value 0.5. The figure shows that greater density dispersions correspond to lower values of α_d, and hence to lower proportions of synthetic records that are not representative of the original record.

To give additional insight, Figure 7 illustrates the trade-off between the proportion of non-representative records and the dispersion around the original value of the records that are representative, for different values of ε. The closer both magnitudes to zero, the better for utility, but the worse for privacy (we need a greater ε to come closer to $(0, 0)$).

5.2 Cardinality and dispersion

It seems natural that the impact of including a record in a data set should be smaller when the size of the data set is large. This is indeed the case in ε-synthetic privacy. Choosing ε and α_d determines the measure of the region U_u that is assigned density α_u. When the original data set contains a single record, such as in the top graph of Figure 8, the higher-probability region U_u is entirely concentrated around this record. On the other hand, if there are several records in the original data set, such as in the bottom graph of Figure 8, the region U_u is distributed around each of the original records; hence, the probability mass around each original record is smaller. By comparing both graphs in the figure, it can be seen that, the larger the cardinality of the original data set, the smaller the dispersion of the synthetic values with respect to the original values (because the higher-probability subregions around the original values become narrower).

For fixed α_d, the dispersion around the original records depends not only on the number n of records but also on the privacy parameter ε. Thus, if ε and n result in a certain dispersion, we can wonder which is the ε' that yields the same dispersion with n' records. The dispersion is proportional to the width of the higher-probability subregions around each original data record, and this width is $m(U_u)$ divided by the number of records. On the other hand, $m(U_u)$ is given by Expression (3). Hence, we use the latter

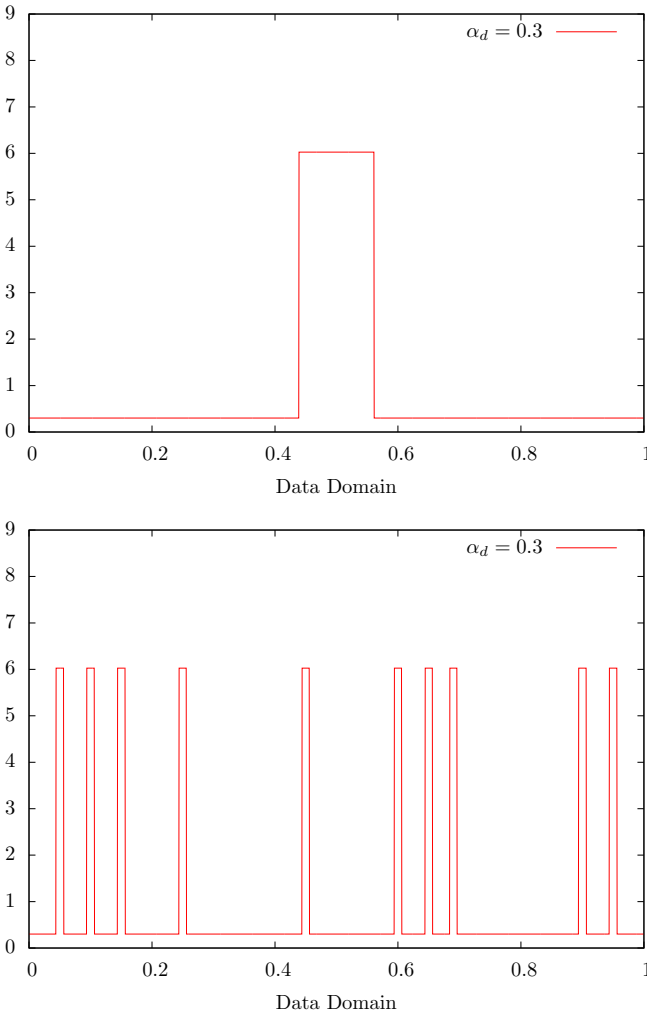

Figure 8: Reduction of the dispersion around the original records as the number of original records increases. Top, original data set with one record; bottom, original data set with 10 records.

expression to write that we want ε' such that

$$\frac{1-\alpha_d m}{n\alpha_d(\exp(\varepsilon)-1)} = \frac{1-\alpha_d m}{n'\alpha_d(\exp(\varepsilon')-1)}.$$

By solving the previous equality for ε' we get

$$\varepsilon' = \ln\left(\frac{(\exp(\varepsilon)-1)n}{n'}+1\right).$$

If we take the dispersion around the original records as a (reciprocal) measure of utility (the less dispersion, the more utility), we can summarize the above argument by saying: increasing the size of the original data set from n to n' is equivalent in terms of utility to decreasing the level of protection from $\varepsilon' = \ln(\frac{(\exp(\varepsilon)-1)n}{n'}+1)$ to ε (it is easy to see that if $n < n'$ then $\varepsilon' < \varepsilon$, that is, changing from ε' to ε means less privacy).

It is interesting to analyze what happens if we take the size of the original data to be infinity. Since U_u is distributed among such a large number of original records, the measure of the higher-probability subregion around each original record tends to zero (assuming that the original records are distant enough as for the subregions of U_u corresponding to different original records not to intersect). Thus, the synthetic data set is generated by repeating n times the following procedure: take a random record in the domain $\mathcal{D}_r$ with probability α_d, and a random original record otherwise.

5.3 Errors in count queries

Count queries (a.k.a. predicate queries or range queries) count the number of records within a region. The error in count queries is a common measure to evaluate the utility of a mechanism generating differentially private data sets [1]. Even though ε-synthetic privacy is not differential privacy, it is similar enough for count queries to remain a sensible utility benchmark.

Coming up with an analytical formula for the error in count queries can be quite difficult, so one usually resorts to simulations. Yet, obtaining a formula may give a better insight on the utility of the data synthesis mechanism, even if it comes at the price of some simplifying assumptions. We now set off to conduct this simplified analytical assessment.

Let $U \subset \mathcal{D}_r$ be a subset of the domain of the records. By construction, the ε-synthetically private distribution has density mass α_u in the vicinity of the original records and α_d all over the rest of $\mathcal{D}_r$. Thus, if $m(U)$ is the measure of U and c is the number of original records within U, we can approximate the probability mass assigned to U by the ε-synthetically private distribution as

$$P(U) = m(U)\alpha_d + \frac{c}{n}(1-\alpha_d m). \tag{4}$$

The rationale of Expression (4) is that the probability mass that remains after assigning density α_d to all points in D_r is $1-\alpha_d m$; this mass must be evenly distributed among the n original records and therefore the c original records in U get a proportion c/n of it. In addition, U has density α_d at all its points, which is an additional mass $m(U)\alpha_d$. This approximation assumes that, for each record contained in U, the region with density α_u surrounding the record is entirely contained in U. This assumption is realistic if the size of this region is small (such as it is when the number of records in the original data set is large). Moreover, the previous approximation also requires that the number c of original records contained in U be not greater than $m(U)/(m(U_u)/n)$, that is, the measure of U divided by the width of the higher-probability subregion around each original record. Beyond this bound, those subregions are not large enough to contain the higher-probability subregion around each original record; in this case, a better approximation for $P(U)$ is

$$P(U) = m(U)\alpha_u.$$

In any case, since the synthetic data set is generated by independently drawing from the adjusted distribution, the number of records contained in U is given by a binomial distribution with the appropriate probability: $Binomial(n, P(U))$. Table 1 shows the absolute value of the expected error of the errors in terms of ε, α_d, $m(U)$ and c, where for the sake of concreteness we assume that $m = 1$ and $n = 100$.

Table 1: Expected error in counts for a data set with $m = 1$ and $n = 100$ in terms of ε, α_d and $m(U)$

		Expected Error					
	$m(U)$	$c = 0$	$c = 5$	$c = 10$	$c = 15$	$c = 20$	$c = 25$
$\varepsilon = 1, \alpha_d = 0.4$	0.05	2	0	4.56	9.56	14.56	19.56
	0.10	4	2	0	4.12	9.12	14.12
	0.15	6	4	2	0	3.69	8.69
	0.20	8	6	4	2	0	3.25
	0.25	10	8	6	4	2	0
	0.30	12	10	8	6	4	2
$\varepsilon = 2, \alpha_d = 0.4$	0.05	2	0	2	4	6	10.22
	0.10	4	2	0	2	4	6
	0.15	6	4	2	0	2	4
	0.20	8	6	4	2	0	2
	0.25	10	8	6	4	2	0
	0.30	12	10	8	6	4	2
$\varepsilon = 2, \alpha_d = 0.2$	0.05	1	0	2.61	7.61	12.61	17.61
	0.10	2	1	0	1	5.22	10.22
	0.15	3	2	1	0	1	2.83
	0.20	4	3	2	1	0	1
	0.25	5	4	3	2	1	0
	0.30	6	5	4	3	2	1

A common approach to generate differentially private data sets is based on generating a histogram by partitioning the data domain in bins and counting the number of records in each of the bins. The absolute error in each bin count is independent of the number of records. As a result, the relative error is small when the number of records in a bin is large but may be large when the number of records in the bin is small. This limits the granularity of the histogram bins: if there is a large number of bins with a small number of records, the total error becomes too high. In contrast, in our approach we do not use a specific partition in bins, and hence we do not need to sacrifice granularity to gain accuracy. Another shortcoming of histogram-based methods is that they compute counts for specific bins, but if a count over a larger set is desired, bin counts need to be added, which may incur a large error if the number of added bin counts is large. This is not an issue in our approach, which directly estimates a count query based on the previously mentioned binomial distribution.

On the other side, a limitation of our approach is that the adjusted distribution considers a minimum and a maximum density (α_d and α_u, respectively). The accuracy of the count in a given region depends on its average *actual* density and the error can only be expected to be low only if the average actual density is within $[\alpha_d, \alpha_u]$. By increasing the width of the $[\alpha_d, \alpha_u]$ interval, the above limitation is mitigated. Since $\alpha_u = \exp(\varepsilon)\alpha_d$, the interval width can be increased by increasing ε, that is, by lowering the privacy level. This is perfectly coherent: the lower the privacy level, the closer the adjusted distribution can be to the original data.

6 CONCLUSIONS AND FUTURE RESEARCH

Generating synthetic data is a well-known approach to disclosure risk limitation. A model for the population data is built (based on the original data) and a new data set is generated from scratch based solely on the model. In this paper we argue about the difficulties of building an accurate model of the population and we propose, instead, to approximate the original data as closely as possible. To do so, we seek to incorporate in the model as much information as possible about the original data. Of course, increasing the amount of information about the original data in the model also increases the risk of disclosure and, thus, can only be done to a limited extent. We need a privacy model that gives an *ex ante* assessment of the disclosure risk incurred by the synthetic data model.

Our first candidate privacy model was differential privacy, which is well-known and provides strong privacy guarantees. However, we found that differential privacy leads to a large utility loss. The reason is that, given any two original data sets that differ in a single record and an arbitrary synthetic data set, differential privacy requires the probabilities of deriving the arbitrary synthetic data set from either original data set to be similar. The fact that this condition must be satisfied even for synthetic data sets that are not representative of the original data sets is what leads to such a large utility loss. We have proposed an alternative privacy model that preserves the strong privacy guarantees of differential privacy but is better suited to synthetic data generation. Given a synthetic data model (adjusted to the original data) and two possible synthetic data sets that differ in a single record, our alternative privacy model requires the probabilities of getting each of those two data sets to be similar.

In terms of data utility, approximating the empirical distribution of the original data may be controversial. By overfitting the synthetic data model to the original data set, we lose predictive power (that is, the ability to predict what would happen for an individual not contained in the original data set). However, as long as we are not interested in making predictions, such a loss is affordable. In fact, it is common in disclosure risk limitation to seek the slightest modification of the original data set that is enough to preserve the privacy of the respondents. Which of the two approaches, *i.e.*

building a model for the population (based on the original data) or directly approximating the original data, is better depends on the situation. If a model for the population data is known in advance, using it seems the most sensible approach. If we view the original data set as the population of interest, approximating the empirical distribution of the original data seems the best option. If the original data set is viewed as a sample of some underlying population for which a model is not known in advance, the confidence in the inferred model is a key factor: a small original data set is unlikely to allow inferring a model with reasonable confidence, so approximating the empirical distribution of the data set is a better option; a large original data set makes both approaches reasonable, because it allows inferring a model with high confidence and it also yields an empirical distribution closer to the underlying population distribution. However, even with a large original data set, an inferred model may fail to capture complex relations or relations between attributes that are only apparent in a small number of records (*e.g.* a relation that is only present for records that have some specific values for some attributes, for example records corresponding to individuals with a certain gender and a certain narrow age range).

In terms of disclosure risk, the more information about the original data we incorporate in the model, the greater the risk. Hence, our proposal to overfit the synthetic data model to the original data set might seem to incur too high a risk of disclosure. However, constraining the data model overfitting so that it remains compliant with ε-synthetic privacy guarantees that disclosure risk is kept under a pre-specified threshold.

Future work will involve extending and refining the methods to attain ε-synthetic privacy. In particular, we need to develop methods to satisfy this privacy model in the case of categorical attributes. Extending/adapting the method given for the continuous case seems a natural avenue.

7 ACKNOWLEDGMENTS

The following funding sources are gratefully acknowledged: European Commission (projects H2020-644024 "CLARUS" and H2020-700540 "CANVAS"), Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2014 SGR 537) and Spanish Government (projects TIN2011-27076-C03-01 "CO-PRIVACY", TIN2014-57364-C2-R "SmartGlacis" and TIN2015-70054-REDC). The views in this paper are the authors' own and do not necessarily reflect the views of UNESCO or any of the funders.

REFERENCES

[1] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *40th Annual ACM Symposium on Theory of Computing (STOC 2008)*, pp. 609–618. ACM, 2008.

[2] J. Burridge. Information preserving statistical obfuscation. *Statistics and Computing*, 13(4):321–327, 2003.

[3] R. A. Dandekar, J. Domingo-Ferrer, and F. Sebé. LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In *Inference Control in Statistical Databases: from Theory to Practice*, LNCS 2316, pp. 153-162. Springer, 2002.

[4] J. Domingo-Ferrer and Ú. González-Nicolás. Hybrid microdata using microaggregation. *Information Sciences*, 180(15):2834–2844, 2010.

[5] J. Domingo-Ferrer, S. Ricci, and J. Soria-Comas. Disclosure risk assessment via record linkage by a maximum-knowledge attacker. In *13th Annual Conference on Privacy, Security and Trust (PST 2015)*, pp. 28–35. IEEE, 2015.

[6] J. Domingo-Ferrer, D. Sanchez, and J. Soria-Comas. *Database Anonymization: Privacy Models, Data Utility, and Microaggregation-Based Inter-Model Connections*. Morgan & Claypool, 2016.

[7] J. Drechsler. *Synthetic Datasets for Statistical Disclosure Control*. Springer, 2011.

[8] C. Dwork. Differential privacy: A survey of results. In *Proceedings of the 5th international conference on theory and aplications of models of computation (TAMC'08)*. pp. 1–19. Springer-Verlag, 2008.

[9] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: privacy via distributed noise generation. In *Proceedings of the 24th annual international conference on The Theory and Applications of Cryptographic Techniques (EUROCRYPT'06)*. pp. 486–503. Springer-Verlag, 2006.

[10] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography (TCC 2006)*, LNCS 3876, pp. 265–284. Springer, 2006.

[11] C. Dwork, and A. Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):135–154, 2009.

[12] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

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

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

[15] K. Muralidhar and R. Sarathy. Generating sufficiency-based non-synthetic perturbed data. *Transactions on Data Privacy*, 1(1):17–33, 2008.

[16] J. Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003.

[17] J. P. Reiter. Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics*, 19:441–462, 2003.

[18] D. B. Rubin. Discussion: statistical disclosure limitation. *Journal of Official Statistics*, 9(2):461–468, 1993.

[19] P. Samarati and L. Sweeney. *Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppression*. Technical report, SRI International, 1998.

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

[21] J. Soria-Comas and J. Domingo-Ferrer. Differential privacy through knowledge refinement. In *4th IEEE International Conference on Privacy, Security, Risk and Trust (PASSAT 2012)*, pp. 702-707. IEEE, 2012.

[22] 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)*, pp. 1–8. IEEE, 2012.

[23] J. Soria-Comas and J. Domingo-Ferrer. Optimal data-independent noise for differential privacy. *Information Sciences*, 250:200–214, 2013.

[24] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez. Enhancing data utility in differential privacy via microaggregation-based *k*-anonymity. *The VLDB Journal*, 23(5):771–794, 2014.

[25] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and D. Megias. Individual differential privacy: a utility-preserving formulation of differential privacy guarantees. *IEEE Transactions on Information Forensics and Security*, 12(6):1418–1429, 2017.

[26] Y. Xiao, L. Xiong, and C. Yuan. Differentially private data release through multidimensional partitioning. In *Secure Data Management*, LNCS 6358, pp. 150–168. Springer, 2010.

[27] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: private data release via bayesian networks. In *2014 ACM SIGMOD International Conference on Management of Data-SIGMOD '14*, pp. 1423–1434, New York, NY, USA, 2014. ACM.