TRANSACTIONS ON DATA PRIVACY 10 (2017) 61–81

Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion

Anna Oganian*, **Josep Domingo-Ferrer****

*National Center for Health Statistics, 3311 Toledo Rd, Hyattsville, MD 20782, USA

**UNESCO Chair in Data Privacy, Universitat Rovira i Virgili, Department of Computer Engineering and Maths, Av. Països Catalans 26, E-43007 Tarragona, Catalonia.

E-mail: annaoganyan7@gmail.com, aoganyan@cdc.gov, josep.domingo@urv.cat

Received 7 March 2016; received in revised form 19 August 2016; accepted 20 November 2016

Abstract. Before releasing databases which contain sensitive information about individuals, data publishers must apply Statistical Disclosure Limitation (SDL) methods to them, in order to avoid disclosure of sensitive information on any identifiable data subject. SDL methods often consist of masking or synthesizing the original data records in such a way as to minimize the risk of disclosure of the sensitive information while providing data users with accurate information about the population of interest. In this paper we propose a new scheme for disclosure limitation, based on the idea of *local synthesis* of data. Our approach is predicated on model-based clustering. The proposed method satisfies the requirements of k-anonymity; in particular we use a variant of the k-anonymity privacy model, namely probabilistic k-anonymity, by incorporating constraints on cluster cardinality. Regarding data utility, for continuous attributes, we exactly preserve means and covariances of the original data, while approximately preserving higher-order moments and analyses on subdomains (defined by clusters and cluster combinations). For both continuous and categorical data, our experiments with medical data sets show that, from the point of view of data utility, local synthesis compares very favorably with other methods of disclosure limitation including the sequential regression approach for synthetic data generation.

Keywords. Statistical Disclosure Limitation (SDL), synthetic data, probabilistic k-anonymity, mixture model, Expectation-Maximization (EM) algorithm.

1 Introduction

When statistical agencies collect their data from individual data providers, they are required to protect individual data records from disclosure of sensitive information these records may contain. Disclosure of such sensitive information can cause serious damage to both individuals and agencies. Legal regulations in many countries, such as the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) in the US, define policies, procedures and guidelines for maintaining the privacy and security of individually identifiable information.

To protect individual data, direct identifiers, such as names, addresses and social security numbers, should be removed. However, some risk of identification still exists, for example, by means of linkage of the released data to external databases. So in addition, released microdata —collections of individual records— are typically modified, in order to make disclosure more difficult. In other words,

statistical disclosure limitation (SDL) methods are applied to the data prior to their release. These methods can be divided in two groups: masking methods, that release a modified version of the original microdata, and synthetic methods, that release artificial records generated from the distribution representing the original data.

Examples of masking methods include: data swapping, in which data values are swapped for selected records; noise addition, in which noise is added to numerical data values to reduce the likelihood of exact matching on key variables or to distort the values of sensitive variables; and microaggregation, a technique similar to data binning, which is briefly reviewed next. See Hundepool et al. (2010, 2012) for more details.

Microaggregation can be viewed as cardinality-constrained clustering which can be applied to numerical and categorical variables (Torra, 2004; Domingo-Ferrer and Torra, 2005). It consists of a partition step and an aggregation step:

- In the partition step, the set of the original records is partitioned into a number of clusters each containing at least k records for some preset integer k and with the aim that the records within each cluster be as homogeneous as possible. For example, for continuous variables, the sum of squares criterion is a common measure of homogeneity in clustering (Ward, 1963; Edwards and Cavalli-Sforza, 1965; Hansen et al., 1998). An important feature of microaggregation is that the number of records per cluster should be at least k, which is a parameter of the method.
- In the aggregation step, an aggregation operator (for example, the mean for continuous data or the median for categorical ordinal data) is computed and used to replace the original records. So, the released masked data set consists of the cluster means/medians and the parameter k is responsible for the utility/risk trade-off.

Regarding synthetic methods, the crux is to obtain a good data generation model. Often synthetic data are generated using sequential modeling strategies, similar to those used for imputation of missing data in Raghunathan et al. (2001, 2003). However, if used alone, this method may not be able to preserve complex relationships between the attributes, such as higher-order interactions and nonlinear relationships between the attributes. This is true especially when the original data consist of records from various subpopulations where variables have particular relationships, different within each subpopulation. For example, in healthcare data patient records form natural groups according to the type of their disease/diagnosis.

Similar problems may be encountered by nonparametric approaches which try to improve on parametric methods. For example Caiola and Reiter (2010) and Reiter (2005) propose methods based on classification and regression trees and also random forests where a separate tree is built for each sensitive attribute. Even if these methods are suited to different types of attributes, they work better for categorical outcomes or truncated continuous variables that are not smooth, due to the discontinuity of partition boundaries. In particular, the CART-based approach (Reiter, 2005) may work very well when synthesis is applied only to a specific subpopulation of individuals (*e.g.* only individuals with incomes greater than $100,000 are synthesized). However, if the tree is built on data that include also records not belonging to that particular subpopulation, the distribution obtained may not accurately capture the properties of the various sub-populations. Hence, if records from different subpopulations need to be synthesized together, the utility of the resulting data may not be very high.

A possible alternative for disclosure limitation for data sets with complex structure is to combine features of masking and synthetic methods in a way to preserve their strengths and neutralize their pitfalls. So, in Dandekar et al. (2002); Muralidhar and Sarathy (2008); Domingo-Ferrer and González-Nicolás (2010) *hybrid* methods of statistical disclosure limitation were proposed which allow the data protector to select the degree of closeness of the protected data to the original data by calibrating the amount of synthesis involved in the disclosure limitation procedure. For example, Muralidhar and Sarathy (2008) uses a regression-like scheme with a term responsible for the "proximity" of the protected data to the original data; Dandekar et al. (2002) relies on latin hypercube techniques, and Domingo-Ferrer and González-Nicolás (2010) proposes to apply microaggregation to the data and

then generate synthetic records for each group. In this paper we present a method which resembles the one in Domingo-Ferrer and González-Nicolás (2010); however, we use a different methodology, so our method provides better utility guaranties and can be applied not only to continuous data but to categorical data as well. To assess the performance of the proposed methods, we need to quantify the amount of distortion of statistical characteristics caused by the method (data utility assessment) and also the risk associated with the release of the resulting data (disclosure risk assessment).

There are different types of utility assessment: analysis-specific utility measures, tailored to specific analyses, and broad measures reflecting global differences between the distributions of original and masked data (see some examples in Domingo-Ferrer et al. (2002); Oganian (2003); Karr et al. (2006); Woo et al. (2009)). In this paper, we will use both types of measures. One is a broad measure proposed in Woo et al. (2009) based on propensity scores; hereafter, referred to as propensity score measure. This measure is suitable for data sets with mixed attributes. It compares favorably with other data utility measures (Woo et al. (2009)) and was adapted and used by the US Census Bureau ((Drechsler, 2011)) We also consider some analysis-specific utility measures, such as the average percentage change in regression coefficients, their standard errors, and average percentage changes in the third and fourth moments (attributed to disclosure limitation procedures).

To quantify disclosure risk we will adopt an approach of enforcing a privacy criterion for the released data that offers *a priori* guarantees of low disclosure risk. Specifically, we will use a variant of the k-anonymity criterion, called *probabilistic k-anonymity* (Soria-Comas and Domingo-Ferrer, 2012; Soria-Comas, 2013) defined below.

Definition 1 (Probabilistic k-anonymity)**.** A published data set T' is said to satisfy probabilistic k-anonymity if, for any non-anonymous external data set E, the probability that an intruder knowing T', E and the anonymization mechanism M correctly links any record in E to its corresponding record (if any) in T' is at most $1/k$.

To contrast the difference between Definition 1 (Probabilistic k-anonymity) and standard k-anonymity criterion (Samarati and Sweeney, 1998; Ciriani et al., 2008), we give its definition below:

Definition 2 (k-anonymity)**.** A microdata set T' is said to satisfy k-anonymity if, for each record $t \in T'$, there are at least $k-1$ other records sharing the same values for all the quasi-identifier attributes (quasi-identifiers are the attributes available in external data sets, these attributes can be used for linkage by intruder).

We realize that the definitions given above may seem rather different to the reader. To help seeing their similarities, first, we want to note that both probabilistic and standard k-anonymity models are anonymity-oriented criteria. Anonymity is one of the aspects of data privacy. It usually means that it is not be possible to re-identify any individual in the published data set. Anonymity can be contrasted by a different aspect of data privacy, namely, confidentiality or secrecy, which means that the released data should not allow an attacker to increase its knowledge about confidential information related to any specific individual (see(Soria-Comas, 2013)).

The meaning of the parameter k in Definitions 1 and 2, however, is not exactly the same. In case of the standard k-anonymity, it refers to the minimal number of quasi-identifiers that should share the same values. This is how standard k-anonymity guaranties the upper bound on the probability of re-identification to be equal to $1/k$. And, while the quasi-identifiers are not mentioned in the definition of probabilistic k-anonymity, and k refers to the reciprocal of the upper bound on the probability of the re-identification, both criteria establish the same limit on the probability of re-identification. Furthermore, the ultimate goal the standard k-anonymity criterion is trying to achieve is not the reinforcement of the indistinguishability of quasi-identifiers per se, but the guarantee that

re-identification of individuals is limited; in particular, that the upper bound on the probability of re-identification is equal to some threshold $(1/k)$. Hence, the indistinguishability of quasi-identifiers is just the way how standard k-anonymity achieves its goal, but it is not the goal by itself. So, from this point view, the probabilistic and standard k-anonymity are just two variants of the same criterion. Probabilistic k-anonymity, however, has several advantages over standard k-anonymity due to the relaxation of the requirement of indistinguishability of quasi-identifiers.

First, from the point of view of utility, it becomes possible to maintain the variability in the released data set as opposed to the standard k-anonymization which reduces the variability in the released data therefore worsening its utility. This is especially important for the cases when the number of quasi-identifiers is big (it is well known that for k-anonymous models the utility degrades rapidly if the number of quasi-identifiers is increased - "the curse of dimensionality" (Soria-Comas (2013) and Aggarwal (2005))).

Second, standard k-anonymity assumes that data protector is capable of discerning between quasi-identifiers and non-quasi-identifier attributes, that is, he/she is supposed to be able to determine which attributes may be available externally for the intruder in a non-anonymised data set. Probabilistic k-anonymity, however, does not make such an assumption.

Third, if we consider an informed intruder who knows some values of the confidential attributes, then such an intruder may use these values for linkage and significantly increase the probability of re-identification. To protect the data against informed intruders confidential attributes should be considered as quasi-identifiers as well, so all the attributes become quasi-identifiers. The indistinguishability requirement of k-anonymity in such a case will have a very significant impact on the utility of the resultant k-anonymous data (Soria-Comas (2013) and Aggarwal (2005))). Several fixes/alternatives to k-anonymity have been proposed, for example, l-diversity (Machanavajjhala et al., 2008), t-closeness (Li et al., 2007), etc. However, they are based on the partitioning of the data set in groups of indistinguishable records and none of those alternatives is free from shortcomings (see Domingo-Ferrer and Torra (2008)).

On the other hand, the probabilistic k-anonymity criterion is a more general framework focused on the probability of re-identification, requiring this probability to be at most $1/k$. So it achieves the same level of protection against re-identification provided by k-anonymity and at the same time allows much more flexibility in data alteration. The data protector can search and apply the method which offers better utility guarantees and it is possible to preserve the variability in quasi-identifiers (see Soria-Comas and Domingo-Ferrer (2012); Soria-Comas (2013)).

Last, while discussing privacy models, we have to mention an important privacy criterion originating in computer science that has received a lot of scientific attention, namely ε-differential privacy (Dwork, 2006; Dwork et al., 2006; Dwork, 2011)) and some of its variants, for example, δ-approximate ε-differential privacy, probabilistic differential privacy (Machanavajjhala et al., 2008), random differential privacy (Hall et al., 2011), etc. As opposed to k-anonymity, ε-differential privacy is solely based on the confidentiality aspect of data privacy. Differential privacy provides a very strong level of privacy guarantees, no matter the intruder's side knowledge, by limiting the influence of any single respondent on the released information. It was originally proposed to anonymize the query answers in query-based systems (interactive setting), rather than to anonymize entire data sets in view of releasing them (non-interactive setting). In fact, Dwork, the author who introduced differential privacy, admits in Dwork (2011) that it is impossible "to generate a "noisy table" in a non-interactive setting that will permit highly accurate answers to be derived for computations that are not specified at the outset". Although different relaxations of ε-differential privacy may lead to improved utility of the resultant data, recent results (Charest, 2012a,b; Fienberg et al., 2010) show that differentially private methods achieve their strong privacy guarantees at great cost in data utility, and even specially designed inferential techniques cannot compensate for that utility loss. Hence, methods for differentially private data releases aim at preserving utility for a certain class of queries: for example, Hardt et al. (2012) presents a differentially private algorithm producing a synthetic data set that preserves utility for any set of linear queries (those that apply a function to each record and sum the result, like

for example count queries). This contrasts with the general-purpose utility preserving data release offered by the k-anonymity model.

Hence, we focus here on achieving probabilistic k-anonymity rather than ε-differential privacy. This will allow us to give much more general utility preservation guarantees.

1.1 Contribution and plan of this paper

In this paper we propose a new disclosure limitation method which is based on the idea of *local synthesis* that satisfies the requirements of probabilistic k-anonymity. The procedure consists of clustering the original data subject to constraints on cluster cardinality and then synthesizing the records within each cluster. The proposed expressions for the constraints aim to prevent too detailed local synthesis. In particular, we want to prevent formation of very small clusters, so we impose lower bound constraints on cluster membership probabilities.

We assert that to obtain good results, the clustering procedure should be model-based. In this case the distribution of the obtained clusters may better conform to a particular type, that is, the type used in the mixture model. When the records are synthesized on the next step, the same type of distribution can be used by the synthesizer. For example, if the clustering algorithm is such that it produces approximately normally distributed clusters, then a normal model can be used in the synthesis step, which can help to reduce information loss.

For continuous data, we analytically show that our method exactly preserves means and covariances; also, it approximately preserves higher-order moments, with the quality of the approximation improving as data within the clusters approach normality. Furthermore, moments are also approximately preserved over subdomains defined by clusters and cluster combinations. Finally, we present an empirical comparison for continuous and categorical real medical data using different types of utility measures. This comparison shows that our method is able to provide better utility guarantees than fully synthetic data obtained via multiple imputation with sequential regressions and other SDL methods as well.

The core idea of the method is described in Section 2. In Section 3 we outline the local synthesis approach for continuous attributes, and in Section 4 we present some utility properties of the method. The results of a numerical experiment with continuous medical data are reported in Section 5. In Section 6 we describe the scheme for mixed categorical and continuous attributes, and in Section 7 we report the results of numerical experiments with mixed continuous and categorical attributes. Finally, in Section 8 we provide a concluding discussion and sketch lines for future work.

Parts of this paper were presented in the conference paper (Oganian and Domingo-Ferrer, 2012), which described just a basic idea of local synthesis for continuous data only and without characterization of disclosure risk. In this paper, we further developed our method. Constraints on the cluster sizes to achieve probabilistic k-anonymity, the extension to mixed categorical and continuous attributes (Section 6) and the experiments with those mixed data (Section 7) are presented for the first time. Furthermore, Sections 1, 2, 3, 5 and 8 have been substantially expanded.

2 Local synthesis using mixture models

As mentioned in the introduction, our idea is to use a model-based clustering algorithm with constraints imposed by the requirements of probabilistic k-anonymity and then synthesize from a mixture model obtained in the previous step. If a mixture model is used to model the data, then the density of the entire data set can be represented as

$$f(\boldsymbol{x}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}|\boldsymbol{\theta}_g), \tag{1}$$

where π_g is the probability that an observation belongs to the g-th cluster ($\pi_g \geq 0$; $\sum_{g=1}^{G} \pi_g = 1$), f_g is the multivariate density function of the g-th cluster and $\boldsymbol{\theta}_g$ are the parameters of f_g. The form of f_g depends on the type of the attribute, *i.e.* continuous or categorical. Cluster membership is the unobserved part of the data, so the "complete data", which include observed and unobserved attributes, can be represented by $\boldsymbol{y}_i = (\boldsymbol{x}_i, \boldsymbol{z}_i)$, where $\boldsymbol{z}_i = (z_{i1}, \cdots, z_{iG})$ is a latent variable, with $z_{ig} = 1$ if $\boldsymbol{x}_i$ belongs to the cluster g and $z_{ig} = 0$ otherwise.

Our choice of model-based clustering is explained by the fact that the mixture model is a very flexible and powerful tool. In particular, if the number of latent classes (*i.e.* clusters) is sufficiently large, the mixture model has the ability to accurately represent the first, second, and higher-order observed moments of the continuous response attributes. For categorical variables, these moments are the univariate distributions, bivariate associations, and the higher-order interactions. Also, by using mixture models, we will be able to preserve important distributional characteristics not only of the overall original data set, but also of its subdomains.

Furthermore, this approach may be particularly suited to healthcare data because in many areas of medical research mixture models are used to classify individuals into disease categories. Hence, preservation of the distributional characteristics within these meaningful classes can be considered as a desirable feature for the user.

Our method can be classified as hybrid data generation because, by varying the number of clusters, the resultant data become more or less similar to the original data. If there is only one cluster, we have a fully synthetic data. If the number of clusters approaches the number of records, then the resultant data become very similar to the original data. In this way, each cluster can be regarded as a constraint on the synthetic data generation, that is, the more constraints, the less freedom there is for generating synthetic data, and the output looks more like the original data. While there is no correspondence between the individual records, there is, however, the correspondence between the clusters in the original and hybrid data set. Such a correspondence between the clusters implies the possibility of re-identification, especially if the cardinality of clusters is not restricted from below. In fact, if the intruder (in the worst case scenario) is able to identify a small group of records (a small cluster) in the hybrid data set that contains the (original) record of his/her interest, then we can say that a form of re-identification has occurred. That is why it is important to reinforce the criteria of probabilistic k-anonymity by limiting from below the number of records each cluster can have in order to hide each record within a group of at least k other records. So, we set the lower bound on cluster cardinality to be equal to k. Since the records are synthesized within clusters, the probability of re-identification is naturally limited by $1/k$.

Finally, we would like to add that model-based clustering methodology was used before in different settings of privacy preserving data mining. For example, Lin et al. (2005) describes a method for performing model-based clustering on distributed data in a secure way. Their goal, however, was not to produce synthetic data that can be released to public for general purpose, but rather to carry out the procedure of model-based clustering on their data in a secure way. On the other hand, a synthetic data generation scheme based on the mixture model is described in Lee (2009). This method, however, doesn't satisfy any privacy criterion and is only applicable for categorical variables. Another example of application of mixture model methodology in the context of privacy preserving data mining can be found in Pathak and Raj (2012). This work describes a discriminatively trained Gaussian mixture model-based classification algorithm that satisfies differential privacy. Similar to Lin et al. (2005), the algorithm proposed in Pathak and Raj (2012) was designed for a specific data use - classification. So, to the best of our knowledge, the methodology of model-based clustering was not used in the context of general-purpose synthetic microdata release that satisfies some privacy criterion.

3 Local synthesis for continuous attributes

The implementation of local synthesis using model-based clustering depends on the type of attribute. We will start with continuous variables. For continuous attributes we will use a Gaussian mixture

model. There are two reasons for such a choice: (1) density estimation theory guarantees that any distribution can be effectively approximated by a mixture of Gaussians (Scott, 1992; Silverman, 1986); (2) synthesizing multivariate normal data (that is, generating data from the multivariate normal mixture) is fast computationally. For a Gaussian mixture, Equation (1) becomes a weighted sum of multivariate normal densities $f_g(\boldsymbol{x}|\boldsymbol{\theta}_g)$, where the distribution parameters $\boldsymbol{\theta}_g$ are represented by the within-cluster mean vector $\boldsymbol{\mu}_g$ and the covariance matrix $\boldsymbol{\Sigma}_g$. Data generated by multivariate normal densities can be represented by groups of ellipsoid clusters centered at mean vectors $\boldsymbol{\mu}_g$. The geometric characteristics of the cluster are determined by the covariance matrices $\boldsymbol{\Sigma}_g$. To speed up the estimation procedure, constraints on the covariance matrix structure can be introduced. This will reduce the number of parameters to be estimated. For example, the following constraints can be used: $\boldsymbol{\Sigma}_g = \lambda \boldsymbol{I}$, where all clusters are spherical and of the same size; or $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, where all clusters have the same covariance and size, but do not need to be spherical. It is of course possible to use an unrestricted covariance matrix $\boldsymbol{\Sigma}_g$, where each cluster may have a different geometry (Fraley and Raftery, 2002; Celeux and Govaert, 1995). In such a case the number of model parameters to be estimated is $G(d + d(d+1)/2 + 1) - 1$, where d is the dimensionality of the data.

The EM algorithm can be used to find maximum likelihood estimates of $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$, and π_g. In particular, in the E step of the EM algorithm a probability of assigning record i to cluster g is estimated for each $i \in \{1, \cdots, n\}$ as

$$\hat{z}_{ig} \leftarrow \frac{\hat{\pi}_g f_g(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_g)}{\sum_{j=1}^{G} \hat{\pi}_g f_j(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_j)} \,. \tag{2}$$

For the M step, estimates of the means $\boldsymbol{\mu}_g$, covariance matrices $\hat{\boldsymbol{\Sigma}}_g$, and probabilities π_g have closed-form expressions and can be found in the literature (Celeux and Govaert, 1995). However, we will not use the expression for π_g exactly as given in Celeux and Govaert (1995), namely $\hat{\pi}_g = \sum_{i=1}^{n} \hat{z}_{ig}/n$, because such $\hat{\pi}_g$'s can be arbitrarily small and the corresponding clusters would have very few records. As we mentioned in section 2, even though there is no correspondence between the individual records in the original data and the resultant synthetic data, there is, however, correspondence between the clusters. If there are less than k records in a cluster the requirements of the probabilistic k-anonymity criterion will not be satisfied.

So, we propose to update the vector π_g after each M step according to the following rule

$$\hat{\pi}_g^{new} = \frac{\hat{\pi}_g + \delta}{\sum_{g=1}^{G}(\hat{\pi}_g + \delta)}, \tag{3}$$

where δ is computed as

$$\delta = \begin{cases} 0, & \text{if } \hat{\pi}_g \geq \frac{k}{n} \text{ for all } g \in \{1, \cdots, G\}, \\ \frac{\frac{k}{n} - \pi_{min}}{1 - \frac{k}{n}G}, & \text{otherwise,} \end{cases} \tag{4}$$

where π_{min} is the minimal cluster membership probability, corresponding to the smallest cluster. It is easy to verify that, by imposing such constraints, $\sum_{g=1}^{G} \hat{\pi}_g^{new} = 1$. Also π_{min}^{new} will be equal to k/n:

$$\pi_{min}^{new} = \frac{\pi_{min} + \delta}{1 + G\delta} = \frac{(1 - \frac{k}{n}G)\pi_{min} + \frac{k}{n} - \pi_{min}}{1 - \frac{k}{n}G + G(\frac{k}{n} - \pi_{min})} = \frac{\frac{k}{n}(1 - \pi_{min}G)}{1 - \pi_{min}G} = \frac{k}{n} \tag{5}$$

Furthermore, if $\pi_1 \leq \pi_2 \leq \cdots \leq \pi_G$ then $\pi_1^{new} \leq \pi_2^{new} \leq \cdots \leq \pi_G^{new}$. In this way all the $\hat{\pi}_g$'s become greater than or equal to k/n. This will satisfy the requirements of probabilistic k-anonymity. In addition to the aforementioned property, this transformation tends to make the cluster membership probabilities (when $\delta \neq 0$) more uniform in distribution. In particular, the π_g's become bigger for small clusters and smaller for large clusters. The π_g's remain unchanged for clusters with $\pi_g = \frac{1}{G}$.

These properties are easy to verify. For example, let $\pi_g = \frac{1}{G} - \epsilon$, where $0 < \epsilon < \frac{1}{G}$ and $\pi_{min} < \frac{k}{n}$, then

$$\pi_g^{new} = \frac{\frac{1}{G} - \epsilon + \frac{k/n - \pi_{min}}{1 - k/nG}}{1 + \frac{G(k/n - \pi_{min})}{1 - \frac{k}{n}G}} = \frac{\frac{1}{G} - \epsilon + \epsilon k/nG - \pi_{min}}{1 - \pi_{min}G} >$$

$$> \frac{\frac{1}{G} - \epsilon + \epsilon \pi_{min} G - \pi_{min}}{1 - \pi_{min}G} = \frac{(\frac{1}{G} - \epsilon)(1 - \pi_{min}G)}{1 - \pi_{min}G} = \pi_g$$

In a similar way it can be shown that $\pi_g^{new} < \pi_g$ for those clusters for which $\pi_g > \frac{1}{G}$ and $\pi_g^{new} = \pi_g$ for the clusters with $\pi_g = \frac{1}{G}$.

The change in cluster membership probability depends on the difference between π_g and the uniform probability $\frac{1}{G}$, and it also depends on the difference between π_{min} and the security constraint $\frac{k}{n}$. Consequetly, the smallest clusters will get the largest increase increase in cluster membership probability. (The larger the difference between π_g and the uniform probability $\frac{1}{G}$ the larger the change.) Similar reasoning holds for the clusters with $\pi_g > \frac{1}{G}$, but their probabilities will decrease. In this way, the protection provided to the records in different clusters becomes more uniform. Furthermore, we would like to add that these are not drastic changes, but rather small increases and decreases in membership probabilities.

Note that this transformation is equivalent to Laplace or additive smoothing with a specific smoothing parameter.

To choose the number of clusters and the parameterization of covariance matrices which define the shape of the clusters, we use the Bayesian Information Criterion, BIC (Schwarz, 1978):

$$BIC_g = 2 \log p(D|\hat{\boldsymbol{\theta}}_g, M_g) - \nu_g \log(n), \tag{6}$$

where D is the data and ν_g is the number of independent parameters to be estimated in model M_g.

The literature on model-based clustering suggests that the model choice based on BIC provides good results from the data utility perspective (Campbell et al., 1997, 1999; Fraley and Raftery, 1998; Stanford and Raftery, 2000). In our experiments, we also noticed that model selection based on BIC led to the creation of well populated clusters, which is good from the disclosure risk perspective. It also means that the algorithm may not need to update the cluster membership probabilities very often using Equations (3) and (4); yet, this, of course, depends on the value of k.

Finally, let us give intuitive reasoning why generating data from a mixture model may be better from the utility point of view than applying a non-model-based clustering method, such as microaggregation, and then synthesizing records within each cluster using any synthesizer as it is done in Domingo-Ferrer and González-Nicolás (2010). If we apply non-model-based clustering to the data and then synthesize the records for each cluster, we would have to decide what models should be used for data synthesis in different clusters. This is a complex task especially if it needs to be repeated for each cluster. So, to make the whole procedure feasible the model for each cluster needs to stay relatively simple. However, distributional properties of clusters obtained as a result of non-model based clustering are generally unknown. Therefore, a simple model for synthesis may not conform well to those clusters. On the other hand, a multivariate normal model may be a good choice for synthesis in each cluster when clustering was model-based, in particular, when the multivariate normal mixture was used. To better understand this, let us consider the clusters that can be obtained from EM for mixture of multivariate normal distributions. Strictly speaking, EM does a "soft" assignment of the records to the clusters by computing the probability z_{ig} of record i belonging to every cluster $g \in \{1, \cdots, G\}$; however, in order to understand the properties of the clusters we have to actually form them, so we will make hard assignments of records to the clusters. A record i will be assigned to the cluster with the largest value of z_{ig}, given by Expression (2). This is to say that it will be assigned to the cluster g for which the expression $\hat{\pi}_g f_g(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_g)$ is maximal. Subject to the cluster size, the record will end up in the cluster where it fits best according to the desired cluster distribution (normal). For

example, a point may be much closer to cluster 1 than to cluster 2 distance-wise; however, $f_1(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_1)$ may be much smaller than $f_2(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}_2)$ even after multiplying by the corresponding $\hat{\pi}_g$, meaning that the point does not really conform to the normal distribution of the cluster 1, so it will be assigned to cluster 2, where it will better fit its corresponding distribution (or disturb it less). Thus a certain tendency towards normality appears within these clusters.

Next, by using BIC we choose a number of clusters and a parameterization of the covariance matrix that maximize the fit of the normal mixture to the data since the formula for BIC includes the log-likelihood function, and this log-likelihood is normal in our case. This may create a tendency towards the formation of normal clusters as well.

In this way when we generate hybrid data using the multivariate normal mixture, the actual distribution of the clusters and the distribution of the synthesizer model might be more similar to each other than the distribution of the clusters obtained by microaggregation and the (multivariate normal) synthesizer model.

4 Analytical properties for continuous attributes

The proposed method preserves the mean vector and the covariance matrix within the clusters. In general, these characteristics are preserved in expectation. It is also possible (and we did it in our experiments) to generate clusters with sample means and sample covariance matrices equal to the corresponding means and covariance matrices of the original clusters. This involves simple transformation based on shifting and scaling. R package command mvrnorm (package MASS) with the option empirical=TRUE was used for that.

Preservation of the mean vector for the overall data set follows from its preservation within each cluster and from the fact that we generate the same number of records for each cluster in the locally synthesized data as in the original data.

The covariance matrix for the overall data set is related to the covariance matrix of the clusters as

$$\boldsymbol{\Sigma} = \sum_{g=1}^{G} \pi_g(\boldsymbol{\Sigma}_g + \mathbf{MDIF}_g), \tag{7}$$

where $\boldsymbol{\Sigma}_g$ is the covariance matrix of the locally synthesized data in the cluster g and $\mathbf{MDIF}_g$ is the following matrix:

$$\begin{pmatrix} (\mu_{g_1} - \mu_1)^2 & \cdots & (\mu_{g_1} - \mu_1)(\mu_{g_d} - \mu_d) \\ (\mu_{g_2} - \mu_2)(\mu_{g_1} - \mu_1) & \cdots & (\mu_{g_2} - \mu_2)(\mu_{g_d} - \mu_d) \\ & \ddots & \\ (\mu_{g_d} - \mu_d)(\mu_{g_1} - \mu_1) & \cdots & (\mu_{g_d} - \mu_d)^2 \end{pmatrix}.$$

In $\mathbf{MDIF}_g$, μ_i is the mean of variable X_i for the overall data set and μ_{g_i} is the mean of variable X_i over cluster g. Because in the original and locally synthesized data the overall means μ_i and the cluster means μ_{g_i} are preserved, matrices $\mathbf{MDIF}_g$ are the same in the original and locally synthesized data. Cluster covariances $\boldsymbol{\Sigma}_g$ are also preserved, so the overall covariance $\boldsymbol{\Sigma}$ will be preserved, too.

We want to note that sometimes a meaningful subpopulation in the data, for example, a group of patients with the same type of disease does not have a normal distribution. In such a case it is often represented by a mixture of normal components. Note that, using the same reasoning as above, we can see that the first two moments of this complex group will be preserved as well. In fact, the first two moments of any union of normal components will be preserved.

Now let us consider higher-order moments (third order and above). Preservation of these moments depends on the distribution within the clusters of the original data. Let us consider a generic central

moment $E[(X_1-\mu_1)^{s_1}(X_2-\mu_2)^{s_2}\cdots(X_d-\mu_d)^{s_d}]$. Let X_{m_i} be the variable i in the locally synthesized data. Since the cluster and overall means are preserved, we will omit the index m (denoting locally synthesized data) in the expressions for the means. Hence, for the locally synthesized data, the moment $E\Big[\prod_{i=1}^{d}(X_{m_i}-\mu_i)^{s_i}\Big]$ is

$$
\begin{aligned}
&\sum_{g=1}^{G}\pi_g E\Big[\prod_{i=1}^{d}(X_{m_{g_i}}-\mu_{g_i}+\mu_{g_i}-\mu_i)^{s_i}\Big]=\\
&=\sum_{g=1}^{G}\pi_g E\Big[\prod_{i=1}^{d}(\sum_{l_i=0}^{s_i}\binom{s_i}{l_i}(X_{m_{g_i}}-\mu_{g_i})^{l_i}(\mu_{g_i}-\mu_i)^{s_i-l_i})\Big]=\\
&=\sum_{g=1}^{G}\pi_g E\Big[\sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\cdots\sum_{l_d=0}^{s_d}\prod_{i=1}^{d}\Big(\binom{s_i}{l_i}(X_{m_{g_i}}-\mu_{g_i})^{l_i}\times\\
&\times(\mu_{g_i}-\mu_i)^{s_i-l_i}\Big)\Big]=\\
&=\sum_{g=1}^{G}\pi_g\sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\cdots\sum_{l_d=0}^{s_d}\prod_{i=1}^{d}\Big(\binom{s_i}{l_i}(\mu_{g_i}-\mu_i)^{s_i-l_i}\Big)\times\\
&\times NormM_g(l_1,l_2,\cdots,l_d),
\end{aligned}
\tag{8}
$$

where $NormM_g(l_1,l_2,\cdots,l_d)=E[\prod_{i=1}^{d}(X_{g_i}-\mu_{g_i})^{l_i}]$ is the normal mixed central moment for $\boldsymbol{X}\sim N(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ over cluster g. The expression for $NormM_g(l_1,l_2,\cdots,l_d)$ can be found in the literature, for example in Phillips (2010).

Note that $NormM_g(l_1,l_2,\cdots,l_d)$ should be computed only for those moments for which $\sum_{i=1}^{d}l_i$ is even, because all other moments are equal to 0.

Taking into account that our method preserves the first two moments, the difference between the corresponding moments computed on the original and locally synthesized data is the following:

$$
\begin{aligned}
&E\Big[\prod_{i=1}^{d}(X_{m_i}-\mu_i)^{s_i}\Big]-E\Big[\prod_{i=1}^{d}(X_{o_i}-\mu_i)^{s_i}\Big]=\\
&=\sum_{g=1}^{G}\pi_g\Big(\Big(\underbrace{\sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\cdots\sum_{l_d=0}^{s_d}}_{\sum l_i>2\text{ and even}}\prod_{i=1}^{d}(\binom{s_i}{l_i}(\mu_{g_i}-\mu_i)^{s_i-l_i})\times\\
&\times\big(NormM_g(l_1,l_2,\cdots,l_d)-M_g(l_1,l_2,\cdots,l_d)\big)\Big)-\\
&-\Big(\underbrace{\sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\cdots\sum_{l_d}^{s_d}}_{\sum l_i>1\text{ and odd}}\prod_{i=1}^{d}\Big(\binom{s_i}{l_i}(\mu_{g_i}-\mu_i)^{s_i-l_i}\Big)\times\\
&\times M_g(l_1,l_2,\cdots,l_d)\Big)\Big),
\end{aligned}
\tag{9}
$$

where subscripts "o" and "m" denote the original and locally synthesized data, respectively, and $M_g(l_1,\cdots,l_d)$ are the moments computed in the g^{th} cluster of the original data.

The difference between the original and locally synthesized moments depends on the non-normal properties of the clusters of the original data. Obviously, if all the clusters in the original data are normally distributed then all the moments will be preserved by our local synthesis scheme. As we noted in section 3, the distribution of the clusters should tend to normality; hence, the first term on

the right-hand side of equation (9), which reflects the difference between "even" moments, should not be very big, and the second term should be close to zero.

We want to note that even though in our experiments some clusters were not normal the overall utility of the resultant data was still high, as shown in Section 5 below. A possible remedy for non-normality in some clusters would be to include a non-normal component/components in the mixture. A multivariate t-distribution can be used for such a component because it provides a longer tailed alternative to the normal distribution.

5 Experimental results with continuous data

The procedure described above was implemented using R software and evaluated on two medical data sets:

- The first data set, called THYROID, was obtained from the UCI Machine Learning Repository (Bache and Lichman, 2013). It contains measurements of the following five continuous attributes: AGE (patient's age), TSH (thyroid-stimulating hormone), $T3$ (triiodothyronine), $T4U$ (thyroxine utilization rate), FTI (free thyroxine index). There are 2,800 records in this data set.
- The second data set, called NEOPLASM, was extracted from the Patient Discharge Data for 2010. This data set can be obtained from the California's Office of Statewide Health Planning and Development (OSHPD, 2010). The following numerical attributes were selected for the patients whose principal diagnosis was some kind of neoplasm: AGE_YRS (age in years), LOS (length of stay from admission to discharge in days), $CHARGE$ (in dollars). There are 19,502 records in this data set.

We applied our method to the THYROID and NEOPLASM data sets with the parameter $k = 60$ corresponding to the minimal number of records per cluster to guarantee k-anonymity. Since the EM algorithm is sensitive to the initial solution, we initialized EM with the result of a model-based hierarchical agglomerative clustering, which approximately maximizes the classification likelihood (as suggested in Fraley and Raftery (2002)).

As for the number of clusters, we considered from 2 to 10 clusters, and computed the BIC for each case. The model for which BIC was maximal was chosen. So, for the THYROID data six clusters (179, 232, 440, 325, 708 and 63 records) and unconstrained covariance matrices was chosen by BIC. For the NEOPLASM data, we obtained a model with 9 clusters with equal shape but different orientation and volume. The sizes of clusters ranged from 93 to 3625 records. From now on we denote our method by Hybrid.

For the sake of comparison we also generated data sets where the clustering step was done by MDAV multivariate microaggregation (Domingo-Ferrer and Torra, 2005), and the data synthesis within each cluster was done using a synthesizer that preserves means and covariance matrices, as described in Domingo-Ferrer and González-Nicolás (2010). Denote this method by Hybridmicro. To achieve a fair comparison with the model-based approach we set the microaggregation parameter k equal to the average cluster size in the Hybrid method for the respective data set. This is necessary because microaggregation implicitly sets an upper bound on the cluster size which is equal to $2k - 1$ records (Domingo-Ferrer and Mateo-Sanz, 2002), thereby producing many more clusters than Hybrid for the same value of k. Thus, for the THYROID data set, the microaggregation parameter k was set to 324 records per cluster and, for the NEOPLASM data set, it was set to $k = 2166$ for Hybridmicro method. In this way, THYROID was divided into 6 clusters and NEOPLASM into 9 (same as in Hybrid).

The obtained hybrid data sets were compared with fully synthetic data. For synthetic data generation we used a method based on multivariate sequential regressions described in Raghunathan et al. (2003); Reiter (2002); Little et al. (2004) and implemented in the free multiple imputation software IVEware (Raghunathan et al., 2011).

Other methods used for comparison were plain multivariate microaggregation MDAV (Domingo-Ferrer and Torra, 2005), denoted by Micro, and noise addition, which are perturbation methods. Multivariate microaggregation was performed with $k = 20$ records per cluster for THYROID and $k = 200$ for NEOPLASM. The choice of k was made empirically to reach a reasonably fair comparison with the other methods. Since Hybrid and HybridMicro restore the variance within the clusters and Micro does not, it would be unfair to compare them with Micro with $k = 2166$ records per cluster for NEOPLASM data, because such microaggregated data would have only 9 distinct records. With $k = 200$ there are 97 different records, which is a much better case. Similar considerations apply to justify the $k = 20$ used in THYROID.

We used the implementation of MDAV microaggregation available in the sdcMicro R package (Templ, 2008) for our method Micro and the first step of HybridMicro. Regarding noise addition, we used a version that preserves the mean vector and the covariance matrix. This method was implemented in the following way

$$\boldsymbol{X}_m = E[\boldsymbol{X}_o] + \frac{(\boldsymbol{X}_o - E[\boldsymbol{X}_o]) + \boldsymbol{E}}{\sqrt{1+c}}, \tag{10}$$

where $\boldsymbol{X}_m$ is the masked data, $\boldsymbol{X}_o$ is the original data, $E[\boldsymbol{X}_o]$ denotes the expectation of $\boldsymbol{X}_o$, $\boldsymbol{E}$ is random noise with $N(\boldsymbol{0}, c\boldsymbol{\Sigma}_o)$, $\boldsymbol{\Sigma}_o$ is the covariance matrix of the original data, and c is the parameter of the method which regulates the amount of the noise added to the data. We used $c = 0.15$, as recommended in the literature (Oganian, 2003; Oganian and Karr, 2006; Woo et al., 2009). We call this method Noise.

To evaluate the data quality provided by these methods, we chose a generic measure of data utility suitable for data with different types of attributes, continuous and categorical, and for a number of analyses: the propensity score measure (Woo et al., 2009). This measure assesses the discrepancy between the distribution of the original and the protected data. It is based on discrimination between the original and the disclosure-protected data: protected data that are difficult to distinguish from the original data have relatively high utility. As noted in Drechsler and Reiter (2009) and Drechsler (2011) it can be quite useful for synthetic methods and has been adapted by the US Census Bureau.

Below are the details of the propensity score utility metric. The propensity score is defined as the probability that a binary variable T (which can take values 0 or 1) is equal to 1, given covariate values $\boldsymbol{x}$. As Rosenbaum and Rubin (1983) shows, T and $\boldsymbol{x}$ are conditionally independent given the propensity score. Thus, when two large groups have the same distribution of propensity scores, the groups should have similar distributions of $\boldsymbol{x}$.

This theory suggests an approach for measuring data utility. First, we merge (by "stacking") the original and masked data sets, adding a variable T that equals one for all records from the masked data set and equals zero for all records from the original data set. If variables have been dropped as part of the masking, they are also dropped in computation of propensity scores. Secondly, for each record in the original and masked data, we compute the probability of being in the *masked* data set—the propensity score. Propensity scores can be estimated via a logistic regression of the "masked/original" variable T on functions of all variables $\boldsymbol{x}$ in the data set. The propensity scores are the predicted probabilities in this logistic regression (Cox and Snell, 1989). In our experiments we used the regression model with all main effects and interactions from first to the third order among all the variables.

Thirdly, we compare the distributions of the propensity scores in the original and the masked data. When those distributions are similar, the distributions of the original and masked data are similar, and so data utility should be relatively high. The similarity of the propensity scores for the masked and original observations can be assessed in different ways. We will use the summary proposed in Woo et al. (2009):

$$U_p = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{p}_i - 1/2\right]^2, \tag{11}$$

where N is the total number of records in the merged data set and $\hat{p}_i$ is the estimated propensity score for unit i. According to Woo et al. (2009), when the original and masked data have the same

Table 1: Continuous data. Propensity score utility (multiplied by $2N$ for better representation) for various methods (lower values mean better utility)

Method	Thyroid	Neoplasm
Hybrid	23.07	68.44
HybridMicro	106.70	142.13
Noise	301.11	119.68
Synthetic	565.56	1390.37
Micro	276.11	105.01

Table 2: Average percentage change in regression coefficients and standard errors (the average is over all regressions on THYROID and NEOPLASM)

Method	Reg. coef	Std. error
Hybrid	0.6%	0.35%
HybridMicro	0.7%	0.3%
Noise	4.6%	1.44%
Synthetic	12.53%	3.05%
Micro	8.94%	2.41%

distribution, the propensity scores of all the records in the merged data set are approximately equal to $1/2$ (if the original and masked file have the same number of records) and hence the above summary U_p should be near zero. Intuitively, this corresponds to the maximum uncertainty in the classification of the records as masked or original.

The results for different methods are shown in Table 1. These are average values of data utility for 30 realizations of protected data sets obtained from the same original data set by the application of Hybrid, HybridMicro, Synthetic and Noise; for Micro a single realization was enough because it is a deterministic method. We see that both hybrid methods, Hybrid and HybridMicro, by far outperform the fully synthetic method Synthetic (smaller values mean better utility). Further, Hybrid is the best method in terms of utility. We noticed that Hybrid performs better than HybridMicro even if we increase the number of clusters for HybridMicro. For example, in the case of the THYROID data set, when we changed the aggregation level from 324 records per cluster to 60 for HybridMicro, thus increasing the number of clusters from 6 to 32, the average utility for HybridMicro was about 40, which is still significantly worse than the utility of Hybrid with only 6 clusters. Recall that reducing the number of clusters without losing utility may be desirable because higher levels of aggregation can be expected to result in lower disclosure risk.

In addition to the propensity score utility metric, which is a generic utility measure, we considered some measures specific to a particular statistical analysis, for example, a linear regression. The Hybrid method preserves very well all the regression coefficients and their standard errors in the resulting protected data. In particular, we performed linear regressions on THYROID and NEOPLASM (by regressing every variable in each data set on all the other variables in the corresponding data set). Table 2 contains the average percent change in regression coefficients and standard errors. Average percentage change here is the ratio $\frac{|r_{orig}-r_{masked}|}{|r_{orig}|}$, where r_{orig} and r_{masked} are the statistics of interest (regression coefficient, std. error) computed on the original and masked data correspondingly.

Hybrid methods appear to be the best ones in terms of preservation of regression coefficients and their standard errors. This result is attributed to the fact that both hybrid methods by design preserve the first two moments.

The advantage of Hybrid method over HybridMicro is apparent when we consider higher-order moments. For example, the average percentage change over the third and fourth order moments for the

Table 3: Average percentage change in 3^{rd} and 4^{th}-order moments (THYROID data)

Data set	3^{rd} order moments	4^{th} order moments
Hybrid	0.6%	1.6%
HybridMicro	1.9%	5.1%
Noise	1.1%	2.5%
Synthetic	17%	33.7%
Micro	1.4%	1.6%

overall protected THYROID data are given in the Table 3. We can see in Table 3 that Hybrid compares very favorably with other methods.

In general, the utility of our approach depends on the value of the parameter k (the parameter of the probabilistic k-anonymity criterion). This value sets the minimal number of records per cluster for our method. For example, for the THYROID data when we increased k from 60 to 200, the propensity score utility measure increased to 77, which was still better than the utility of other methods, including HybridMicro. Estimates of regression coefficients and their standard errors did not change. When we increased k to 400, the propensity score utility increased to 95, which was still better than the score of other methods. A similar tendency was observed for the NEOPLASM data.

Concerning the disclosure risk, Hybrid and HybridMicro both satisfy the requirements of probabilistic k-anonymity. Noise does not satisfy the requirements of probabilistic k-anonymity. Micro satisfies them, but the value of k was set much smaller than the value of k for the hybrid methods, in order to achieve a fair utility comparison with other methods. As a result, the microaggregated data set has higher risk compared to Hybrid and HybridMicro. The Synthetic method obviously satisfies the requirements of probabilistic k-anonymity with the highest possible value for k ($k = n$). However, as we can see from the tables above, it does not compare very well with the other methods in terms of utility.

6 Local synthesis for continuous and categorical attributes

Our approach for hybrid data generation can be extended to data with continuous and categorical attributes. To incorporate categorical variables in the scheme, we will use a version of the modified latent class model (Lazarsfeld, 1950).

In Latent Class Analysis (LCA), an unobserved "clustering variable" (with categories corresponding to specific clusters) is a "latent" variable. It is assumed that this variable "explains" all the relationships between the observed categorical attributes. Hence, conditional upon the values of the latent variable, the responses to all the observed categorical attributes are assumed to be statistically independent —a so-called "local" independence assumption in LCA.

Let us introduce some notation. Suppose that for individual respondents $i = 1, \cdots, N$, we observe J polytomous categorical attributes, with the j-th attribute having C_j possible outcomes. For $i = 1, \cdots, N$, $j = 1, \cdots, J$ and $c = 1, \cdots, C_j$, let Y_{ijc} be a binary attribute such that $Y_{ijc} = 1$ if respondent i gives the c-th response for the j-th attribute, and $Y_{ijc} = 0$ otherwise.

The parameters that are estimated by the latent class model are the proportions p_g of observations in each class g (with $\sum_g p_g = 1$) and the probabilities π_{jcg} of category c for the attribute j conditional on latent class g. The probability that an individual i has a particular set of categorical responses is

$$P(Y_i) = \sum_{g=1}^{G} p_g \prod_{j=1}^{J} \prod_{c=1}^{C_j} (\pi_{jcg})^{Y_{ijc}}.$$

To generate multivariate data with mixed continuous and categorical attributes which are not necessarily independent, we propose a two-step approach. First, we will generate continuous attributes using the mixture model as described in Section 3. Then, we will estimate the parameters of categorical attributes using a generalization of the latent class model, called the latent class regression model (Dayton and Macready, 1988; Hagenaars and McCutcheon, 2002), where the mixing proportions p_g depend on the values of the continuous attributes. For such a model the log-likelihood function is

$$l = \sum_{i=1}^{N} \ln \sum_{g=1}^{G} p_g(\boldsymbol{X}_i) \prod_{j=1}^{J} \prod_{c=1}^{C_j} (\pi_{jcg})^{Y_{ijc}}, \tag{12}$$

where X_i are the observed continuous attributes for individual i and $p_g(\boldsymbol{X}_i)$ are the latent class membership priors for the categorical variables that are dependent on continuous attributes. $p_g(\boldsymbol{X}_i)$ can be computed using multinomial logistic regression. Details can be found in the literature (Dayton and Macready, 1988; Hagenaars and McCutcheon, 2002).

To guarantee that the requirements of probabilistic k-anonymity are satisfied, we propose to incorporate constraints on cluster cardinality (as with continuous variables). First, we note that Equation (12) has the property that, ignoring covariates, the likelihood function has the latent class form

$$l = \sum_{i=1}^{N} \ln \sum_{g=1}^{G} p_g^* \prod_{j=1}^{J} \prod_{c=1}^{C_j} (\pi_{jcg})^{Y_{ijc}}. \tag{13}$$

p_g^* here represents the mean prevalence of the g-th class by averaging over the distribution of continuous variables X.

As in the latent class model, estimates of p_g^* can be computed in the M-step of the EM algorithm as

$$p_g^* = \sum_{i=1}^{n} \hat{z}_{ig}/n. \tag{14}$$

where $\hat{z}_{ig}$ are the posterior class membership probabilities, estimated on the E-step (see the formula in Bandeen-Roche et al. (1997)).

Hence, similar to the case of continuous variables, we can update p_g^* using Equations (3) and (4). Denote by p_g^{*old} the old version and by p_g^{*new} the updated version. If p_g^{*new} is different from p_g^{*old}, then we will recalculate z_{ig} as

$$z_{ig}^{new} = z_{ig} + (p_g^{*new} - p_g^{*old})/n. \tag{15}$$

It is easy to show that the vector z_g^{new} is the closest to z_g according to the two-norm.

The E and M-steps are repeated until convergence by assigning the new parameter estimates to the old ones. Then categorical variables are generated using the values of the parameters estimated above.

In regard to computational complexity of the algorithm, we want to note that the computational burden of the proposed approach is mainly due to the model estimation part, that is, the E and M steps of the EM algorithm. Each iteration requires $O(Gnd)$ computations. The total run time of the algorithm depends on the number of iterations required for convergence, and the latter depends on the convergence threshold used. In our experiments we observed that the number of iterations until convergence was not greater than twelve. The total run time for our data sets was less than two minutes on a MAC with a 2.7 GHz processor and a 16GB RAM.

For very big data sets (with many thousands of records and hundreds of variables), some adjustments can be done to increase the computational speed. In particular, methods of dimensionality reduction, such as principal components and subsampling can be used prior to the application of the method

(Fraley et al., 2005; Wehrens et al., 2004). For high-dimensional data we can also impose restrictions on the cluster models; for example, for continuous variables we may consider only spherical or diagonal models thereby reducing the number of parameters that need to be estimated for the cluster covariance matrices.

Nonetheless, in the case of non-interactive one-time data release, which is the focus of this paper, the amount of time spent on model estimation does not have such a crucial importance as in interactive query-based systems. The increase in the run time for finding an adequate model for data generation can be justified when the utility of the resultant data is high.

7 Experimental results with continuous and categorical attributes

The procedure described in Section 6 was implemented and applied to the THYROID data set with both continuous and categorical variables. In addition to the continuous variables AGE, TSH, T3, T4U, FTI described in Section 5 we also included fifteen categorical attributes, which were available on the UCI Machine Learning Repository. These variables are: Sex, On Thyroxine, Query On Thyroxine, On Antithyroid Medication, Sick, Pregnant, Thyroid Surgery, I131 Treatment, Query Hypothyroid, Query Hyperthyroid, Lithium, Goitre, Tumor, Hypopituitary, Psych (Quinlan, 1987). All these variables are dichotomous. As we mentioned in Section 5, application of our Hybrid method led to creation of six clusters for the continuous attributes, so we used a model with six latent classes here as well, and parameter $k = 60$ (minimal number of records per cluster). For the sake of comparison we also generated a fully synthetic data set based on the THYROID file. We used a sequential regression method where categorical and continuous attributes were generated, respectively, using logistic and normal linear regression.

To compare the Synthetic and Hybrid methods, we used the propensity-based information loss measure. To compute propensity scores, we used a model that includes all main effects and interactions between continuous variables, continuous - categorical interactions and some categorical interactions (we didn't include all the interactions in order not to get overparameterized model).

The resulting data utility measure for Hybrid was 121 and for Synthetic it was 690.55, showing that the Hybrid method significantly outperforms the Synthetic method. The advantage of Hybrid over Synthetic was always present, no matter what models we used to compute propensity scores, or whether the models included only main effects or many different interactions.

8 Concluding discussion and future work

Model-based clustering followed by generation of synthetic records using parameters estimated in the clustering step seems to be quite a promising and flexible approach to achieve k-anonymous releases of healthcare data.

In our experiments with continuous and categorical variables, this approach outperformed other disclosure limitation methods which were considered for comparison including the fully synthetic data generator based on sequential regressions. This suggests that global synthesis of data sets with complex structure may not perform very well in terms of data utility. In contrast, *local synthesis* may be the best option, provided that clustering is model-based. Indeed, a proper combination of clustering and synthesis may capture the properties of the data which are hard to model on the global data set.

Our method is also flexible in the sense that, by increasing or decreasing the number of clusters, we can obtain data that resemble the original data more or less closely.

Finally, we want to add that the availability of the diagnosis variable in many medical data sets can be informative for the data protector about the model; in particular, the number of diagnosis/disease categories can give an idea about the possible number of mixture components. We also believe that the use of model-based clustering to classify individuals into disease categories in many areas of medical research should make our method more appealing to the potential data users.

In the future we plan to continue investigating a hybrid approach for disclosure limitation. Some of the directions of our future work are the following:

- Investigate and compare different mixture models with different component distributions;
- For disclosure limitation of categorical variables, investigate the possibility of relaxing the local independence assumption from the point of view of data utility and disclosure risk;
- Explore local masking, that is, the combination of clustering and within-cluster masking, in a way parallel to the local synthesis proposed in this paper.

Acknowledgments and disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Center for Health Statistics or the Centers for Disease Control and Prevention.

The second author is with the UNESCO Chair in Data Privacy, but the views expressed in this paper do not necessarily reflect the position of UNESCO nor commit that organization. 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 project "DwB". The second author is partially supported as an ICREA Acadèmia researcher.

References

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *31st International Conference on Very Large Data Bases - VLDB'05*, pp. 901–909.

Bache, K. and M. Lichman (2013). Uci machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. `http://archive.ics.uci.edu/ml`.

Bandeen-Roche, K., D. Miglioretti, S. Zeger, and P. Rathouz (1997). Latent variable regression for multiple discrete outcomes. *Journal of American Statistical Association 92*(440), 1375–1386.

Caiola, G. and J. Reiter (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy 3*(1), 27–42.

Campbell, J., C. Fraley, F. Murtagh, and A. Raftery (1997). Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters 18*, 1539–1548.

Campbell, J., C. Fraley, D. Stanford, and A. Raftery (1999). Model-based methods for real-time textile fault detection. *International Journal of Imaging Systems and Technology 10*, 339–346.

Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition 28*, 781–793.

Charest, A. (2012a). *Creation and Analysis of Differentially-Private Synthetic Datasets*. Ph. D. thesis, Carnegie-Mellon University.

Charest, A. (2012b). Empirical evaluation of statistical inference from differentially-private contingency tables. In J. Domingo-Ferrer and I. Tinnirello (Eds.), *Privacy in Statistical Databases-PSD 2012, Lecture Notes in Computer Science 7556*, pp. 257–272. Springer-Verlag.

Ciriani, V., S. De Capitani di Vimercati, S. Foresti, and P. Samarati (2008). k-anonymous data mining: a survey. In P. Yu and C. Aggarwal (Eds.), *Privacy-Preserving Data Mining: Models and Algorithms*, pp. 105–136. Springer-Verlag.

Cox, D. R. and E. J. Snell (1989). *Analysis of Binary Data* (2nd Ed. ed.). Chapman and Hall.

Dandekar, R., M. Cohen, and N. Kirkendall (2002). Sensitive microdata protection using latin hypercube sampling tecchnique. In *Inference Control in Statistical Databases, Lecture Notes in Computer Science*, Volume 2316, Springer, Berlin Heidelberg, pp. 245–253.

Dayton, C. and G. Macready (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association 83*(401), 173–178.

Domingo-Ferrer, J. and U. González-Nicolás (2010). Hybrid microdata using microaggregation. *Information Sciences 180*, 2834–2844.

Domingo-Ferrer, J. and J. Mateo-Sanz (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering 14*(1), 189–201.

Domingo-Ferrer, J., A. Oganian, and V. Torra (2002). Information-theoretic disclosure risk measures in statistical disclosure control of tabular data. In *14th International Conference on Scientific and Statistical Database Management - SSDBM 2002*, pp. 227–231. Los Alamitos CA: IEEE Computer Society.

Domingo-Ferrer, J. and V. Torra (2005, 2). Ordinal, continuous and heterogeneous k-anonimity through microaggregation. *Data Mining and Knowledge Discovery 11*, 195–212.

Domingo-Ferrer, J. and V. Torra (2008). A critique of k-anonymity and some of its enhancements. In *Third International Conference on Availability, Reliability and Security, ARES'08*, Washington DC, USA, pp. 990–993. IEEE Computer Society.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Springer.

Drechsler, J. and J. Reiter (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey. *Journal of Official Statistics 25*, 589–603.

Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming - ICALP 2006*, Lecture Notes in Computer Science 4052, pp. 1–12. Springer-Verlag.

Dwork, C. (2011). A firm foundation for private data analysis. *Communications of the ACM 54*(1), 86–95.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography - TCC 2006*, Lecture Notes in Computer Science 3876, pp. 265–284. Springer-Verlag.

Edwards, A. and L. Cavalli-Sforza (1965). A method for cluster analysis. *Biometrics 21*, 362–375.

Fienberg, S., A. Rinaldo, and X. Yang (2010). Differential privacy and the risk-utility tradeoff for multidimensional contingency tables. In J. Domingo-Ferrer and E. Magkos (Eds.), *Privacy in Statistical Databases - PSD 2010*, Lecture Notes on Computer Sscience 6344, pp. 187–199. Springer-Verlag.

Fraley, C. and A. Raftery (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal 41*, 578–588.

Fraley, C. and A. Raftery (2002, June). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Fraley, C., A. Raftery, and R. Wehrens (2005). Incremental model-based clustering for large datasets with small clusters. *Journal of Computational and Graphical Statistics 14*, 1–18.

Hagenaars, J. and A. McCutcheon (2002). *Applied Latent Class Analysis*. Cambridge University Press.

Hall, R., A. Rinaldo, and L. Wasserman (2011). Random differential privacy. arXiv:1112.2680.

Hansen, P., B. Jaumard, and N. Mladenovic (1998). Minimum sum of squares clustering in a low dimensional space. *Journal of Classification 15*, 37–55.

Hardt, M., K. Ligett, and F. McSherry (2012). A simple and practical algorithm for differentially private data release. In *26th Annual Conference on Neural Information Processing Systems - NIPS 2012*, pp. 2348–2356.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, and P.-P. DeWolf (2010). *Handbook on Statistical Disclosure Control (version 1.2)*. ESSNET SDC project. `http://neon.vb.cbs.nl/casc`.

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

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. Sanil (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician 60*(3), 224–232.

Lazarsfeld, P. (1950). The logical and mathematical foundations of latent structure analysis. In S. e. a. Stouffer (Ed.), *Measurement and Prediction*, pp. 362–412. John Wiley & Sons.

Lee, A. (2009). Generating synthetic microdata from published marginal tables and confidential files. *Official Statistics Research Series 5*. `http://www.statisphere.govt.nz/further-resources-and-info/official-statistics-research/series/volume-5-2009.aspx#3`.

Li, N., T. Li, and S. Venkatasubramanian (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In R. Chirkova, A. Dogac, M. Ozsu, and T. Sellis (Eds.), *ICDE*, pp. 106–115. IEEE.

Lin, X., C. Clifton, and M. Zhu (2005). Privacy preserving clustering with distributed em mixture modeling. *Knowledge and Information Systems 8*, 68–81.

Little, R., F. Liu, and T. Raghunathan (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. Meng (Eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pp. 141–152. Wiley.

Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber (2008). Privacy: theory meets practice on the map. In *24th Intl. Conf. on Data Engineering - ICDE 2008*, pp. 277–286. IEEE.

Muralidhar, K. and R. Sarathy (2008). Generating sufficiency-based non-synthetic perturbed data. *Transactions on Data Privacy 1*(1), 17–33. `http://www.tdp.cat/issues/tdp.a005a08.pdf`.

Oganian, A. (2003). *Security and Information Loss in Statistical Database Protection.* Ph. D. thesis, Universitat Politecnica de Catalunya.

Oganian, A. and J. Domingo-Ferrer (2012). Hybrid microdata via model-based clustering. In *Privacy in Statistical Databases-PSD 2012, Lecture Notes in Computer Science 7556*, pp. 103–115. Springer-Verlag.

Oganian, A. and A. F. Karr (2006). Combinations of SDC methods for microdata protection. In *Privacy in Statistical Databases-PSD 2006, Lecture Notes in Computer Science 4302*, pp. 102–113. Springer-Verlag.

OSHPD (2010). Patient discharge data. `http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html`.

Pathak, M. and B. Raj (2012). Large margin gaussian mixture models with differential privacy. *IEEE Transactions on dependable and secure computing 9*(4), 463–469.

Phillips, K. (2010). R functions to symbolically compute the central moments of the multivariate normal distribution. *Journal of Statistical Software, Code Snippets 33"*(1), 1–14. `http://www.jstatsoft.org/v33/c01`.

Quinlan, R. (1987). Thyroid desease data set. `https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease`.

Raghunathan, T., J. Reiter, and D. Rubin (2003). Multivariate imputation for statistical disclosure limitation. *Journal of Official Statistics 19*(1), 1–16.

Raghunathan, T. E., J. M. Lepkowski, J. van Hoewyk, and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a series of regression models. *Survey Methodology 27*, 85–96.

Raghunathan, T. E., P. Solenberger, and J. van Hoewyk (2011). Iveware. Imputation and Variance Estimation software. `http://www.isr.umich.edu/src/smp/ive/`.

Reiter, J. (2005). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics 21*, 441– 462.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics 18*, 531–544.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Samarati, P. and L. Sweeney (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Scott, D. (1992). *Multivariate Density Estimation*. New York: John Wiley & Sons.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

Soria-Comas, J. (2013). *Improving data utility in differential privacy and k-anonymity*. Ph. D. thesis, Universitat Rovira i Virgili, Tarragona, Catalonia.

Soria-Comas, J. and J. Domingo-Ferrer (2012). Probabilistic k-anonymity through micro aggregation and data swapping. In *IEEE International Conference on Fuzzy Systems - FUZZ IEEE 2012*, pp. 1–8.

Stanford, D. and A. Raftery (2000). Finding curvilinear features in spatial point patterns: Principle curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Inteligence 22*(6), 601–609.

Templ, M. (2008). Statistical disclosure control for microdata using the R-package sdcMicro. *Transactions on Data Privacy 1*(2), 67–85.

Torra, V. (2004). Microaggregation for categorical variables: a median based approach. In *Privacy in Statistical Databases – PSD 2004, Lecture Notes in Computer Science*, Volume 3050, Springer, Berlin Heidelberg, pp. 162–174.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. *Journal of American Statistical Association 58*, 236–244.

Wehrens, R., L. Buydens, C. Fraley, and A. Raftery (2004). Model-based clustering for image segmentation and large datasets via sampling. *Journal of Classification 21*, 231–253.

Woo, M.-J., J. Reiter, A. Oganian, and A. Karr (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality 1*(1), 111–124.