Synthetic Data Generation via the Permutation Paradigm with Optional k-Anonymity

Josep Domingo-Ferrer, *Fellow, IEEE*, Krishnamurty Muralidhar, and Sergio Martínez

Abstract—Most methods in the literature on synthetic microdata (individual records) generation are parametric, that is, they require knowing or estimating the joint or the conditional distribution of the original microdata. This may be a significant hurdle unless the original microdata are multivariate normal. We propose a rank-based approach to generating synthetic microdata based on the permutation paradigm. We present three different methods and we analyze the utility and the confidentiality they afford. The third method is actually an extension of the second method that adds k-anonymity protection against reidentification to the confidentiality against attribute disclosure offered by the first two methods. Our algorithms only require the identification of the *marginal* distributions of attributes and yield synthetic attributes that replicate the relationships between the original attributes exclusively based on ranks. This proposal is especially attractive for non-normal or multi-type microdata.

Index Terms—Privacy, data protection, synthetic data, permutation paradigm, disclosure risk assessment, anonymization

1 Introduction

In recent years, there has been a renewed interest in using synthetic data for analysis purposes in many areas, including healthcare, machine learning training, policy making, etc. If we focus on microdata —tables where columns are attributes and rows are records on individual respondents—, synthetic microdata conveying some analytical properties of the confidential original microdata may be regarded as a convenient approach to make data dissemination compatible with respondent privacy [19], where privacy means both protection against *reidentification disclosure* (anonymity) and protection against *attribute disclosure* (confidentiality). A distinction is sometimes made between *partially* synthetic microdata, where only some of the attributes or attribute values are synthesized, while the rest are released without confidentiality protection, and *fully* synthetic microdata, where all released values are synthetic. In what follows we will refer to fully synthetic microdata, and we will use data to mean microdata.

A key difference between traditional data masking and synthetic data is that the design of masking methods involves a one-to-one relationship between original and masked records, whereas no such relationship is used when generating synthetic data. Indeed, each masked value can be viewed as the result of adding some noise to a certain corresponding original value. In contrast, synthetic data are generated or simulated based on the *aggregate* characteristics of the original data.

J. Domingo-Ferrer and S. Martínez are with CYBERCAT-Center for Cybersecurity Research of Catalonia, Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia, e-mail {josep.domingo, sergio.martinezl}@urv.cat.
K. Muralidhar is with the Price College of Business, University of Oklahoma, 307 West Brooks, Adams Hall Room 10, Norman OK 73019-4007, USA, e-mail krishm@ou.edu.

However, the fact that no one-to-one relationship is employed in data synthesis can give a false sense of data protection. A concern is that, by itself, data synthesis does not necessarily give privacy guarantees: indeed, if too many analytical properties of the original data are to be preserved, one might end up generating overfitted synthetic data that are too similar to the original data. To remedy that problem, generation algorithms have been proposed that satisfy some privacy model by design, such as k-anonymity or differential privacy (DP).

Even though the origin of synthetic data generation is normally associated with multiple imputation and paper [37], an earlier proposal by [25] described a method to sample data from a fitted model, which amounts to computing synthetic data. There is a substantial body of research on synthetic data generation by multiple imputation [37], [35], [15]. More recently, several proposals to general DP synthetic data have appeared, based on the sampling approach [7], [17], [36], [27], [30], [26], [46], [41].

On their side, [32] have proposed to generate partially synthetic data by sampling a distribution conditional on the non-confidential attributes that preserves certain statistics. Other partially synthetic data approaches include [5] and [28]. Also based on conditional distributions are Monte Carlo techniques surveyed in [21]. The more recent paper [1] follows the sampling approach as well, and it leverages both sampling from a fitted distribution and sampling from a conditional distribution.

Further, there are approaches that use machine learning, based on diffusion models [18], [40], generative adversarial networks (GANs, [45]), or normalizing flows [16], [22]. Some GAN approaches produce DP synthetic data [20], [42], [44]. The majority of those methods focus on synthesizing images. For the specific case of synthetic microdata, see the practical guide to the

literature in [4].

Contribution and plan of this paper

Most existing approaches to generating synthetic microdata are based on modeling the characteristics of the original data, that is, they assume knowledge of the joint or conditional distribution of the original data, and they use such a distributional model to generate synthetic data. For example, if using CART-based synthetic data generation [36], synthetic attributes $Y_1, Y_2, \ldots$ are generated in sequence. Hence, it will be necessary to draw from $Y_2|Y_1$, then from $Y_3|Y_1, Y_2$, etc.

However, unless the original data are multivariate normal, it may be difficult to estimate the joint or the conditional distribution of the attributes. Generating mixed numerical and categorical attributes in this way can be even more difficult.

In this paper, we leverage the permutation paradigm introduced in [10] to propose a *rank-based* approach to generation of *synthetic microdata*. Our permutation-based methods use ranks, only require the *marginal* distributions of attributes, and perform *univariate* data generation. Our synthetic attributes replicate relationships between the original attributes exclusively based on the rank order of data. For non-normal data this is a significant advantage.

In Section 2 we give background on the permutation paradigm and on generic utility and confidentiality metrics. In Section 3, we describe two algorithms to generate synthetic data via the permutation paradigm, and a third algorithm which is a variant of the second one that achieves k-anonymity; the utility and confidentiality offered by each algorithm are also characterized. In Section 4 we report empirical work on the three proposed algorithms. Section 5 discusses the advantages of our permutation-based generation approach vs the state of the art. Finally, conclusions and future research lines are gathered in Section 6.

2 Background

In this section we first recall the permutation paradigm. Then we describe generic utility and confidentiality metrics proposed in [11] that will be used in the empirical section.

2.1 The permutation paradigm

In [10], the permutation model of anonymization was introduced. Consider an original attribute $X = \{x_1, x_2, \cdots, x_n\}$ and its anonymized version $Y = \{y_1, y_2, \cdots, y_n\}$. Assume X and Y can be ranked (even categorical nominal attributes can be ranked, using a semantic distance [13]). For $i = 1$ to n: compute $j = \text{Rank}(y_i)$ and let $z_i = x_{[j]}$, where $x_{[j]}$ is the value of X of rank j. Then call attribute $Z = \{z_1, z_2, \cdots, z_n\}$ the *reverse-mapped* version of X. For example, if original value $x_1 \in X$ is anonymized as $y_1 \in Y$, and y_1 is, say, the 3rd smallest value in Y, then we take z_1 to be the 3rd smallest value in X.

If there are several attributes in the original data set $\mathbf{X}$ and anonymized data set $\mathbf{Y}$, the previous reverse-mapping procedure is conducted for each attribute; call $\mathbf{Z}$ the data set formed by reverse-mapped attributes. Note that: (i) a reverse-mapped attribute Z is a permutation of the corresponding original attribute X; (ii) the rank order of Z is the same as the rank order of Y. Therefore, *any microdata anonymization technique is functionally equivalent to permutation* (from $\mathbf{X}$ into $\mathbf{Z}$) *followed by residual noise addition* (from $\mathbf{Z}$ into $\mathbf{Y}$). The added noise is residual, because the ranks of $\mathbf{Z}$ and $\mathbf{Y}$ are the same.

An important consequence of this permutation paradigm is that it shows that, *even if the anonymized data are synthetic,* an adversary could use reverse mapping to induce a linkage between the original and the synthetic data [11].

2.2 Utility metric based on propensity scores

We will use two general utility metrics for data sets with any number of attributes: a metric based on propensity scores and another metric based on covariances.

The authors of [39] suggest using the metric based on propensity scores proposed in [43] to assess the utility of synthetic data. This is a general utility metric that can be computed as follows:

1) Stack the n original records and the n synthetic records to create a data set with $2n$ records.
2) Add an indicator attribute I that labels original records as 0 and synthetic records as 1.
3) Fit a classifier to predict I based on the other attributes in the original and synthetic data sets.
4) For each record i, with $1 \leq i \leq 2n$, predict its indicator as $\hat{p}_i$.
5) Obtain the utility statistic as

$$U = \frac{1}{2n} \sum_{i=1}^{2n} (\hat{p}_i - 1/2)^2.$$

The intuition of the above metric is that the utility of the synthetic data is good if the classifier cannot tell the original from the synthetic records. In this case, the classifier will tend to predict $1/2$ for every record (because $1/2$ is the proportion of synthetic records in the stacked data set), rather than neatly classify as original (prediction 0) or synthetic (prediction 1). Thus, the better the utility, the closer U to 0.

Note that the classifier used to compute U should not overfit the data. An overfitted classifier would be able to memorize every original record and every synthetic record, and hence it would be trivially able to decide whether a record is original or synthetic. To avoid overfitting, simple models such as regression with only low-order interactions are preferable to deep learning models.

2.3 The UM utility metric

In [11], another general utility metric named UM is proposed that assesses the similarity between the covariance matrix $\mathbf{C}_{XX}$ of the rank matrix of the original data set $\mathbf{X}$ and the covariance matrix $\mathbf{C}_{YY}$ of the rank matrix of the anonymized data set $\mathbf{Y}$. The rationale is that covariance preservation is the most relevant utility feature for analyses aimed at discovering relationships between attributes. The metric $UM(\mathbf{X}, \mathbf{Y})$ is bounded between 0 and 1, and higher values indicate higher utility.

Let a data set $\mathbf{X}$ with m attributes be masked as $\mathbf{Y}$. Consider the ranks of values in both data sets, rather than the values themselves. If all attributes are numerical and sparse, one might choose to work on values rather than ranks in order to capture utility more closely.

In terms of covariances, maximum utility occurs when $\mathbf{C}_{XX} = \mathbf{C}_{YY}$, in which case the (second-order) relationships between attributes in the original data set are exactly preserved in the masked data set. To compare how similar $\mathbf{C}_{XX}$ and $\mathbf{C}_{YY}$ are, a rough procedure is to compare the magnitudes of their respective eigenvalues and the directions of the corresponding eigenvectors.

Let $\lambda_1^X, \ldots, \lambda_m^X$, resp. $\lambda_1^Y, \ldots, \lambda_m^Y$, be the eigenvalues of $\mathbf{C}_{XX}$, resp. $\mathbf{C}_{YY}$, in non-increasing order. Let $\mathbf{v}_1^X, \ldots, \mathbf{v}_m^X$, resp. $\mathbf{v}_1^Y, \ldots, \mathbf{v}_m^Y$, be the corresponding eigenvectors of $\mathbf{C}_{XX}$, resp. $\mathbf{C}_{YY}$.

Then we have

$$\lambda_j^X = (\mathbf{v}_j^X)^T \mathbf{C}_{XX} \mathbf{v}_j^X, \;\; j = 1, \ldots, m.$$

Now consider

$$\lambda_j^{Y|X} = (\mathbf{v}_j^X)^T \mathbf{C}_{YY} \mathbf{v}_j^X, \;\; j = 1, \ldots, m.$$

It is now possible to define the utility metric as

$$UM(\mathbf{X}, \mathbf{Y}) = \begin{cases} 1 \text{ if } \hat{\lambda}_j^X = \hat{\lambda}_j^{Y|X} = 1/m \text{ for } j = 1, \ldots, m; \\ 1 - \min\left(1, \frac{\sum_{j=1}^m (\hat{\lambda}_j^X - \hat{\lambda}_j^{Y|X})^2}{\sum_{j=1}^m (\hat{\lambda}_j^X - 1/m)^2}\right) \text{ otherwise.} \end{cases} \quad (1)$$

The first case in Expression (1) covers the (very exceptional) situation in which each of the original and the anonymized data sets is perfectly uncorrelated internally, which means there is no utility loss. Regarding the second case, it can be shown that the sum in the numerator is bounded within $[0, 2]$, whereas the sum in the denominator is bounded within $[0, (m-1)/m]$. In fact, this latter sum represents the maximum utility loss, which occurs when the covariances of $\mathbf{X}$ are completely lost in $\mathbf{Y}$ (that is, when any of the m eigenvectors of $\mathbf{C}_{XX}$ explains a fraction $1/m$ of the variance of $\mathbf{Y}$). By using the minimum in the second case, we make sure $UM(\mathbf{X}, \mathbf{Y})$ is bounded between 0 and 1. Thus we have:

- Top utility $UM(\mathbf{X}, \mathbf{Y}) = 1$ is reached when information loss is zero, which occurs when $\hat{\lambda}_j^X = \hat{\lambda}_j^{Y|X}$ for $j = 1, \ldots, m$.
- Zero utility $UM(\mathbf{X}, \mathbf{Y}) = 0$ occurs if $\hat{\lambda}_j^X$ and $\hat{\lambda}_j^{Y|X}$ differ at least as much as $\hat{\lambda}_j^X$ and the eigenvalues of an uncorrelated data set.

2.4 The $CM3$ confidentiality metric

In [11], a confidentiality metric $CM3$ based on canonical correlations between an original data set $\mathbf{X}$ and an anonymized data set $\mathbf{Y}$ was proposed. This metric does not need to know which anonymized record corresponds to which original record, and can therefore be used also when $\mathbf{Y}$ is synthetic. $CM3$ is bounded within $[0, 1]$ and higher values correspond to higher confidentiality.

$CM3$ is computed using the following algorithm:

1) For $j = 1$ to m do:
   a) Sort the original data set by its j-th attribute and let $\mathbf{X}^{-j}$ be the projection of the sorted data set on all attributes except the j-th one.
   b) Sort the anonymized data set by its j-th attribute and let $\mathbf{Y}^{-j}$ be the projection of the sorted data set on all attributes except the j-th one.
2) Let $CM3(\mathbf{X}, \mathbf{Y}) = \min_{1 \le j \le m} CM2(\mathbf{X}^{-j}, \mathbf{Y}^{-j})$, where

$$CM2(\mathbf{X}, \mathbf{Y}) = \prod_{i=1}^{m} (1 - \rho_i^2) \left[= e^{-I(\mathbf{X};\mathbf{Y})}\right]. \quad (2)$$

In Expression (2), ρ_i ($i = 1, \ldots, m$) are the canonical correlations between $\mathbf{X}$ and $\mathbf{Y}$, and $I(\mathbf{X}; \mathbf{Y})$ is the mutual information between $\mathbf{X}$ and $\mathbf{Y}$. Also, $CM2(\mathbf{X}, \mathbf{Y}) = 1$ is reached when the two data sets in the arguments tell nothing about each other, whereas $CM2(\mathbf{X}, \mathbf{Y}) = 0$ if at least one of the canonical correlations is 1, that is, at least one attribute in $\mathbf{X}$ is disclosed when releasing $\mathbf{Y}$.

Thus, $CM3$ measures confidentiality as the minimum value of $CM2$ over all possible sortings of the original and synthetic data sets by a single attribute. This circumvents the need to know the mapping between records in both data sets.

3 Generating Synthetic Data with the Permutation Paradigm

In this section, we present three different algorithms for synthetic data generation. They follow the principle of sampling from a fitted distribution to generate fully synthetic data [25].

Our algorithms take as input an original data set $\mathbf{X}$ with m attributes and n records (where each record can be construed as containing a respondent's answer), and they generate a synthetic data set $\mathbf{Y}$ with the same number of attributes and records. We use X_j to denote the j-th attribute and x_{ij} to denote the value of X_j in the i-th record.

3.1 Permutation through sampling with fixed correlations

Our first algorithm (Algorithm 1) uses sampling with fixed correlations to attain definite utility and confidentiality guarantees. We generate a rank matrix $\mathbf{Z}^*$ whose correlation is asymptotically the same as that of the rank matrix $\mathbf{Z}$ of the original data. At first glance,

this may not seem like a permutation approach. Yet, at a closer look, we observe that $\mathbf{Z}^*$ is simply a permuted version of $\mathbf{Z}$, where the permutation is achieved by sampling. This is one of the major advantages of the permutation paradigm: it can describe any anonymization method.

Algorithm 1:

1) Identify the marginal distribution of each attribute X_j of the original data set $\mathbf{X}$.
2) Let $\mathbf{Z} = (z_{ij})$ be the rank matrix of the original data set $\mathbf{X}$, where $1 \leq i \leq n$, $1 \leq j \leq m$, and z_{ij} is the rank of value x_{ij} within the values of attribute X_j.
3) Generate synthetic values of size n for each attribute *independently*. Generation can be done by sampling a distribution fitted to the original values of the attribute, for example using the acceptance-rejection method (see Section 8.2.4 of [24]). Let $\mathbf{A} = (a_{ij})$ be the matrix of generated values, with $1 \leq i \leq n$, $1 \leq j \leq m$.
4) Compute the rank order (Spearman) correlation matrix $\mathbf{R} = (r_{ij})$ of $\mathbf{Z}$, with $1 \leq i, j \leq m$.
5) Convert $\mathbf{R}$ to a product moment (Pearson) correlation matrix $\mathbf{P} = (\rho_{ij})$ using the following relation (Expression (6.4) from [23]):
$$\rho_{ij} = 2 \sin \left[\frac{\pi}{6} r_{ij} \right].$$
Product-moment correlations are needed for the Cholesky-based sampling in the next step.
6) Sample a random multivariate standard normal data set $\mathbf{B} = (b_{ij})$ of size $n \times m$ with correlation matrix $\mathbf{P}$ by using the Cholesky transformation. Let $\mathbf{Z}^* = (z^*_{ij})$ represent the rank matrix of $\mathbf{B}$.
7) Let the synthetic data $\mathbf{Y} = \{y_{ij}\}$, with $1 \leq i \leq n$, $1 \leq j \leq m$, be such that $y_{i,j} = a_{[z^*_{ij}],j}$, where $a_{[r],j}$ stands for the value in $\mathbf{A}$ with rank r with respect to attribute X_j.

The last two steps of Algorithm 1 correspond to the NORTA procedure in [6]. The procedure described here is similar to copula-based perturbation approaches presented in [38] and [33]. We can state the following propositions regarding the utility and the confidentiality afforded by the synthetic data output by Algorithm 1. We characterize utility as the preservation of the correlation matrix of the original data set by the synthetic data set, whereas we characterize confidentiality as the degree of independence between the ranks of the original record values and the synthetic record values (higher confidentiality means that seeing the synthetic attribute values discloses less about the original attribute values).

Proposition 1 (Utility): The rank order correlation matrix of the synthetic data $\mathbf{Y}$ is asymptotically the same as the rank order correlation matrix of the original data $\mathbf{X}$.

Proof: Subject to the sampling error of data set $\mathbf{B}$, the rank order correlations of $\mathbf{B}$ are asymptotically the same as those of data set $\mathbf{X}$. On the other hand, by construction $\mathbf{B}$ and $\mathbf{Y}$ have the same rank matrix $\mathbf{Z}^*$. Hence the proposition follows. □

Proposition 2 (Confidentiality): The rank matrix of $\mathbf{Y}$ is independent of the rank matrix of $\mathbf{X}$.

Proof: Since the rank matrices of $\mathbf{B}$ and $\mathbf{Y}$ are the same, it suffices to prove that the rank matrices of $\mathbf{X}$ and $\mathbf{B}$ are independent. Now, according to the NORTA procedure, each value b_{ij} in $\mathbf{B}$ is obtained by randomly sampling a distribution. Thus, in general the rank of b_{ij} within the j-th attribute of $\mathbf{B}$ is independent from the rank of x_{ij} within the j-th attribute of $\mathbf{X}$. □

Hence, an adversary that only sees $\mathbf{Y}$ does not obtain information on any specific record of $\mathbf{X}$.

3.2 Controlled permutation

Our second algorithm relies on direct permutation of the rank order matrix $\mathbf{Z}$ of the original data. It allows for controlled levels of permutation, from no permutation all the way to maximum permutation. The sampling procedure to generate synthetic data and the notations are the same as in the previous algorithm, unless otherwise stated.

Algorithm 2:

1) Let k be a positive integer parameter such that $k \leq n$, whose purpose is to control the permutation level.
2) Identify the marginal distribution of each attribute X_j of the original data set $\mathbf{X}$.
3) Let $\mathbf{Z} = (z_{ij})$ be the rank matrix of the original data set $\mathbf{X}$.
4) Generate synthetic values of size n for each attribute *independently*. Let $\mathbf{A} = (a_{ij})$ be the matrix of generated values.
5) Group the ranks $\{1, \ldots, n\}$ into the following set of rank subsets:
$$\{\{1, \ldots, k\}, \{k+1, \ldots, 2k\}, \{2k+1, \ldots, 3k\}, \ldots, \{\lfloor n/k - 1 \rfloor k + 1, \ldots n\}\},$$
where each rank subset contains k consecutive ranks, except the last one, which contains between k and $2k-1$ consecutive ranks.
6) Randomly permute each column of $\mathbf{Z}$ in such a way that permutation is carried out *within* each rank subset.
7) Let $\mathbf{Z}^* = (z^*_{ij})$ be the permuted rank matrix. Let the synthetic data $\mathbf{Y} = \{y_{ij}\}$ be such that $y_{i,j} = a_{[z^*_{ij}],j}$, where $a_{[r],j}$ stands for the value in $\mathbf{A}$ with rank r with respect to attribute X_j.

It can be seen that k does its job of controlling the permutation level. In particular, if $k = 1$, the values within each column of $\mathbf{Z}$ are not permuted, whereas if $k = n$, they are maximally permuted (with a random permutation on the overall set of ranks $\{1, \ldots, n\}$). In

general, the maximum rank difference caused by permutation will be $k-1$.

Note that Algorithm 2 is a generalization of data swapping and other similar procedures [8], [29].

Proposition 3: When using rank subsets of size k in Algorithm 2 with a data set of n records, the entropy of the rank of an original value given the synthetic data set lies within

$$[\log_2 k, \log_2(n - \lfloor n/k - 1 \rfloor k)] \text{ bits.} \quad (3)$$

Proof: Let Z be a random variable representing the rank of an original attribute value before it is actually known. If rank subsets of size k are used in Algorithm 2, the Shannon entropy of Z is $H(Z) = \log_2 k$ bits, except if the attribute value falls within the last rank subset, in which case $H(Z) = \log_2(n - \lfloor n/k - 1 \rfloor k) \geq log_2 k$ bits. □.

Corollary 1 (Confidentiality): The confidentiality of the original data set $\mathbf{X}$ in Algorithm 2 grows with k. In particular, it is maximum when $k = n$ and minimum when $k = 1$.

Proof: When $k = n$ in Expression (3), the entropy of the rank of an original value is $\log_2 n$ bits and hence maximum. Hence, an adversary's uncertainty on the rank of the original value is maximum: the rank of the original value is independent from the rank of the synthetic value.

On the other hand, when $k = 1$ in Expression (3), the entropy of the rank of an original value given the synthetic data set is $\log_2 1 = 0$ bits and hence minimum. Since the adversary knows the synthetic data set, she knows the rank of the original value (which is the same). Although the adversary does not know the actual original value, she can approximate it by the synthetic value having the same rank.

When $1 < k < n$, the adversary's uncertainty on the rank of the original value falls within the interval in Expression (3), whose bounds both grow with k. □.

Corollary 2 (Utility): The utility of the synthetic data set $\mathbf{Y}$ generated by Algorithm 2 decreases as k grows, where utility is understood as preservation of the rank correlation matrix.

Proof: With $k = n$, the ranks of the original and synthetic values *for each attribute* are independent, which means the rank correlation matrix of the original data set is not preserved at all by the synthetic data set. With $k = 1$, the ranks of the original and synthetic values are the same, which means the rank correlation matrix of the original data set is completely preserved by the synthetic data set. Further, the difference between an original attribute value and the synthetic attribute value in corresponding records must be small enough to preserve the rank. When $1 < k < n$, the rank correlation matrix is partially preserved, with preservation improving as k decreases. □

Algorithm 1 and Algorithm 2 provide confidentiality against attribute disclosure, but they do not deal with the risk of reidentification disclosure. The latter risk is relevant if the adversary is able to link the synthetic records to some external identified data source. The following variant of Algorithm 2 can be devised to achieve k-anonymity, which is a convenient privacy model here, because it focuses on the risk of reidentification. In this way, we add protection against reidentification to the protection against attribute disclosure inherited from Algorithm 2. Regarding the DP alternative, see Section 5 for a discussion. In this variant, we make a distinction between *quasi-identifier attributes* (those that are not identifiers but that can jointly lead to reidentification of the subject to whom the record corresponds, like Age, Job, Birthplace, Gender, etc.) and non-quasi-identifier attributes (those that are neither identifiers not quasi-identifiers, and include confidential attributes like salary, health condition, etc.). As usual in k-anonymity, quasi-identifiers are assumed known to the adversary, whereas non-quasi-identifier attributes are not. If the adversary's knowledge cannot be established, a safe worst-case option is to consider that all attributes are quasi-identifiers.

Algorithm 3: (Variant of Algorithm 2 with k-anonymity)

1) Let k be a positive integer parameter such that $k \leq n$, whose purpose is to control the permutation level.
2) Identify the marginal distribution of each attribute X_j of the original data set $\mathbf{X}$.
3) Let $\mathbf{Z} = (z_{ij})$ be the rank matrix of the original data set $\mathbf{X}$.
4) Generate synthetic values of size n for each attribute *independently*. Let $\mathbf{A} = (a_{ij})$ be the matrix of generated values.
5) Group the ranks $\{1, \ldots, n\}$ into the following set of rank subsets:
$$\{\{1, \ldots, k\}, \{k+1, \ldots, 2k\}, \{2k+1, \ldots, 3k\},$$
$$\ldots, \{\lfloor n/k - 1 \rfloor k + 1, \ldots n\}\},$$
where each rank subset contains k consecutive ranks, except the last one, which contains between k and $2k-1$ consecutive ranks.
6) Randomly permute each column of $\mathbf{Z}$ corresponding to a *non-quasi-identifier attribute* in such a way that permutation is carried out *within* each rank subset.
7) Use multivariate microaggregation with fixed size k (*e.g.*, the MDAV algorithm [14]) to jointly microaggregate the columns of $\mathbf{Z}$ corresponding to *quasi-identifier attributes* into clusters of rank tuples of size k, except perhaps the last cluster, which contains between k and $2k-1$ tuples.
8) Randomly and jointly permute the columns of $\mathbf{Z}$ corresponding to *quasi-identifier attributes* in such a

way that permutation is carried out within each cluster.

9) Let $\mathbf{Z}^* = (z^*_{ij})$ be the permuted rank matrix. Let the synthetic data $\mathbf{Y} = \{y_{ij}\}$ be such that $y_{i,j} = a_{[z^*_{ij}],j}$, where $a_{[r],j}$ stands for the value in $\mathbf{A}$ with rank r with respect to attribute X_j.

Proposition 4 (k-Anonymity): Algorithm 3 applied to the projection of records on quasi-identifier attributes achieves k-anonymity.

Proof. When microaggregation with parameter k is performed over the projection of records on the quasi-identifier attributes, the probability of reidentification is at most $1/k$ and k-anonymity is attained (see [14]). □

As mentioned above, *Algorithm 3 offers a protection stronger than k-anonymity, because it also protects against confidential attribute disclosure.* Specifically, in addition to achieving a probability of reidentification at most $1/k$, it replaces the values of the confidential attributes (typically a subset of the non-quasi-identifier attributes) by permuted synthetic values. Hence, the adversary cannot be certain whether an inference on the protected confidential attributes also holds on the original confidential attributes.

Regarding utility, like for Algorithm 2, the smaller k, the higher is utility preservation in Algorithm 3.

4 Empirical results

We conducted the experimental evaluation using both artificial and real-world original data sets. In the first category, we created an original data set $\mathbf{X}$ with 1000 records and 5 attributes X_1, X_2, X_3, X_4, et X_5. The values for each of those attributes were obtained by sampling as follows:

- X_1 was sampled from a Gamma distribution with parameters $\alpha = 1$ and $\beta = 100$, that is, an exponential distribution with parameter $\lambda = 1/100$.
- X_2 was sampled from a Uniform $[0, 1000]$ distribution.
- X_3 was sampled from a Gaussian distribution with mean 1000 and standard deviation 200.
- X_4 was sampled from a discrete distribution over $\{1, 2, 3, 4, 5\}$ such that $\Pr(1) = 0.3$, $\Pr(2) = 0.25$, $\Pr(3) = 0.20$, $\Pr(4) = 0.15$, and $\Pr(5) = 0.10$.
- X_5 was sampled from a binary distribution with $\Pr(0) = 0.6$ and $\Pr(1) = 0.4$.

We also used two real-world data sets for our analysis. The first is *Adult* [2]. After removing missing values, this data set comprises 45,222 records of census income information including both numerical and categorical values. The second is *Census* [3], which contains 1,080 records and includes 13 numerical attributes obtained using the Data Extraction System of the U.S. Bureau of the Census. We only used 12 among the 13 attributes in *Census*, because attribute PEARNVAL is just the difference between two other attributes (PTOTVAL and INTVAL).

Although no other method for synthetic microdata generation in the literature (*e.g.*, see [4]) allows the same precise control of the utility-confidentiality trade-off as our Algorithms 2 and 3, we also compared our proposal with *synthpop* [34], a method for generating synthetic versions of sensitive microdata for statistical disclosure control. This method synthesizes numerical or categorical attributes one by one using sequential modeling, and it is extensively used in the literature.

We used Visual Basic for Applications to implement Algorithm 1. We used Python on Jupyter Notebooks to implement Algorithms 2, 3, and the synthpop method. We used SAS version 9.4 to implement the propensity scores utility metric of Section 2.2 and the canonical correlation confidentiality metric of Section 2.4. The covariance-based utility metric of Section 2.3 was readily computable in Excel. No single run of any of our three algorithms took longer than 3 seconds of runtime on a MacBook Air with M2 chip and 16 GB RAM.

To facilitate reproducibility, the above original data set, the synthetic data computed with the acceptance-rejection method, and our codes are available at https://github.com/guialmons/permutation_synthetic

4.1 Utility: correlation matrices and scatterplots

This section presents utility results obtained with the three algorithms on the above-described artificial original data set $\mathbf{X}$. Its small number of attributes allows presenting correlations and scatterplots in a reasonable space. The (upper triangular) Pearson correlation matrix for $\mathbf{X}$ is

$$\mathbf{P}_{orig} = \begin{pmatrix} 1 & -0.686 & 0.535 & 0.296 & 0.148 \\ & 1 & -0.684 & -0.408 & -0.073 \\ & & 1 & 0.265 & 0.168 \\ & & & 1 & 0.291 \\ & & & & 1 \end{pmatrix}. \quad (4)$$

Figure 1 gives pairwise scatterplots of the original numerical attributes X_1, X_2, and X_3 in data set $\mathbf{X}$.

With Algorithm 1 (permutation through sampling), we obtained a synthetic data set also with 1000 records and attributes Y_1, Y_2, Y_3, Y_4, and Y_5. The Pearson correlation matrix among the generated synthetic attributes was

$$\mathbf{P}_{sam} = \begin{pmatrix} 1 & -0.695 & 0.556 & 0.249 & 0.157 \\ & 1 & -0.682 & -0.392 & -0.052 \\ & & 1 & 0.262 & 0.144 \\ & & & 1 & 0.215 \\ & & & & 1 \end{pmatrix}.$$

Figure 2 gives pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 1. It can be seen that these scatterplots are pretty similar to those of Figure 1.

We then tried Algorithm 2 (controlled permutation). We took different values of k:

- $k = 1$. This means that no permutation was performed. The Pearson correlation matrix among the

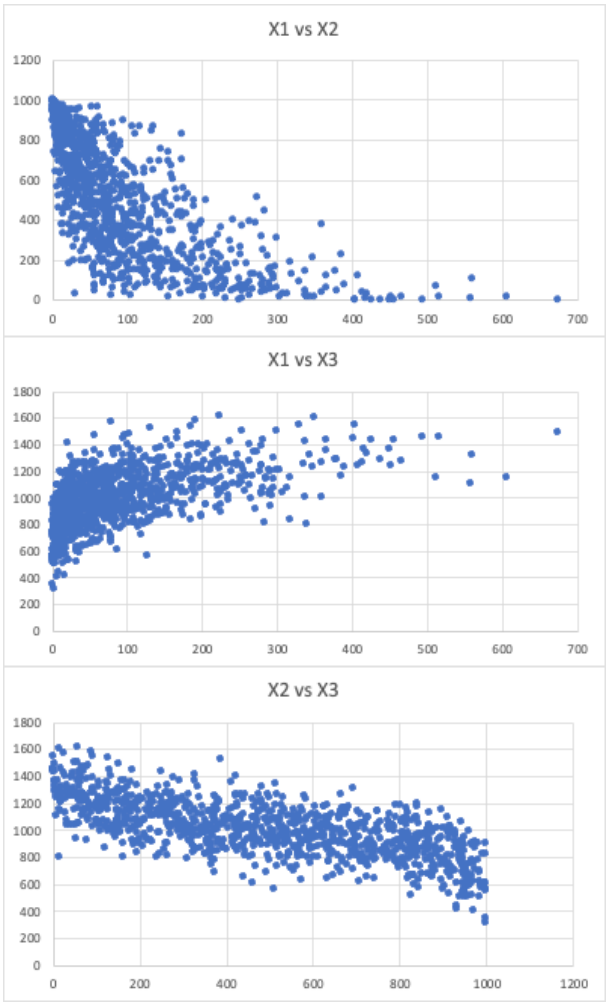


Fig. 1. Artificial data set. Pairwise scatterplots of the original numerical attributes X_1, X_2 and X_3.

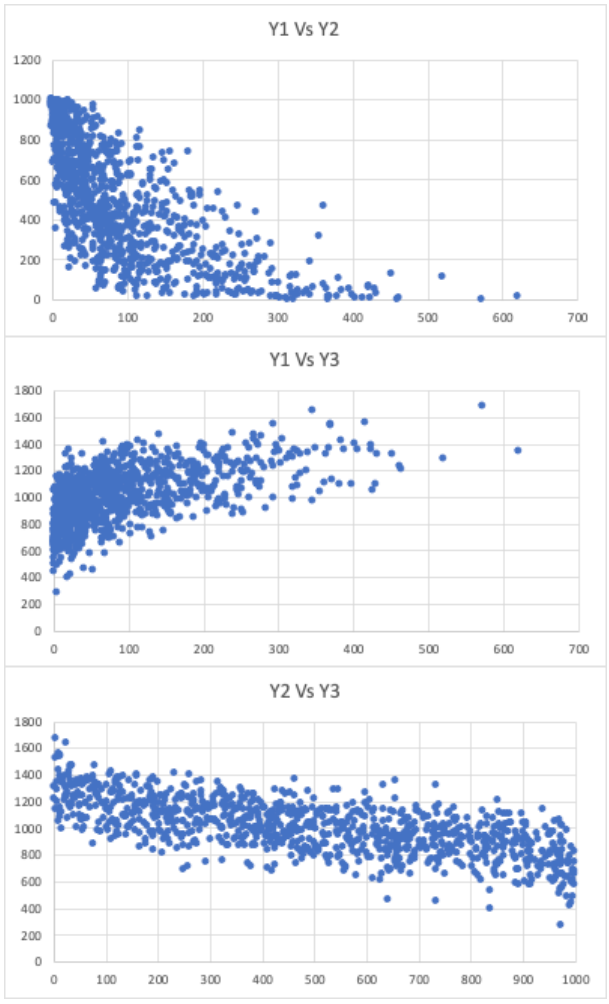


Fig. 2. Artificial data set. Pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 1 (permutation through sampling).

synthetic attributes obtained in this case was:

$$\mathbf{P}_{k=1} = \begin{pmatrix} 1 & -0.701 & 0.540 & 0.335 & 0.181 \\ & 1 & -0.687 & -0.460 & -0.126 \\ & & 1 & 0.347 & 0.206 \\ & & & 1 & 0.399 \\ & & & & 1 \end{pmatrix}.$$

Figure 3 gives pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 2 for $k = 1$. Whereas utility was good, because the rank orders of $\mathbf{X}$ and $\mathbf{Y}$ were the same, $\mathbf{P}_{k=1}$ is very similar to $\mathbf{P}_{orig}$, and the scatterplots of Figure 3 are similar to those of Figure 1, confidentiality was poor. If we look at the discrete attributes: the values of the pair (Y_4, Y_5) coincided with those of the pair (X_4, X_5) for 899 records out of the total 1000 records. For the continuous attributes, there were no exact coincidences but very similar corresponding values.

- $k = 50$. This means that groups of $k = 50$ records were formed and permutations took place within each group (maximum rank difference 49). Since $n = 1000$, this was a low permutation amount. The following is the (upper triangular) Pearson correlation matrix among the synthetic attributes, which is still pretty similar to that of the original data set (Expression (4)):

$$\mathbf{P}_{k=50} = \begin{pmatrix} 1 & -0.691 & 0.518 & 0.328 & 0.182 \\ & 1 & -0.677 & -0.446 & -0.132 \\ & & 1 & 0.334 & 0.193 \\ & & & 1 & 0.256 \\ & & & & 1 \end{pmatrix}.$$

Figure 4 gives pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 2 for $k = 50$. Similarity with respect to Figure 1 is still well preserved.

- $k = 250$. Here values whose ranks differed by up to 249 could be permuted. This was a relatively high permutation amount. The following Pearson correlation matrix was obtained for the synthetic

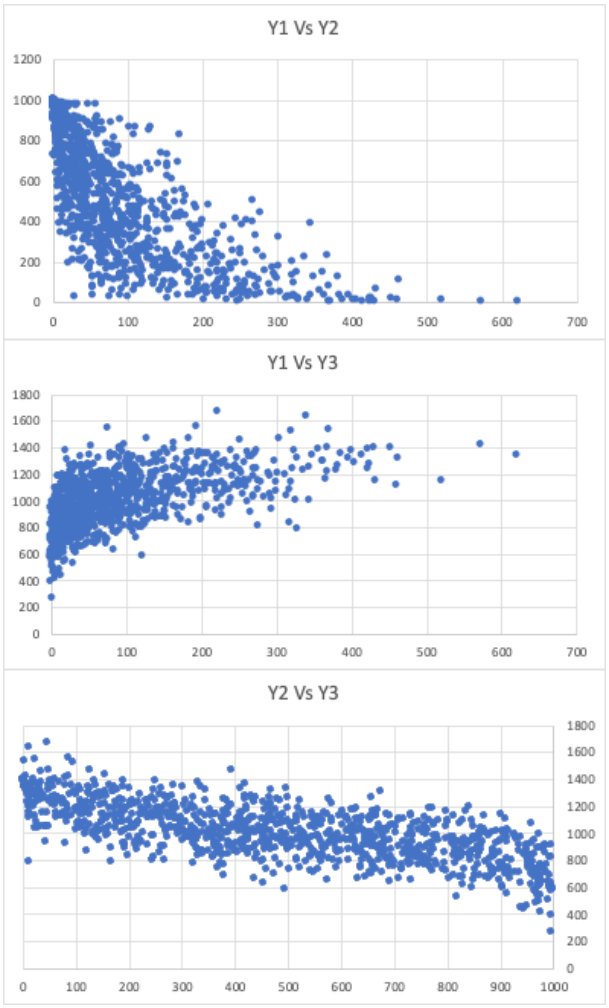


Fig. 3. Artificial data set. Pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 2 for $k = 1$ (no permutation).

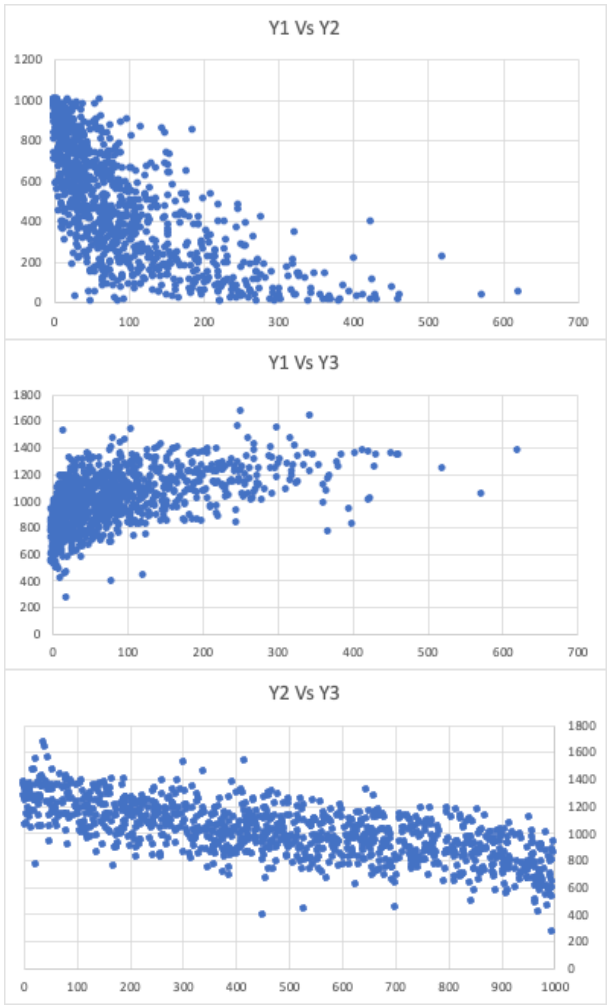


Fig. 4. Artificial data set. Pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 2 for $k = 50$ (low permutation).

attributes:

$$\mathbf{P}_{k=250} = \begin{pmatrix} 1 & -0.576 & 0.406 & 0.268 & 0.185 \\ & 1 & -0.558 & -0.403 & -0.221 \\ & & 1 & 0.321 & 0.225 \\ & & & 1 & 0.546 \\ & & & & 1 \end{pmatrix}.$$

It can be seen that $\mathbf{P}_{k=250}$ differs more than $\mathbf{P}_{k=50}$ from $\mathbf{P}_{orig}$. Figure 5 gives pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 2 for $k = 250$. Differences with respect to Figure 1 are already quite apparent.

- $k = 1000$ (maximum permutation). For each attribute, a random permutation involving all the values of the attribute was used. The Pearson correlation matrix obtained for the synthetic attributes was as follows:

$$\mathbf{P}_{k=1000} = \begin{pmatrix} 1 & -0.033 & 0.038 & -0.004 & -0.036 \\ & 1 & 0.015 & 0.049 & -0.033 \\ & & 1 & 0.034 & -0.017 \\ & & & 1 & 0.037 \\ & & & & 1 \end{pmatrix}.$$

It can be seen that the synthetic attributes obtained with the maximum permutation are practically independent. Hence, all the correlation structure (and thus a lot of the analytical utility) of the original attributes has been sacrificed in exchange for confidentiality. In the case of numerical attributes Y_1, Y_2 and Y_3, this independence is confirmed by the scatterplots of Figure 6.

Comparing the utility preserved by an anonymization method with the utility preserved by another anonymization method using correlation matrices or scatterplots is not always easy, especially when there are many attributes. For that reason, we will use the more compact utility metrics U and UM described in Section 2.

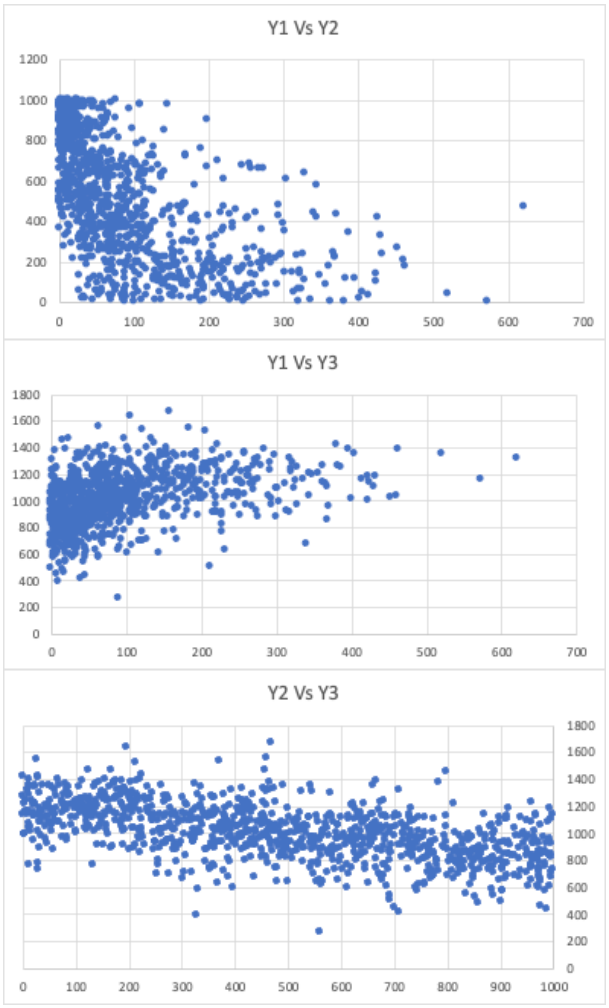


Fig. 5. Artificial data set. Pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 2 for $k = 250$ (high permutation).



Fig. 6. Artificial data set. Pairwise scatterplots of the synthetic numerical attributes Y_1, Y_2, and Y_3 obtained with Algorithm 2 for $k = 1000$ (maximum permutation).

4.2 Results with U, UM, and $CM3$

In the case of Algorithm 3, a distinction is needed between quasi-identifier and non-quasi-identifier attributes; the former are jointly microaggregated by the algorithm. For the artificial data set, we took X_4 and X_5 as quasi-identifiers, and the rest as non-quasi-identifiers. For *Adult*, we took all attributes as quasi-identifiers except Hours-per-Week and the class attribute (Income), which were taken as non-quasi-identifiers. For *Census*, we took all attributes as quasi-identifiers except WSALVAL and ERNVAL, which were taken as non-quasi-identifiers.

We computed U by linearly regressing I on all attributes in each data set plus their two-way interactions (in order to allow the classifier to take correlation preservation into account when labeling records).

The figures in the tables mentioned in this section are the average over five runs of the algorithms.

The purpose of the artificial data set was to study the behavior of the proposed permutation-based algorithms for known population parameters. For this reason, we did not use synthpop on this data set (in addition, its small size would have caused synthpop to overfit it). Table 1 summarizes the results. Columns U and UM show the results for both metrics computed using Algorithm 1, Algorithm 2 with $k = 1$, 50, 100, 250, 500, and 1000, and Algorithm 3 with the same values of k. Column $CM3$ reports the confidentiality metric.

Columns U, UM, and $CM3$ of Table 2, resp. Table 3, show analogous results for *Adult*, resp. *Census*. Additionally, Tables 2 and 3 display the general utility and confidentiality metrics when the synthpop method is applied to the two real-world data sets.

Regarding utility, the figures in Table 1 for the artificial data set basically agree with what could be seen when comparing correlation matrices. With both general utility metrics, Algorithm 1 gives a utility between that of Algorithm 2 with $k = 50$ and with $k = 100$. Also, Algorithm 2 provides less utility as k grows. Algorithm 3 behaves very similarly to Algorithm 2, with a slightly lower utility for $k \leq 100$ (price paid for enforcing

joint microaggregation of the quasi-identifier attributes in order to achieve k-anonymity). For $k \geq 250$, the utility loss is higher for both Algorithms 2 and 3 and the k-anonymity price is not noticeable.

Regarding confidentiality, Table 1 shows that Algorithm 1 offers a level of $CM3$ confidentiality close to that of Algorithms 2 and 3 with $k = 50$.

For *Adult* (Table 2), synthpop provides the best utility based both on U and UM, but the worst confidentiality based on $CM3$; this may suggest that synthpop could be overfitting *Adult*. Our proposed methods have the advantage of allowing a controlled trade-off between utility and confidentiality, with confidentiality being prioritized as k increases. The situation is similar for *Census* (Table 3), except that Algorithm 1 beats synthpop both regarding UM utility and confidentiality.

All in all, synthpop offers the highest utility in most cases, but also provides the lowest confidentiality. Algorithm 1 offers utility that is very close to that of synthpop and provides better confidentiality. If higher confidentiality is desired, then Algorithms 2 or 3 with $k \geq 100$ are the right choice, at the expense of some utility loss. Although Algorithms 2 and 3 perform similarly in terms of utility and confidentiality (with a very slight utility advantage for Algorithm 2), Algorithm 3 has the advantage of achieving k-anonymity.

TABLE 1
Artificial data set. General utility metrics U and UM, and confidentiality metric $CM3$ for the proposed methods. U is based on propensity scores, whereas UM is based on covariance matrices. $CM3$ is based on canonical correlations. Values are bounded within $[0, 1]$. Lower values of U indicate higher utility. Higher values of UM indicate higher utility. Higher values of $CM3$ indicate higher confidentiality. The results correspond to the average of five runs.

Method	U	UM	$CM3$
Algorithm 1	0.00153621	0.99963197	0.46535679
Algorithm 2, $k = 1$	0.00109379	0.99997113	0.00000002
Algorithm 2, $k = 50$	0.00147800	0.99989857	0.48408073
Algorithm 2, $k = 100$	0.00225551	0.99930648	0.53753237
Algorithm 2, $k = 250$	0.00847751	0.98814777	0.65423529
Algorithm 2, $k = 1000$	0.05460976	0.63360468	0.97545976
Algorithm 3, $k = 1$	0.00110857	0.99990723	0.00000002
Algorithm 3, $k = 50$	0.00259849	0.99997500	0.46999510
Algorithm 3, $k = 100$	0.00266053	0.99965373	0.54214920
Algorithm 3, $k = 250$	0.00444821	0.99098253	0.67808630
Algorithm 3, $k = 1000$	0.05067025	0.66263375	0.98147624

TABLE 2
Adult data set. General utility metrics U and UM, and confidentiality metric $CM3$ for synthpop and the proposed methods. The results correspond to the average of five runs.

Method	U	UM	$CM3$
Synthpop	0.000199989	0.999902558	0.797535572
Algorithm 1	0.007129566	0.998596158	0.960459070
Algorithm 2, $k = 1$	0.020879682	0.996055364	0.983165823
Algorithm 2, $k = 50$	0.041967283	0.993026892	0.998250227
Algorithm 2, $k = 100$	0.043502418	0.993359679	0.997865830
Algorithm 2, $k = 250$	0.044041416	0.992999594	0.997805878
Algorithm 2, $k = 1000$	0.044339659	0.992308455	0.998516833
Algorithm 3, $k = 1$	0.020946206	0.995634062	0.983956477
Algorithm 3, $k = 50$	0.041774397	0.993500011	0.998079390
Algorithm 3, $k = 100$	0.042494886	0.992882056	0.998469916
Algorithm 3, $k = 250$	0.042494265	0.992414163	0.997901739
Algorithm 3, $k = 1000$	0.042900398	0.993309618	0.998530825

TABLE 3
Census data set. General utility metrics U and UM, and confidentiality metric $CM3$ for synthpop and the proposed methods. The results correspond to the average of five runs.

Method	U	UM	$CM3$
Synthpop	0.010051625	0.998745279	0.801924806
Algorithm 1	0.037299448	0.999983539	0.913472878
Algorithm 2, $k = 1$	0.102331585	0.977082526	0.781739833
Algorithm 2, $k = 50$	0.131320286	0.971613936	0.915486768
Algorithm 2, $k = 100$	0.134189995	0.971019689	0.924330056
Algorithm 2, $k = 250$	0.139306149	0.970141533	0.928046441
Algorithm 2, $k = 1000$	0.145426214	0.970637199	0.935189071
Algorithm 3, $k = 1$	0.120443159	0.986359689	0.849924501
Algorithm 3, $k = 50$	0.153133227	0.975905774	0.931187499
Algorithm 3, $k = 100$	0.153450959	0.9755788	0.933510557
Algorithm 3, $k = 250$	0.154779774	0.977711213	0.933829972
Algorithm 3, $k = 1000$	0.157350232	0.976561628	0.944952241

5 Discussion

The permutation paradigm has been demonstrated above to be useful to generate synthetic data in a *non-parametric* manner. This has several advantages:

- Our approach can be used for all types of data: continuous, ordinal, nominal, or binary. In this sense, it offers a flexibility analogous to that offered by swapping, shuffling or microaggregation in the case of masking. It is true that some parametric procedures, like those based on CART [36], offer such flexibility as well.
- With modeling-based approaches, one needs to model the entire original data set to obtain a statistical model of the original data from which the synthetic data can be sampled. The extent to which the resulting synthetic data preserve the characteristics and the structure of the original data entirely depends on the accuracy of the modeling of the original data. It is well known that constructing an accurate model of a large, complex data set is notoriously difficult. Any estimation errors end up degrading the quality of the generated synthetic data. By contrast, our permutation-based approach requires no modeling. We just need the marginal distributions of original attributes and we obtain synthetic attributes that replicate the relationships between original attributes exclusively based on ranks. This is much simpler than estimating joint or conditional distributions as in CART-based methods.

- Synthetic data generation methods in the literature not adhering to a particular privacy model (*e.g.*, non-DP or non-k-anonymous methods) follow the maximum utility approach. That is, their objective is to generate a synthetic data set that replicates all attribute relationships that were present in the original data set. This approach is problematic because it may entail a significant risk of disclosure. In particular, this problem is especially acute if using deep learning techniques to generate synthetic data, because such very powerful models are likely to overfit the original data. Mitigating overfitting with DP-noise is a possibility, but it is hard to interpret what specific utility-confidentiality trade-off is achieved by a certain ϵ value. Furthermore, as pointed out in [12], [31], releasing DP microdata with meaningful utility requires using high ϵ values, which makes it hard to understand how much privacy is being offered and how trustworthy the results obtained on the data are. With our approach, Algorithm 2 allows selecting a specific trade-off between utility and confidentiality by choosing k in the bounded range $\{1, \ldots, n\}$ (highest utility and lowest confidentiality for $k = 1$ and vice versa for $k = n$), and, on top of attribute confidentiality, Algorithm 3 adds k-anonymity protection against reidentification.

6 Conclusions and Future Research

We have used the permutation paradigm to propose an approach to synthetic data generation based on ranks. Specifically, we have presented three different methods and we have analyzed the utility and the confidentiality they afford. Our third method satisfies the k-anonymity privacy model, but in fact provides stronger protection against attribute disclosure.

Our methods only require the *marginal* distributions of attributes and yield synthetic attributes that replicate the relationships between the original attributes exclusively based on ranks. This rank-based proposal is especially attractive for non-normal or multi-type data.

Acknowledgments and Disclaimer

This work has received partial support from the European Commission (project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2021 SGR 00115), MCIN/AEI/ 10.13039/501100011033 and "ERDF A way of making Europe" (grant PID2021-123637NB-I00 "CURLING"), and INCIBE and European Union NextGenerationEU/PRTR (project "HERMES" and INCIBE-URV Cybersecurity Chair).

References

[1] Awan, J., and Cai, Z. 2022. One step to efficient synthetic data. arXiv math preprint. https://arxiv.org/abs/2006.02397

[2] Becker, B., and Kohavi, R. 1996. *Adult Income Data Set*. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20

[3] Brand, R., Domingo-Ferrer, J., and Mateo-Sanz, J. M. 2002. 2002. *Reference Data Sets to Test and Compare SDC Methods for Protection of Numerical Microdata*. IST-2000-25069 CASC Project. https://research.cbs.nl/casc/CASCtestsets.htm

[4] Burnett-Isaacs, K., Girard, C., Sallier, K., Raab, G., Ramsden, A., Slokom, M., Task, C., Salamh, M., Wardlaw, A., Ghashim, E., Ostic, R., Belanger, P. *et al.*. 2022. *Synthetic Data for Official Statistics: A Starter Guide*. United Nations Economic Commission for Europe, Geneva.

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

[6] Cario, M. C., and Nelson, B. L. 1997. *Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix*. Technical Report, Dept. of Industrial Engineering and Management Sciences, Northwestern University.

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16:321-357.

[8] DePersio, M., Lemmons, M., Ramanayake, K. A., Tsay, J., and Zayatz, L. 2012. n-Cycle swapping for the American Community Survey. In: *Privacy in Statistical Databases (PSD 2012)*. Springer, pp. 143-164.

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

[10] Domingo-Ferrer, J., and Muralidhar, K. 2016. New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. *Information Sciences* 337-338:11-24.

[11] Domingo-Ferrer, J., Muralidhar, K., and Bras-Amorós, M. 2021. General confidentiality and utility metrics for privacy-preserving data publishing based on the permutation model. *IEEE Transactions on Dependable and Secure Computing* 18(5):2506-2517.

[12] Domingo-Ferrer, J., Sánchez, D., and Blanco-Justicia, A. 2021. The limits of differential privacy (and its misuse iin data release and machine learning). *Communications of the ACM*, 84(7):33-35.

[13] Domingo-Ferrer, J., Sánchez, D., and Rufian-Torrell, G. 2013. Anonymization of nominal data based on semantic marginality. *Information Sciences* 242:35-48.

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

[15] Drechsler, J. 2011. *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*. Springer.

[16] Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. 2019. Neural spline flows. In *NeurIPS 2019*, 32.

[17] Hall, R., Rinaldo, A., and Wasserman, L. 2013. Differential privacy for functions and functional data. *Journal of Machine Learning Research* 14:703-727.

[18] Ho, J., Jain, A., and Abbeel, P. 2020. Denoising diffusion probabilistic models. In *NeurIPS 2020*, 33:6840-6851.

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

[20] Jordon, J., Yoon, J.,and van der Schaar, M. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations (ICLR 2019)*.

[21] Karwa, V., and Slavković, A. 2013. Conditional inference given partial information in contingency tables using Markov bases. *Wiley Interdisciplinary Reviews: Computational Statistics* 5(3):207-218.

[22] Kingma, D. P., and Dhariwal, P. 2018. Glow: generative flow with invertible 1x1 convolutions. In *NeurIPS'2018*, 31.

[23] Kruskal, W. H. 1958. Ordinal measures of association. *Journal of the American Statistical Association* 53(284):814-861.

[24] A. M. Law. 2015. *Simulation Modeling and Analysis, Fifth Edition*. McGraw Hill.

[25] Liew, C. K., Choi, U. J., and Liew, C. J. 1985. A data distortion by probability distribution. *ACM Transactions on Database Systems* 10(3):395-411.

[26] Liu, F. 2021. Model-based differentially private data synthesis and statistical inference in multiply synthetic differentially private data. arXiv preprint arXiv:1606.08052. https://arxiv.org/pdf/1606.08052.pdf

[27] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. 2008. Privacy: theory meets practice on the map. In *2008 IEEE 24th International Conference on Data Engineering*. IEEE, pp. 227-286.
[28] Mateo-Sanz, J. M., Martínez-Ballesté, A., and Domingo-Ferrer, J. 2004. Fast generation of accurate synthetic microdata. In *Privacy in Statistical Databases (PSD 2004)*. Springer, pp. 298-306.
[29] McCaa, R., Muralidhar, K., Sarathy, R., Comerford, M., and Esteve-Palòs, A. 2014. Controlled shuffling, statistical confidentiality and microdata utility: a successful experiment with a 10% household sample of the 2011 Population Census of Ireland for the IPUMS-International Database. In *Privacy in Statistical Databases (PSD 2014)*. Springer, pp. 326-337.
[30] McClure, D., and Reiter, J. P. 2012. Differential privacy and statistical disclosure risk measures: an investigation with binary synthetic data. *Transactions on Data Privacy* 5(3):535-552.
[31] Mervis J. 2020. Researchers finally get access to data on Facebook's role in political disclosure. *Science*, Feb. 13. https://bit.ly/3ynS7Kj
[32] Muralidhar, K., and Sarathy, R. 2003. A theoretical basis for perturbation methods. *Statistics and Computing* 13(4):329-335.
[33] Muralidhar, K., and Sarathy, R. 2006. Data shuffling – A new masking approach for numerical data. *Management Science* 52(2):658-670.
[34] Nowok B, Raab GM, Dibben C. 2016. synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software* 74(11):1-26.
[35] Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. 2003. Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* 19(1):1-16.
[36] Reiter, J. P. 2005. Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics* 21(3):441-462.
[37] Rubin, D. B. 1993. Statistical disclosure limitation. *Journal of Official Statistics* 9(2):461-468.
[38] Sarathy, R., Muralidhar, K., and Parsa, R. 2003. Perturbing non-normal confidential attributes: the copula approach. *Management Science*, 48(12):1517-1644.
[39] Snoke, J., Raab, G. M., Nowok, B., Dibben, C., and Slavkovic, A. 2018. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3):663-688.
[40] Song, J., Meng, C., and Ermon, S. 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502. https://arxiv.org/abs/2010.02502
[41] Tai, B.C., Li, S.C., and Huang, Y. 2017. K-Aggregation: Improving Accuracy for Differential Privacy Synthetic Dataset by Utilizing K-Anonymity Algorithm. IEEE 31st International Conference on Advanced Information Networking and Applications (AINA), pp. 772-779.
[42] Triastcyn, A., and Faltings, B. 2018. Generating differentially private data sets using GANs. arXiv preprint arXiv:1803.03148. https://arxiv.org/abs/1803.03148
[43] Woo, M.-J., Reiter, J. P., Oganian, A., and Karr, A. F. 2009. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1:111-124.
[44] Xu, C., Ren, J., Zhang, D., Zhang, Y., Qin, Z., and Ren, K. 2019. GANobfuscator: Mitigating information leakage under GAN via differential privacy. *IEEE Transactions on Information Forensics and Security*, 14(9):2358-2371.
[45] Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. 2019. Modeling tabular data using conditional GAN. In *NeurIPS 2019*, 32:7335-45.
[46] Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. 2017. PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems* 42(4):1-41.



Josep Domingo-Ferrer (Fellow, IEEE) is a Distinguished Professor of Computer Science and an ICREA-Acadèmia Researcher at Universitat Rovira i Virgili, Tarragona, Catalonia, where he leads CYBERCAT. He received BSc-MSc and PhD degrees in Computer Science from the Autonomous University of Barcelona. He also earned a BSc-MSc degree in Mathematics from U.N.E.D. (Madrid) and an MA in Philosophy from the University of Paris-Nanterre. His research interests are in data privacy, data security and ethics-by-design in information technology. More information on him can be found at http://crises-deim.urv.cat/jdomingo



Krishnamurty Muralidhar is a Professor of Marketing & Supply Chain Management and the Research Director for the Center for Business of Healthcare at the University of Oklahoma. He received a BSc from the University of Madras, India, an MBA from Sam Houston State University and a PhD from Texas A&M University. His main research interests are in data privacy.



Sergio Martínez is a senior researcher at Universitat Rovira i virgili (URV) in Tarragona, Catalonia. He earned an MSc in Intelligent Systems in 2010 and a PhD in Computer Science in 2013, both awarded by URV. His research interests include artificial intelligence, semantic data management, data security, and privacy preservation.