ϵ-Differential Privacy for Microdata Releases Does Not Guarantee Confidentiality (Let Alone Utility)

Krishnamurty Muralidhar[1], Josep Domingo-Ferrer[2](✉), and Sergio Martínez[2]

[1] Department of Marketing and Supply Chain Management,
University of Oklahoma, 307 West Brooks, Adams Hall Room 10,
Norman, OK 73019, USA
krishm@ou.edu

[2] Department of Computer Engineering and Mathematics,
Universitat Rovira i Virgili, CYBERCAT-Center for Cybersecurity Research
of Catalonia UNESCO Chair in Data Privacy,
Av. Països Catalans 26, 43007 Tarragona, Catalonia
{josep.domingo,sergio.martinezl}@urv.cat

Abstract. Differential privacy (DP) is a privacy model that was designed for interactive queries to databases. Its use has then been extended to other data release formats, including microdata. In this paper we show that setting a certain ϵ in DP does not determine the confidentiality offered by DP microdata, let alone their utility. Confidentiality refers to the difficulty of correctly matching original and anonymized data, and utility refers to anonymized data preserving the correlation structure of original data. Specifically, we present two methods for generating ϵ-differentially private microdata. One of them creates DP synthetic microdata from noise-added covariances. The other relies on adding noise to the cumulative distribution function. We present empirical work that compares the two new methods with DP microdata generation via prior microaggregation. The comparison is in terms of several confidentiality and utility metrics. Our experimental results indicate that different methods to enforce ϵ-DP lead to very different utility and confidentiality levels. Both confidentiality and utility seem rather dependent on the amount of permutation performed by the particular SDC method used to enforce DP. Thus suggests that DP is not a good privacy model for microdata releases.

Keywords: Anonymized microdata · Differential privacy · Synthetic data · Confidentiality · Analytical utility

1 Introduction

Traditional anonymization by national statistical institutes consists of applying a statistical disclosure control (SDC) method with a heuristic choice of parameters

J. Domingo-Ferrer and K. Muralidhar (Eds.): PSD 2020, LNCS 12276, pp. 21–31, 2020.
https://doi.org/10.1007/978-3-030-57521-2_2

and then assessing the disclosure risk and the analytical utility of the anonymized data. If the risk is deemed too high, the SDC method is run again with more stringent parameters, which is likely to reduce the risk and the utility as well.

Privacy models, originated in the computer science community, take a different view of anonymization. A privacy model is an *ex ante* parameterized condition that is meant to guarantee a pre-specified level of disclosure protection—that is, confidentiality—regardless of the impact on utility. If the utility loss is deemed too high, then the privacy model parameter must be made less strict. Privacy models are enforced using SDC methods whose parameters depend on the privacy model parameter. The earliest privacy model instance was k-anonymity [10], and the most talked about privacy model these days is differential privacy (DP, [4]). Privacy models are usually enforced by using one or more SDC methods: in the case of k-anonymity, one uses generalization, local suppression or microaggregation. In the case of DP, there are several options, the most usual being Laplace noise addition.

The initial formulation of DP was for the interactive setting. A randomized query function κ (that returns the query answer plus some noise) satisfies ϵ-DP if for all data sets D_1 and D_2 that differ in one record and all $S \subset Range(\kappa)$, it holds that $\Pr(\kappa(D_1) \in S) \leq \exp(\epsilon) \times \Pr(\kappa(D_2) \in S)$. In plain English, the presence or absence of any single record must not be noticeable from the query answers, up to an exponential factor ϵ (called the privacy budget). The smaller ϵ, the higher the protection. The most usual SDC method employed to enforce differential privacy is Laplace noise addition. The amount of noise depends on ϵ (the smaller ϵ, the more noise is needed) and, for fixed ϵ, it increases with the global sensitivity of the query (defined as the maximum variation of the query output when one record in the data set is changed, added or suppressed).

Differential privacy offers a neat privacy guarantee for interactive queries, at least for small values of ϵ. Unlike k-anonymity, its privacy guarantee does not require any assumptions on the intruder's background knowledge. Very soon, researchers proposed extensions of DP for the non-interactive setting, that is, to produce DP microdata sets that could be used for any analysis, rather than for a specific query. Based on that, Google, Apple and Facebook are currently using DP to anonymize microdata collection from their users, although in most cases with ϵ values much larger than 1 [6] (which is against the recommendations of [5]).

Unfortunately, as noted in [9], generating DP microdata is a very challenging task. A DP microdata set can be viewed as a collection of answers to identity queries, where an identity query is about the content of a specific record (*e.g.* tell me the content of the i-th record in the data set). Obviously, the sensitivity of an identity query is very high: if one record is changed the value of each attribute in the record can vary over the entire attribute domain. This means that a lot of noise is likely to be needed to produce DP microdata, which will result in poor utility. This should not be surprising, because by design, DP attempts to make the presence or absence of any single original record undetectable in the DP output, in this case, the DP microdata set.

The usual approach to obtain DP microdata is based on histogram queries [12,13]. In [11], a method to generate DP microdata that uses a prior microaggregation step was proposed. Microaggregation replaces groups of similar records by their average record. Since the average record is less sensitive than individual original records, if one takes the microaggregation output as the input to DP, the amount of Laplace noise required to enforce a certain ϵ is smaller than if taking the original data set as input. This yields DP microdata with higher utility than competing approaches.

1.1 Contribution and Plan of This Paper

Our aim in this paper is to demonstrate that a certain privacy budget ϵ can result in very different levels of confidentiality and utility. In fact, we show that the achieved confidentiality and utility depend on the particular SDC methods used to enforce DP.

Specifically, we present two new methods for generating DP microdata. One of them creates DP synthetic microdata from noise-added covariances. The other relies on adding noise to the cumulative distribution function (CDF). We present empirical work that compares the two new methods with the microaggregation-based method [11]. The comparison is in terms of several confidentiality and utility metrics. It becomes apparent that different methods to enforce ϵ-DP lead to very different utility and confidentiality levels.

Section 2 describes the synthetic data method. Section 3 describes the CDF-based method. Empirical work comparing the two new methods among them and with the microaggregation-based method is reported in Sect. 4. Conclusions and future research issues are gathered in Sect. 5.

2 A Method for Generating Synthetic DP Microdata

In this method, a DP synthetic microdata set is generated based on the original data. The approach is to add Laplace noise to: i) the sum of each attribute; ii) the sum of squared values of each attribute; and iii) the sum of the product of each pair of attributes. This allows obtaining DP versions of the attribute means and covariances. Finally, the synthetic microdata are obtained by sampling a multivariate normal distribute with parameters the DP mean vector and the DP covariance matrix. Therefore, *the synthetic data thus obtained are DP by construction.*

If the original data set has m attributes, there are m sums of attribute values, m sums of squared attribute values and $m(m-1)/2$ sums of products of pairs of attributes. Hence, the privacy budget ϵ must be divided among the total $2m + m(m-1)/2$ sums. Let $\epsilon^* = \epsilon/(2m + m(m-1)/2)$.

Let x_{ij}, for $i = 1, \dots, n$ and $j = 1, \dots, m$, represent the value of the j-th attribute in the i-th record of the original data set. Let μ_j and σ_{jj} denote, respectively, the mean and the variance of the j-th attribute, and let σ_{jk}, for $j \neq k$, represent the covariance between the j-th and the k-th attributes. On the other hand, let Δf_j represent the global sensitivity of the j-th attribute. Then Algorithm 1 formalizes the above-sketched method to generate DP synthetic microdata.

Algorithm 1. Method 1: DP synthetic microdata generation

```
Input: Original data set $\{x_{ij} : i = 1, \dots, n; j = 1, \dots, m\}$
Output: DP data set $\{x^*_{ij} : i = 1, \dots, n; j = 1, \dots, m\}$
for $j = 1$ to $m$ do                         /* DP-perturb means and variances */
    $R_j = \sum_{i=1}^{n} x_{ij} + \text{Laplace}\left(0, \frac{\Delta f_j}{\epsilon^*}\right)$;
    $S_j = \sum_{i=1}^{n} x_{ij}^2 + \text{Laplace}\left(0, \frac{\Delta f_j^2}{\epsilon^*}\right)$;
    $\mu^*_j = R_j / n$;
    $\sigma^*_{jj} = \frac{S_j - \frac{R_j^2}{n}}{n-1}$;
for $j = 1$ to $m$ do                                 /* DP-perturb covariances */
    for $k = j + 1$ to $m$ do
        $T_{jk} = \sum_{i=1}^{n} (x_{ij} x_{ik}) + \text{Laplace}\left(0, \frac{\Delta f_j \times \Delta f_k}{\epsilon *}\right)$;
        $\sigma^*_{jk} = \frac{T_{jk} - \frac{R_j R_k}{n}}{n-1}$;
        $\sigma^*_{kj} = \sigma^*_{jk}$;
for $i = 1$ to $n$ do              /* Draw DP synthetic data from DP-normal */
    $(x^*_{i1}, \dots, x^*_{im}) = \text{Sample}(N(\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*))$, where $\boldsymbol{\mu}^* = (\mu^*_1, \dots, \mu^*_m)$ and
    $\boldsymbol{\sigma}^* = [\sigma^*_{ij}]_{i,j=1,\dots m}$.
```

Note 1. This method is problematic unless the number of attributes is really small. Indeed, for fixed ϵ the Laplace perturbation added to the variances and covariances quadratically grows with m, because this perturbation has privacy budget $\epsilon/(2m + m(m-1)/2)$. Thus, m does not need to be very large to risk getting a perturbed covariance matrix $\boldsymbol{\sigma}^*$ that is no longer positive definite, and hence not valid as a covariance matrix.

3 A CDF-Based Method to Obtain DP Microdata

This method follows the inspiration of [7] in that it anonymizes by sampling a distribution adjusted to the original data. Yet, unlike [7], we DP-perturb the distribution. In this way, rather than drawing from a multivariate normal distribution with DP-perturbed parameters as in Algorithm 1, we obtain DP microdata by: i) for each attribute, obtaining DP attribute values by sampling from a *univariate* normal distribution with DP-perturbed mean and variance; ii) replacing each original attribute value with a DP-attribute value whose rank is a DP-perturbed version of the rank of the original attribute value. The DP-perturbation of attribute values ensures that the original attribute values are unnoticeable in the DP microdata, whereas the DP-perturbation of ranks ensures that the rank correlation of attribute values within each record is altered enough for the resulting multivariate data set to be DP.

If we normalize ranks by dividing them by n, adding or suppressing one record will at most change the CDF (normalized rank) of any other record by $1/n$, and hence the global sensitivity of the CDF is $1/n$. Since records are assumed independent of each other, there is no sequential composition among records in the DP sense, and therefore the privacy budget ϵ does not need to be divided among the number of records. If there are m attributes, ϵ must just be divided among the m sums of attribute values, the m sums of squared attribute values and the m empirical CDFs of the attributes. This yields a budget $\epsilon^* = \epsilon/3m$ for the required Laplace perturbations.

Algorithm 2 formalizes the CDF-based method to generate DP microdata. Note that this approach is not synthetic, because each record in the DP data set results from a specific record in the original data set. For each attribute j, each original attribute value x_{ij} is replaced by a DP attribute value whose rank is DP-perturbed version of the rank of x_{ij}.

Algorithm 2. METHOD 2: CDF-BASED DP MICRODATA GENERATION

```
Input: Original data set $\{x_{ij} : i = 1, \dots, n; j = 1, \dots, m\}$
Output: DP data set $\{x^*_{ij} : i = 1, \dots, n; j = 1, \dots, m\}$
for $j = 1$ to $m$ do                    /* DP-perturb means and variances */
    $R_j = \sum_{i=1}^{n} x_{ij} + \text{Laplace}\left(0, \frac{\Delta f_j}{\epsilon^*}\right)$;
    $S_j = \sum_{i=1}^{n} x_{ij}^2 + \text{Laplace}\left(0, \frac{\Delta f_j^2}{\epsilon^*}\right)$;
    $\mu_j^* = R_j/n$;
    $\sigma_{jj}^* = \frac{S_j - \frac{R_j^2}{n}}{n-1}$;
for $j = 1$ to $m$ do
    for $i = 1$ to $n$ do
        $y_i = \text{Sample}(N(\mu_j^*, \sigma_{jj}^*))$;        /* Generate DP attribute values */
        $c_i = \left(\frac{\text{Rank}(x_{ij})}{n}\right) + \text{Laplace}\left(0, \frac{1}{n\epsilon^*}\right)$;     /* Convert the rank of the
        original attribute value to real and DP-perturb it */
    for $i = 1$ to $n$ do          /* Replace original attribute values by DP
    attribute values with DP-perturbed ranks */
        $x_{ij}^* = y_{[\text{Rank}(c_i)]}$, where $y_{[k]}$ stands for the value in $\{y_1, \dots, y_n\}$ with
        rank $k$.
```

The following holds.

Proposition 1. *If the number of attributes is m is more than 3, for a given privacy budget ϵ the method of Algorithm 2 perturbs means and variances less than the method of Algorithm 1.*

Proof. In both algorithms, the perturbations of means and variances are directly proportional to the perturbations used to obtain R_j and S_j. On the other hand, the latter perturbations are inversely proportional to ϵ^*. In Algorithm 1 we have $\epsilon^* = \epsilon/(2m + m(m-1)/2)$, whereas in Algorithm 2 we have $\epsilon^* = \epsilon/3m$. Now $(2m + m(m-1)/2) > 3m$ if and only if $m > 3$.

Note that Proposition 1 does not necessarily imply that for $m > 3$ the utility of the output DP microdata is better in Algorithm 2 than in Algorithm 1, because the ways in which perturbed means and variances are used in both algorithms differ.

4 Empirical Work

We implemented the two proposed methods and we measured the analytical utility and the confidentiality they provide. Note that, although DP is a privacy model specifying an *ex ante* privacy condition with the ϵ budget, absolute unnoticeability of any particular record only holds when $\epsilon = 0$. For any other value of ϵ it makes sense to measure how protected against disclosure are the data, that is, what is the confidentiality level being achieved.

Further, to compare the two proposed methods against the state of the art, we included in the comparison the microaggregation-based DP microdata generation method [11].

4.1 Utility Metrics

We considered two metrics for generic analytical utility, which do not require assumptions on specific data uses.

The first one is the sum of squared errors SSE, defined as the sum of squares of attribute distances between records in the original data set and their versions in the DP data set. That is,

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - x^*_{ij})^2. \tag{1}$$

We took the squared Euclidean distance between x_{ij} and x^*_{ij} because our in experiments all attributes were numerical. For a version of SSE that works also with categorical data, see [11]. On the other hand, SSE needs to know which DP attribute value x^*_{ij} corresponds to each original attribute value x_{ij}. For that reason, SSE cannot be used to measure the utility of Method 1, because in that method the DP data are synthetic, which means that no correspondence can be established between original attribute values and DP attribute values.

The second utility metric is the one proposed in [2]:

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

In Expression (2), $\mathbf{X}$ is the original microdata set, $\mathbf{Y}$ is the DP microdata set, $\hat{\lambda}_j^X$ are the eigenvalues of the covariance matrix $\mathbf{C}_{XX}$ of $\mathbf{X}$ scaled so that they add to 1, and $\hat{\lambda}_j^{Y|X}$ are scaled versions of

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

where $\mathbf{C}_{YY}$ is the covariance matrix of $\mathbf{Y}$ and $\mathbf{v}_j^X$ is the j-th eigenvector of $\mathbf{C}_{XX}$.

The rationale of UM is as follows. Each eigenvalue $\hat{\lambda}_j^X$ represents the proportion of the variance of the attributes in $\mathbf{X}$ explained by the corresponding eigenvector $\mathbf{v}_j^X$. On the other hand each $\lambda_j^{Y|X}$ represents the proportion of the variance of the attributes in $\mathbf{Y}$ explained by $\mathbf{v}_j^X$. Then we have:

– The highest level of utility ($UM(\mathbf{X}, \mathbf{Y}) = 1$) occurs when $\hat{\lambda}_j^X = \hat{\lambda}_j^{Y|X}$ for $j = 1, \ldots, m$, which occurs when $\mathbf{C}_{XX} = \mathbf{C}_{YY}$.
– The lowest level of 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 (which are $1/m$).

Note that an advantage of UM over SSE is that the former also applies for synthetic data, and hence for Method 1. However, both metrics view utility as the preservation of variability, more precisely as the preservation of the correlation structure of the original data set.

4.2 Confidentiality Metrics

We used four confidentiality metrics. First, the share of records in the original data set that can be correctly matched from the DP data set, that is, the proportion of correct record linkages

$$RL = \frac{\sum_{\mathbf{x}_i \in \mathbf{X}} \Pr(\mathbf{x}_i^*)}{n}, \tag{3}$$

where $\Pr(\mathbf{x}_i^*)$ is the correct record linkage probability for the i-th DP record $\mathbf{x}_i^*$. If the original record $\mathbf{x}_i$ from which $\mathbf{x}_i^*$ originates is not at minimum distance from $\mathbf{x}_i^*$, then $\Pr(\mathbf{x}_i^*) = 0$; if $\mathbf{x}_i$ is at minimum distance, then $\Pr(\mathbf{x}_i^*) = 1/M_i$, where M_i is the number of original records at minimum distance from $\mathbf{x}_i^*$.

The three other confidentiality metrics $CM1$, $CM2$ and $CM3$ are those proposed in [2], based on canonical correlations. We have:

$$CM1(\mathbf{X}, \mathbf{Y}) = 1 - \rho_1^2, \tag{4}$$

where ρ_1^2 is the largest canonical correlation between the *ranks* of attributes in $\mathbf{X}$ and $\mathbf{Y}$. The rationale is that:

– Top confidentiality ($CM1(\mathbf{X}, \mathbf{Y}) = 1$) is reached when the ranks of the attributes in $\mathbf{X}$ are independent of the ranks of attributes in $\mathbf{Y}$, in which case anonymization can be viewed as a random permutation.

- Zero confidentiality ($CM1(\mathbf{X},\mathbf{Y}) = 0$) is achieved then the ranks are the same for *at least* one original attribute X^j and one DP attribute X^j. Note that this notion of confidentiality is quite strict: leaving a single attribute unprotected brings the confidentiality metric down to zero.

The next confidentiality metric is similar $CM1$ but it considers all canonical correlations:

$$CM2(\mathbf{X},\mathbf{Y}) = \prod_{i=1}^{m}(1-\rho_i^2)\left[= e^{-I(\mathbf{X};\mathbf{Y})}\right]. \tag{5}$$

The second equality between brackets in Expression (5) can only be guaranteed if the collated data sets $(\mathbf{X},\mathbf{Y})$ follow an elliptically symmetrical distribution (a generalization of the multivariate Gaussian), in which case Expression (5) can be rewritten in terms of the mutual information $I(\mathbf{X};\mathbf{Y})$ between the original and the DP data sets.

Regardless of the distributional assumptions, $CM2(\mathbf{X},\mathbf{Y})$ can be computed from the canonical correlations and the following holds:

- Top confidentiality $CM2(\mathbf{X},\mathbf{Y}) = 1$ is reached when the anonymized data set and the original data sets tell nothing about each other, which is the same as saying that mutual information between them is $I(\mathbf{X};\mathbf{Y}) = 0$.
- Zero confidentiality $CM2(\mathbf{X},\mathbf{Y}) = 0$ occurs if at least one of the canonical correlations is 1. This occurs if at least one original attribute is disclosed when releasing $\mathbf{Y}$. Since ρ_1 is the largest correlation, this means that we have $CM2(\mathbf{X},\mathbf{Y}) = 0$ if and only if $\rho_1 = 1$, in which case we also have that the metric of Expression (4) is $CM1(\mathbf{X},\mathbf{Y}) = 0$.

Note that RL, $CM1$ and $CM2$ cannot be applied to the DP synthetic data produced by Method 1, because the three metrics need to know the mapping between original and DP records. The last metric $CM3$ that we use is mapping-free and is intended for synthetic data (yet it should not be used when the mapping between original and DP records is known). Specifically, $CM3$ is derived from $CM2$ as follows:

$$CM3(\mathbf{X},\mathbf{Y}) = \min_{1\le j\le m} CM2(\mathbf{X}^{-j},\mathbf{Y}^{-j}). \tag{6}$$

where $\mathbf{X}^{-j}$, resp. $\mathbf{Y}^{-j}$, is obtained from $\mathbf{X}$, resp. $\mathbf{Y}$, by sorting $\mathbf{X}$, resp. $\mathbf{Y}$, by its j-th attribute and suppressing the values of this attribute in the sorted data set.

The common principle of RL, $CM1$, $CM2$ and $CM3$ is to view confidentiality as permutation. The farther the anonymized values from the original values in value (for RL) or in rank (for the other metrics), the higher the confidentiality.

4.3 Results for Data Sets with Two Attributes

We considered five data sets with two numerical attributes X_1, X_2 and 10,000 records. In each data set, X_1 was drawn from a $N(50, 10)$ distribution and X_2

was also drawn from a $N(50, 10)$ distribution but in such a way that the expected correlation between X_1 and X_2 was 0.5.

For each data set, we ran the microaggregation-based DP microdata generation method of [11] with $\epsilon = 1$ (higher values are not recommended in [5]) and microaggregation group sizes $k = 250$, 500, 1000, 2000 and 3000. Since this method is not synthetic, for each resulting DP microdata set we computed utility metrics SSE and UM, and confidentiality metrics RL, $CM1$ and $CM2$. In Table 1 we report the values of those metrics for each value k averaged over the five data sets.

We then ran Method 1 for each data set with $\epsilon = 1$. Since it is a synthetic method, we computed utility metric UM and confidentiality metric $CM3$. In Table 1 we display those metrics averaged over the five data sets.

Finally, we ran Method 2 for each data set with $\epsilon = 1$. Since this method is not synthetic, we computed the same metrics as for the microaggregation-based method. Table 1 reports the averages for the five data sets.

Table 1. Empirical comparison of microaggregation-based DP generation, Method 1 and Method 2. In all cases $\epsilon = 1$ and all results are averages over five original data sets with the same distribution. "Micro*" denotes microaggregation-based DP microdata generation with $k = *$.

	SSE	UM	RL	CM1	CM2	CM3
Micro250	26833.22	0.647230517	0.00028	0.639798854	0.57703286	N/A
Micro500	3446.60	0.972413957	0.0008	0.226067842	0.112192366	N/A
Micro1000	2160.91	0.984149182	0.00096	0.164139854	0.057491616	N/A
Micro2000	2855.62	0.959011191	0.0005	0.214820038	0.10679245	N/A
Micro3000	3980.19	0.502650927	0.0003	0.197760886	0.197698071	N/A
Method1	N/A	0.992537652	N/A	N/A	N/A	0,935883394
Method2	49.74	0.981339047	0.1212	0.000492087	7.43603E-07	N/A

One thing that stands out in Table 1 is that utility metrics SSE and UM are consistent with each other. Higher values of SSE translate into lower values of UM, meaning less utility. Also, lower values of SSE result in higher values for the UM, meaning more utility. Thus, they capture the same utility notion and it is enough for us to consider one utility metric in what follows. We choose UM because it can be computed both for non-synthetic and synthetic data.

In terms of UM, we see in Table 1 that Method 1 achieves the highest utility, while offering a confidentiality metric $CM3$ that is also high, being close to 1. Thus, Method 1 seems the best performer.

Method 2 also offers high utility UM, but extremely low confidentiality in terms of $CM1$ and $CM2$. The DP data it produces turn out to be very similar to the original data.

The microaggregation-based DP microdata generation method can be seen to offer intermediate performance regarding the trade-off between utility and confidentiality. Whatever the choice of k, it achieves better confidentiality metrics $CM1$ and $CM2$ than Method 2, but its utility UM only beats Method 2 for $k = 1000$. Thus, microaggregation-based DP generation for $k = 1000$ is the second best performer.

The microaggregation-based DP method offers poorer utility for extreme values of k. The explanation is that for smaller k (250, 500) the prior microaggregation step does not reduce the sensitivity of the data as much as $k = 1000$ and hence still needs sustantial Laplace noise to attain DP with $\epsilon = 1$. On the other hand, for large $k = 2000, 3000$, averaging over such large groups causes a lot of information loss.

On the other hand, we can see that setting the same $\epsilon = 1$ for all methods can lead to very different confidentiality and utility levels.

4.4 Results for Data Sets with 10 Attributes

To check what is said in Note 1 and Proposition 1, we also tested Methods 1 and 2 for data sets with $m = 10$ attributes. We generated five data sets with normally distributed attributes and we took $\epsilon = 1$ as above. We kept running Method 1 for the five data sets until we got positive definite DP covariance matrices. The results were:

– As expected, the average utility achieved by Method 1 was extremely low, namely $UM = 0.00733781$. In contrast, the average confidentiality was high, $CM3 = 0.99752121$, even higher than for the two-attribute data sets.
– Method 2 yielded an extremely high average utility $UM = 0.99999618$. In contrast, confidentiality was as small as in the two-attribute case, with $CM1 = 0.00076754$ and $CM2$ =2.3331E-13.

5 Conclusions

We have compared three methods for generating DP microdata, two of them new. The three of them leverage different principles to generate DP microdata with $\epsilon = 1$. However, the confidentiality and utility levels they achieve for that value of ϵ are extremely different. Hence, setting a certain value of ϵ does *not* guarantee a certain level of confidentiality, let alone utility. The actual confidentiality and utility offered depend on the specific method used to enforce ϵ-DP. Our results complement those obtained in [8] for ϵ-DP synthetic data. In that paper, DP-synthetic data were generated with a single method but using several values of ϵ; it turned out that ϵ determined neither the protection against disclosure nor the utility of the synthetic data.

In our experiments, the methods that result in higher confidentiality seem to be those that operate a stronger permutation in terms of the permutation model of SDC [1]. Specifically Method 1, being synthetic, can be viewed as a random permutation of ranks, whereas Method 2 yields ranks for masked data that are very close to the ones of original data; this would explain the high confidentiality offered by Method 1 and the low confidentiality of Method 2.

In conclusion, the fact that parameter ϵ does not give any specific confidentiality guarantee for microdata releases suggests that DP should not be used to anonymize microdata. This adds to the arguments given in [3] in that sense.

Acknowledgments. Partial support to this work has been received from the European Commission (project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prize to J. Domingo-Ferrer and grant 2017 SGR 705), and from the Spanish Government (project RTI2018-095094-B-C21 "Consent"). The second and third authors are with the UNESCO Chair in Data Privacy, but the views in this paper are their own and are not necessarily shared by UNESCO.

References

1. Domingo-Ferrer, J., Muralidhar, K.: New directions in anonymization: permutation paradigm, verifiability by subjects and intruders, transparency to users. Inf. Sci. **337–338**, 11–24 (2016)
2. Domingo-Ferrer, J., Muralidhar, K., Bras-Amorós, M.: General confidentiality and utility metrics for privacy-preserving data publishing based on the permutation model. IEEE Trans. Dependable Secure Comput. (2020). To appear
3. Domingo-Ferrer, J., Sánchez, D., Blanco-Justicia, A.: The limits of differential privacy (and its misuse in data release and machine learning). Commun. ACM. To appear
4. Dwork, C.: Differential privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006). https://doi.org/10.1007/11787006_1
5. Dwork, C.: A firm foundation for private data analysis. Commun. ACM **54**(1), 86–95 (2011)
6. Greenberg, A.: How one of Apple's key privacy safeguards falls short. Wired, 15 September 2017. https://www.wired.com/story/apple-differential-privacy-shortcomings/
7. Liew, C.K., Choi, U.J., Liew, C.J.: A data distortion by probability distribution. ACM Trans. Database Syst. **10**(3), 395–411 (1985)
8. McClure, D., Reiter, J.P.: Differential privacy and statistical disclosure risk measures: an investigation with binary synthetic data. Trans. Data Privacy **5**(3), 535–552 (2012)
9. Ruggles, S., Fitch, C., Magnuson, D., Schroeder, J.: Differential privacy and census data: implications for social and economic research. AEA Papers Proc. **109**, 403–408 (2019)
10. Samarati, P., Sweeney, L.: Protecting Privacy When Disclosing Information: k-Anonymity and Its Enforcement Through Generalization and Suppression. Technical report, SRI International (1998)
11. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Martínez, S.: Enhancing data utility in differential privacy via microaggregation-based k-anonymity. VLDB J. **23**(5), 771–794 (2014). https://doi.org/10.1007/s00778-014-0351-4
12. Xiao, Y., Xiong, L., Yuan, C.: Differentially private data release through multidimensional partitioning. In: Proceedings of the 7th VLDB Conference on Secure Data Management-SDM 2010, pp. 150–168 (2010)
13. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G.: Differentially private histogram publication. In: IEEE International Conference on Data Engineering-ICDE 2012, pp. 32–43 (2012)