Differentially Private Data Publishing via Cross-Moment Microaggregation

Javier Parra-Arnau, Josep Domingo-Ferrer, *Fellow, IEEE*, and Jordi Soria-Comas

Abstract

Differential privacy is one of the most prominent privacy notions in the field of anonymization. However, its strong privacy guarantees very often come at the expense of significantly degrading the utility of the protected data. To cope with this, numerous mechanisms have been studied that reduce the sensitivity of the data and hence the noise required to satisfy this notion. In this paper, we present a generalization of classical microaggregation, where the aggregated records are replaced by the group mean and additional statistical measures, with the purpose of evaluating it as a sensitivity reduction mechanism. We propose an anonymization methodology for *numerical* microdata in which the target of protection is a data set microaggregated in this generalized way, and the disclosure risk limitation is guaranteed through differential privacy via record-level perturbation. Specifically, we describe three anonymization algorithms where microaggregation can be applied to either entire records or groups of attributes independently. Our theoretical analysis computes the sensitivities of the first two central cross moments; we apply fundamental results from matrix perturbation theory to derive sensitivity bounds on the eigenvalues and eigenvectors of the covariance and coskewness matrices. Our extensive experimental evaluation shows that data utility can be enhanced significantly for medium to large sizes of the microaggregation groups. For this range of group sizes, we find experimental evidence that our approach can provide not only higher utility but also higher privacy than traditional microaggregation.

Index Terms

Microaggregation, Differential privacy, Data utility.

1 Introduction

In the past few decades, we have witnessed revolutionary changes related to data collection and usage. In particular, triggered by the development of the Internet, we have transitioned from data scarcity to data abundance. This data abundance and the development of novel data analysis methodologies have made research and decision-making processes markedly data-centric. As a result, demands to access external data have increased and data sharing has become an emerging trend. The open data initiative [21] is a paradigmatic example of this direction.

When a data set is to be shared, the privacy of the individuals therein must be taken into account. That is precisely what privacy-preserving data publishing is about. While there is a large variety of techniques to limit disclosure risk, the privacy guarantees offered by most of them [10], [31], [26], [25] depend on assumptions on the side knowledge available to intruders. With the current data abundance, however, it is increasingly difficult to make such assumptions accurately. In this context, privacy models whose privacy guarantees are strong and side-knowledge independent gain relevance. Differential privacy (DP) [12] is the most prominent representative of this class of models[1].

DP was proposed as a privacy model to limit disclosure risk in database queries. Soon after its inception, numerous mechanisms to generate DP data sets began to appear. Nonetheless, except for the simplest data domains, generating useful DP data sets remains a highly challenging task.

Two main approaches have been followed to generate DP data sets: histograms and record perturbation. In the first case, the data set is generated by partitioning the data domain into a number of bins and giving DP counts of records within each bin. In the second case, the DP data set is generated by perturbing the original records appropriately. Due to the limitations of the histogram approach in complex data domains (*e.g.* for a fixed accuracy, the number of required bins grows exponentially with the dimension of the domain), we focus on record perturbation.

Naively generating a DP data set by perturbing the original records according to their sensitivity is definitely unfeasible —the typically large sensitivity of individual records in a database would render it useless. Consequently, there is an evident need to apply some sensitivity reduction mechanism. In [36], [37], [32], [33], [35], microaggregation is leveraged precisely with that purpose. Specifically, a protected data set is generated by collecting DP centroids, rather than records. However, although the microaggregation step is essential in those works to keep perturbation under control, the fact that the records within each cluster are replaced by a single statistic (i.e., the sample mean of all records in the cluster) has a

• *The authors are with Universitat Rovira i Virgili, Department of Computer Science and Mathematics, UNESCO Chair in Data Privacy, 43007 Tarragona, Catalonia,*
E-mail: {javier.parra,josep.domingo,jordi.soria}@urv.cat

Manuscript prepared February, 2018.

1. For conciseness, throughout this work we use the acronym DP to refer to both "differential privacy" and its adjective form "differentially private".

side effect on the variability of the microaggregated data, which may severely limit the utility of the released data set. This is more so because, in general, the sizes of the microaggregation clusters need to be quite large to effectively reduce the DP noise.

1.1 Contribution and Plan of this Paper

In this work, we focus on anonymization of numerical data and generalize the classical microaggregation algorithms, which are based on a single measure of central tendency within clusters, to include additional statistical measures such as dispersion and dependency between attributes. We refer to this generalization as *cross-moment microaggregation* and we will use it as a preliminary sensitivity reduction step in view of attaining differential privacy, while delivering more utility (several moments are reported rather than just the within-cluster centroid).

Our study of cross-moment microaggregation as a sensitivity reduction factor contemplates two cases: one in which the statistical measures are given as matrix coefficients, and another where this information is available as spectral decompositions. We investigate for each case the sensitivities of several central cross moments in a mathematically systematic fashion, drawing upon the methodology of engineering optimization. Next, we summarize the major contributions of this work:

- We propose a microdata anonymization methodology that capitalizes on a generalization of microaggregation. Our methodology relies on cross-moment microaggregation only to diminish data sensitivity, since the disclosure risk limitation comes from the enforcement of DP. Accordingly, we describe three anonymization algorithms in which microaggregation can be applied to either entire records or groups of attributes independently.
- We compute the sensitivities of the first two central cross moments and apply fundamental results from matrix perturbation theory to derive sensitivity bounds on the eigenvalues and eigenvectors of the covariance and coskewness matrices.
- We conduct an extensive, thorough experimental evaluation in which we analyze the impact on data utility caused by 26 different configurations of the proposed anonymization methods, including previous work. A variety of empirical results show that data utility can be enhanced significantly for medium to large sizes of the microaggregation clusters, which demonstrates the suitability of our approach. Interestingly, for this range of group sizes, our proposal may provide not only higher utility but also higher privacy than traditional microaggregation.

The remainder of this paper is organized as follows. Section 2 establishes the notation and recalls key aspects of DP and statistics. Section 3 reviews the state of art relevant to this work. Section 4 introduces cross-moment microaggregation, investigates the sensitivities of several central cross moments and describes our approach to generate DP data sets through record perturbation. Section 5 conducts an experimental evaluation of the proposed anonymization algorithms. Finally, conclusions are drawn in Section 6.

2 Notation and Background

This section describes necessary background on microaggregation, DP and statistics. Prior to that, we deal with notation. We shall adopt the same notation for vectors and matrices used in [39]. Specifically, we delimit matrices with parentheses, with the components separated by space, and use parentheses as well to construct column vectors from comma-separated lists. Occasionally, we shall use the notation $x^{\mathrm{T}}y$ to indicate the standard inner product on $\mathbb{R}^n$, $\sum_{i=1}^{n} x_i y_i$, and $\|x\|_1$ and $\|x\|_2$ to denote the L1- and L2-norms of a vector $x \in \mathbb{R}^n$, i.e., $\|x\|_1 = \sum_{i=1}^{n} |x|$ and $\|x\|_2 = (x^{\mathrm{T}}x)^{1/2}$.

For matrices $A = [\alpha_{i,j}] \in \mathbb{R}^{d\times d}$, recall that the Frobenius norm is defined as

$$\|A\|_{\mathrm{F}} = \left(\sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_{ij}^2 \right)^{1/2},$$

and the 2-norm as

$$\|A\|_2 = \max_{\substack{x\in\mathbb{R}^d \\ \|x\|_2 \leqslant 1}} \|A\,x\|_2.$$

It can be shown that $\|A\|_2 = \max_{1\leqslant i\leqslant d} \sqrt{\lambda_i}$, where $\lambda_1, \ldots, \lambda_d$ represent the eigenvalues of A.

In our analysis, databases will be modeled as matrices of dimensions n records by d attributes. If attribute j takes values in the interval $[0, \Lambda_j]$, denote by Λ the column vector $(\Lambda_1, \ldots, \Lambda_d)$ and by Ω the square matrix with all d columns equal to Λ. Occasionally, we shall use a set notation and hence databases will also be construed as sets of records. Throughout this work, we shall use the terms data set and database interchangeably.

2.1 Microaggregation

In essence, microaggregation [9] is a family of anonymization algorithms for data sets that operates in two stages:

- First, the set of records in a data set is clustered in such a way that (i) each cluster contains at least k records, and (ii) records within a cluster are as similar as possible.
- Secondly, records within each cluster are replaced by a representative of the cluster, typically the centroid record.

Clearly, when microaggregation is applied to the projection of records on their quasi-identifier attributes, the resulting data set is k-anonymous.

Microaggregation methods can be classified into *univariate* and *multivariate*:

- Univariate methods deal with multi-attribute data sets by microaggregating one attribute at a time. IR microaggregation [8] falls within this category. First, records are sorted by the first attribute; second, clusters of k consecutive values of the first attribute are created; and, third, all values within each cluster are replaced by the cluster representative (e.g., centroid). The same procedure is repeated for the rest of attributes. In general, microaggregation algorithms of this category provide low levels of disclosure risk limitation.
- Multivariate methods deal with several attributes at a time. If we define optimal microaggregation as finding a partition in clusters of size at least k such that within-clusters homogeneity is maximum, it turns out that optimal multivariate microaggregation is NP-hard [29]. This motivates the use of heuristic multivariate microaggregation algorithms such as MDAV [10].

2.2 Differential Privacy

DP was originally proposed by [12] as a privacy model in the interactive setting, that is, to protect the outcomes of queries to a database. The assumption is that an anonymization mechanism sits between the user submitting queries and the database answering them.

Central to DP is the notion of *neighbor* databases, which can be interpreted in two different ways. On the one hand, the *unbounded* case assumes one entry is either removed or added. On the other hand, the *bounded* notion considers the replacement of one record by another. An important difference is that the former case assumes the size n of the database to be publicly known, whereas the latter assumes this parameter is private. Nonetheless, the two notions of neighborhood are very related and mechanisms satisfying one can be adapted to meet the other. For the sake of mathematical simplicity, we use the latter definition.

Definition 1 (L1-sensitivity). Let $\mathcal{D}$ be the class of possible data sets. The L1-sensitivity or global sensitivity of a query function $f\colon \mathcal{D} \to \mathbb{R}^d$ is defined as

$$\mathrm{GS}(f) = \max_{X,X' \in \mathcal{D}} ||f(X) - f(X')||_1,$$

where X, X' are any two neighbor databases in the sense described above.

In the sequel, we shall often use just "sensitivity" to refer to global sensitivity. Also, to keep notation as light as possible, in some cases we shall also use Δ_f as a synonym of $\mathrm{GS}(f)$ or just Δ if there is no possible confusion about which is the function f.

Definition 2 (ϵ-Differential privacy). A randomized mechanism $\mathcal{M}$ on a query function f satisfies ϵ-differential privacy with $\epsilon \geqslant 0$ if, for all pairs of neighbor databases X, X' and for all $\mathcal{O} \subseteq \mathit{range}(\mathcal{M})$,

$$\frac{\mathrm{P}\{\mathcal{M}(f(X)) \in \mathcal{O}\}}{\mathrm{P}\{\mathcal{M}(f(X')) \in \mathcal{O}\}} \leqslant \exp(\epsilon).$$

Throughout this paper, the primary approach to design DP mechanisms is through the addition of Laplace noise [12]. The Laplace mechanism $\mathcal{M}_{\mathcal{L}}$, as this approach is known, operates as follows: first, it computes the real value $f(X)$ of the response to a certain user query f, then it masks $f(X)$ by adding noise $L(X)$ distributed according to a Laplace distribution, and finally it returns the randomized response $\mathcal{M}_{\mathcal{L}}(f(X)) = f(X) + L(X)$. It can be shown [12] that the Laplace mechanism preserves ϵ-DP when the Laplace distribution has zero mean and is scaled to $\mathrm{GS}(f)/\epsilon$.

2.3 Individual Differential Privacy

DP is a popular privacy model within the research community because of the strong privacy guarantee it offers. However, enforcing this strict guarantee may distort data significantly. Several relaxations of DP have been proposed that trade off privacy guarantees for improved data utility. Among these, individual DP (iDP) [38] is of particular interest since it preserves the privacy guarantees that DP gives to individuals. Namely, the presence or absence of any individual in a data set does not significantly influence the results of analyses on the data set.

Specifically, iDP argues that the formalization of DP (see Definition 2) is more stringent than required by the underlying intuitive view (the presence or absence of any record should not be noticeable). In particular, it argues that, to attain such intuitive guarantees, it is sufficient to require indistinguishability of the responses to queries between the *actual* data set and its neighbor data sets.

Definition 3 (Local sensitivity). Let $\mathcal{D}$ be the class of possible data sets. Given a data set $X \in \mathcal{D}$, the local sensitivity of a query function $f\colon \mathcal{D} \to \mathbb{R}^d$ is defined as

$$\mathrm{LS}(f, X) = \max_{X' \in \mathcal{D}} ||f(X) - f(X')||_1,$$

where the maximum is taken over all databases X' that are neighbors of X.

Clearly, $\mathrm{LS}(f, X) \leqslant \mathrm{GS}(f)$ for any query function f and input data set X. In practice, the local sensitivity is usually small except for especially ill-conditioned databases.

Definition 4 (ϵ-Individual differential privacy). Given a data set X, a randomized mechanism $\mathcal{M}$ on a query function f satisfies ϵ-individual differential privacy (or ϵ-iDP) if, for any data set X' that is a neighbor of X and any $S \subseteq \mathit{range}(\mathcal{M})$, we have

$$\exp(-\epsilon)\Pr\{\mathcal{M}(f(X')) \in \mathcal{O}\} \leqslant \Pr\{\mathcal{M}(f(X)) \in \mathcal{O}\} \leqslant \exp(\epsilon)\Pr\{\mathcal{M}(f(X')) \in \mathcal{O}\}.$$

Remarkably, in the iDP model we can employ the mechanisms originally designed for DP and apply the properties of sequential and parallel composition [28], with the sole difference that global sensitivity must be replaced with local sensitivity.

2.4 Statistical Quantities

This section recalls several statistical quantities that will appear in the remainder of this work. We adopt the same notation used in [11].

The sample (unbiased) covariance matrix of a database X is a $d \times d$ matrix $Q_X = [q_{ij}]$ with entries

$$q_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \tag{1}$$

for $n \geqslant 2$, where $\bar{x}_i$ is the sample mean of the attribute i, that is, $\sum_{k=1}^{n} x_{ki}/n$.

The Fisher (unbiased) sample coskewness estimator [16] of a database X is a $d \times d^2$ matrix $\Phi_X = [\phi_{ijk}]$, with entries

$$\phi_{ijk} = \frac{n}{(n-1)(n-2)}\sum_{m=1}^{n}(x_{mi} - \bar{x}_i)(x_{mj} - \bar{x}_j)(x_{mk} - \bar{x}_k) \tag{2}$$

for $n \geqslant 3$. Notice that permuting the indices i, j, k does not alter the value of ϕ_{ijk}. This property is called supersymmetry [4] and, as a result of it, Φ_X only has $d(d+1)(d+2)/6$ unique elements.

3 State of the Art

In statistical disclosure control (SDC) ones face two conflicting goals: to preserve the privacy of the individuals on the one hand, and on the other to keep the protected data as useful as possible. Several approaches have been investigated in the quest for a good balance between these conflicting interests.

3.1 Utility-first vs Privacy Models

In the early times of SDC, the methods used to limit disclosure risk in microdata releases took always a utility-first approach: under this approach, data are only mildly masked (to keep the desired level of data utility), and then the risk of disclosure is assessed on the generated data set (e.g., via record linkage). If the risk of disclosure is deemed too high, the protected data set is discarded and a new version is generated using stricter masking. Masking methods used include microaggregation, recoding and data swapping, among others [20].

The first attempt to provide formal privacy guarantees came with k-anonymity [31] and its extensions (e.g., l-diversity [26] and t-closeness [25]). In particular, k-anonymity seeks to prevent re-identification of the individuals to whom the records in the protected data set correspond. The intruder is assumed to be able to perform record linkage between the protected data set and some external identified data set by using a subset of attributes called quasi-identifiers. To hinder linkage, k-anonymity makes sure that each combination of values of quasi-identifiers is shared by at least k records in the protected data set.

In k-anonymity and its extensions, assumptions on the side knowledge available to the intruder are necessary: for example, which external identified data sets he can access and which attributes he can use as quasi-identifiers for linkage. However, in the current landscape, where data collection is pervasive, the need for such assumptions on the intruder can be seen as a flaw. It seems difficult to guarantee that a certain attribute is not externally available and, thus, that it does not need to be taken as a quasi-identifier. Privacy models that avoid making assumptions about the externally available information seem more appropriate. DP [12], which aims at making unnoticeable the presence or absence of any individual in a data set, is the most prominent among them.

DP attains its intended guarantees by making the outcome of queries similar between data sets that differ in one record. There is a general consensus about the strong privacy protection offered by DP. However, criticisms related to the utility of the generated data are common [34] and several relaxations have been proposed: for example, (ϵ, δ)-DP [13], (ϵ, δ)-probabilistic DP [27] and ϵ-individual DP [38].

Even if DP was initially proposed to protect query responses, mechanisms to generate DP data sets appeared soon after its inception. There are two main approaches: histograms and record masking.

The histogram approach takes advantage of the low sensitivity of counting queries over a partition of the data domain [15]. The naive application of this mechanism becomes problematic as the complexity of the data domain increases.

Note that, for fixed accuracy, the cardinality of the partition (number of bins) grows exponentially with the number of attributes. This has important effects on the computational cost and the accuracy of the protected data. The time and space complexity of counting the number of records in each bin is proportional to the cardinality of the partition. The state-of-the-art DP histogram-based method [23] runs in time $\mathcal{O}(n^3 \log n)$ for histograms of n bins. For histogram sizes in the orders of tens of thousands, this method may take months to complete. Additionally, since the effect of the noise in the various bins accrues, when the cardinality of the partition is large even a small noise may be too much. Works such as [24], [41], [18] suffer from these issues. Some mitigation strategies have been proposed. In [17], given a partition, the authors propose an algorithm that minimizes the error for a given family of counting queries. In [7], data summarization techniques are used to reduce the time and space complexity. In particular, time and space are made proportional to the number of non-empty cells in the summarized data set. An alternative way to tackle the issues caused by data dimensionality is to apply dimensionality reduction techniques. In [42], the dependency between attributes is modeled to generate the DP data set from a set of low-order marginals.

The alternative to obtain DP data sets based on record masking avoids partitioning the data domain. Instead, the DP data set is generated by masking the original records. Masking each record by adding a Laplace-distributed noise with magnitude proportional to the record sensitivity is not a feasible solution, however. Since the purpose of DP is to hide the presence of any single record, such a naive approach inescapably needs to introduce too much noise, thereby producing significant utility damage. Specifically, suppose we want to generate an ϵ-DP version of a data set X. Denote by X_ϵ the anonymized data set. For $r = 1, \ldots, n$, define $I_r(X)$ as the function that returns the attribute values of the r-th record of X. Since X can be interpreted as the collected answers to the queries $I_r(X)$ for all records r, an intuitive way to generate X_ϵ is collecting ϵ-DP responses to $I_r(X)$ for all r. If the responses to the queries I_r satisfy ϵ-DP, then, since each query refers to a different record, it follows from the parallel composition property of DP that X_ϵ also meets the desired ϵ-DP requirement. In short, with this methodology, X_ϵ is generated by providing a DP response to the queries asking for the values of all attributes in each record. Although this record-level perturbation methodology does not make any assumptions on the uses of the output data, unfortunately it may come at the expense of a huge information loss. Since each query I_r refers to a *single* individual, its L1-sensitivity is large and so is the masking required to attain ϵ-DP. The upshot is a database X_ϵ with very limited utility.

Thus, some strategy is needed to reduce the sensitivity of the records before masking them to achieve DP. The state of the art on this matter is discussed next.

3.2 Microaggregation-based Differentially Private Data Publishing

In [36], [37], [32], [33], [35], DP data sets are generated via record masking. In these works, microaggregation is employed to reduce the sensitivity of the queries used to generate the DP data sets: rather than querying each original record, in these works representatives of the microaggregation clusters are queried. Since a cluster representative is an aggregation of the records in the cluster, it is less sensitive to changes than any single record. The amount of sensitivity reduction depends on how such representative values are computed.

Specifically, in [36], [37], multivariate microaggregation is used to partition the original data set into clusters of k or more records. The DP data set is derived by querying the centroid (arithmetic average) of each cluster. Since multivariate microaggregation with minimal cluster size k over all the attributes ensures k-anonymity, we can regard this approach as combining k-anonymity and DP. A novel type of microaggregation called insensitive microaggregation had to be designed for this approach. The main reason is that, in standard microaggregation algorithms, changing one value in one record may yield a completely unrelated set of clusters, which would not reduce sensitivity; although it is certainly unlikely that changing one record value changes all clusters, to guarantee DP we need to consider the worst case. In insensitive microaggregation we require changes in a single record to produce a set of clusters that differ (at most) in one record. In this way, the sensitivity of a centroid divides the sensitivity of the original records by the corresponding cluster size. The downside of insensitive microaggregation is that it yields worse within-cluster homogeneity than standard microaggregation and, hence, higher information loss. Furthermore, the minimum cluster size k grows with the data set size, as we show next. Let n be the number of records in a data set. To obtain a DP version of it, n/k centroids are released, each one computed on a cluster of cardinality k and having sensitivity Δ/k, where Δ is the sensitivity of the original records. Hence, the sensitivity of the whole data set to be released is $n/k \times \Delta/k$. Thus, for numerical data sets, Laplace noise with scale parameter $(n/k \times \Delta/k)/\epsilon$ must be added to each centroid to obtain an ϵ-DP output. For this approach to reduce the amount of noise required to attain ϵ-DP, the inequality $n/k \times \Delta/k \leqslant \Delta$ must hold. Equivalently, one needs $k \geqslant \sqrt{n}$.

To circumvent the previous problems of [36], [37], an alternative approach was proposed in [32], [33] based on individual ranking (IR) microaggregation. Since IR microaggregates only one attribute at a time, it achieves insensitivity while avoiding the information loss penalty of multivariate insensitive microaggregation. On the other hand, there is no lower bound on k: when using IR microaggregation, we can bound the sensitivity of the set of centroids by Δ/k (to be compared with $n/k \times \Delta/k$ in the case multivariate insensitive microaggregation); hence, no matter the value of k, the sensitivity is reduced.

The seeming superiority of the IR microaggregation approach is, however, severely limited by the fact that it can only deal with data sets that contain a single attribute; combining the DP versions of several attributes via sequential

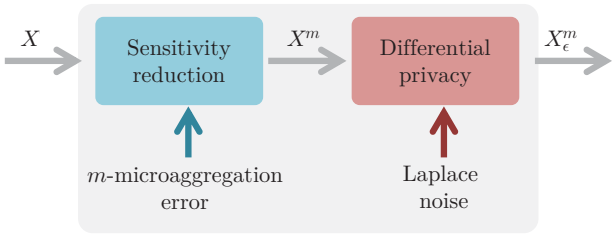


Fig. 1: Our methodology relies on a generalized form of microaggregation (m-microaggregation) as an intermediate step to reduce query sensitivity. Thus, there are two sources of error: one due to m-microaggregation itself, and another due to Laplace noise.

composition does not yield a DP data set. A way to avoid this issue was proposed in [35]. The difference with respect to [32], [33] is that the target of DP protection is the microaggregated data set, rather than the original data set. That is, one protects with DP a data set that consists of the centroids separately computed for each attribute via IR. Note that, since the microaggregated data set contains less information than the original data set, no additional privacy risk is incurred when shifting the target of protection from the original to the microaggregated data set. The main advantage of this transition is that one can combine through sequential composition the DP versions of the attributes obtained via IR microaggregation to obtain a single DP data set.

4 Differentially Private Data Sets via Cross-Moment Microaggregation

In this section, we describe our methodology to publish DP data sets through record-level perturbation. Our proposal is motivated by the fact mentioned in Section 3.1 that the prevailing approach for publishing DP microdata, based on histograms, suffers from serious limitations in terms of computational efficiency and data utility. But, as also noted in Section 3.1, plain independent masking of the records in the original data set may degrade utility severely.

To make record-level masking viable to generate DP data sets, there is an evident need to reduce the sensitivity of the query function to be used. One way to achieve this is to query aggregate information on records rather than individual records.

4.1 Sensitivity Reduction through Cross-Moment Microaggregation

Like [36], [37], [32], [33], [35], our methodology capitalizes on microaggregation as an intermediate anonymization step. In spite of microaggregation being in its own right a well-known statistical disclosure control technique, we use it here with the sole purpose of diminishing the sensitivity of the queries. The disclosure risk limitation comes from the enforcement of DP. This change of purpose carries along a change in the traditional way of thinking about microaggregation.

Unlike the works cited in the previous paragraph, however, we assume a more general interpretation of microaggregation. As explained in Section 2.1, standard microaggregation splits the data set into clusters of at least k records and then replaces the records in each cluster by the first raw moment, or cluster centroid, where the minimum value k prevents the centroid from representing too closely any particular record in the cluster. In this work, we consider microaggregation algorithms in which each cluster is replaced not only by the first raw moment, but also by central cross moments like the covariance, coskewness or cokurtosis. More specifically, we define *m-microaggregation*, with $m \geqslant 0$, as an algorithm that proceeds in these three steps:

1) partition the data set into clusters of at least k records;
2) compute the sample mean and the first m central cross moments of each cluster;
3) and replace the records in the cluster by the computed statistics.

Clearly, any standard microaggregation algorithm can be adapted to comply with the definition of *m-microaggregation*: it is just a question of accompanying the centroid with more statistics about the cluster, an information that is easily computable once the data set is clustered (first step above).

Apart from assuming a more general class of microaggregation algorithms, our anonymization methodology also differs from those of [36], [37], [32], [33] using standard microaggregation, that is, 0-microaggregation. In particular, following the approach suggested by [35] we consider that the data set to be protected is the *microaggregated one*, rather than the original one. Put another way, given an original data set X, we generate X^m by m-microaggregation of the records in X and drop X. Consequently, the aim is to publish X^m_ϵ, a DP version of X^m.

In the evaluation of the utility of X^m_ϵ, we need to consider then two sources of error. On the one hand, the error due to m-microaggregation, that is, the error caused by using X^m instead of X; and on the other hand, the noise introduced to attain ϵ-DP. The advantage of the proposed method lies in the fact that the former error is likely to be more than compensated by the reduction in the noise required to attain DP; the benchmark is the noise that would be required to attain DP directly from the original data set X. Figure 1 provides a conceptual depiction of these two sources of error.

To reduce the error introduced by microaggregation, we shall require algorithms that generate clusters that are as homogeneous as possible. For the sake of generality, throughout this work we do not favor any particular strategy to create the microaggregation clusters: they can all have the same or different cardinalities, they can be optimal (maximally homogeneous) or not, randomized or deterministic, etc.

As to the second type of error, the purpose of cross-moment microaggregation in this work is precisely to reduce the sensitivity of the query functions and thus the amount of Laplace noise to be added. As mentioned in Section 3.2, some sensitivity reduction was also achieved by 0-microaggregation: replacing original records with centroids divides the sensitivity by the minimum cluster size k.

4.2 Sensitivity of Cross-Moment Microaggregated Data Sets

In this subsection we present the theoretical analysis of the sensitivities of several statistical quantities central to cross-moment microaggregation. Specifically, we analyze the sensitivity of the coefficients of the covariance and coskewness matrices, as well as the sensitivity of the eigenvalues and eigenvectors of these matrices. As we shall elaborate later in Section 4.3, the statistical information provided by an m-microaggregated data set can be given in the form of either matrix coefficients or eigenvalues and eigenvectors.

The computation of the sensitivities of several statistical quantities, especially the covariance matrix, has been approached by various works seeking to provide DP versions of algorithms (e.g., principal component analysis) where those statistics are relevant. In this section, we find the sensitivity of each coefficient of the covariance matrix and derive tighter bounds on the sensitivity of the entire matrix. On the other hand and to the best of our knowledge, our work is the first to investigate the sensitivities of the coskewness matrix and the eigenvalues and eigenvectors of the covariance and coskewness matrices.

Before proceeding, we introduce some definitions that will facilitate presenting the results provided in this section. Let $\mathscr{X} = \{1, \dots, k\}$. For $a, b \in \{1, \dots, d\}$ and a fixed $r \in \mathscr{X}$, define $\mathcal{A}_a = \sum_{l \in \mathscr{X} \setminus \{r\}} x_{la}$ and $\mathcal{A}_{ab} = \sum_{l \in \mathscr{X} \setminus \{r\}} x_{la} x_{lb}$.

Proposition 5 (Mean). Denote by $f_{\bar{x}}$ the query function that returns the sample mean of a cluster $\mathcal{C} \subset \mathcal{D}$ of k records with d attributes. The L1-sensitivity of this function is $\mathrm{GS}(f_{\bar{x}}) = \mathbf{1}^{\mathrm{T}} \Lambda / k$.

Proof: The proof is immediate from the definition of sample mean. ■

Proposition 6 (Covariance). Let $\mathcal{C} \subset \mathcal{D}$ be a cluster of k records with d attributes and $f_{q_{ij}}$ be the query function that returns the coefficient q_{ij} of the covariance matrix $Q_{\mathcal{C}}$. The L1-sensitivity of this function, denoted by $\mathrm{GS}(f_{q_{ij}})$, is $\Lambda_i \Lambda_j / k$.

Proof: For the sake of brevity, we only show the case $i \neq j$. Consider two neighbor clusters $\mathcal{C}$ and $\mathcal{C}'$. For any $r \in \mathscr{X}$, denote by $x_r = (x_{ri}, x_{rj})$ and $x'_r = (x'_{ri}, x'_{rj})$ the respective values of the different record in either cluster, and note that $\bar{x}_l = (\mathcal{A}_l + x_{rl})/k$ and $\bar{x}'_l = \bar{x}_l + (x'_{rl} - x_{rl})/k$. Write Expression (1) as

$$q_{ij} = \frac{k}{k-1} \left(\frac{1}{k} \sum_{l=1}^{k} x_{li} x_{lj} - \bar{x}_i \bar{x}_j \right),$$

and accordingly compute the difference between the query responses q'_{ij} and q_{ij} as

$$q'_{ij} - q_{ij} = \frac{k}{k-1} \left(\bar{x}_i \bar{x}_j - \bar{x}'_i \bar{x}'_j + \frac{x'_{ri} x'_{rj} - x_{ri} x_{rj}}{k} \right),$$

which after simple manipulation yields

$$\frac{1}{k} \left(x'_{ri} x'_{rj} - x_{ri} x_{rj} - \frac{\mathcal{A}_i (x'_{rj} - x_{rj})}{k-1} - \frac{\mathcal{A}_j (x'_{ri} - x_{ri})}{k-1} \right). \tag{3}$$

Although the sensitivity is a maximum in absolute terms, we shall first find the maximum value of Expression (3) over the compact set $\mathcal{F} = \{0 \leqslant x_{ri}, x'_{ri} \leqslant \Lambda_i, 0 \leqslant x_{rj}, x'_{rj} \leqslant \Lambda_j\}$ with the rest of records $\mathcal{C} \setminus \{x_r\}$ fixed. To this end, define $g \colon \mathbb{R}^2 \to \mathbb{R}$ as the function $(\alpha, \beta) \mapsto \alpha\beta - \frac{\mathcal{A}_i}{k-1}\beta - \frac{\mathcal{A}_j}{k-1}\alpha$ and note that

$$\max_{x_r, x'_r \in \mathcal{F}} q'_{ij} - q_{ij} = \frac{1}{k} \max_{x_r, x'_r \in \mathcal{F}} g(x'_r) - g(x_r), \tag{4}$$

which allows us to write the maximization problem in question into the separable objective function g and separable constraints on x_r, x'_r. For convenience, we shall interpret Expression (4) as a function of all but the r-th record, $\mathcal{C} \setminus \{x_r\}$, and denote it by h. Next, we proceed to find the maximum and minimum values of g over $\mathcal{F}$.

First, we note that $(\alpha, \beta) = \left(\frac{\mathcal{A}_i}{k-1}, \frac{\mathcal{A}_j}{k-1} \right)$ is the only stationary point of g. Since $\nabla^2 g$ at this point has eigenvalues of different sign, the Hessian is indefinite and therefore the minimum and maximum are attained on the boundary of $\mathcal{F}$.

We shall only check the case $\beta = \Lambda_j$, for which g becomes the linear function

$$g(\alpha, \beta) = \left(\Lambda_j - \frac{\mathcal{A}_j}{k-1} \right) \alpha - \frac{\Lambda_j \mathcal{A}_i}{k-1}.$$

Since $\Lambda_j (k-1) \geqslant \mathcal{A}_j$, it immediately follows that the minimum and maximum occur at $\alpha = 0$ and $\alpha = \Lambda_i$, respectively. Systematic study of the three remaining boundaries permits us to derive the maximum and minimum of g over $\mathcal{F}$, and the difference between them, namely,

$$h(\mathcal{C}\setminus\{x_r\}) = \frac{1}{k(k-1)} \times \begin{cases} \Lambda_j\mathcal{A}_i, & \text{if } \Lambda_i\Lambda_j(k-1) \leqslant \Lambda_i\mathcal{A}_j + \Lambda_j\mathcal{A}_i \text{ and } \mathcal{A}_i > \mathcal{A}_j \quad (5)\\ \Lambda_i\mathcal{A}_j, & \text{if } \Lambda_i\Lambda_j(k-1) \leqslant \Lambda_i\mathcal{A}_j + \Lambda_j\mathcal{A}_i \text{ and } \mathcal{A}_i \leqslant \mathcal{A}_j \quad (6)\\ \Lambda_i\Lambda_j(k-1) - \Lambda_i\mathcal{A}_j, & \text{if } \Lambda_i\Lambda_j(k-1) > \Lambda_i\mathcal{A}_j + \Lambda_j\mathcal{A}_i \text{ and } \mathcal{A}_i > \mathcal{A}_j \quad (7)\\ \Lambda_i\Lambda_j(k-1) - \Lambda_j\mathcal{A}_i, & \text{if } \Lambda_i\Lambda_j(k-1) > \Lambda_i\mathcal{A}_j + \Lambda_j\mathcal{A}_i \text{ and } \mathcal{A}_i \leqslant \mathcal{A}_j. \quad (8) \end{cases}$$

We proceed to derive $\mathrm{GS}(f_{q_{ij}})$ from Expressions (5) to (8), by observing that

$$\begin{aligned} \mathrm{GS}(f_{q_{ij}}) &= \max_{\mathcal{C},\mathcal{C}'} |q'_{ij} - q_{ij}| \\ &= \max_{\mathcal{C}\setminus\{x_r\}} \max_{x_r,x'_r\in\mathcal{F}} \{q'_{ij} - q_{ij}, q_{ij} - q'_{ij}\} \\ &= \max_{\mathcal{C}\setminus\{x_r\}} \max_{x_r,x'_r\in\mathcal{F}} q'_{ij} - q_{ij}. \end{aligned}$$

In plain words, we just need to find the maximum of h over all possible clusters of $k-1$ records. It is straightforward to verify that Expressions (5) and (6) attain the maximum at $\mathcal{A}_i = \Lambda_i(k-1)$ and $\mathcal{A}_j = \Lambda_j(k-1)$, or equivalently when $(x_{li})_{l\neq r} = \Lambda_i$ and $(x_{lj})_{l\neq r} = \Lambda_j$, respectively. On the other hand, Expressions (7) and (8) are maximized at $\mathcal{A}_j = 0$ and $\mathcal{A}_i = 0$, or equivalently when $(x_{li})_{l\neq r} = 0$ and $(x_{lj})_{l\neq r} = 0$, respectively. In either case, the maximum value is $\Lambda_i\Lambda_j/k$. ■

The conclusions that can be drawn from Proposition 6 are rather straightforward. First, the global sensitivity of the coefficients of a covariance matrix is directly proportional to the width Λ, and inversely proportional to the size of the cluster, which shows the appropriateness of basing our anonymization methodology on cross-moment microaggregation as a means to reducing sensitivity. Secondly, since the covariance matrix is symmetric, it follows that $\mathrm{GS}(f_{Q_c}) \leqslant \sum_{j\geqslant i}^{d} \Lambda_i\Lambda_j/k$.

As mentioned at the beginning of this section, the problem of estimating the sensitivity of key statistical quantities, including the covariance matrix, has been tackled previously [14], [40], [6]. However, the bounds achieved by [14] are admittedly crude, and the ones obtained by [40] are looser. Our result is also more general than [6] since we do not make any assumption on the database or cluster. The result of [6] requires the database to be centered and essentially coincides with ours. But our sensitivity bound is in fact smaller because we consider just the unique elements of the matrix.

Next, we examine coskewness. The exact L1-sensitivity of the coefficients of a coskewness matrix will be derived from the following result, Lemma 7, which finds the solution to three related nonconvex quadratic problems.

Lemma 7. Consider the functions $\mathcal{B}_{ij}, \mathcal{C}_{ij}, \mathcal{D}_{ij} \colon \mathbb{R}^{2(k-1)} \to \mathbb{R}$,

$$\mathcal{B}_{ij}(x_i, x_j) = k\sum_{c=1}^{k-1} x_{ci}x_{cj} - 2\sum_{c=1}^{k-1} x_{ci}\sum_{c=1}^{k-1} x_{cj},$$

$\mathcal{C}_{ij}(x_i, x_j) = \mathcal{B}_{ij}(x_i, x_j) + \Lambda_i(k-2)\mathcal{A}_j$ and $\mathcal{D}_{ij}(x_i, x_j) = \mathcal{C}_{ij}(x_i, x_j) + \Lambda_j(k-2)\mathcal{A}_i - \Lambda_i\Lambda_j(k-1)(k-2)$. The following optimization problems in the variables x_i and x_j yield the same maximum value,

$$\max_{\substack{0\preccurlyeq x_i\preccurlyeq\Lambda_i,\\ 0\preccurlyeq x_j\preccurlyeq\Lambda_j}} |\mathcal{B}_{ij}(x_i, x_j)| = \max_{\substack{0\preccurlyeq x_i\preccurlyeq\Lambda_i,\\ 0\preccurlyeq x_j\preccurlyeq\Lambda_j}} |\mathcal{C}_{ij}(x_i, x_j)| = \max_{\substack{0\preccurlyeq x_i\preccurlyeq\Lambda_i,\\ 0\preccurlyeq x_j\preccurlyeq\Lambda_j}} |\mathcal{D}_{ij}(x_i, x_j)| = \Lambda_i\Lambda_j(k-1)(k-2).$$

Proof: For brevity, we only show the maximum of $|\mathcal{B}_{ij}|$. The derivation of the solution of the other two nonconvex quadratic problems proceeds along the same lines and is omitted.

The proof consists of three steps. We first find the maximum of $\mathcal{B}_{ij}$, then its minimum, and finally check the maximum of the two in absolute terms.

The existence of the solution is a consequence of the fact that we maximize and minimize a continuous function over a compact set. Assume $X_i = \Lambda_i l + v$ and $X_j = \Lambda_j m + w$ for some integers $0 \leqslant l, m \leqslant k-1$ and real numbers satisfying $0 \leqslant v < \Lambda_i$ and $0 \leqslant w < \Lambda_j$. The maximization problem can be equivalently recast as

$$\max_{\substack{l,m\\ v,w}} g(X_i, X_j), \tag{9}$$

where

$$\begin{aligned} g(X_i, X_j) = \text{maximize} \quad & \mathcal{B}_{ij}(x_i, x_j) \\ \text{subject to} \quad & 0 \preccurlyeq x_i \preccurlyeq \Lambda_i, \quad 0 \preccurlyeq x_j \preccurlyeq \Lambda_j, \\ & \sum_{c=1}^{k-1} x_{ci} = X_i, \sum_{c=1}^{k-1} x_{cj} = X_j. \end{aligned}$$

Recall from [19] that

$$x_{ci} = \begin{cases} \Lambda_i, & c = 1,\ldots,l \\ v\ , & c = l+1 \\ 0\ , & c = l+2,\ldots,k-1 \end{cases} \quad \text{and } x_{cj} = \begin{cases} \Lambda_j, & c = 1,\ldots,m \\ w\ , & c = m+1 \\ 0\ , & c = m+2,\ldots,k-1 \end{cases} \tag{10}$$

is a global maximizer of $\mathcal{B}_{ij}$. A couple of observations follow. First, note that the tuple (x_i, x_j) satisfying $x_{ci} = \Lambda_i, x_{cj} = \Lambda_j$ for an index $c = 1, \ldots, k-1$ and $x_{ci} = x_{cj} = 0$ for the rest of indices is not necessarily a maximizer of $\mathcal{B}_{ij}$ but yields

$$\mathcal{B}_{ij}(x_i, x_j) = \Lambda_i \Lambda_j (k-2) > 0,$$

by virtue of $k \geqslant 3$. Secondly, observe that if $X_i = \Lambda_i(k-1)$, then $v = 0$ and $g(X_i, X_j) = \Lambda_i X_j (2-k) < 0$ for any X_j; and similarly, if $X_j = \Lambda_j(k-1)$, then $g(X_i, X_j)$ is negative. The consequence of these three observations is that, if $X_i = \Lambda_i(k-1)$ or $X_j = \Lambda_j(k-1)$, the corresponding maximizer (x_i^*, x_j^*) of $\mathcal{B}_{ij}$ satisfies

$$\mathcal{B}(x_i^*, x_j^*) < \max_{\substack{0 \leqslant X_i < \Lambda_i(k-1) \\ 0 \leqslant X_j < \Lambda_j(k-1)}} g(X_i, X_j),$$

which indicates we only need to consider the cases $0 \leqslant l, m \leqslant k-2$.

That said, there exist three possibilities in Expression (10), namely, $l < m$, $l = m$ and $l > m$. However, the symmetry of the problem with respect to x_i and x_j and the fact that

$$g(\Lambda_i l + v, \Lambda_j(l+1) + w) > g(\Lambda_i l + v, \Lambda_j(l+\tau) + w)$$

for $\tau = 2, 3, \ldots$, allow us to dispense with one of the two inequality cases above and the case $m > l+1$. Accordingly, we shall examine

$$g(X_i, X_j) = \begin{cases} k(\Lambda_i \Lambda_j l + \Lambda_j v) - 2(\Lambda_i l + v)(\Lambda_j(l+1) + w), & \text{if } l = m-1 \\ k(\Lambda_i \Lambda_j l + vw) - 2(\Lambda_i l + v)(\Lambda_j l + w), & \text{if } l = m. \end{cases}$$

We proceed with the case $l = m-1$. Note that Expression (9) can be viewed as equivalent to maximizing the function

$$h(l, m) = \max_{\substack{0 \leqslant v < \Lambda_i \\ 0 \leqslant w < \Lambda_j}} g(X_i, X_j), \tag{11}$$

subject to $0 \leqslant l, m \leqslant k-2$. For any given l and m, observe that Expression (11) is a bilinear optimization problem and, as such, is maximized at the vertices of the feasible set. Before we continue with each case, note that, when $v = \Lambda_i$ or $w = \Lambda_j$, we just have to subtract one from l or m to ensure the consistency of the right-hand inequalities of Expression (11). Having said this, verify that, if $l > k/2 - 1$ ($l \leqslant k/2 - 1$), then $g(\Lambda_i l, \Lambda_j(l+1))$ is greater (smaller) than $g(\Lambda_i(l+1), \Lambda_j(l+1))$, and hence

$$h(l) = \begin{cases} \Lambda_i \Lambda_j l \left(k - 2(l+1)\right), & \text{if } l > k/2 - 1 \\ \Lambda_i \Lambda_j (l+1) \left(k - 2(l+1)\right), & \text{if } l \leqslant k/2 - 1. \end{cases} \tag{12}$$

Denote by $[\cdot]$ the nearest integer function. Accordingly, we easily find that the maximum of Expression (12) is

$$\Lambda_i \Lambda_j \left[\frac{k}{4}\right] \left(k - 2\left[\frac{k}{4}\right]\right), \tag{13}$$

which is attained at $l = \left[\frac{k}{4}\right] - 1$ and $v = \Lambda_i, w = 0$.

Now we turn to the case $l = m$ and again, for any l, we check that the maximizer of (11) is one of the vertices of $[0, \Lambda_i] \times [0, \Lambda_j]$. Straightforward manipulation shows

$$h(l) = \begin{cases} \Lambda_i \Lambda_j l(k - 2l), & \text{if } l \geqslant \frac{k-2}{4} \\ \Lambda_i \Lambda_j (l+1)(k - 2(l+1)), & \text{if } l < \frac{k-2}{4}. \end{cases}$$

The case $l < \frac{k-2}{4}$ has been examined for $m = l+1$, but the corresponding maximum value of Expression (13) is not attained here. On the other hand, it is easy to verify that $l = \left[\frac{k}{4}\right]$ maximizes $h(l)$ when $l \geqslant \frac{k-2}{4}$. In closing, we have $m = \left[\frac{k}{4}\right]$ and $v = w = 0$, that is, the same solution and objective value as those of case $m = l+1$. Consequently, the maximum of $\mathcal{B}_{ij}$ is given by Expression (13).

Following a similar, systematic approach, next we proceed to find the minimum of $\mathcal{B}_{ij}$. With this aim, we define $g(X_i, X_j)$ now as the minimum of this function for a given $X_i = \Lambda_i l + v$ and $X_j = \Lambda_j(k-m) + w$, with $l = 0, \ldots, k-1$ and $m = 1, \ldots, k$.

Intuitively,

$$x_{ci} = \begin{cases} \Lambda_i, & c = 1, \ldots, l \\ v\ , & c = l+1 \\ 0\ , & c = l+2, \ldots, k-1 \end{cases} \quad \text{and } x_{cj} = \begin{cases} 0\ , & c = 1, \ldots, k-m-2 \\ w\ , & c = k-m-1 \\ \Lambda_j, & c = k-m, \ldots, k-1 \end{cases}$$

is a global minimizer of $\mathcal{B}_{ij}$. Accordingly, we may represent the original problem equivalently as the minimization of

$$h(l, m) = \min_{\substack{0 \leqslant v < \Lambda_i \\ 0 \leqslant w < \Lambda_j}} \begin{cases} -2(\Lambda_i l + v)(\Lambda_j(k-m) + w), & \text{if } m > l+2 \\ k\, v\, w - 2(\Lambda_i l + v)(\Lambda_j(k-l-2) + w), & \text{if } m = l+2 \\ k\left(\Lambda_i \Lambda_j (l+1-m) + \Lambda_i w + \Lambda_j v\right) - 2(\Lambda_i l + v)(\Lambda_j(k-m) + w), & \text{if } m < l+2 \end{cases}$$

subject to $0 \leqslant l \leqslant k-1$ and $1 \leqslant m \leqslant k$. Any of the three cases pose again a bilinear problem in the variables v and w. The solution, in any case, is found in one of the vertices of the feasible set.

We begin with the case $m \leqslant l+1$ and the vertex $v = w = 0$. For brevity, we only show this vertex. In this case,

$$\begin{aligned}\min_{\substack{0\leqslant X_i<\Lambda_i(k-1)\\0\leqslant X_j<\Lambda_j(k-1)}} g(X_i,X_j) &= \min_{\substack{0\leqslant l\leqslant k-1\\1\leqslant m\leqslant k\\ l\geqslant m-1}} \Lambda_i\Lambda_j\big(l(2m-k)-k(m-1)\big)\\ &= \min_{1\leqslant m\leqslant k}\begin{cases} 2\Lambda_i\Lambda_j(m-k)(m-1), & \text{if } 2m\geqslant k\\ \Lambda_i\Lambda_j(m-k)(k-2) \quad , & \text{if } 2m<k.\end{cases}\end{aligned} \tag{14}$$

Clearly, if $2m \geqslant k$, then $g(X_i, X_j)$ is minimized at $m = \left[\frac{k+1}{2}\right]$ and $l = m-1$, and yields

$$2\Lambda_i\Lambda_j\left(\left[\frac{k+1}{2}\right]-k\right)\left(\left[\frac{k+1}{2}\right]-1\right).$$

On the other hand, the solution for $2m < k$ is $m = 1$ and $l = k-1$, and the corresponding minimum objective value is

$$-\Lambda_i\Lambda_j(k-1)(k-2). \tag{15}$$

Now we assume $m > l+2$. It is easy to check that the solution to the problem $h(l,m)$ is, in this case, the vertex $(v,w) = (\Lambda_i, \Lambda_j)$, which implies

$$\min_{\substack{0\preccurlyeq x_i\preccurlyeq\Lambda_i\\0\preccurlyeq x_j\preccurlyeq\Lambda_j}} \mathcal{B}_{ij}(x_i,x_j) = \min_{1\leqslant m\leqslant k}\ \min_{\substack{0\leqslant l\leqslant k-1\\ l<m-2}} -2\Lambda_i\Lambda_j(l+1)(k-m+1). \tag{16}$$

Then, since $k-m+1 \geqslant 0$, the minimum of the inner optimization problem occurs at $l = m-3$, and accordingly, the outer problem is minimized at $m = \left[\frac{k+3}{2}\right]$.

Lastly, we examine the case $m = l+2$. The solution to the problem $h(l,m)$ is one of the vertices $(\Lambda_i, 0)$, $(\Lambda_i, 0)$ and (Λ_i, Λ_j), and the corresponding objective values are

$$h(l) = \begin{cases} -2\Lambda_i\Lambda_j(l+1)(k-l-2), & \text{if } (v,w) = (\Lambda_i,0) \quad (17)\\ -2\Lambda_i\Lambda_j l(k-l-1), & \text{if } (v,w) = (0,\Lambda_j) \quad (18)\\ k\Lambda_i\Lambda_j - 2\Lambda_i\Lambda_j(l+1)(k-l-1), & \text{if } (v,w) = (\Lambda_i,\Lambda_j). \quad (19)\end{cases}$$

The minimizers in each case are $\left[\frac{k-3}{2}\right]$, $\left[\frac{k-1}{2}\right]$ and $l = \left[\frac{k}{2}\right]-1$, respectively.

Next, we show that Expression (15) is the global minimum. Substitute the nearest integer function of the solution of Expression (14) when $2m \geqslant k$, and of Expressions(16), (17) and (18), for the real value of each fraction; in other words, remove the nearest integer function of each of those five minimizers. Evaluating these solutions in (9) yields the minimum value $-\Lambda_i\Lambda_j(k-1)^2/2$. Analogously with (19), we have the minimum is $-\Lambda_i\Lambda_j k(k-2)/2$. It is then immediate to check that

$$2(k-1)(k-2) \geqslant (k-1)^2 \geqslant k(k-2),$$

for $k \geqslant 3$.

Finally, it remains to show $|\min \mathcal{B}_{ij}| \geqslant |\max \mathcal{B}_{ij}|$. To this end, note that the maximizer of Expression (12), without the constraint that l must be an integer, is $l = k/4$. From this, we verify

$$(k-1)(k-2) \geqslant \frac{k}{4}\left(k-2\frac{k}{4}\right),$$

on account of $k \geqslant 3$, which completes the proof. ■

Proposition 8 (Coskewness). Let $\mathcal{C} \subset \mathcal{D}$ be a cluster of k records with d attributes and $f_{\phi_{ijl}}$ be the query function that returns the element ϕ_{ijl} of the coskewness matrix $\Phi_{\mathcal{C}}$. The L1-sensitivity of this function is $\mathrm{GS}(f_{\phi_{ijl}}) = \Lambda_i\Lambda_j\Lambda_l/k$.

Proof: For brevity, we only show the case in which i, j, l are all different. We start by computing the difference $\phi'_{ijl} - \phi_{ijl}$. To this end, we adopt a notation similar to that of Proposition 6 and denote by $x_r = (x_{ri}, x_{rj}, x_{rl})$ and $x'_r = (x'_{ri}, x'_{rj}, x'_{rl})$ the record $r \in \mathscr{X}$ that is different in the two neighbor clusters $\mathcal{C}$ and $\mathcal{C}'$, respectively.

For $k > 2$, let $g\colon \mathbb{R}^3 \to \mathbb{R}$ be the function

$$g(\alpha,\beta,\gamma) = \frac{\alpha\beta\gamma}{k} - \frac{\mathcal{A}_i\beta\gamma + \mathcal{A}_j\alpha\gamma + \mathcal{A}_l\alpha\beta}{k(k-1)} - \frac{\alpha(\mathcal{A}_{jl}k - 2\mathcal{A}_j\mathcal{A}_l) + \beta(\mathcal{A}_{il}k - 2\mathcal{A}_i\mathcal{A}_l) + \gamma(\mathcal{A}_{ij}k - 2\mathcal{A}_i\mathcal{A}_j) - 2\mathcal{A}_i\mathcal{A}_j\mathcal{A}_l}{k(k-1)(k-2)}. \tag{20}$$

Furthermore, define $h\colon \mathbb{R}^{k-1} \to \mathbb{R}$ as

$$h(\mathcal{C} \setminus \{x_r\}) = \max_{x_r, x'_r \in \mathcal{F}} g(x'_r) - g(x_r).$$

It follows from the definition of the Fisher sample coskewness estimator in Expression (2) that

$$\phi'_{ijl} - \phi_{ijl} = g(x'_r) - g(x_r),$$

and accordingly that

$$\max_{\mathcal{C},\mathcal{C}'} |\phi'_{ijl} - \phi_{ijl}| = \max_{\mathcal{C}\setminus\{x_r\}} h(\mathcal{C}\setminus\{x_r\}). \tag{21}$$

We now proceed to find the maximum and minimum of g over $\mathcal{F} = \{0 \leqslant x_{ri}, x'_{ri} \leqslant \Lambda_i, 0 \leqslant x_{rj}, x'_{rj} \leqslant \Lambda_j, 0 \leqslant x_{rl}, x'_{rl} \leqslant \Lambda_l\}$. Define $\mathcal{R}_{ab} = \mathcal{A}_{ab}(1-k) + \mathcal{A}_a\mathcal{A}_b$, where a, b take on values on the set $\{i, j, l\}$. We find and study the stationary points by solving $\nabla g(\alpha^*, \beta^*, \gamma^*) = (0, 0, 0)$ and by computing the Hessian of g, namely,

$$\alpha^* = \frac{\mathcal{A}_l\mathcal{A}_{ij} - \mathcal{A}_i\mathcal{A}_{jl}}{\mathcal{R}_{jl}} + \gamma^*\frac{\mathcal{R}_{ij}}{\mathcal{R}_{jl}}, \qquad \beta^* = -\frac{\mathcal{A}_j\mathcal{A}_{il} - \mathcal{A}_l\mathcal{A}_{ij}}{\mathcal{R}_{il}} + \gamma^*\frac{\mathcal{R}_{ij}}{\mathcal{R}_{il}}, \qquad \gamma^* = \frac{\mathcal{A}_l \pm \left(\frac{k\mathcal{R}_{il}\mathcal{R}_{jl}}{(2-k)\mathcal{R}_{ij}}\right)^{1/2}}{k-1}, \tag{22}$$

and

$$\nabla^2 g(\alpha,\beta,\gamma) = \frac{1}{k}\begin{pmatrix} 0 & \gamma - \frac{A_l}{k-1} & \beta - \frac{A_j}{k-1} \\ \gamma - \frac{A_l}{k-1} & 0 & \alpha - \frac{A_i}{k-1} \\ \beta - \frac{A_j}{k-1} & \alpha - \frac{A_i}{k-1} & 0 \end{pmatrix}.$$

Note that $\gamma^* = \mathcal{A}_l/(k-1)$ if, and only if, $\mathcal{R}_{il} = 0$ or $\mathcal{R}_{jl} = 0$. Also, observe from Expression (22) that, if any of these two latter equalities hold, then β^* or α^* go/es to infinity and the two stationary points are infeasible. Consequently, for $\alpha^*, \beta^*, \gamma^*$ to be feasible we require $\gamma^* \neq \mathcal{A}_l/(k-1)$. However, if this condition is satisfied, the second-order principal submatrix obtained by removing the third row and third column of $\nabla^2 g(\alpha,\beta,\gamma)$ has negative determinant, and hence the Hessian is neither positive- nor negative- semidefinite. Since the first leading principal is zero, it follows then that Expression (22) corresponds to saddle points. Since $\mathcal{F}$ is nonempty and compact and g is continuous, a global minimum and a global maximum of g over $\mathcal{F}$ will be found at the boundary of this set.

We only show the case $\gamma = 0$. It is easy to check $g(\alpha, \beta, 0)$ has one single stationary point. However, if $\mathcal{A}_l \neq 0$, then

$$\nabla^2 g(\alpha,\beta,0) = \frac{1}{k}\begin{pmatrix} 0 & -\frac{A_l}{k-1} \\ -\frac{A_l}{k-1} & 0 \end{pmatrix}$$

is indefinite and the point turns out to be a saddle point. On the other hand, if $\mathcal{A}_l = 0$, g becomes linear with α and β, which ultimately means that the maximum and minimum are reached on one of the edges $\alpha = 0, \alpha = \Lambda_i,\ \beta = 0, \beta = \Lambda_j$.

Again, for the sake of brevity we only show the case $\alpha = 0$. From Expression (20), we immediately see that $g(0, \beta, 0)$ is maximized (minimized) at $\beta = 0$ ($\beta = \Lambda_j$) or $\beta = \Lambda_j$ ($\beta = 0$) depending on whether $\mathcal{A}_{il}k \geqslant 2\mathcal{A}_i\mathcal{A}_l$ or $\mathcal{A}_{il}k < 2\mathcal{A}_i\mathcal{A}_l$. Specifically,

$$\frac{1}{k(k-1)(k-2)} \max_{0\leqslant\beta\leqslant\Lambda_j} g(0,\beta,0) = \begin{cases} 2\mathcal{A}_i\mathcal{A}_j\mathcal{A}_l, & \text{if } \mathcal{A}_{il}k \geqslant 2\mathcal{A}_i\mathcal{A}_l \\ 2\mathcal{A}_i\mathcal{A}_j\mathcal{A}_l - (\mathcal{A}_{il}k - 2\mathcal{A}_i\mathcal{A}_l)\Lambda_j, & \text{if } \mathcal{A}_{il}k < 2\mathcal{A}_i\mathcal{A}_l \end{cases}$$

and

$$\frac{1}{k(k-1)(k-2)} \min_{0\leqslant\beta\leqslant\Lambda_j} g(0,\beta,0) = \begin{cases} 2\mathcal{A}_i\mathcal{A}_j\mathcal{A}_l - (\mathcal{A}_{il}k - 2\mathcal{A}_i\mathcal{A}_l)\Lambda_j, & \text{if } \mathcal{A}_{il}k \geqslant 2\mathcal{A}_i\mathcal{A}_l \\ 2\mathcal{A}_i\mathcal{A}_j\mathcal{A}_l, & \text{if } \mathcal{A}_{il}k < 2\mathcal{A}_i\mathcal{A}_l. \end{cases}$$

In light of this, it is straightforward to verify then that

$$h(\mathcal{C}\setminus\{x_r\})|_{\substack{x_{ri}=x'_{ri}=0 \\ x_{rl}=x'_{rl}=0}} = \max_{x_{rj},x'_{rj}} g(0,x'_{rj},0) - g(0,x_{rj},0) = \Lambda_j\frac{|\mathcal{A}_{il}k - 2\mathcal{A}_i\mathcal{A}_l|}{k(k-1)(k-2)},$$

which, as a function of all cluster records except the r-th, is one of the candidates to be $\mathrm{GS}(f_{\phi_{ijl}})$.

An analogous study of the remaining faces and edges of $\mathcal{F}$ leads to

$$\mathrm{GS}(f_{\phi_{ijl}}) = \frac{1}{k(k-1)(k-2)} \max_{\substack{\mathcal{C}\setminus\{x_r\} \\ \kappa,\lambda,\mu\in\{i,j,l\} \\ \kappa\neq\lambda,\lambda\neq\mu,\kappa\neq\mu}} \begin{cases} \Lambda_\kappa|\mathcal{A}_{\lambda\mu}k - 2\mathcal{A}_\lambda\mathcal{A}_\mu| \\ \Lambda_\kappa|\mathcal{A}_{\lambda\mu}k - 2\mathcal{A}_\lambda\mathcal{A}_\mu + \Lambda_\lambda(k-2)\mathcal{A}_\mu| \\ \Lambda_\kappa|\mathcal{A}_{\lambda\mu}k - 2\mathcal{A}_\lambda\mathcal{A}_\mu + \Lambda_\lambda(k-2)\mathcal{A}_\mu + \Lambda_\mu(k-2)\mathcal{A}_\lambda - \Lambda_\lambda\Lambda_\mu(k-1)(k-2)|. \end{cases}$$

Observe that the three arguments of the maximum function above correspond, except for the factor Λ_κ, to the three objective functions of Lemma 7. From this lemma, it follows then the global sensitivity value given in the statement of the proposition. ■

In the sequel, we shall investigate the sensitivities of the eigenvalues and eigenvectors of the covariance and coskewness matrices. Unlike the previous results, however, we shall only be able to obtain upper bounds on those sensitivity values. In the case of eigenvectors, besides, we shall derive just their local sensitivities.

For notational convenience, we assume the eigenvalues of both matrices, $\lambda(Q)$ and $\lambda(\Phi)$, are sorted in nonincreasing order and so are their corresponding eigenvectors, $v(Q)$ and $v(\Phi)$. Accordingly, we define $\gamma^{-1} = (\gamma_1^{-1}, \ldots, \gamma_d^{-1})$, where

$$\gamma_i = \min_i\{|\lambda_{i+1} - \lambda_i|, |\lambda_i - \lambda_{i-1}|\}$$

is the *minimum eigengap* of the eigenvalue λ_i [39]. Consistently, for $i = 1$, $\gamma_i = \lambda_1 - \lambda_2$, and for $i = d$, $\gamma_i = \lambda_{d-1} - \lambda_d$. Furthermore, for any cluster $\mathcal{C} \subset \mathcal{D}$ of k records with d attributes, we define $\Lambda(Q_\mathcal{C})$ as the matrix whose elements are the L1-sensitivities of the entries of $Q_\mathcal{C}$,

$$\Lambda(Q_\mathcal{C}) = \frac{1}{k}\Lambda\Lambda^{\mathrm{T}} = \frac{1}{k}\begin{pmatrix} \Lambda_1^2 & \Lambda_1\Lambda_2 & \dots & \Lambda_1\Lambda_d \\ \Lambda_2\Lambda_1 & \Lambda_2^2 & \dots & \Lambda_2\Lambda_d \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_d\Lambda_1 & \Lambda_d\Lambda_2 & \dots & \Lambda_d^2 \end{pmatrix}.$$

The following two propositions provide bounds on the L1-sensitivities of the eigenvalues of a covariance and a coskewness matrix, as well as bounds on the local sensitivities of the corresponding eigenvectors. To compute such bounds, we resort to the theory of matrix perturbation [39], a field that tackles, among other problems, those of eigenvalue perturbation. Informally, the problem at hand can be stated as follows: given a symmetric matrix A and a perturbation E of A, how are the spectra of A and $A + E$ related?

As we shall see next, Propositions 9 and 10 will apply some fundamental results from matrix perturbation analysis to derive our sensitivity bounds. However, the fact that we cannot analytically bound the minimum eigengap of a covariance and coskewness matrix will preclude us from obtaining the L1-sensitivities of their eigenvectors.

Proposition 9 (Eigenvalues and eigenvectors of covariance). Let $\mathcal{C} \subset \mathcal{D}$ be a cluster of k records with d attributes. Denote by $f_{\lambda(Q_\mathcal{C})}$ the query function that returns the eigenvalues of the covariance matrix $Q_\mathcal{C}$, and by $f_{v_i(Q_\mathcal{C})}$ the query function that returns the eigenvector corresponding to the i-th largest eigenvalue of $Q_\mathcal{C}$. It holds that

$$\mathrm{GS}\left(f_{\lambda(Q_\mathcal{C})}\right) \leqslant \frac{\sqrt{d}}{k}\Lambda^{\mathrm{T}}\Lambda, \qquad \mathrm{LS}\left(f_{v_i(Q_\mathcal{C})}, \mathcal{C}\right) \leqslant \frac{\sqrt{d}\,\gamma_i^{-1}}{k}\Lambda^{\mathrm{T}}\Lambda.$$

Proof: We apply two important results from the field of matrix perturbation theory. On the one hand, recall [39, §IV, Corollary 4.13] that, given two symmetric matrices $A, E \in \mathbb{R}^{d\times d}$, the eigenvalues of A and $A' = A + E$, denoted by $\lambda = (\lambda_1, \dots, \lambda_d)$ and $\lambda' = (\lambda'_1, \dots, \lambda'_d)$ and sorted in nonincreasing order, satisfy

$$\|\lambda' - \lambda\|_2 \leqslant \|A' - A\|_{\mathrm{F}}.$$

On the other hand, denote by v_i and v'_i the i-th eigenvectors of A and A', and recall [39, §V, Theorem 2.8] that

$$\|v'_i - v_i\|_2 \leqslant \gamma_i^{-1}\|A' - A\|_2.$$

Let $Q_\mathcal{C}$ and $Q_{\mathcal{C}'}$ be the covariance matrices of the neighbor clusters $\mathcal{C}$ and $\mathcal{C}'$, and denote by λ and λ' their corresponding tuples of eigenvalues. Accordingly,

$$\frac{1}{\sqrt{d}}\|\lambda' - \lambda\|_1 \leqslant \|\lambda' - \lambda\|_2 \overset{\text{(a)}}{\leqslant} \|Q_{\mathcal{C}'} - Q_\mathcal{C}\|_{\mathrm{F}} \overset{\text{(b)}}{\leqslant} \|\Lambda(Q_\mathcal{C})\|_{\mathrm{F}},$$

where (a) follows from the cited former result, and (b) follows from Proposition 6 on account of $|Q_{\mathcal{C}'} - Q_\mathcal{C}| \preccurlyeq \Lambda(Q_\mathcal{C})$. Let v_i and v'_i denote the i-th eigenvectors of $Q_\mathcal{C}$ and $Q_{\mathcal{C}'}$. Then,

$$\frac{1}{\sqrt{d}}\|v'_i - v_i\|_1 \leqslant \|v'_i - v_i\|_2 \overset{\text{(a)}}{\leqslant} \gamma_i^{-1}\|Q_{\mathcal{C}'} - Q_\mathcal{C}\|_2 \overset{\text{(b)}}{\leqslant} \gamma_i^{-1}\|\Lambda(Q_\mathcal{C})\|_2,$$

where (a) follows from the cited latter result, and (b) follows from Proposition 6.

The proof is completed by checking that $\mathrm{rank}(\Lambda(Q_\mathcal{C})) = 1$, which implies $\|\Lambda(Q_\mathcal{C})\|_{\mathrm{F}} = \|\Lambda(Q_\mathcal{C})\|_2$, and by noting that the largest singular value of $\Lambda(Q_\mathcal{C})$ is $\sum_{l=1}^{d}\Lambda_l^2/k$. ■

To compute the sensitivity of the eigenvalues and eigenvectors of the coskewness matrix, we shall interpret Φ as d matrices Φ^i of size $d \times d$. For notational conciseness, we define Ψ as the matrix whose rows are the inverse eigengaps γ^{-1} corresponding to Φ^i.

Proposition 10 (Eigenvalues and eigenvectors of coskewness). Let $\mathcal{C} \subset \mathcal{D}$ be a cluster of k records with d attributes. Denote by $f_{\lambda(\Phi_\mathcal{C})}$ the query function that returns the eigenvalues of the coskewness matrix $\Phi_\mathcal{C}$, and by $f_{v_i(\Phi^j_\mathcal{C})}$ the query function that returns the eigenvector corresponding to the i-th largest eigenvalue of $\Phi^j_\mathcal{C}$. Then,

$$\mathrm{GS}\left(f_{\lambda(\Phi_\mathcal{C})}\right) \leqslant \tfrac{\sqrt{d}}{k}\,\Lambda^{\mathrm{T}}\Omega\,\Lambda, \qquad \mathrm{LS}\left(f_{v_i(\Phi^j_\mathcal{C})}, \mathcal{C}\right) \leqslant \tfrac{\sqrt{d}}{k}\,\gamma_i^{-1}\Lambda_j\,\Lambda^{\mathrm{T}}\Lambda.$$

Proof: Let λ^i denote the eigenvalues of Φ^i. The proof is analogous to that of Proposition 9 with the particularity that $|\Phi^{i\prime} - \Phi^i| \preccurlyeq \Lambda_i\Lambda(Q_\mathcal{C})$ from Proposition 8, and that $\mathrm{GS}\left(f_{\lambda(\Phi_\mathcal{C})}\right) = \sum_{l=1}^{d}\|\lambda^{l\prime} - \lambda^l\|_1$. ■

4.3 Anonymization Methods

Recall that our aim is to protect the m-microaggregated data set X^m, that is, we aim to publish a DP version of X^m, denoted by X^m_ϵ. In this section, we specify how to generate X^m_ϵ for two families of m-microaggregation algorithms. In the former family, the cross-moment microaggregation algorithm provides the matrix coefficients of the first m central cross moments, while in the latter, the algorithm publishes the eigenvalues and eigenvectors of the corresponding matrices. We refer to these two families as *coefficient-based* and *eigenvalue-based* microaggregation and denote the corresponding m-microaggregated data sets by $X^{m,c}$ and $X^{m,e}$, respectively.

In the previous subsection, we found the global sensitivities of the statistical data output by coefficient-based microaggregation for $m = 0, 1, 2$. However, for eigenvalue-based microaggregation, we just derived bounds on the *local* sensitivities of the eigenvalues and eigenvectors of the covariance and coskewness matrices.

The fact that we operate with local sensitivities for eigenvalue-based microaggregation (instead of global sensitivities) unfortunately prevents us from applying DP to generate X^m_ϵ. To cope with this and thus compare coefficient-based and eigenvalue-based microaggregation in terms of their impact on data utility, we resort to iDP [38]. iDP is an alternative formalization of DP that relies on local sensitivity to offer exactly the same privacy guarantees (for numerical queries) as standard DP to individuals. As mentioned in Section 2.2, iDP has the same properties of parallel and sequential composition for Laplace noise addition and it is more utility-preserving than standard DP because it uses local sensitivity.

Thus, we rely on the Laplace mechanism to guarantee ϵ-DP for coefficient-based microaggregated data sets and ϵ-iDP for eigenvalue-based microaggregated data sets. To apply this mechanism, however, we require estimating the sensitivity of the cross-moment microaggregation functions. Next, we derive upper bounds on these sensitivities from the sensitivity results obtained in Section 4.2 for each matrix coefficient, eigenvalue and eigenvector.

Theorem 11 (1,2-Cross-moment microaggregation). Let $f_{1,c} = (f_{\bar{x}}, f_Q)$ and $f_{2,c} = (f_{\bar{x}}, f_Q, f_\Phi)$ be the query functions that return the elements of the first and the second central cross moments of a cluster $\mathcal{C} \subset \mathcal{D}$ of k with d attributes. Analogously, denote by $f_{1,e} = (f_{\bar{x}}, f_{\lambda(Q)}, f_{v(Q)})$ and $f_{2,e} = (f_{\bar{x}}, f_{\lambda(Q)}, f_{v(Q)}, f_{\lambda(\Phi)}, f_{v(\Phi)})$ the queries returning the eigenvalues and eigenvectors of the first and the second central cross moments of that cluster. Define

$$\hat{\Lambda}(f_{1,c}) = \sum_{i=1}^{d} \Lambda_i + \sum_{j \geqslant i}^{d} \Lambda_i \Lambda_j, \qquad \hat{\Lambda}(f_{2,c}) = \hat{\Lambda}(f_{1,c}) + \sum_{l \geqslant j \geqslant i}^{d} \Lambda_i \Lambda_j \Lambda_l,$$
$$\hat{\Lambda}(f_{1,e}) = \mathbf{1}^{\mathrm{T}} \Lambda + \sqrt{d}\left(\Lambda^{\mathrm{T}} \Lambda + \Lambda^{\mathrm{T}} \Omega\, \gamma^{-1}\right), \quad \hat{\Lambda}(f_{2,e}) = \hat{\Lambda}(f_{1,e}) + \sqrt{d}\left(\Lambda^{\mathrm{T}} \Omega\, \Lambda + \Lambda^{\mathrm{T}} \Lambda\, \Lambda^{\mathrm{T}} \Psi\, \mathbf{1}\right).$$

(i) The L1-sensitivities of $f_{1,c}$ and $f_{2,c}$ satisfy $\mathrm{GS}(f_{1,c}) \leqslant \hat{\Lambda}(f_{1,c})/k$ and $\mathrm{GS}(f_{2,c}) \leqslant \hat{\Lambda}(f_{2,c})/k$.

(ii) The local sensitivities of $f_{1,e}$ and $f_{2,e}$ satisfy $\mathrm{LS}(f_{1,e}, \mathcal{C}) \leqslant \hat{\Lambda}(f_{1,e})/k$ and $\mathrm{LS}(f_{2,e}, \mathcal{C}) \leqslant \hat{\Lambda}(f_{2,e})/k$.

Proof: We only show the upper bound of $\mathrm{GS}(f_{2,c})$. The proof is immediate from the sensitivity results derived in Propositions 5, 6 and 8, by observing that

$$\begin{aligned}
\mathrm{GS}(f_{2,c}) &= \max_{\mathcal{C},\mathcal{C}'} \|f'_{2,c} - f_{2,c}\|_1 \\
&\leqslant \max_{\mathcal{C},\mathcal{C}'} \|\bar{x}' - \bar{x}\|_1 + \max_{\mathcal{C},\mathcal{C}'} \|Q'_{\mathcal{C}} - Q_{\mathcal{C}}\|_1 + \max_{\mathcal{C},\mathcal{C}'} \|\Phi'_{\mathcal{C}} - \Phi_{\mathcal{C}}\|_1 \\
&\leqslant \mathrm{GS}(f_{\bar{x}}) + \frac{1}{k} \sum_{j \geqslant i}^{d} \Lambda_i \Lambda_j + \frac{1}{k} \sum_{l \geqslant j \geqslant i}^{d} \Lambda_i \Lambda_j \Lambda_l,
\end{aligned}$$

where the summations are over the unique elements of $Q_{\mathcal{C}}$ and $\Phi_{\mathcal{C}}$. The rest of the claims given in the statement follow analogously. ■

Having upper-bounded the sensitivities of 1,2-microaggregation, ϵ-DP and ϵ-iDP can now be attained just by adding zero-mean Laplace noise with scale $\hat{\Lambda}(f_{m,\cdot})/(k\,\epsilon)$. Since clusters are non-overlapping, parallel composition ensures that, by adding Laplace noise independently to each cluster, the list of cross-moment microaggregation statistics is ϵ-DP and ϵ-iDP, respectively.

As with traditional microaggregation, m-microaggregation replaces each record within a cluster with the corresponding estimators of the mean (centroid), covariance, coskewness, cokurtosis and so on. That is, the same tuple of statistics is repeated as many times as there are records in the cluster. A central aspect of our anonymization method is that the Laplace noise to be added to these estimators is the same within a cluster. The reason is to avoid canceling the benefits of microaggregation. To see this for coefficient-based microaggregation, suppose the opposite case, i.e., we generate k different samples drawn from a Laplace distribution. In this case, we would obtain k non-independent DP outcomes each of which with a noisy component scaled to $\hat{\Lambda}(f_{m,c})/k$. By the sequential composition property, the estimated sensitivity upper bound corresponding to the list of repetitions of statistics in the cluster would then become $\hat{\Lambda}(f_{m,c})$, which would render microaggregation useless. Consequently, to keep the sensitivity of the statistics repetitions at $\hat{\Lambda}(f_{m,c})/k$, we must have a single DP value of the cluster statistics, that is, we need to add exactly the same noise to all repetitions of a given tuple of statistics. An entirely analogous argument applies to eigenvalue-based microaggregation by replacing DP with iDP and $\hat{\Lambda}(f_{m,c})/k$ with $\hat{\Lambda}(f_{m,e})/k$. Figure 2 illustrates the proposed methodology.

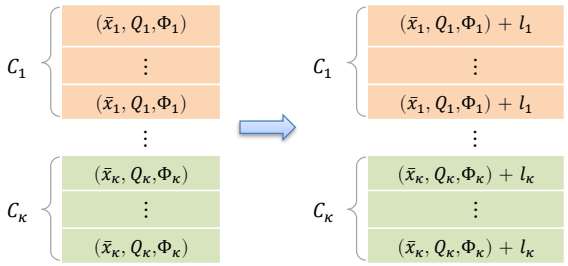


Fig. 2: We generate an ϵ-DP (ϵ-iDP) data set X_ϵ^m from an m-microaggregated data set X^m, by adding the same noise sample to each tuple of statistics within a given cluster.

Algorithm 1: Anonymization method that generates an ϵ-DP (ϵ-iDP) version of an m-microaggregated data set, with $m = 1, 2$.

Input: A data set $X^m = \{x_1^m, \ldots, x_n^m\}$ obtained from the m-microaggregation of an original data set $X \in \mathbb{R}^{n \times d}$; the mapping Π between clusters and records in X^m; the desired level of protection ϵ

Output: Either an ϵ-DP data set X_ϵ^m if X^m is coefficient-based, or an ϵ-iDP data set X_ϵ^m if X^m is eigenvalue-based.

1 **Let** $\mathcal{C}_1, \ldots, \mathcal{C}_\kappa$ be the set of clusters of X^m
2 **for** $i = 1, \ldots, \kappa$ **do**
3 **if** X^m *is coefficient-based* **then**
4 Set $r = 7d/6 - d^3/6 + d\,m\,(d+1)(d+2)/6$
5 **else**
6 Set $r = d(d\,m + d^2\,m - d^2 + 2)$
7 Compute γ^{-1} and Ψ from X^m
8 Let k_i be the cardinality of $\mathcal{C}_i$
9 Generate a tuple l_i of r random draws from a Laplace distribution with zero mean and scale $\hat{\Lambda}(f_{m,\cdot})/(k_i\,\epsilon)$
10 Select any record x_j^m that belongs to $\mathcal{C}_i$
11 Compute the noisy, representative record of $\mathcal{C}_i$ as $\tilde{c}_{\epsilon,i}^m = x_j^m + l_i$
12 **Process** $\tilde{c}_{\epsilon,i}^m$ to obtain $c_{\epsilon,i}^m$
13 **end**
14 **for** $j = 1, \ldots, n$ **do**
15 Let $\mathcal{C}_i := \Pi(x_j^m)$
16 Set $x_{\epsilon,j}^m = c_{\epsilon,i}^m$
17 **end**
18 **return** $X_\epsilon^m = \{x_{\epsilon,1}^m, \ldots, x_{\epsilon,n}^m\}$

We formally describe the procedure in Algorithm 1. The input parameters of our algorithm are the microaggregated data set X^m, or more specifically, the *clustered data set* whose records are to be replaced by up to m cross moments; the correspondence Π between records in X^m and clusters; and the privacy budget ϵ. Assuming clusters include all attributes, the algorithm proceeds as follows. From Theorem 11, we compute the sensitivity bounds of each tuple of statistics, which depend on the cardinality of each cluster and obviously Λ. For each cluster $\mathcal{C}_i$, then (i) we generate a single draw l_i from a Laplace distribution with zero mean and scale calibrated to the sensitivity bounds of Theorem 11 (and divided by ϵ); and (ii) use that noise sample to perturb the $|\mathcal{C}_i|$ occurrences of such statistics.

An important point addressed by Algorithm 1 is that, after adding noise in line 11, the perturbed statistics may not satisfy the properties they are expected to meet. For example, in the case $m \geqslant 1$ the perturbed sample covariance matrix may cease to be positive semi-definite. To tackle this kind of issues, we use the generic method *process* in line 12, which, in the particular example given, would compute the nearest symmetric positive semi-definite matrix in some matrix norm and would truncate any coefficient q_{ij} exceeding $\Lambda_i \Lambda_j / 4$ [2]. Similarly, our algorithm would check the properties of additional central cross moments are met.

Finally, an aspect that is not reflected in Algorithm 1 is the possibility of publishing a sampled version of an m-microaggregated data set. Although database users would in principle be interested in the explicit statistical knowledge provided by X_ϵ^m, a database containing only attributes values —like the original X— may be more manageable, especially for large m, and practical, since the sampled data can retain such statistical knowledge to some extent, depending on the minimum cluster size chosen. Also, from the standpoint of a database curator, a sampled version of X_ϵ^m, denoted henceforth by $\tilde{X}_\epsilon^m$, allows evaluating the utility loss of our anonymization algorithm with respect to the original data set.

2. An analogous processing would be required for eigenvalue-based microaggregation. In the case of coefficient-based microaggregation, the given upper bound follows easily from the Cauchy-Schwarz inequality and the fact that $\sigma_i \leqslant \Lambda_i/2$.

Algorithm 2: Anonymization method that produces an ϵ-DP (ϵ-iDP) version of an m-microaggregated data set for $m = 1, 2$ by perturbing each statistic independently, in accordance with a predefined masking strategy.

Input: A data set $X^m = \{x_1^m, \ldots, x_n^m\}$ obtained from the m-microaggregation of the original data set $X \in \mathbb{R}^{n \times d}$; the mapping Π between clusters and records in X^m; a masking strategy $(\epsilon_1, \ldots, \epsilon_r)$ such that $\sum_{i=1}^{r} \epsilon_i = \epsilon$
Output: An ϵ-DP data set X_ϵ^m

Let $\mathcal{C}_1, \ldots, \mathcal{C}_\kappa$ be the set of clusters of X^m
for $i = 1, \ldots, \kappa$ **do**
  for $j = 1, \ldots, r$ **do**
    Let GS(f_j) denote the L1-sensitivity of the j-th component of the query tuple $(f_{\bar{x}}, f_{q_{11}}, \ldots)$
    Generate a random draw l_{ij} from a Laplace distribution with zero mean and scale GS(f_j)/ϵ_j
  end
  Select any record x_j^m that belongs to $\mathcal{C}_i$
  Compute the noisy representative record of $\mathcal{C}_i$ as $\tilde{c}_{\epsilon,i}^m = x_j^m + l_i$
  Process $\tilde{c}_{\epsilon,i}^m$ to obtain $c_{\epsilon,i}^m$
end
for $j = 1, \ldots, n$ **do**
  Let $\mathcal{C}_i := \Pi(x_j^m)$
  Set $x_{\epsilon,j}^m = c_{\epsilon,i}^m$
end
return $X_\epsilon^m = \{x_{\epsilon,1}^m, \ldots, x_{\epsilon,n}^m\}$

4.3.1 Variants of the Proposed Anonymization Method

This subsection describes two variants of Algorithm 1. One of them proposes masking statistics independently, while the other is a generalization that permits operating with microaggregation algorithms such as IR [8]. Both variants can be applied to coefficient-based and eigenvalue-based microaggregation.

The first variant stems from the observation that the sensitivities of the coefficients of the first two cross moments grow squarely and cubicly with Λ, in contrast to that of the sample mean—which grows linearly with this parameter. Clearly, depending on the minimum cluster size and the statistical properties of the data set in question, this difference may cause the sensitivity of such coefficients to dominate the sensitivity of $f_{m,\cdot}$, which would result in the sample mean being added more noise than necessary. A significant reduction in data utility would ensue, since utility largely hinges on the masking of the sample mean.

To avoid this potential effect, we propose the following publication strategy. Instead of using the *joint* query function $f_{m,\cdot}$, which simultaneously queries the first raw moment and the first m central cross moments, we suggest querying such statistics *separately*. In the case of coefficient-based microaggregation and $m = 2$, this means using $f_{\bar{x}}$, $f_{q_{11}}, \ldots, f_{q_{ddd}}$ and $f_{\Phi_{111}}, \ldots, f_{\Phi_{ddd}}$ and operating with the sensitivity of each of these statistics independently, rather than the joint sensitivity of $f_{2,c}$.

More specifically, we consider an ϵ_i for each of those statistics, including the sample mean too; in the example above, we would assign ϵ_1 to $f_{\bar{x}}$, ϵ_2 to $f_{q_{11}}$ and so on. Accordingly, we propose that the Laplace noise to be added to each statistic is scaled to its own ϵ_i and its own sensitivity bound. Let r be the total number of statistical coefficients, or eigenvalues and eigenvectors in the case of eigenvalue-based microaggregation. Consider a masking strategy $(\epsilon_1, \ldots, \epsilon_r)$, possibly satisfying $\epsilon_1 \gg \epsilon_i$ to cause less distortion on the sample mean than on the rest of statistics. Under the assumption that $\sum_{i=1}^{r} \epsilon_i = \epsilon$, the sequential composition property ensures that the noisy cross-moment microaggregation statistics of any cluster is ϵ-DP. Then again, as each cluster contains disjoint records, parallel composition guarantees that the list of cross-moment microaggregation statistics is ϵ-DP. This procedure is described more formally in Algorithm 2, although to avoid repetition and unnecessary duplication we only show the case of coefficient-based microaggregation.

We now introduce the second variant of Algorithm 1. Recall that Algorithm 1 assumed the clustering step of m-microaggregation (step 1 in Section 4.1) was conducted for entire records. That is, we implicitly assumed that either the data set had a single attribute or multivariate clustering was applied over all attributes. Our second variant assumes, per contra, that in the process of generating the m-microaggregation clusters the algorithm is not applied to all attributes simultaneously. This is the case of, for example, IR, which operates over each attribute individually. Thus, we describe here a more general case in which the clustering is carried out independently for several individual attributes or disjoint subsets of attributes.

Roughly speaking, Algorithm 3 works as follows. We first group the attributes into disjoint subsets, then apply Algorithm 1 independently to each subset, and finally use the sequential composition property to compute the overall level of protection. More formally, denote by $S = \{G_1, \ldots, G_{|S|}\}$ the set of disjoint groups of attributes resulting from the clustering process in the aforementioned step 1. Recall from the sequential composition property that the independent use of ϵ_i-DP mechanisms for $i = 1, 2...$, when taken together, accumulates the level of protection to satisfy $(\epsilon_1 + \epsilon_2 + ...)$-DP. Consider a masking strategy $(\epsilon_1, \ldots, \epsilon_{|S|})$ such that $\sum_{i=1}^{|S|} \epsilon_i = \epsilon$. Since we apply Algorithm 1 independently to each subset

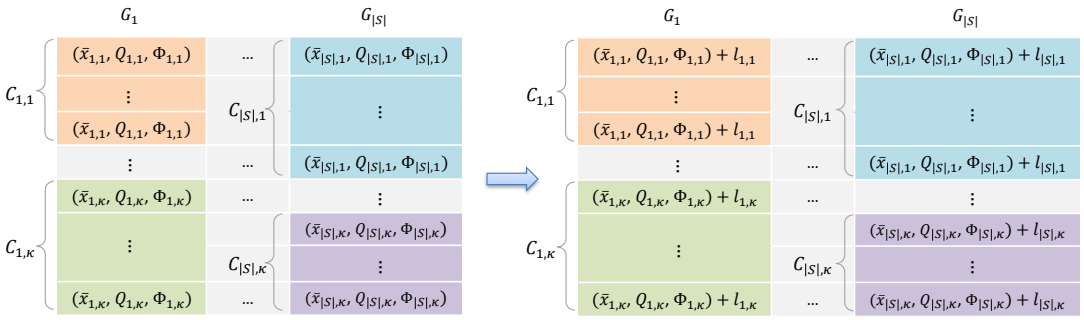


Fig. 3: We generate an ϵ-DP (ϵ-iDP) version of a data set that has been m-microaggregated on disjoint subsets of attributes independently.

Algorithm 3: Anonymization method that generates an ϵ-DP (ϵ-iDP) version of a data set that has been m-microaggregated independently on disjoint groups of attributes.

Input: A set $S = \{G_1, \ldots, G_{|S|}\}$ of disjoint groups of attributes; a data set X^m, which has been m-microaggregated by restricting X to each of the groups of S; the mappings $\Pi_1, \ldots, \Pi_{|S|}$ between clusters in the group G_i and records in X^m; a masking strategy $(\epsilon_1, \ldots, \epsilon_{|S|})$ satisfying $\sum_{i=1}^{|G|} \epsilon_i = \epsilon$

Output: Either an ϵ-DP data set X^m_ϵ if X^m is coefficient-based, or an ϵ-iDP data set X^m_ϵ if X^m is eigenvalue-based

for $i = 1, \ldots, |S|$ **do**
  Let X^m_i be the m-microaggregation of X when restricted to the attribute/s of G_i
  Execute Algorithm 1 with parameters X^m_i, Π_i and ϵ_i
end
return X^m_ϵ

of attributes, the publication of the list of cross-moment microaggregation statistics is ϵ-DP. An analogous reasoning applies to eigenvalue-based microaggregation and iDP. We illustrate this procedure in Figure 3.

5 Experimental Evaluation

In this section, we evaluate experimentally the anonymization method proposed in Section 4.3 as well as the two variants described in Section 4.3.1, and compare them with existing solutions relying on record-level masking. The ultimate purpose is to show that our approach, which builds on m-microaggregation to reduce data sensitivity, may in fact diminish the amount of noise required to attain ϵ-DP and ϵ-iDP. The empirical analysis provided in this section has been conducted in its entirety with Matlab 2017a.

5.1 Anonymization Methods under Evaluation

As mentioned in Section 4.1, the proposed anonymization methods are not tied to any particular strategy to generate the microaggregation clusters (step 1 in that section). However, it is clear that the selected clustering approach will have an impact on the utility of the resulting DP (iDP) data sets, and hence on our empirical analysis.

In this work, we evaluate the utility provided by a set of anonymization methods in the special case when the clustering step is performed with MDAV [10] and IR-MDAV. The former is a heuristic multivariate microaggregation algorithm, whereas the latter operates by running independent univariate MDAV microaggregations on each attribute. The reason for this choice is twofold. First, both are well-known algorithms in the literature of database anonymization; and secondly, both have been used with the same aim of enhancing the utility of DP data sets via record masking [37], [33], [35].

Figure 4 shows the anonymization methods evaluated in this section, which include the two families of cross-moment microaggregation algorithms (coefficient-based and eigenvalue-based) as sensitivity reduction procedures to generate DP and iDP data sets. However, since IR-MDAV operates on a per-attribute basis, eigenvalue-based microaggregation is only tested with MDAV.

In our series of experiments we investigate m-microaggregation for $m = 0, 1, 2$. For each m, each microaggregation clustering and each family of cross-moment microaggregation, we test the Algorithms 1, 2 and 3 described in Section 4.3. For the former two algorithms, although only in the case of coefficient-based microaggregation, we consider two possibilities:

- X^m contains all coefficients of covariance and coskewness;
- X^m contains only the variances and the marginal coefficients of coskewness.

We respectively refer to these two possibilities as *pure* and *variances* in Figure 4.

We evaluate two masking strategies for Algorithm 2. Recall that this algorithm scales Laplace noise to the sensitivity of each statistical element independently. With the aim of assessing the impact of noise addition on each statistic, we test one strategy that distributes the privacy budget evenly among the sample mean and the coefficients of covariance, and another strategy that, instead, gives priority (i.e., larger ϵ) to the first raw moment; this latter strategy is justified by the fact that the sample mean may be more informative than the other moments. The normalized weights of these strategies are $(0.50, 0.50)$

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
m	0	1	1	1	1	1	1	2	2	2	2	2	2	1	1	1	2	2	2	0	1	1	1	2	2	2
Coefficient/**E**igenvalue-based	-	C	C	C	C	C	C	C	C	C	C	C	C	E	E	E	E	E	E	-	C	C	C	C	C	C
MDAV/IR-MDAV	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	M	I	I	I	I	I	I	I
Joint/**S**eparate (Algorithm 1/2)	-	J	J	S	S	S	S	J	J	S	S	S	S	J	S	S	J	S	S	-	J	S	S	J	S	S
Pure/**V**ariances	-	P	V	P	P	V	V	P	V	P	P	V	V	P	P	P	P	P	P	-	V	V	V	V	V	V
Strategy **1/2**	-	-	-	1	2	1	2	-	-	1	2	1	2	-	1	2	-	1	2	-	-	1	2	-	1	2

Fig. 4: Anonymization methods evaluated in our experiments. Methods 1 and 20 correspond to the anonymization algorithms proposed in [35].

Method	B1	B2	B3	B4	B5	B6	B7	B8
m	0	1	1	2	2	0	1	2
MDAV/IR-MDAV	M	M	M	M	M	I	I	I
Pure/**V**ariances	-	P	V	P	V	-	V	V

Fig. 5: As baseline methods, we examine the standalone MDAV and IR-MDAV microaggregation algorithms for $m = 0, 1, 2$. By standalone we mean there is no Laplace noise addition.

and $(0.75, 0.25)$. Analogously for $m = 2$, we use these two distributions of weights: $(0.50, 0.30, 0.20)$ and $(0.75, 0.15, 0.10)$. The same distributions are employed for eigenvalue-based microaggregation. We refer to the more balanced distribution of weights as *strategy 1*, and the other as *strategy 2*.

In our experimental evaluation, we compare the above-mentioned methods with several baseline methods from the literature, listed in Figure 5 and explained next:

- The anonymization algorithm proposed in [35] for multivariate microaggregation of entire records. This algorithm relies on traditional MDAV microaggregation ($m = 0$) to generate DP data sets via Laplace noise addition. It corresponds to method 1 in Figure 4.
- The method proposed in [35] for univariate microaggregation. It uses traditional IR-MDAV microaggregation and the Laplace mechanism to produce DP data sets. This algorithm corresponds to method 20 in Figure 4.
- Plain Laplace noise without m-microaggregation. This method, described at the beginning of Sec. 4, adds Laplace noise directly to unaggregated records. Although it is a rather naive strategy, it will allow us to assess the benefits of m-microaggregation and derive worst-case bounds on the utility loss.
- Plain m-microaggregation with MDAV and IR-MDAV as clustering algorithms, but without the addition of Laplace noise to satisfy ϵ-DP or ϵ-iDP. Although these methods do not provide DP or iDP, we want to study the contribution of the m-microaggregation step to the utility loss.

5.2 Post-Processing and Sampling Distributions

We emphasized in Section 4.3 the need of post-processing the statistics masked by the Laplace noise. This section succinctly describes some of the processing tasks carried out in our experiments and provides several implementation notes on the sampling of m-microaggregated data sets. For brevity, we just list them:

- We rely on the quadratically convergent Newton method proposed by [30] to compute the positive semi-definitive matrix nearest to a perturbed covariance matrix[3].
- The implementation of Algorithm 2 for eigenvalue-based microaggregation computes the corresponding separate sensitivity bounds from the results obtained in [39, §IV, Corollary 4.10]. Particularly, in the notation of Section 4.2, the sensitivity of a *single* eigenvalue is upper-bounded by $\|\Lambda(Q_{\mathcal{C}})\|_2$.
- For $m = 1$, we use the normal distribution to sample the noisy statistics of X^m_ϵ. For $m = 2$, we employ the univariate and multivariate skew-normal distributions proposed by Azzalini [1], [2]. In the univariate case, Azzalini's skew-normal distribution is one of the few tractable classes of distributions that include the Gaussian one as a proper one, not just as a limit case. The multivariate version, on the other hand, provides a parametric family in which the marginal densities are scalar skew-normal.
- For $m = 2$, Azzalini's multivariate skew-normal distribution takes as input the marginal coefficients of coskewness $\frac{\phi_{iii}}{\sigma^{3/2}}$, instead of the whole coskewness matrix. This means that the methods evaluated for $m = 2$ only mask d coefficients of this matrix.

5.3 Data Sets

To conduct our experimental evaluation, we use two reference data sets in the field of statistical disclosure control (SDC), namely, Census and EIA; both were employed for the first time in the "CASC" European project [5] to test and compare SDC methods. The Census and EIA data sets contain respectively 1 080 and 4 096 records, and 13 and 11 numerical attributes. For the sake of comparison with [35], we restrict our analysis to the four attributes employed by the cited work for Census, and the same number of attributes for EIA[4]. The selected attributes are shown in Figure 6.

3. We use the Matlab code available at http://www.math.nus.edu.sg/~matsundf/
4. Census was the only data set evaluated in [35].

Data set	# of records	# of attributes	Selected numerical attributes
Census [5]	1,080	13	FICA (social security retirement payroll deduction), FEDTAX (federal income tax liability), INTVAL (amount of interest income), POTHVAL (total other persons income)
EIA [5]	4,092	11	RESREVENUE (revenue from sales to residential consumers), RESSALES (sales to residential consumers) COMREVENUE (revenue from sales to commercial consumers), COMSALES (sales to commercial consumers)

Fig. 6: Overview of the data sets used in our experiments.

In our analysis, we consider the projections of both data sets onto 2, 3 and 4 such attributes, starting with the first attribute listed in Figure 6 and ending with the last one. That is, we consider a data set Census 2 taking the first two attributes (FICA and FEDTAX), another data set Census 3 taking the first three attributes, and finally the full data set that we call Census 4. For EIA, we define data sets EIA 2, EIA 3 and EIA 4 in an analogous way. The reason for this multiple data set analysis lies in the well-known curse of dimensionality, which is about the information loss caused by clustering growing with the number of attributes. In other words, with this series of data sets we also want to analyze the impact of m-microaggregation on data utility.

5.4 Utility Metric and Privacy Parameters

We use the sum of squared errors (SSE) to evaluate the impact on data utility caused by anonymization. The SSE is a measure of overall information loss that is frequently employed in the evaluation of SDC methods, particularly in traditional microaggregation. The SSE between X and $\tilde{X}_\epsilon^m$ is given by

$$\mathrm{SSE}(X, \tilde{X}_\epsilon^m) = \sum_{i=1}^{n}\sum_{j=1}^{d}(r_{ij} - r_{\epsilon_{ij}}^m)^2,$$

where r_{ij} is the j-th attribute value of original record r_i, and $r_{\epsilon_{ij}}^m$ is the j-th attribute value of record $r_{\epsilon_i}^m$ (of $\tilde{X}_\epsilon^m$) that corresponds to r_i. Since the SSE results provided in our experiments only involve the databases X and $\tilde{X}_\epsilon^m$, we shall omit the database names and write just SSE.

On the other hand, recall that the anonymization methods under study aim to provide, respectively, DP and iDP versions of coefficient-based and eigenvalue-based microaggregated data sets. In this section, we conduct our series of experiments for levels of privacy protection in the interval $\epsilon \in [1, 2]$, which cover the usual range of values observed in the literature [3], [22], [37], [32], [33], [35].

Lastly, we note that the attributes shown in Figure 6 are not naturally upper-bounded. Since the sensitivity bounds derived in Section 4.2 are essentially proportional to the length of the intervals in which these attributes take values, we need to delimit the domain of each attribute. For the sake of comparison, we follow the methodology described in [37], [32], [33], [35] and upper-bound the domain of an attribute to be 1.5 times the maximum value of this attribute in the data set.

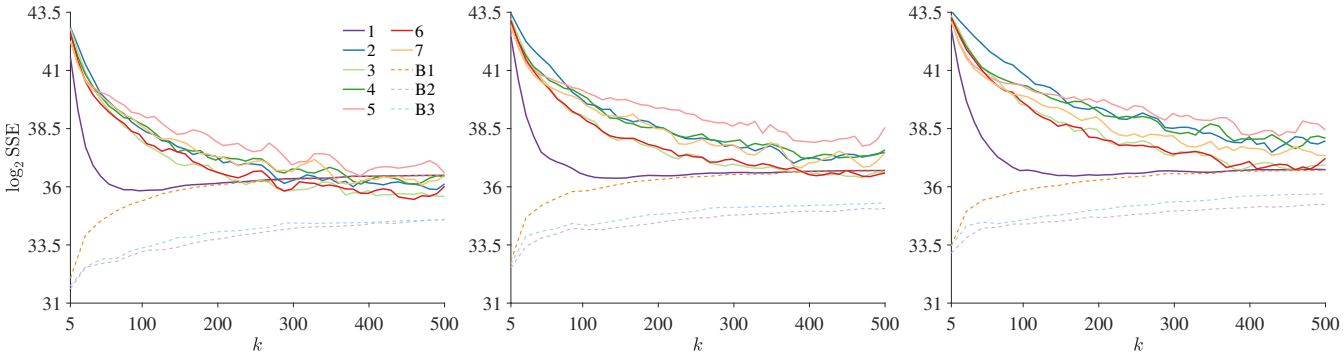
5.5 Results

We start examining coefficient-based microaggregation and assuming MDAV is applied to all attributes.

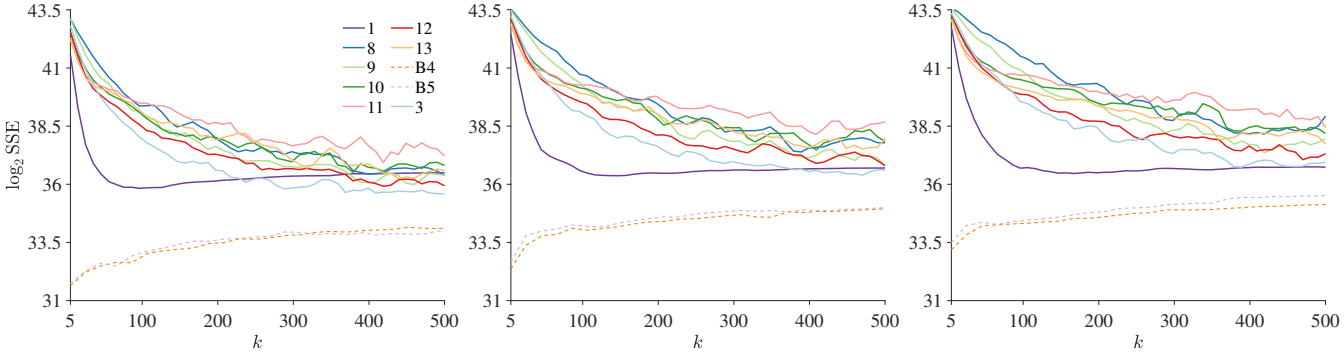
Figure 7 shows the impact caused by the anonymization methods 1 to 13 on the utility of the Census data set for $\epsilon = 1$. The top-row figures represent the cases in which the sample mean ($m = 0$), and the sample mean together with the sample covariance ($m = 1$), are included in each cluster. The figures from (d) to (f), on the other hand, show the case $m = 2$ where the second central cross moment is added to the previous statistics. For the sake of comparison among different values of m, the bottom-row figures also depict the case $m = 0$ and the best anonymization method for $m = 1$.

Recall that the baseline methods summarized in Figure 5 represent lower bounds on minimum distortion; the degradation in data utility comes just from cross-moment microaggregation, since no Laplace noise is injected. In Figure 7, we use dotted lines to plot such methods and observe that their graphs are roughly monotonically increasing. The reasoning is the following: as the number of records within a cluster increases, so does the potential variability of the data and the less the records can be accurately represented them with just a mean or a covariance matrix. This is obviously under the assumption that, in a real data set, the data are likely to be multimodal and not necessarily normal or skew-normal, which is the case here.

In the cases $m = 1, 2$, we see that the distortion caused by all anonymization methods substantially decreases with the cluster size. Essentially, what this means is that the reduction of Laplace noise due to increase of k is greater than the microaggregation error. In fact, the difference in SSE between an anonymization algorithm and its corresponding baseline method is the distortion caused by the DP protection. When this difference vanishes, as we observe for $m = 0$ and the

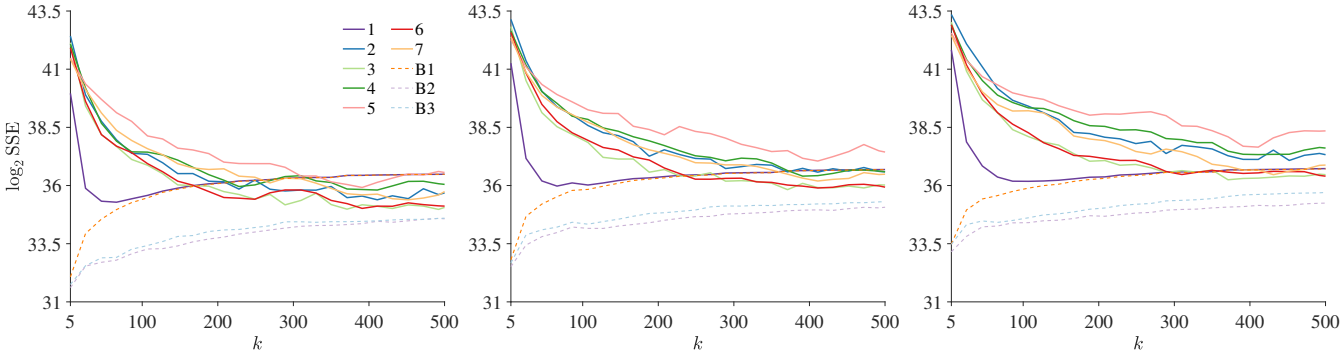


(a) Census 2, $m = 1$. (b) Census 3, $m = 1$. (c) Census 4, $m = 1$.

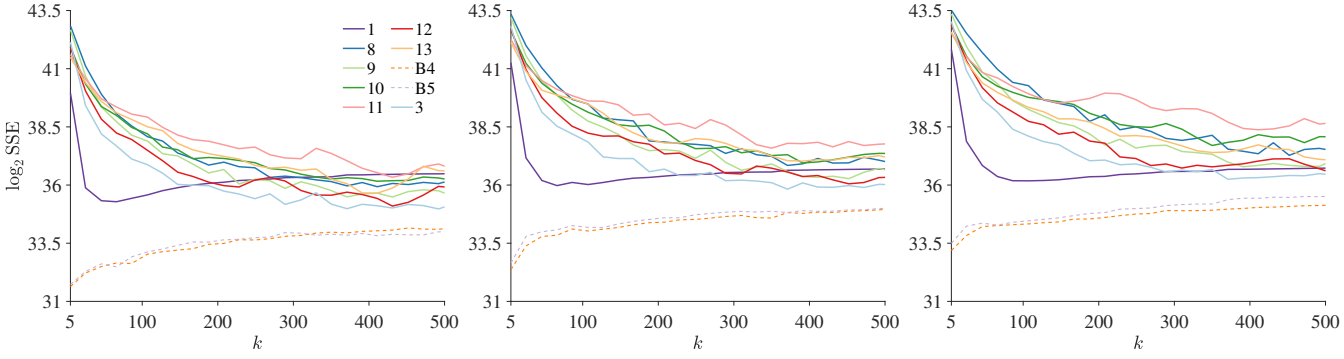


(d) Census 2, $m = 2$. (e) Census 3, $m = 2$. (f) Census 4, $m = 2$.

Fig. 7: Information loss in the Census data sets for coefficient-based microaggregation, MDAV and $\epsilon = 1$.



(a) Census 2, $m = 1$. (b) Census 3, $m = 1$. (c) Census 4, $m = 1$.



(d) Census 2, $m = 2$. (e) Census 3, $m = 2$. (f) Census 4, $m = 2$.

Fig. 8: Information loss in the Census data sets for coefficient-based microaggregation, MDAV and $\epsilon = 2$.

range of values of k assumed in these experiments, the generation of a DP version of the cross-moment microaggregated data set comes at no additional cost in terms of data utility.

TABLE 1: Information loss ($\log_2 SSE$) incurred by plain Laplace noise addition.

ϵ	Census 2	EIA 2	Census 3	EIA 3	Census 4	EIA 4
1	47.68	60.28	49.13	61.45	49.97	62.87
1.2	47.16	59.75	48.60	60.93	49.44	62.34
1.4	46.71	59.30	48.16	60.48	49.00	61.89
1.6	46.33	58.92	47.77	60.09	48.61	61.50
1.8	45.99	58.58	47.44	59.76	48.27	61.16
2	45.68	58.27	47.14	59.46	47.97	60.86

In addition to this decreasing trend, Figure 7 confirms the intuition that there must exist a value of k beyond which leveraging one or several central cross moments causes less distortion than just using the mean. Formally, we denote by $k_{\min}^{1,2}$ the cluster size beyond which 1,2-microaggregation is better than 0-microaggregation in terms of SSE.

An important conclusion that can be drawn from Figures 7(a-c) is that, for the Census data set and a level of protection of $\epsilon = 1$, methods 3 and 6 are the most utility-preserving. We note that the errors incurred by both methods are in fact very similar (for this reason, we only plot method 3) and that, respectively for Census 2, Census 3 and Census4, $k_{\min}^{1}$ takes approximate values 283, 400 and 510. The reason for this behavior is due to two factors. On the one hand, the curse of dimensionality. On the other hand, the larger the number of attributes, the larger the number of matrix coefficients, and the smaller the privacy budget assigned to each of them. The fact that ϵ is distributed among more coefficients implies that the most relevant statistics (here, the mean and the variances) are added more noise. However, since there are fewer coefficients in the "variance-only" methods (compared with the "pure" ones), we consistently observe greater differences in SSE between these two types of anonymization procedures as the number of attributes grows.

One of the most remarkable results of the bottom-row figures is that the best anonymization methods for $m = 1$ outperform those for $m = 2$. The result can be explained as follows. First, a higher number of statistics reduces the privacy budget of each statistical coefficient and hence increases the scale of the Laplace noise to be injected. Secondly, unless the distribution of the data within a cluster is very skewed, the benefits of providing the marginal coefficients of coskewness may not outweigh the redistribution of the weights ϵ_i of a masking strategy. As we shall see later in Figure 10, however, a combination of few attributes, large k (1 000) and large ϵ (2) will lead to smaller microaggregation errors and Laplace noise, which will result in less distortion for $m = 2$ than for $m = 1$.

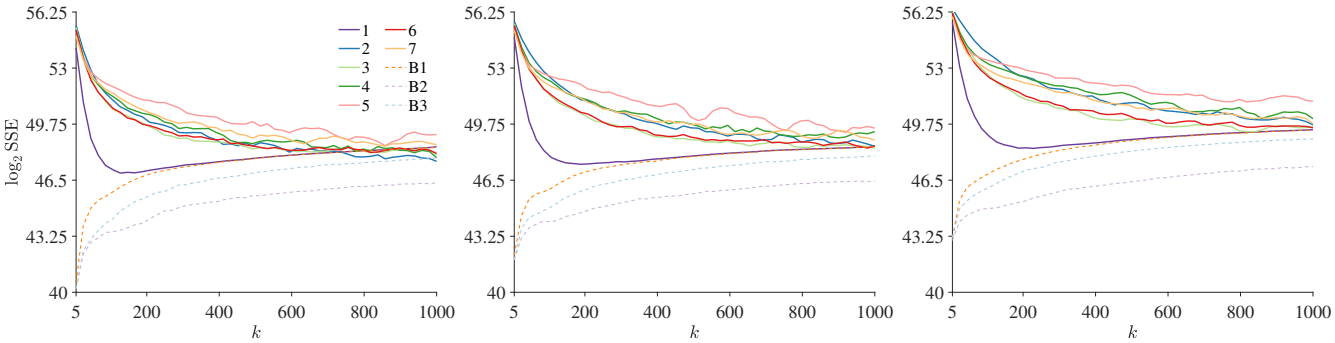
The most interesting observation of this first series of experiments is probably the win-win effect of $1, 2$-microaggregation over $m = 0$ and standalone microaggregation. Although privacy and utility are two confronting aspects in database anonymization, cross-moment microaggregation is shown to improve both of them for certain values of k. We illustrate this effect next with Figure 8(a).

Recall, first, that multivariate microaggregation with minimal cluster size k over entire records yields k-anonymity. In Figure 8(a), B1 represents the case in which a database curator microaggregates Census 2 with $m = 0$ to satisfy k-anonymity. Since our anonymization methodology is not applied subsequently by B1, the measure of protection of the microaggregated data set is just k. In the present work, however, the target of protection is the microaggregated data set, which we safeguard with DP. That said, suppose the database curator chooses $k = 100$, which gives a distortion 35.4, approximately. If the curator had selected $m = 1$ instead and applied a higher level of protection, say $k = 390$ (in the k-anonymity sense), the additional application of our methodology to meet 2-DP would have yielded $SSE = 34.9$. This procedure corresponds to method 6 and shows that cross-moment microaggregation, when used as a sensitivity reduction mechanism to attain DP, can provide higher utility and higher privacy than classical microaggregation. Note that higher privacy, in this case, comes not only because the microaggregation parameter k is greater but also because 2-DP is enforced on top of microaggregation.

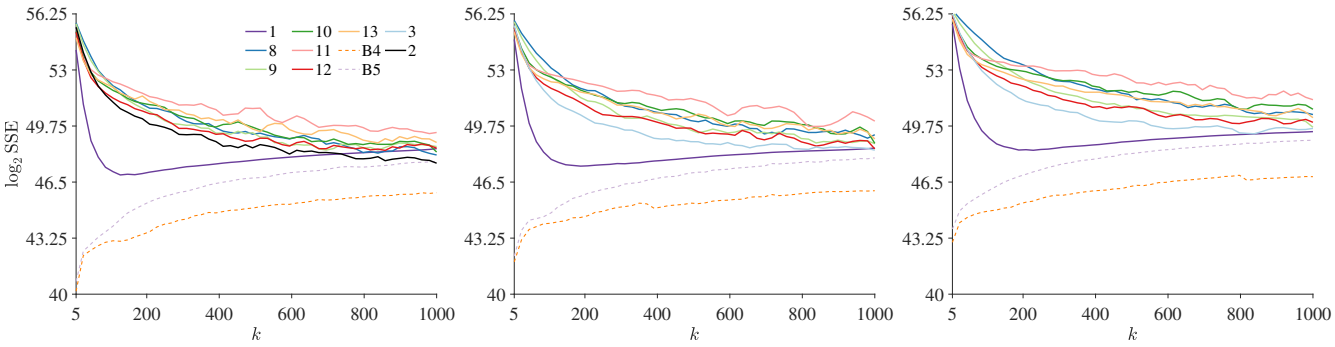
A couple of additional observations follow from Figure 7. On the one hand, we observe that, among the six anonymization methods that include covariance, the one that provides only variances and masks the statistical coefficients separately gives the smallest impact on data utility. This is in stark contrast with the best algorithm for $m = 1$, which relies on Algorithm 1 and applies the masking jointly. On the other hand, compared with the plain Laplace noise strategy, we see that cross-moment microaggregation offers substantial reductions in data distortion. Specifically, the largest SSE value observed for our anonymization algorithms (43.5) represents a 8.39% diminution with respect to direct noise addition. In the best-case scenario, in contrast, SSE is reduced by 25.67%. Table 1 shows the performance of this naive strategy for the six data sets considered in our experiments.

Figure 8 shows the same graphs of Figure 7 but for $\epsilon = 2$. Clearly, the reduction in SSE we observe with respect to the previous figure is due to the smaller scale of the Laplace noise added, as the error associated to microaggregation remains unaltered. As a result, $k_{\min}^{1}$ is, respectively, approximately 167, 239 and 350 for Census 2, Census 3 and Census 4, which represents, respectively, 40.99%, 40.25% and 31.37% reductions compared with $\epsilon = 1$.

Most of the conclusions drawn from Figure 7 also apply to $\epsilon = 2$. Among the most relevant results, it can be seen that methods 3 and 6 have the least impact on data utility for $m = 1, 2$, and that method 12 is the most utility-preserving of its class. Unlike the previous figure, however, we observe that $m = 2$ outperforms $m = 0$ in Census 3 and 4 for k greater than 370 and 493 respectively, whereas this only occurred in Census 2 for $\epsilon = 1$.

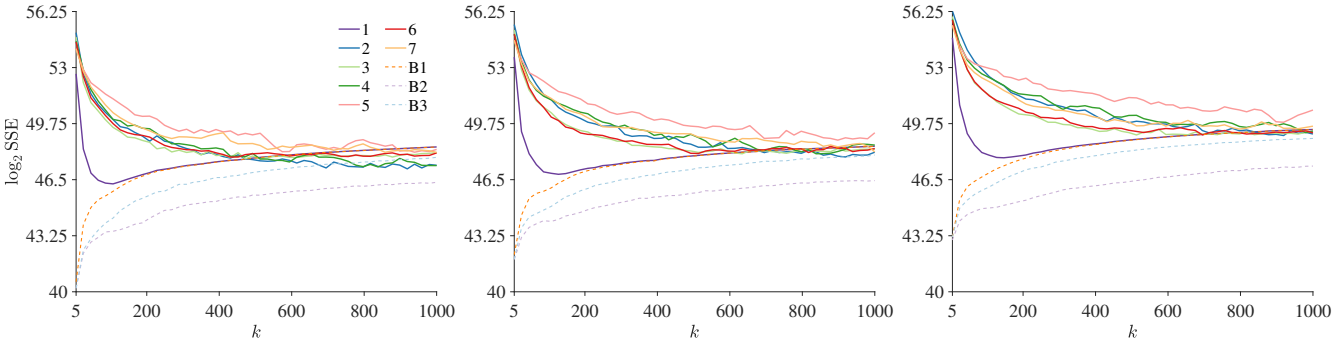


(a) EIA 2, $m = 1$. (b) EIA 3, $m = 1$. (c) EIA 4, $m = 1$.

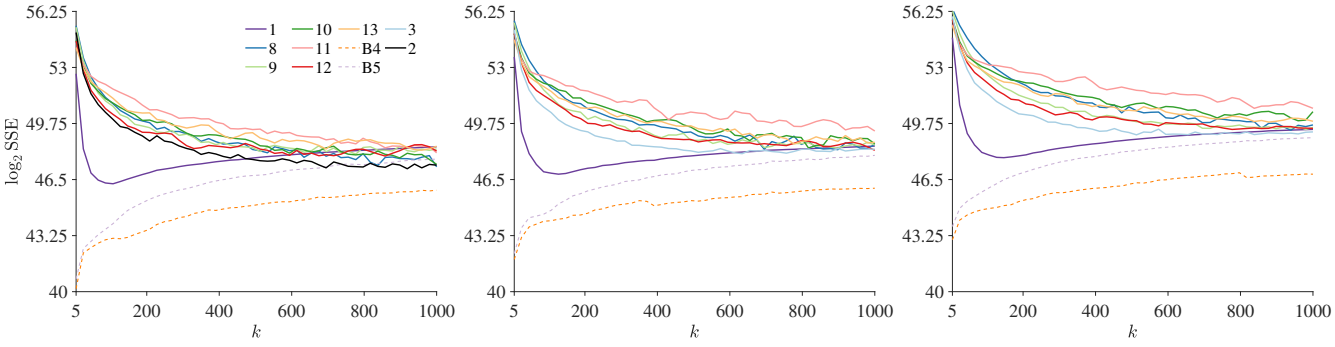


(d) EIA 2, $m = 2$. (e) EIA 3, $m = 2$. (f) EIA 4, $m = 2$.

Fig. 9: Information loss in the EIA data sets for coefficient-based microaggregation, MDAV and $\epsilon = 1$.



(a) EIA 2, $m = 1$. (b) EIA 3, $m = 1$. (c) EIA 4, $m = 1$.



(d) EIA 2, $m = 2$. (e) EIA 3, $m = 2$. (f) EIA 4, $m = 2$.

Fig. 10: Information loss in the EIA data sets for coefficient-based microaggregation, MDAV and $\epsilon = 2$.

In Figure 9, we test the EIA data set with $\epsilon = 1$. Here, we see a much larger error due to microaggregation than in the Census data set, which could be expected because EIA has approximately 4 times more records and SSE is an absolute measure of error. We also observe important differences in SSE between the standalone methods B2 and B3 on the one hand, and between B4 and B5 on the other, which reflects a stronger dependency among attributes in EIA; in the Census data set, the baseline methods showed extremely similar results.

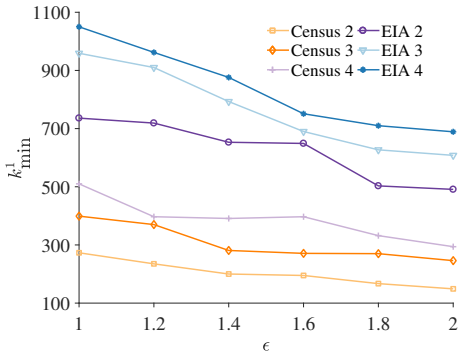


Fig. 11: Dependency of $k^1_{\min}$ on the value of the DP parameter ϵ. As ϵ increases, the performance of $m = 0$ is matched with smaller cluster sizes.

In comparison with Census, we notice that the SSE graphs of all methods decay much more slowly. Specifically, $k^1_{\min}$ is approximately 736, 959 and 1050 for EIA 2, EIA 3 and EIA 4, respectively, which corresponds to 39.40%, 41.71% and 48.57% increases with respect to Census. The result is that larger cluster sizes are required to effectively reduce the contribution of Laplace noise to the overall SSE.

Again, method 3 is the 1-microaggregation algorithm causing the least impact on EIA 3 and EIA 4. In EIA 2, however, this method is outperformed by method 2 for $k \geqslant 469$, which is consistent with the dependencies among attributes mentioned above. Note that the latter method differs from the former in that it incorporates the covariance coefficient q_{12}.

Finally, we check that none of the methods using coskewness improve on the best algorithms of $m = 1$. In this case, method 12 is again the most utility-preserving one, but it only beats $m = 0$ in EIA 2 and for $k \approx 1\,000$.

Figure 10 shows the case $\epsilon = 2$ for the same data set. Essentially, less privacy protection results in less distortion and this is roughly what we observe in each of the six figures. In particular, we note $k^1_{\min}$ now yields 491, 608 and 689 for the 2, 3 and 4-attribute versions of EIA, respectively, which are more practical cluster sizes. In addition, we check for the first time in these experiments that the methods corresponding to $m = 2$ are comparable to those of $m = 1$. Specifically, for $k \approx 1\,000$, methods 8 and 12 are slightly better than 2 and 3. Apart from these observations, the same conclusions drawn from Figure 9 apply here.

In Figure 11, we show $k^1_{\min}$ for different values of DP protection and check that, as ϵ increases, smaller microaggregation clusters are required to reach the performance of $m = 0$. The most important remark is that the ranges of values of k are reasonably practical in microaggregation. We must acknowledge, however, that the adequacy of medium to large cluster sizes will ultimately depend on the level of utility to be guaranteed, which in turn hinges on the database in question, particularly on the number of records and attributes, the dependencies among the latter, and the variability of the data.

Next, we evaluate the anonymization algorithms based on IR-MDAV. We only examine their impact on Census 4 and EIA 4 since the number of attributes has no effect here on microaggregation.

Figure 12 shows the information loss incurred by methods 20 to 26 for $\epsilon = 1$. As with the previous figures, the top-row results correspond to the case $m = 1$ and the bottom-row ones to $m = 2$. For the sake of comparison, the bottom-row figures also represent the performances of the best method for $m = 1$ and $m = 0$.

In Census 4, we observe that the only method outperforming $m = 0$ is number 21, which operates with joint sensitivities as described in Algorithm 1. This method performs similarly and slightly better, however, for k greater than 364. The other methods rely on Algorithm 2 and are shown to cause much greater impact on utility. The scenario in EIA 4 is rather similar and the only difference is that method 21 only beats method 20 for k close to 1 000.

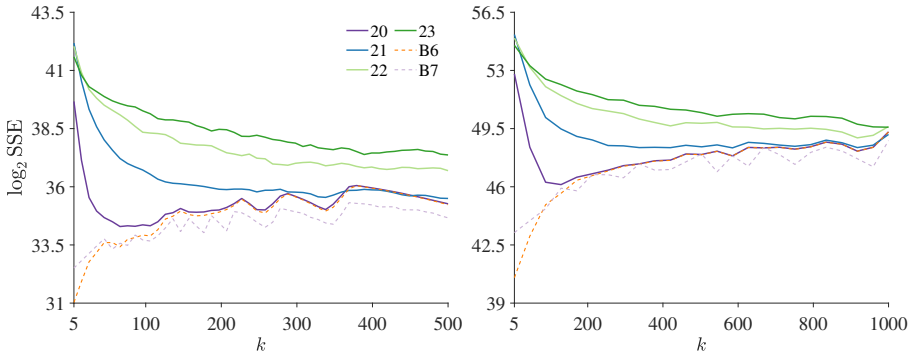
Figures 12(c) and (d) show a clear advantage of the methods that use the sample mean and variance, with respect to those that, in addition, employ skewness. Nonetheless, the baseline methods for Census 4 indicate that there is room for improvement and that total Laplace noise reduction might be attained for larger data sets.

On the other hand, we show in Figure 13 the case $\epsilon = 2$. Reducing the scale of the Laplace noise leads to smaller values of $k^{1,2}_{\min}$. In particular, we see that $k^1_{\min} = 351$ and $k^2_{\min} = 364$ in Census 4, and that the best methods for $m = 1, 2$ in EIA 4 cause less impact than $m = 0$ for values of k around 800.

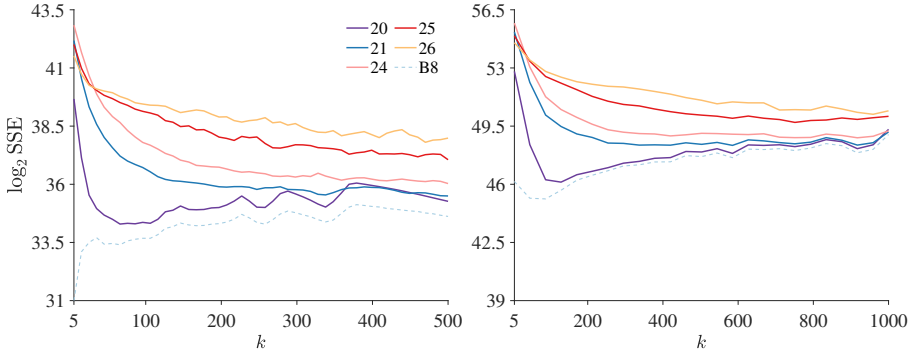
In our last series of experiments, we proceed to examine eigenvalue-based microaggregation. Recall that, for this class of algorithms, we use iDP rather than DP.

We analyze a single level of protection, namely, $\epsilon = 2$. As we shall show next, the results are so unfavorable even for this limited protection —in comparison with coefficient-based microaggregation—, that a single value of ϵ is sufficient to discard this class of microaggregation algorithms; taking a smaller ϵ would only lead to worse results.

Figure 14 shows the impact of using eigenvalue-based microaggregation as a sensitivity reduction mechanism. The main observation is that method 1, which is designed to protect coefficient-based microaggregated data sets and is plotted here as a reference, causes less distortion than any eigenvalue-based method for nearly all values of k and data sets. What is interesting about this observation is that all such methods (from 14 to 19) use local sensitivity to provide iDP. That is, their performance in fact represents a best-case scenario in terms of Laplace noise addition —when compared with those relying on coefficient-based microaggregation—, since the latter adjust noise to the global sensitivity; this implies that any conceivable protection of eigenvalue-based microaggregated data sets with DP would cause at least the same distortion



(a) Census 4, $m = 1$. (b) EIA 4, $m = 1$.



(c) Census 4, $m = 2$. (d) EIA 4, $m = 2$.

Fig. 12: Information loss for coefficient-based microaggregation, IR-MDAV and $\epsilon = 1$.

than with iDP. Hence, we can conclude that, for the data sets of our study, coefficient-based microaggregation comes at the cost of less information loss than eigenvalue-based microaggregation.

Having said this, we observe the best method is number 15, which operates with the sensitivity of each statistical coefficient separately (see Algorithm 2) and uses the sample mean and covariance matrix. In fact, all "separate" methods produce much less distortion than those utilizing joint sensitivities (see Algorithm 1). Interestingly, these latter methods (14 and 17) yield the worst results in our experiments.

In Figure 14, we also notice that strategy 1 causes less information loss than strategy 2 for $m = 1$ and $m = 2$ and almost all values of k. This means that a more balanced distribution of the privacy budget towards the covariance matrix produces better results. Finally, we note that, as the number of attributes grows, so does the gap between the reference method 1 and the methods using eigenvalue-based microaggregation.

6 Conclusions

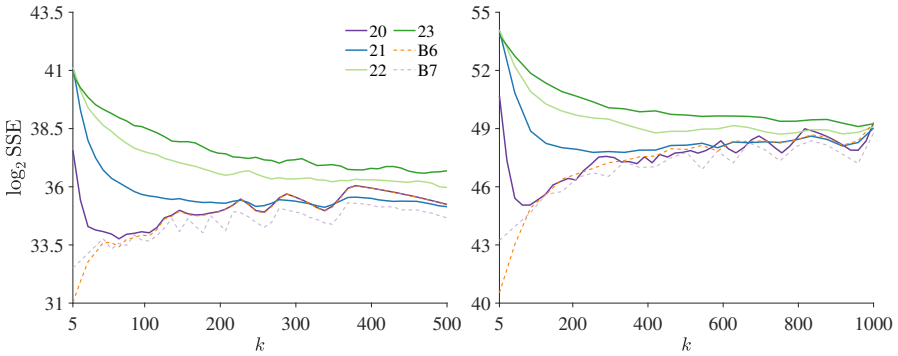
With the advent of the Internet and the development of sophisticated data analytics, the availability of massive amounts of information has increased the demand for data sharing. Before data can be shared among different parties, however, their holders must ensure the privacy of data subjects is protected.

DP is popular among the scientific community working in data anonymization, mainly because it does not make any assumption on the side knowledge available to attackers. While it offers strong privacy guarantees, its main practical limitation is the severe degradation in data utility it typically causes.

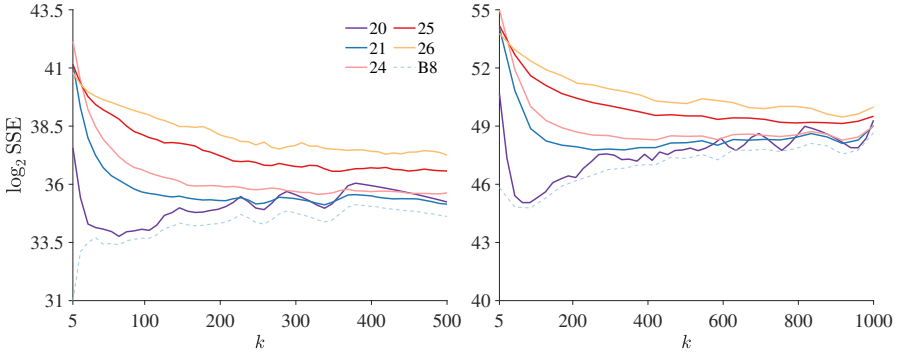
A variety of mechanisms, including microaggregation, have been investigated in order to reduce the sensitivity of the data and hence the noise required to enforce DP. However, although microaggregation may certainly keep perturbation under control, replacing the original records in each cluster by just their centroid attenuates the variability of data and thus the utility of the released data set.

In this work, we have proposed a generalization of classical microaggregation as a means of reducing sensitivity. Our generalization replaces the original records in a cluster not only by their centroid but also by additional statistics that reflect their variability and the dependencies among attributes more accurately. We have investigated two families of cross-moment microaggregation algorithms, one in which these statistics are given as matrix coefficients and another where they are provided as spectral decompositions.

Following the approach suggested by [35], we assume the target of protection is a data set microaggregated either of those two ways. We emphasize, however, that the disclosure risk limitation comes from the enforcement of DP via record-level perturbation. Accordingly, we have proposed three anonymization algorithms where microaggregation deals with all attributes together or with subsets of attributes separately. Our theoretical analysis has computed the sensitivity of



(a) Census 4, $m = 1$. (b) EIA 4, $m = 1$.



(c) Census 4, $m = 2$. (d) EIA 4, $m = 2$.

Fig. 13: Information loss for coefficient-based microaggregation, IR-MDAV and $\epsilon = 2$.

the coefficients of the covariance and coskewness matrices, and has applied fundamental results from matrix perturbation theory to derive sensitivity bounds on the eigenvalues and eigenvectors of such matrices.

Our extensive experimental evaluation demonstrates the suitability of utilizing cross-moment microaggregation to generate high-utility DP data sets. For the data sets under study, the results provide confirmatory evidence that coefficient-based microaggregation causes substantially less impact on data utility than eigenvalue-based microaggregation. We show that, for medium to large cluster sizes, 1,2-microaggregation is more utility-preserving than previous work relying on traditional microaggregation as a sensitivity reduction mechanism. Remarkably enough, for this range of sizes, we observe that our approach may provide higher utility and higher privacy than classical microaggregation. We acknowledge, however, that the suitability of such cluster sizes will depend on the level of utility to be ensured by the database curator, which in turn depends on the database itself, particularly on the number of records, attributes, dependencies among them and variability of the data.

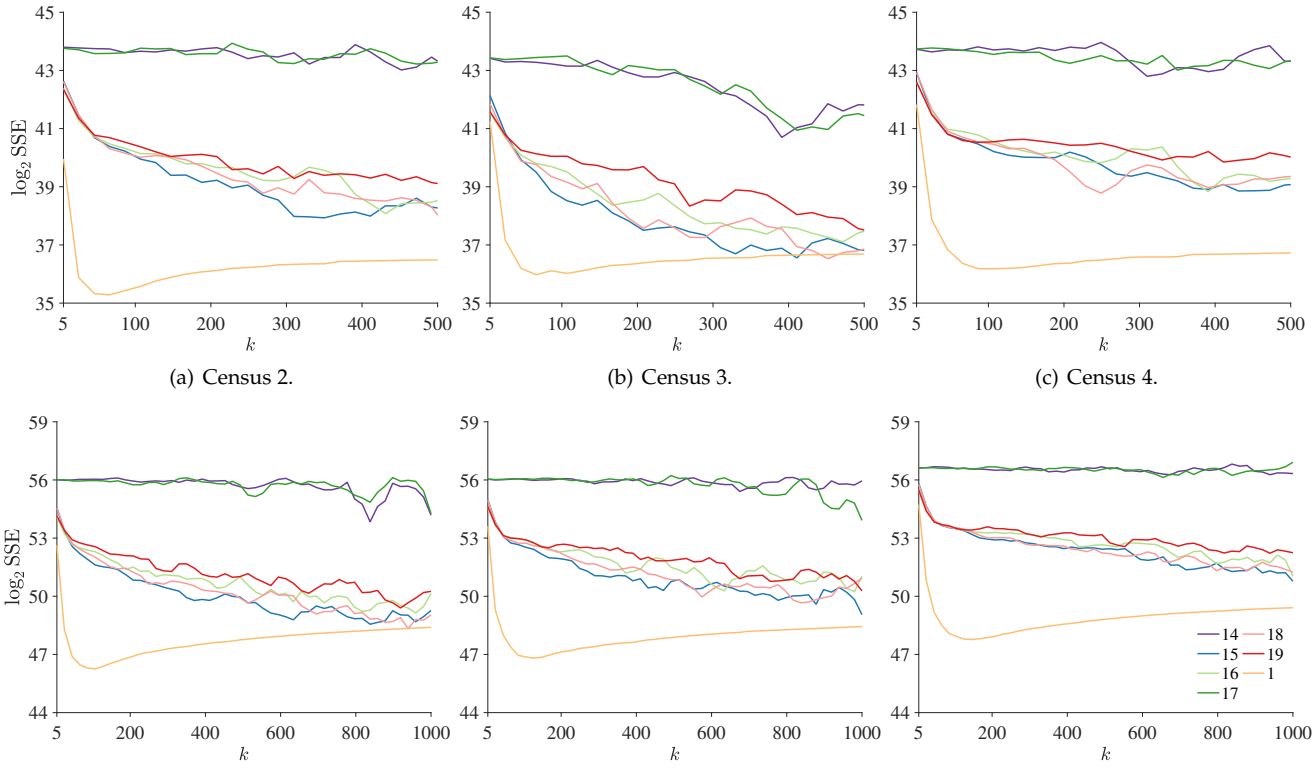
In the EIA data set, where there seem to be strong dependencies among attributes, we observe that publishing all covariance coefficients may cause less distortion than giving just the variances. In the Census data sets, where those dependencies are weaker, we note the opposite effect. Our findings also show that, for large k, few attributes and large ϵ, the methods exploiting coskewness are better than those not using it. In the other cases, on the contrary, $m = 1$ is essentially better than $m = 2$.

Acknowledgments and Disclaimer

The authors are thankful to A. Azzalini for his clarifications on the sampling of multivariate skew-normal distributions. Partial support to this work has been received from the European Commission (projects H2020-644024 "CLARUS" and H2020-700540 "CANVAS"), the Government of Catalonia (ICREA Academia Prize to J. Domingo-Ferrer), and the Spanish Government (projects TIN2014-57364-C2-1-R "Smart-Glacis" and TIN2016-80250-R "Sec-MCloud"). J. Parra-Arnau is the recipient of a Juan de la Cierva postdoctoral fellowship, FJCI-2014-19703, from the Spanish Ministry of Economy and Competitiveness. The 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] A. Azzalini, "A class of distributions which includes the normal ones," *Scand. J. Stat.*, vol. 12, pp. 171–178, 1985.
[2] A. Azzalini and A. D. Valle, "The multivariate skew-normal distribution," *Biometrika*, vol. 83, pp. 715–726, 1996.
[3] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc., Data Min. (KDD)*. ACM, 2010, pp. 503–512.



(a) Census 2. (b) Census 3. (c) Census 4.

(d) EIA 2. (e) EIA 3. (f) EIA 4.

Fig. 14: Information loss of eigenvalue-based microaggregated data sets that are protected with 2-individual DP.

[4] K. Boudt, D. Cornilly, and T. Verdonck, "A coskewness shrinkage approach for estimating the skewness of linear combinations of random variables," 2016. [Online]. Available: https://ssrn.com/abstract=2839781

[5] R. Brand, J. Domingo-Ferrer, and J. M. Mateo-Sanz, "Computational Aspects of Statistical Confidentiality project," European project IST-2000-25069 CASC, Tech. Rep., 2001–2004. [Online]. Available: http://neon.vb.cbs.nl/casc/CASCtestsets.htm

[6] T. Chanyaswad, C. Liu, and P. Mittal, "Coupling dimensionality reduction with generative model for non-interactive private data release," in *arXiv: 1709.00054v1*, Aug. 2017.

[7] G. Cormode, C. M. Procopiuc, D. Srivastava, and T. T. L. Tran, "Differentially private publication of sparse data," *CoRR*, vol. abs/1103.0825, 2011. [Online]. Available: http://arxiv.org/abs/1103.0825

[8] D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: The small aggregates method," in *Proc. Symp. Design, Anal. Longitudinal Surveys, Stat. Canada*, Ottawa, Canada, 1993, pp. 195–204.

[9] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 189–201, 2002.

[10] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogenerous k-anonymity through microaggregation," *Data Min., Knowl. Disc.*, vol. 11, no. 2, pp. 195–212, 2005.

[11] R. M. Dudley, "Mathematical statistics," 2003. [Online]. Available: http://ocw.mit.edu

[12] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Lang., Program.* Springer-Verlag, 2006, pp. 1–12.

[13] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings*, 2006, pp. 486–503.

[14] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. Conf.* Springer-Verlag, 2006, pp. 265–284.

[15] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, 2006, pp. 265–284.

[16] R. A. Fisher, "Moments and product moments of sampling distributions," *Proc. London Math. Soc.*, vol. s2-30, no. 1, pp. 199–238, 1930.

[17] M. Hardt, K. Ligett, and F. McSherry, "A simple and practical algorithm for differentially private data release," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'12. USA: Curran Associates Inc., 2012, pp. 2339–2347. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999325.2999396

[18] M. Hay, V. Rastogi, G. Miklau, and D. Suciu, "Boosting the accuracy of differentially private histograms through consistency," *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1021–1032, Sep. 2010. [Online]. Available: http://dx.doi.org/10.14778/1920841.1920970

[19] R. Horst and H. Tuy, *Global Optimization*, 3rd ed. New York, PA: Springer-Verlag, 1996.

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

[21] R. Kitchin, *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

[22] J. Lee and C. Clifton, "How much is enough? choosing ϵ for differential privacy," in *Proc. Int. Inform. Secur. (ISC)*. Springer-Verlag, 2011, pp. 325–340.

[23] C. Li, M. Hay, G. Miklau, and Y. Wang, "A data-and workload-aware algorithm for range queries under differential privacy," *VLDB J.*, vol. 7, no. 5, pp. 341–352, 2014.

[24] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor, "Optimizing linear counting queries under differential privacy," in *Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, ser. PODS '10. New York, NY, USA: ACM, 2010, pp. 123–134. [Online]. Available: http://doi.acm.org/10.1145/1807085.1807104

[25] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Istanbul, Turkey, Apr. 2007, pp. 106–115.

[26] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian, "l-Diversity: Privacy beyond k-anonymity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Atlanta, GA, Apr. 2006, p. 24.

[27] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber, "Privacy: theory meets practice on the map," in *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE 2008)*, 2008, pp. 277–286.

[28] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," ser. Proc. ACM SIGMOD Int. Conf. Manage. Data. ACM, 2009, pp. 19–30.

[29] A. Oganian and J. Domingo-Ferrer, "On the complexity of optimal microaggregation for statistical disclosure control," *UNECE Stat. J.*, vol. 18, no. 4, pp. 345–354, Apr. 2001.

[30] H. Qi and D. Sun, "A quadratically convergent newton method for computing the nearest correlation matrix," *SIAM J. Matrix Anal., Appl. (SIMAP)*, vol. 28, no. 2, pp. 360–385, 2006.

[31] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-Anonymity and its enforcement through generalization and suppression," SRI Int., Tech. Rep., 1998.

[32] D. Sánchez, J. Domingo-Ferrer, and S. Martínez, "Improving the utility of differential privacy via univariate microaggregation," in *Priv. Stat. Databases (PSD)*, vol. 8744. Springer-Verlag, 2014, pp. 130–142.

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

[34] R. Sarathy and K. Muralidhar, "Evaluating laplace noise addition to satisfy differential privacy for numeric data," *Transactions on Data Privacy*, vol. 4, no. 1, pp. 1–17, Apr. 2011.

[35] J. Soria-Comas and J. Domingo-Ferrer, "Differentially private data sets based on microaggregation and record perturbation," in *Proc. Int. Conf. Model. Decisions Artif. Intell.*, Oct. 2017, pp. 119–131.

[36] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Improving the utility of differentially private data releases via k-anonymity," in *Proc. IEEE Int. Conf. Trust, Security, Priv. Comput., Commun. (TrustCom)*, Jul. 2013, pp. 372–379.

[37] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based k-anonymity," *VLDB J.*, vol. 23, no. 5, pp. 771–794, Oct. 2014.

[38] J. Soria-Comas, J. Domingo-Ferrer, D. Snchez, and D. Megas, "Individual differential privacy: a utility-preserving formulation of differential privacy guarantees," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 6, pp. 1418–1429, Jun. 2017.

[39] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, 1st ed. Academic-Press, 1990.

[40] B. Xi, M. Kantarcioglu, and A. Inan, "Mixture of gaussian models and bayes error under differential privacy," in *Proc. ACM Conf. Data, Appl. Secur., Priv. (CODASPY)*. ACM, 2011, pp. 179–190.

[41] Y. Xiao, L. Xiong, and C. Yuan, "Differentially private data release through multidimensional partitioning," in *Secure Data Management*, ser. Lecture Notes in Computer Science, W. Jonker and M. Petković, Eds. Springer Berlin Heidelberg, 2010, vol. 6358, pp. 150–168.

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