VLDB Journal manuscript No.
(will be inserted by the editor)

Enhancing Data Utility in Differential Privacy via Microaggregation-based k-Anonymity

Jordi Soria-Comas · Josep Domingo-Ferrer · David Sánchez · Sergio Martínez

Received: date / Accepted: date

Abstract It is not uncommon in the data anonymization literature to oppose the "old" k-anonymity model to the "new" differential privacy model, which offers more robust privacy guarantees. Yet, it is often disregarded that the utility of the anonymized results provided by differential privacy is quite limited, due to the amount of noise that needs to be added to the output, or because utility can only be guaranteed for a restricted type of queries. This is in contrast with k-anonymity mechanisms, which make no assumptions on the uses of anonymized data while focusing on preserving data utility from a general perspective. In this paper, we show that a synergy between differential privacy and k-anonymity can be found: k-anonymity can help improving the utility of differentially private responses to arbitrary queries. We devote special attention to the utility improvement of differentially private published data sets. Specifically, we show that the amount of noise required to fulfill ε-differential privacy can be reduced if noise is added to a k-anonymous version of the data set, where k-anonymity is reached through a specially designed microaggregation of all attributes. As a result of noise reduction, the general analytical utility of the anonymized output is increased. The theoretical benefits of our proposal are illustrated in a practical setting with an empirical evaluation on three data sets.

Keywords Privacy-preserving data publishing, Differential privacy, k-Anonymity, Microaggregation, Data utility

J. Domingo-Ferrer, J. Soria-Comas, D. Sánchez, S. Martínez
UNESCO Chair in Data Privacy, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Av. Països Catalans 26, E-43007 Tarragona, Catalonia.
E-mail: {josep.domingo, jordi.soria, david.sanchez, sergio.martinezl}@urv.cat

1 Introduction

Publishing microdata (*e.g.*, responses to polls, census information, healthcare records) collected by organizations such as statistical agencies is of great interest for the data analysis community. At the same time, microdata may contain confidential information about individuals. To overcome this privacy threat, data should be anonymized before making them available for secondary use [11, 58, 32].

In the last two decades, several models for data anonymization have been proposed in the literature. One of the best-known and widely used is k-anonymity [51], which aims at making each record indistinguishable from, at least, $k-1$ other records. The usual computational procedure to reach k-anonymity is a combination of global or local attribute recoding and local suppressions [50, 57, 1, 29]. An alternative procedure, especially suitable for attributes with no obvious generalization hierarchy (like the numerical ones), is microaggregation [19, 17]. Whatever the computational procedure, k-anonymity assumes that identifiers are suppressed from the data to be released and it focuses on masking quasi-identifier attributes; these are attributes (*e.g.*, Age, Gender, Zipcode and Race) that may enable re-identifying the respondent of a record because they are linkable to analogous attributes available in external identified data sources (like electoral rolls, phone books, etc.). k-Anonymity does not mask confidential attributes (*e.g.*, salary, health condition, political preferences, etc.) unless they are also quasi-identifiers. While k-anonymity has been shown to provide reasonably useful anonymized results, especially for small k, it is also vulnerable to attacks based on the possible lack of diversity of the non-anonymized confidential attributes or on additional background knowledge available to the attacker [39,59,35,20]. Several refinements of the k-anonymity model, including l-diversity [39] and t-closeness [35], have been proposed to prevent this kind of attacks.

On the other hand, ε-differential privacy [23] is a more recent and rigorous privacy model that makes no assumptions about the attacker's background knowledge. In a nutshell, it guarantees that the anonymization output is insensitive (up to a factor dependent on ε) to modifications of individual input records. In this way, the privacy of an individual is not compromised by her presence in the data set, which is a much more robust guarantee than the one offered by the k-anonymity model. To do so, ε-differential privacy requires adding an amount of noise to obtain the anonymized output that depends on the variability of the actual non-anonymized values. ε-Differential privacy was originally proposed for the *interactive* scenario, in which, instead of releasing an anonymized version of the data set, the anonymizer returns noise-added answers to interactive queries. The accuracy of the response to a query depends on the sensitivity of the query and the intended level of privacy. Given a query f and an intended level ε of differential privacy, there are two possibilities to improve the utility of the response: using a noise addition strategy that provides a better adjustment to the query under consideration, or replacing f by a modified query f' that approximates f and has less sensitivity. In this paper, we take the second approach, as explained below.

In summary, we can conclude that k-anonymity enables general-purpose data publication with reasonable utility regardless of the data uses, at the cost of some privacy weaknesses. On the contrary, ε-differential privacy offers a very robust privacy guarantee at the cost of substantially limiting the utility of the anonymized outputs.

1.1 Contribution and plan of this paper

We show here that a synergy between k-anonymity and ε-differential privacy can be found in order to achieve more accurate and general-purpose ε-differential privacy: k-anonymity can help increasing the utility of differentially private query outputs while making as few assumptions on the type of queries as k-anonymity does. Specifically, we show that the amount of noise required to fulfill ε-differential privacy can be greatly reduced if the query is run on a k-anonymous version of the data set obtained through microaggregation of all attributes (instead of running it on the raw input data). In other words, we propose to replace any query f by $f \circ M$, where M is a microaggregation function achieving k-anonymity. This implies that, instead of running f on the actual data set, we run it on a microaggregated version of the data set.

The rationale is that the microaggregation performed to achieve k-anonymity helps reducing the sensitivity of the input versus modifications of individual records; hence, it helps reducing the amount of noise to be added to achieve ε-differential privacy. As a result, data utility can be improved without renouncing the strong privacy guarantee of ε-differential privacy.

Section 2 reviews related work and background on k-anonymity and ε-differential privacy. Section 3 discusses the use of a k-anonymous microaggregation step prior to the evaluation of a query function as a means to reduce the query sensitivity, thereby reducing the noise required to attain differential privacy. Section 4 proposes a general algorithm for generating ε-differentially private data sets that employs the k-anonymous microaggregation procedure described earlier. Implementation details for data sets with numerical and categorical attributes are given. Section 5 reports an empirical evaluation of the differentially private outputs obtained from three data sets via k-anonymous microaggregation; the output is compared against standard k-anonymity and ε-differential privacy mechanisms regarding data utility and disclosure risk. Section 6 presents the conclusions and proposes some lines for future research.

This paper is an extension of the preliminary research described in [56]. Section 2.1 in the present paper is new. Most of Section 3 is new as well, namely the extension of our approach to arbitrary queries (rather than just the identity query needed for data publishing) and the theoretical results on insensitive microaggregation. Also new are Sections 4.4 (extension of our approach for categorical attributes), 4.5 (semantic distance for categorical values) and 4.6 (integration of heterogeneous attribute types). Finally, the section on empirical results has been expanded very substantially: in addition to the numerical data set of [56], we now feature two more data sets, one of them categorical and the other one mixed numerical-categorical. Finally, Sections 5.3 and 5.4 are new.

2 Related work and background

2.1 Related work

Compared to the general-purpose data publication offered by k-anonymity, which makes no assumptions on the uses of published data, the interactive scenario of

ε-differential privacy severely limits data analysis, because it only allows answering a limited number of queries. This limitation can be overcome by generating a differentially private data set. Such a data set can be released and then used to answer an unlimited number of queries. The usual approach to release differentially private microdata sets is based on histogram queries [62, 63]; that is, on approximating the data distribution by partitioning the data domain and counting the number of records in each partition set. To prevent the counts from leaking too much information they are computed in a differentially private manner. Apart from the counts, partitioning can also reveal information. One way to prevent partitioning from leaking information consists in using a predefined partition that is independent of the actual data under consideration (*e.g.* by using a grid [40]). The accuracy of the approximation obtained via histogram queries depends on the number of records contained in each of the histogram bins: the more records, the less relative error. For data sets with sparsely populated regions, using a predefined partition may be problematic. Several strategies have been proposed to improve the accuracy of differentially private count (histogram) queries, which we next review. In [31] consistency constraints between a set of queries are exploited to increase accuracy. In [61] a wavelet transform is applied to the data and the noise is added in the frequency domain. In [63,36] the histogram bins are adjusted to the actual data. In [9], the authors consider differential privacy of attributes whose domain is ordered and has moderate to large cardinality (*e.g.* numerical attributes); the attribute domain is represented as a tree, which is decomposed in order to increase the accuracy of answers to count queries (multi-dimensional range queries). Our approach is similar to [61] in that we apply a prior transformation to the original data. However, our proposal differs from all previous ones in that it is not limited to histogram queries and it allows dealing with any type of attributes (ordered or not ordered).

As we apply a microaggregation step before adding the random noise to satisfy differential privacy, we can think of it as constructing a differentially private data set out of a k-anonymous data set. However, it is important to remark that we do not aim at finding any relation between k-anonymity and differential privacy (beyond the fact that the microaggregation step lets us reduce the magnitude of the noise required to satisfy differential privacy). Readers interested in comparative studies discussing the merits, problems and possible bridges between k-anonymity and differential privacy may refer to [8,10]; these recent references conclude that there is room both for syntactic privacy models (k-anonymity-like) and differential privacy. In this sense, a practical method is presented in [38], in which it is shown that a constrained k-anonymization (which does not depend on the input data) preceded by random sampling can satisfy a relaxed form of differential privacy. On the contrary, we use k-anonymity as a means to reduce the information loss caused by standard differential privacy.

2.2 Background on k-anonymity and microaggregation

As mentioned above, k-anonymity [51,50,57] attempts to thwart re-identification. It can be defined as follows.

Definition 1 (k-Anonymity) A data set is said to satisfy k-anonymity for an integer $k > 1$ if, for each combination of values of quasi-identifier attributes, at least k records exist in the data set sharing that combination.

Several criticisms have been raised against k-anonymity since it appeared. Although k-anonymity is able to prevent identity disclosure (re-identification is only possible with probability $1/k$), it may not protect against attribute disclosure. For example, let a medical data set contain quasi-identifier attributes Age, Gender, Zipcode and Race, and confidential attribute AIDS (whose values can be Yes or No). Imagine that we 3-anonymize this data set, but a group of three records sharing a certain combination of quasi-identifier attribute values also shares the confidential attribute value AIDS=Yes. In this case, if the intruder can establish that her target respondent's record is within that group (because it is the only group with compatible Age, Gender, Zipcode and Race), the intruder learns that the target respondent suffers from AIDS. Several fixes/alternatives to k-anonymity also based on the idea of data set partitioning have appeared: l-diversity [39], t-closeness [35], (α, k)-anonymity [59], etc. However, none of those alternatives is free from shortcomings, see [20] for a critical survey.

In [19], it is shown how to achieve k-anonymity via microaggregation. Microaggregation [17] is a family of anonymization algorithms for data sets that works 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; ii) records within a cluster are as similar as possible.
- Second, each record within each cluster is 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,

Algorithm 1 Maximum distance to average record (MDAV)

```
let $X$ be the original data set
let $k$ be the minimal cluster size

while $|X| \geq 3k$ do
    $\overline{x}$ ←average record of $X$
    $x_1$ ←most distant record to $\overline{x}$ in $X$
    $x_2$ ←most distant record to $x_1$ in $X$
    Form a cluster with $x_1$ and its $k-1$ closest records
    Form a cluster with $x_2$ and its $k-1$ closest records
    Remove the clustered records from $X$
end while

if $|X| \geq 2k$ then
    $\overline{x}$ ←average record of $X$
    $x_1$ ←most distant record to $\overline{x}$ in $X$
    Form a cluster with $x_1$ and its closest $k-1$ records
    Remove the clustered records from $X$
end if

Form a new cluster with the remaining records.

In each record within each formed cluster, replace the value
of each quasi-identifier attribute with the average value of the
attribute over the cluster.
```

the resulting data set is k-anonymous. In [19] a simple microaggregation heuristic called MDAV is described, in which all clusters have exactly k records, except the last one, which has between k and $2k-1$ records. As the internals of MDAV will be required in Section 3, we recall the MDAV algorithm (see Algorithm 1).

2.3 Background on differential privacy

Differential privacy was originally proposed by [23] 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.

Definition 2 (ε-Differential privacy) A randomized function κ gives ε-differential privacy if, for all data sets X_1, X_2 such that one can be obtained from the other by modifying a single record, and all $S \subset Range(\kappa)$, it holds

$$P(\kappa(X_1) \in S) \leq \exp(\varepsilon) \times P(\kappa(X_2) \in S). \tag{1}$$

The computational mechanism to attain ε-differential privacy is often called ε-differentially private *sanitizer.* A usual sanitization approach is noise addition: first, the real value $f(X)$ of the response to a certain user query f is computed, and then a random noise, say $Y(X)$, is added to mask $f(X)$, that is, a randomized response $\kappa(X) = f(X) + Y(X)$ is returned. To generate $Y(X)$, a common choice is to use a Laplace distribution with zero mean and $\Delta(f)/\varepsilon$ scale parameter, where:

- ε is the differential privacy parameter;
- $\Delta(f)$ is the L_1-sensitivity of f, that is, the maximum variation of the query function between neighbor data sets, *i.e.*, sets differing in at most one record.

Specifically, the density function of the Laplace noise is

$$p(x) = \frac{\varepsilon}{2\Delta(f)} e^{-|x|\varepsilon/\Delta(f)}.$$

Notice that, for fixed ε, the higher the sensitivity $\Delta(f)$ of the query function f, the more Laplace noise is added: indeed, satisfying the ε-differential privacy definition (Definition 2) requires more noise when the query function f can vary strongly between neighbor data sets. Also, for fixed $\Delta(f)$, the smaller ε, the more Laplace noise is added: when ε is very small, Definition 2 almost requires that the probabilities on both sides of Equation (1) be equal, which requires the randomized function $\kappa(\cdot) = f(\cdot) + Y(\cdot)$ to yield very similar results for all pairs of neighbor data sets; adding a lot of noise is a way to achieve this.

Differential privacy was also proposed for the non-interactive setting in [3, 25, 30, 7]. Even though a non-interactive data release can be used to answer an arbitrarily large number of queries, in all these proposals, this is obtained at the cost of offering utility guarantees only for a restricted class of queries [3], typically count queries. This contrasts with the general-purpose utility-preserving data release offered by the k-anonymity model.

In [45], an ε-differentially private sanitizer based on generalization is proposed for the non-interactive setting. The method first converts the microdata file into a contingency table by accumulating in each table cell the count of records that share a combination of categories of certain attributes (classification attributes). It then generalizes the contingency table by using coarser categories for the classification attributes; this results in higher counts for the table cells, which are much larger than the noise that needs to be added to reach differential privacy. The limitations of this method are that: its analytical utility is restricted to (coarsened) count queries; the aggregations it performs are constrained by the generalization hierarchies of the selected classification attributes. In contrast, we use a free microaggregation only constrained by the k-anonymity requirement, which yields differentially private microdata that can be used for any type of queries.

3 Differential privacy through k-anonymous microaggregation

Differential privacy and microaggregation offer quite different disclosure limitation guarantees. Differential privacy is introduced in a query-response environment and offers probabilistic guarantees that the contribution of any single individual to the query response is limited, while microaggregation is used to protect microdata releases and works by clustering groups of individuals and replacing them by the group centroid, regardless of the query types the user will be interested in. When applied to the quasi-identifier attributes, microaggregation achieves k-anonymity. In spite of those differences, we can leverage microaggregation masking to decrease the amount of random noise required to attain differentially private outputs; besides, these outputs will be as independent of their subsequent uses as the outputs of k-anonymous mechanisms.

Let X be a data set with attributes $A_1, \ldots, A_m$, and $\overline{X}$ be a microaggregated X with minimal cluster size k. Let M be a microaggregation function that takes as input a data set, and outputs a microaggregated version of it: $M(X) = \overline{X}$. Let f be an arbitrary query function for which an ε-differentially private response is requested. A typical differentially private mechanism takes these steps: capture the query f, compute the real response $f(X)$, and output a masked value $f(X) + N$, where N is a random noise whose magnitude is adjusted to the sensitivity of f.

To improve the utility of an ε-differentially private response to f, we seek to minimize the distortion introduced by the random noise N. Two main approaches are used in the literature. In the first approach, a random noise is used that allows for a finer calibration to the query f under consideration. For instance, if the variability of the query f is highly dependent on the actual data set X, using a data-dependent noise (such as in [46]) would probably reduce the utility loss. In the second approach, the query function f is modified so that the new query function is less sensitive to modifications of a record in the data set (paper [45], mentioned in Section 2, exemplifies this approach).

Our proposal falls in the second approach above: we replace the original query function f by $f \circ M$, that is, we run the query f on the microaggregated data set $\overline{X}$. Notice that, to improve the utility of differentially private outputs and unlike related works on differential privacy, we make no assumptions on the type of queries the user will be interested in. In differentially private mechanisms there is usually a single source of error: the noise introduced to comply with Inequality (1). With the approximation of the query function f by $f \circ M$ we introduce a second source of error: the error due to the approximation of f by $f \circ M$ (that is, the error due to computing f over $\overline{X}$ instead of over X). We expect the sum of the two errors in the computation of $f \circ M$ to be smaller than the error we had for f. It can be shown that the sensitivity of $f \circ M$ is smaller than the sensitivity of f: the actual reduction of the sensitivity depends on f (as part of the application to the generation of differentially private data sets, we show that for a query that returns an individual's data the sensitivity is divided by the cluster size used in the microaggregation). Expressing the error due to the approximation of f by $f \circ M$ in terms of equivalent noise under differential privacy is not possible in general, as the former error depends not only on the query function, but also on the actual data and the microaggregation algorithm used.

Since the k-anonymous data set $\overline{X}$ is formed by the centroids of the clusters (*i.e.*, the average records), for the sensitivity of the queries $f \circ M$ to be effectively reduced the centroid must be stable against modifications of one record in the original data set X. This means that modification of one record in the original data set X should only slightly affect the centroids in the microaggregated data set. Although this will hold for most of the clusters yielded by any microaggregation algorithm, we need it to hold for *all* clusters in order to effectively reduce the sensitivity.

Not all microaggregation algorithms satisfy the above requirement; for instance, if the microaggregation algorithm could generate a completely unrelated set of clusters after modification of a single record in X, the effect on the centroids could be large. As we are modifying one record in X, the best we can expect is a set of clusters that differ in one record from the original set of clusters. Microaggregation algorithms with this property lead to the greatest reduction in the query sensitivity; we refer to them as *insensitive* microaggregation algorithms.

Definition 3 (Insensitive microaggregation) Let X be a data set, M a microaggregation algorithm, and let $\{C_1, \ldots, C_n\}$ be the set of clusters that result from running M on X. Let X' be a data set that differs from X in a single record, and $\{C'_1, \ldots, C'_n\}$ be the clusters produced by running M on X'. We say that M is insensitive to the input data if, for every pair of data sets X and X' differing in a single record, there is a bijection between the set of clusters $\{C_1, \ldots, C_n\}$ and the set of clusters $\{C'_1, \ldots, C'_n\}$ such that each pair of corresponding clusters differs at most in a single record.

Since for an insensitive microaggregation algorithm corresponding clusters differ at most in one record, bounding the variability of the centroid is simple. For instance, for numerical data, when computing the centroid as the mean, the maximum change for each attribute equals the size of the range of the attribute divided by k. If the microaggregation was not insensitive, a single modification in X might lead to completely different clusters, and hence to large variability in the centroids.

The output of microaggregation algorithms is usually highly dependent on the input data. On the positive side, this leads to greater within-cluster homogeneity and hence less information loss. On the negative side, modifying a single record in the input data may lead to completely different clusters; in other words, such algorithms are not insensitive to the input data as per Definition 3. We illustrate this fact for MDAV. Figure 1 shows the clusters generated by MDAV for a toy data set X consisting of 15 records with two attributes, before and after modifying a single record. In MDAV, we use the Euclidean distance and $k = 5$. Two of the clusters in the original data set differ by more than one record from the respective most similar clusters in the modified data set. Therefore, no mapping between clusters of both data sets exists that satisfies the requirements of Definition 3. The centroids of the clusters are represented by a cross. A large change in the centroids between the original and the modified data sets can be observed.

We want to turn MDAV into an insensitive microaggregation algorithm, so that it can be used as the microaggregation algorithm to generate $\overline{X}$. MDAV depends on two parameters: the minimal cluster size k, and the distance function d used to measure the distance between records. Modifying k does not help making MDAV insensitive: similar examples to the ones in Figure 1 can easily be proposed for any $k > 1$; on the other hand, setting $k = 1$ does make MDAV insensitive, but it is equivalent to not performing any microaggregation at all. Next, we see that MDAV is insensitive if the distance function d is consistent with a total order relation.

Definition 4 A distance function $d : X \times X \to \mathbb{R}$ is said to be consistent with an order relation $\leq_X$ if $d(x, y) \leq d(x, z)$ whenever $x \leq_X y \leq_X z$.

Proposition 1 *Let X be a data set equipped with a total order relation $\leq_X$. Let $d : X \times X \to \mathbb{R}$ be a distance function consistent with $\leq_X$. MDAV with distance d satisfies the insensitivity condition (Definition 3).*

Proof. When the distance d is consistent with a total order, MDAV with cluster size k reduces to iteratively taking sets with cardinality k from the extremes, until less than k records are left; the remaining records form the last cluster. Let $x_1, \ldots, x_n$ be the elements of X sorted according to $\leq_X$. MDAV generates a set clusters of the form:

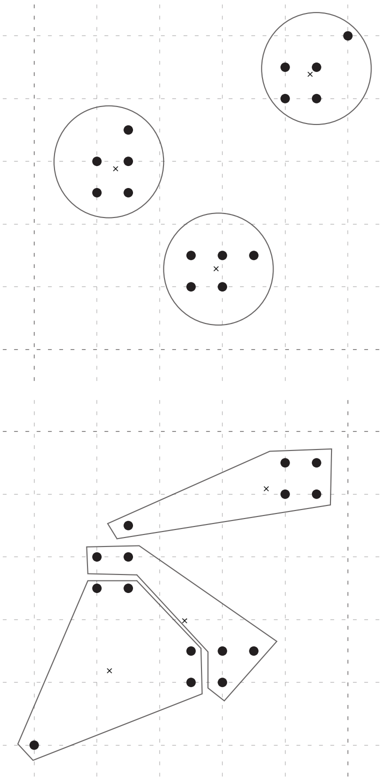

Fig. 1 MDAV clusters and centroids with $k = 5$. Top, original data set X; bottom, data set after modifying one record in X.

$$\{x_1, \ldots, x_k\}, \ldots, \{x_{n-k+1}, \ldots, x_n\}.$$

We want to check that modifying a single record of X leads to a set of clusters that differ in at most one element. Suppose that we modify record x by setting it to x', and let X' be the modified data set. Without loss of generality, we assume that $x \leq_X x'$; the proof is similar for the case $x' \leq_X x$.

Let C be the cluster of X that contains x, and C' the cluster of X' that contains x'. Let m be the minimum of the elements in C, and let M be the maximum of the elements in C'. As MDAV takes groups of k records from the extremes, the clusters of X whose elements are all inferior to m, or all superior to M remain unmodified in X'. Therefore, we can assume that x belongs to the leftmost cluster of X, and x' belongs to the rightmost cluster in X'.

Let $C_1, \ldots, C_m$ and $C'_1, \ldots, C'_m$ be, respectively, the clusters of X and X', ordered according to $\leq_X$. Let x^i_1 and $x^i_{j_i}$ be the minimum and the maximum of the elements of C_i: $C_i = \{z \in X | x^i_1 \leq z \leq x^i_{j_i}\}$. Cluster C'_1 contains the same elements as C_1 except for x that has been removed from C'_1 and for x^2_1 that has

Algorithm 2 General form of a microaggregation algorithm with fixed cluster size

```
let $X$ be the original data set
let $k$ be the minimal cluster size

set $i := 0$
while $|X| \geq 2k$ do
    $C_i \leftarrow k$ smallest elements from $X$ according to $\leq_i$
    $X := X \setminus C_i$
    $i := i + 1$
end while
$\overline{X} \leftarrow$Replace each record $r \in X$ by the centroid of its cluster
return $\overline{X}$
```

been added to C'_1, $C'_1 = (C_1 \cup \{x_1^2\}) \setminus \{x\}$. Clusters $C'_2, \ldots, C'_{m-1}$ contain the same elements as the respective cluster $C_2, \ldots, C_{m-1}$, except for x_1^i that has been removed from C'_i and x_1^{i+1} that has been added to C'_i. Cluster C'_m contains the same elements as C_m except for x_1^m that has been removed from $C'_m$and x' that has been added to C'_m. Therefore, clusters C_i and C'_i differ in a single record for all i, which completes the proof. □

We have seen that, when the distance function is consistent with a total order relation, MDAV is insensitive. Now, we want to determine the necessary conditions for an arbitrary microaggregation algorithm to be insensitive. Algorithm 2 describes the general form of a microaggregation algorithm with fixed cluster size k. Essentially it keeps selecting groups of k records, until less than $2k$ records are left; the remaining records form the last cluster, whose size is between k and $2k-1$. Generating each cluster requires a selection criterion to prioritize some elements over the others. We can think of this prioritization as an order relation $\leq_i$, and the selection criterion for constructing the cluster C_i to be "select the k smallest records according to $\leq_i$". Note that the prioritization used to generate different clusters need not be the same; for instance, MDAV selects the remaining element that is farthest from the average of remaining points, and prioritizes based on the distance to it.

Let X and X' be a pair of data sets that differ in one record. For Algorithm 2 to be insensitive, the sequence of orders $\leq_i$ must be constant across executions of the algorithm; to see this, note that, if one of the orders $\leq_i$ changed, we could easily construct data sets X and X' such that cluster C_i in X would differ by more than one record from its corresponding cluster in X', and hence the algorithm would not be insensitive.

Another requirement for Algorithm 2 to be insensitive is that the priority assigned by $\leq_i$ to any two different elements must be different. If there were different elements sharing the same priority, we could end up with clusters that differ by more than one record. For instance, assume that the sets X and X' are such that $X' = (X \setminus \{x\}) \cup \{x'\}$, and assume that x belongs to cluster C_i and x' belongs to cluster C'_i. Clusters C_i and C'_i already differ in one element, so for the clustering to be insensitive all the other records in these clusters must be equal. If there was a pair of elements, $y \neq y'$, with the same priority, and if only one of them was included in each of the clusters C_i and C'_i, then, as there is no way to discriminate between y and y', we could, for instance, include y in C_i, and y' in C'_i. In that case, the clusters C_i and C'_i would differ by more than one record. Therefore, for the microaggregation to be insensitive, $\leq_i$ must assign a different priority to each element; in other words, $\leq_i$ must be a total order.

A similar argument to the one used in Proposition 1 can be used to show that when the total order relation is the same for all the clusters —in other words, when $\leq_i$ and $\leq_j$ are equal for any i and j—, then Algorithm 2 is insensitive to the input data. However, we want to show that even when the total orders $\leq_i$ are different, insensitivity still holds. In fact, Proposition 2 provides a complete characterization of insensitive microaggregation algorithms of the form of Algorithm 2.

Proposition 2 *Algorithm 2 is insensitive to input data if and only if $\{\leq_i\}_{i\in\mathbb{N}}$ is a fixed sequence of total order relations defined over the domain of X.*

Proof. In the discussion previous to Proposition 2 we have already shown that if Algorithm 2 is insensitive, then $\{\leq_i\}_{i\in\mathbb{N}}$ must be a fixed sequence of total order relations. We show now that the reverse implication also holds: if $\{\leq_i\}_{i\in\mathbb{N}}$ is a fixed sequence of total order relations, then Algorithm 2 is insensitive to input data.

Let X and X' be, respectively, the original data set and a data set that differs from X in one record. Let C_i and C'_i be, respectively, the clusters generated at step i for the data sets X and X'. We want to show, for any i, that C_i and C'_i differ in at most one record.

An argument similar to the one in Proposition 1 shows that the clusters C_0 and C'_0 that result from the first iteration of the algorithm differ in at most one record. To see that Algorithm 2 is insensitive, it is enough to check that the sets $X \setminus C_0$ and $X' \setminus C'_0$ differ in at most one record; then, we could apply the previous argument to $X \setminus C_0$ and $X' \setminus C'_0$ to see that C_1 and C'_1 differ in one record, and so on.

Let $x_1, \ldots, x_n$ be the elements of X ordered according to $\leq_0$, so that $C_0 = \{x_1, \ldots, x_k\}$. Assume that X' has had element x replaced by x': $X' = \{x_1, \ldots, x_n, x'\} \setminus \{x\}$. We have the following four possibilities. (i) If neither x belongs to C_0 nor x' belongs to C'_0, then

C_0 and C_0' must be equal; therefore, $X \setminus C_0$ and $X' \setminus C_0'$ differ, at most, in one record. (ii) If both x belongs to C_0 and x' belongs to C_0', then $X \setminus C_0$ and $X' \setminus C_0'$ are equal. (iii) If x belongs to C_0 but x' does not belong to C_0', we can write C_0' as $\{x_1, \ldots, x_{k+1}\} \setminus \{x\}$; the set $X' \setminus C_0'$ is $\{x_{k+2}, \ldots, x_n, x'\}$, which differs in one record from $X \setminus C_0 = \{x_{k+1}, \ldots, x_n\}$. And (iv) if x is not in C_0 but x' is in C_0', we can write C_0' as $\{x_1, \ldots, x_{k-1}, x'\}$; the set $X' \setminus C_0'$ is $\{x_k, \ldots, x_n\} \setminus \{x\}$, which differs in one record from $X \setminus C_0 = \{x_{k+1}, \ldots, x_n\}$. Therefore, we have seen that $X \setminus C_0$ and $X' \setminus C_0'$ differ in at most one record, which completes the proof. □

Using multiple order relations in Algorithm 2, as allowed by Proposition 2, in contrast with the single order relation used to turn MDAV insensitive in Proposition 1, allows us to increase the within-cluster homogeneity achieved in the microaggregation (see Section 5 for an empirical evaluation).

The modification of the query function f to $f \circ M$ by introducing a prior microaggregation step is intended to reduce the sensitivity of the query function, regardless of its type. Assume that the microaggregation function f computes the centroid of each cluster as the mean of its components. We analyze next how the microaggregation affects the L_1-sensitivity of the query function f.

Definition 5 (L_1-Sensitivity) Let $\mathcal{D}$ be the class of possible data sets. The L_1-sensitivity of a function $f : \mathcal{D} \to \mathbb{R}^d$ is the smallest number $\Delta(f)$ such that for all $X, X' \in \mathcal{D}$ which differ in a single entry,

$$\|f(X) - f(X')\|_1 \leq \Delta(f).$$

The L_1-sensitivity of f, $\Delta(f)$, measures the maximum change in f that results from modification of a single record in X. Essentially, the microaggregation step M in $f \circ M$ distributes the modification suffered by a single record in X among multiple records in $M(X)$. Consider, for instance, the data sets X and X' depicted in Figure 2. The record at the top right corner in X has been moved to the bottom left corner in X'; all the other records remain unmodified. In the microaggregated data sets $M(X)$ and $M(X')$ —the crosses represent the centroids— we observe that all the centroids have been modified but the magnitude of the modifications is smaller: the modification suffered by the record at the top right corner of X has been distributed among all the records in $M(X)$.

When computing the centroid as the mean, we can guarantee that the maximum variation in any centroid is at most $1/k$ of the variation of the record in X. Therefore, we can think of the L_1-sensitivity of $f \circ M$ as the maximum change in f if we allow a variation in each record that is less than $1/k$ times the maximal variation. In fact, this is a very rough estimate, as only a few centroids can have a variation equaling $1/k$ of the maximal variation in X, but it is useful to analyze some simple functions such as the identity. The identity function returns the exact contents of a specific record, and is used extensively in later sections to construct ε-differentially private data sets. The sensitivity of the identity functions depends only on the maximum variation that the selected record may suffer; therefore, it is clear that distributing the variation among several records lowers the sensitivity. This is formalized in the following proposition.

Proposition 3 *Let $X \in \mathcal{D}$ be a data set with numerical attributes only. Let M be a microaggregation function with minimal cluster size k that computes the centroid by taking the mean of the elements of each cluster. Let $I_r()$ be the function that returns the attribute values corresponding to the r-th record of X, for r between 1 and n (the number of records in X). Then $\Delta(I_r \circ M) \leq \Delta(I_r)/k$.*

Proof. The function $I_r \circ M$ returns the centroid of $M(X)$ that corresponds to the r-th record in X. It was shown in the discussion that precedes the proposition that, for a data set that contains only numerical attributes, if the centroid is computed as the mean of the records in the cluster, then the maximum change in any centroid is, at most, $\Delta(I_r)/k$; that is, $\Delta(I_r \circ M) \leq \Delta(I_r)/k$. □

4 Differentially private data sets through *k*-anonymity

Assume that we have an original data set X with n records and that we want to generate a data set X_ε —an anonymized version of X— that satisfies ε-differential privacy. Even if differential privacy was not introduced with the aim of generating anonymized data sets, we can think of a data release as the collected answers to successive queries for each record in the data set. Let $I_r()$ be as defined in Proposition 3. We generate X_ε, by querying X with $I_r(X)$, for $r = 1$ to n. If the responses to the queries $I_r()$ satisfy ε-differential privacy, then, as each query refers to a different record, by the parallel composition property [43] we have that X_ε also satisfies ε-differential privacy. That is, the differentially private data set X_ε is generated by providing a differentially private response to the queries that ask for the values of all the attributes in each record.

The proposed approach for generating X_ε is general (since it does not make any assumptions on the uses of output data) but naive. As each query $I_r()$ refers to a single individual, its sensitivity is large (the diameter of the domain of the records in the data set); therefore, the masking required to attain ε-differential privacy is quite significant, and thus the utility of such a X_ε very limited.

To improve the utility of X_ε, we introduce a microaggregation step as discussed in Section 3: (i) from the original data set X, we generate a k-anonymous data set $\overline{X}$ —by using a microaggregation algorithm with minimum cluster size k, like MDAV, and assuming that all attributes are quasi-identifiers—, and (ii) the ε-differentially private data set X_ε is generated from the k-anonymous data set $\overline{X}$ by taking an ε-differentially private response to the queries $I_r(\overline{X})$, for $r = 1, \cdots, n$ (both X and $\overline{X}$ have the same number n of records, and the k-anonymized version of the r-th record in X is the r-th record in $\overline{X}$).

By constructing the k-anonymous data set $\overline{X}$, we stop thinking in terms of individuals, to start thinking in terms of groups of k individuals. Now, the sensitivity of the queries $I_r(\overline{X})$ used to construct X_ε reflects the effect that modifying a single record in X has on the groups of k records in $\overline{X}$. The fact that each record in $\overline{X}$ depends on k (or more) records in X is what leads to the reduced sensitivity of the cluster centroids in comparison to the sensitivity of the original records. See Proposition 3 above.

Seeing that the centroids have reduced sensitivity in comparison to the original records is not enough. We have to check whether the sensitivity of the set of queries $I_r(\overline{X})$ is smaller than the sensitivity of the set of queries $I_r(X)$, for $r = 1, \cdots, n$ (which is, by parallel composition, equal to the sensitivity of any single query $I_r(X)$). The modification of a single record in X may lead to multiple modifications in $\overline{X}$; thus, parallel composition cannot be used to compute the sensitivity of the set of queries $I_r(\overline{X})$, for $r = 1, \cdots, n$. We take a different approach instead. If k is the cluster size used in the microaggregation, we have seen that the sensitivity of each individual query $I_r(\overline{X})$ is upper bounded by $\Delta I_r(X)/k$. As there are n/k *different* queries in $\overline{X}$ (because the n records in $\overline{X}$ are clustered in n/k clusters such that all records within each cluster are identical), the sensitivity of $I_r(\overline{X})$, for $r \in \overline{X}$, is upper bounded by $n/k \times \Delta I_r(X)/k$.

We want $n/k \times \Delta I_r(X)/k$ to be smaller than $\Delta I_r(X)$. To that end, we adjust the cluster size k. Increasing the cluster size has two effects: it reduces the contribution of each record to the cluster centroid (thus reducing the sensitivity of the centroid), and it reduces the number of generated clusters (thus reducing the number of different queries). For $n/k \times \Delta I_r(X)/k$ to be smaller than $\Delta I_r(X)$, we need $k \geq \sqrt{n}$.

As stated in the previous section, even though the prior k-anonymous microaggregation also incurs a loss of utility, we hypothesize that this loss is more than compensated by the benefits brought by the reduction of the sensitivity when constructing differentially private outputs. This is motivated by the ability of microaggregation to exploit the underlying structure of data to reduce sensitivity with relatively little utility loss.

Algorithm 3 details the procedure for generating the differentially private data set X_ε.

Algorithm 3 Generation of an ε-differentially private data set X_ε from X via microaggregation

```
let $X$ be an original data set with $n$ records
let $M$ be an insensitive microaggregation algorithm with minimal cluster size $k$
let $S_\varepsilon()$ be an $\varepsilon$-differentially private sanitizer
let $I_r()$ be the query for the attributes of the $r$-th record

$\overline{X} \leftarrow$ microaggregated data set $M(X)$
for $r = 1$ to $n$ do
    $x_\varepsilon \leftarrow S_\varepsilon(I_r(\overline{X}))$
    insert $x_\varepsilon$ into $X_\varepsilon$
end for

return $X_\varepsilon$
```

4.1 Achieving differential privacy with numerical attributes

For a data set consisting of numerical attributes only, generating the ε-differentially private data set X_ε as previously described is quite straightforward.

Let X be a data set with m numerical attributes: A_1, ..., A_m. The first step to construct X_ε is to generate the k-anonymous data set $\overline{X}$ via an insensitive microaggregation algorithm. As we have seen in Section 3, the key point of insensitive microaggregation algorithms is to define a total order relation over $Dom(X)$, the domain of the records of the data set X. The domain of X contains all the possible values that make sense, given the semantics of the attributes. In other words, the domain is not defined by the actual records in X but by the set of values that make sense for each attribute and by the relation between attributes.

Microaggregation algorithms use a distance function, $d : Dom(X) \times Dom(X) \rightarrow \mathbb{R}$, to measure the distances between records and generate the clusters. We assume that such a distance function is already available and we define a total order the distance is consistent with as follows:

Definition 6 Given a reference point R, we define a total order according to the distance to R so that, for a pair of elements $x, y \in Dom(X)$, we say that $x \leq y$ if $d(R, x) \leq d(R, y)$.

To construct a total order, we still need to define the relation between elements that are equally distant from R. As we assume that the data set X consists of numerical attributes only, we can take advantage of the fact that individual attributes are equipped with a total order —the usual numerical order— and sort the records that are equally distant from R by means of the alphabetical order: given $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$, with $d(x, R) = d(y, R)$, we say that $x \leq y$ if $(x_1, \ldots, x_m) \leq (y_1, \ldots, y_m)$ according to the alphabetical order.

Proposition 3 shows that, as a result of the insensitive microaggregation, one has $\Delta(I_r \circ M) = \Delta(I_r)/k$; therefore, ε-differential privacy can be achieved by adding to $\overline{X}$ an amount of Laplace noise that would only achieve $k\varepsilon$-differential privacy if directly added to X.

4.2 Insensitive MDAV

According to Proposition 1, to make MDAV insensitive we must define a total order among the elements in $Dom(X)$. According to the previous discussion, this total order is constructed by selecting a reference point. To increase within-cluster homogeneity, MDAV starts by clustering the elements at the boundaries. For our total order to follow this guideline, the reference point R must be selected among the elements of the boundary of $Dom(X)$. For instance, if the domain of A_i is $[a_b^i, a_t^i]$, we can set R to be the point $(a_b^1, \ldots, a_b^m)$.

Figure 2 illustrates the insensitive microaggregation obtained by using MDAV with the total order defined above. The original data set X and the modified data set X' are the same of Figure 1. We also use $k = 5$ and the Euclidean distance for insensitive MDAV. Let us take as the reference point for the above defined total order the point R at the lower left corner of the grids. Note that now clusters C_1,C_2, and C_3 in X differ in a single record from C_1',C_2', and C_3' in X', respectively. By comparing Figures 1 and 2, we observe that the standard (non-insensitive) MDAV results in a set of clusters with greater within-cluster homogeneity; however, in exchange for the lost homogeneity, insensitive MDAV generates sets of clusters that are more stable when one record of the data set changes.

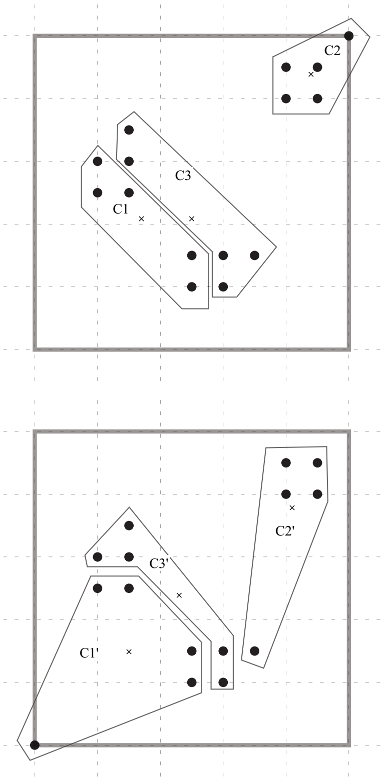

Fig. 2 Insensitive MDAV microaggregation with $k = 5$. Top, original data set X; bottom, data set after modifying one record in X.

4.3 General insensitive microaggregation

It was seen in Section 3 that each clustering step within microaggregation can use a different total order relation, as long as the sequence of order relations is kept constant. The advantage of using multiple total order relations is that it allows the insensitive microaggregation algorithm to better mimic a standard non-insensitive microaggregation algorithm, and thus increase the within-cluster homogeneity.

The sequence of total orders is determined by a sequence of reference points R_i. In the selection of R_i we try to match the criteria used by non-insensitive microaggregation algorithms to increase within-cluster homogeneity: start clustering at the boundaries, and generate a cluster that is far apart from the previously generated cluster.

Let the domain of A_i be $[a_b^i, a_t^i]$. Define the set $\mathcal{R}$ of candidate reference points at those points in the boundaries of $Dom(X)$, that is:

$$\mathcal{R} = \{(a_{v_1}^1, \ldots, a_{v_m}^m) | v_i \in \{b, t\} \text{ for } 1 \leq i \leq m\}.$$

The first reference point R_1 is arbitrarily selected from $\mathcal{R}$; for instance, $R_1 = (a_b^1, \ldots, a_b^m)$. Once a point R_i has been selected, R_{i+1} is selected among the still unselected points in $\mathcal{R}$ so that it maximizes the Ham-

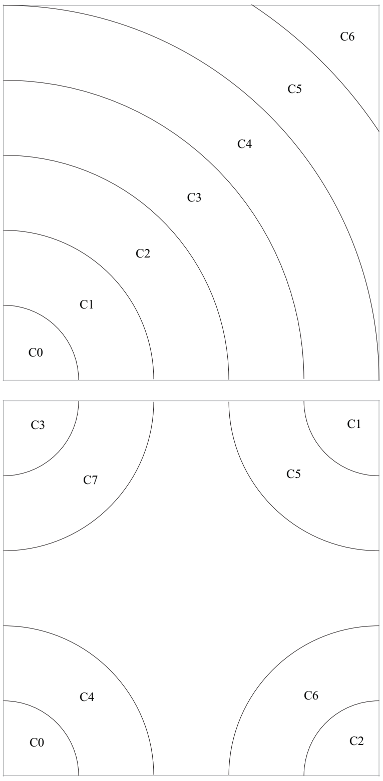

Fig. 3 Cluster formation. Top, using a single reference point; bottom, taking each corner of the domain as a reference point.

ming distance to R_i —if $R_1 = (a_b^1, \dots, a_b^m)$, then $R_2 = (a_t^1, \dots, a_t^m)$—. If several unselected points in $\mathcal{R}$ maximize the Hamming distance to R_i, we select the one among them at greatest distance from R_{i-1}, and so on; that is, similar to other microaggregation heuristics (like MDAV, [19]), in order to minimize interference in cluster formation we prioritize having the greatest possible distance to the most recently selected reference points.

Figure 3 shows the form of the clusters for a data set containing two numerical attributes. The top graphic is for a single reference point —this is also the form of the clusters obtained by insensitive MDAV, which uses a single total order relation—. The bottom graphic uses four reference points, one for each edge of the domain, which are selected in turns as described above.

4.4 Achieving differential privacy with categorical attributes

Many data sets contain attributes with categorical values, such as Race, Country of birth, or Job [27]. Unlike continuous-scale numerical attributes, categorical attributes take values from a finite set of categories for which the arithmetical operations needed to microaggregate and add noise to the outputs do not make sense. In the sequel, we detail alternative mechanisms that are suitable for categorical attributes in order to achieve differential privacy as detailed above.

Let X be a data set with m categorical attributes: A_1,. . ., A_m. The first challenge regards the definition of $Dom(X)$. Unlike for numerical attributes, the universe of each categorical attribute can only be defined by extension, listing all the possible values. This universe can be expressed either as a flat list or it can be structured in a hierarchic/taxonomic way. The latter scenario is more desirable, since the taxonomy implicitly captures the semantics inherent to conceptualizations of categorical values (*e.g.*, disease categories, job categories, sports categories, etc.). In this manner, further operations can exploit this taxonomic knowledge to provide a semantically coherent management of attribute values [42].

Formally, a taxonomy τ can be defined as an upper semilattice $\leq_\varsigma$ on a set of concepts ς with a top element $root_\varsigma$. We define the taxonomy $\tau(A_i)$ associated to an attribute A_i as the lattice on the minimum set of concepts that covers all values in $Dom(A_i)$. Notice that $\tau(A_i)$ will include all values in $Dom(A_i)$ (*e.g.*, “skiing”, “sailing”, “swimming”, “soccer”, etc., if the attribute refers to sport names) and, usually, some additional generalizations that are necessary to define the taxonomic structure (*e.g.*,“winter sports”, “water sports”, “field sports”, and “sport” as the *root* of the taxonomy).

If A_1, . . ., A_m are independent attributes, $Dom(X)$ can be defined as the ordered combination of values of each $Dom(A_i)$, as modeled in their corresponding taxonomies $\tau(A_1)$, . . ., $\tau(A_m)$. If $A_1, \cdots, A_m$ are not independent, value tuples in $Dom(X)$ may be restricted to a subset of valid combinations.

Next, a suitable distance function $d : Dom(X) \times Dom(X) \to \mathbb{R}$ to compare records should be defined. To tackle this problem, we can exploit the taxonomy $\tau(A_i)$ associated to each A_i in X and the notion of *semantic distance* [55]. A semantic distance δ quantifies the amount of semantic differences observed between two terms (*i.e.*, categorical values) according to the knowledge modeled in a taxonomy. Section 4.5 discusses the adequacy of several semantic measures in the context of differential privacy. By composing semantic distances δ for individual attributes A_i, each one computed from the corresponding taxonomy $\tau(A_i)$, we can define the required distance $d : Dom(X) \times Dom(X) \to \mathbb{R}$.

To construct a total order that yields insensitive and within-cluster homogeneous microaggregation as detailed in Section 4.3, we need to define the boundaries of $Dom(X)$, from which records will be clustered. Unlike in the numerical case, this is not straightforward, since most categorical attributes are not ordinal and, hence, a total order cannot be trivially defined even for individual attributes. However, since the taxonomy $\tau(A_i)$ models the domain of A_i, boundaries of $Dom(A_i)$, that is, $[a_b^i, a_t^i]$, can be defined as the most distant and opposite values from the "middle" of $\tau(A_i)$. From a semantic perspective, this notion of centrality in a taxonomy can be measured by the *marginality model* [14]. This model determines the central point of the taxonomy and how far each value is from that center, according to the semantic distance between value pairs.

The *marginality* $m(\cdot,\cdot)$ of each value a_j^i in A_i with respect to its domain of values $Dom(A_i)$ is computed as

$$m(Dom(A_i), a_j^i) = \sum_{a_l^i \in Dom(A_i) - \{a_j^i\}} \delta(a_l^i, a_j^i) \tag{2}$$

where $\delta(\cdot,\cdot)$ is the semantic distance between two values. The greater $m(Dom(A_i), a_j^i)$, the more marginal (*i.e.*, the less central) is a_j^i with regard to $Dom(A_i)$.

Hence, for each A_i, one boundary a_b^i of $Dom(A_i)$ can be defined as the most marginal value of $Dom(A_i)$:

$$a_b^i = \arg \max_{a_j^i \in Dom(A_i)} m(Dom(A_i), a_j^i). \tag{3}$$

The other boundary a_t^i can be defined as the most distant value from a_b^i in $Dom(A_i)$:

$$a_t^i = \arg \max_{a_j^i \in Dom(A_i)} \delta(a_j^i, a_b^i). \tag{4}$$

By applying the above expressions to the set of attributes $A_1, \cdots, A_m$ in X, the set $\mathcal{R}$ of candidate reference points needed to define a total order according to the semantic distance can be constructed as described in Section 4.3.

If no taxonomic structure is available, other centrality measures based on data distribution can be used (*e.g.*, by selecting the modal value as the most central value [19]). However, such measures omit data semantics and result in significantly less useful anonymized results [41].

Similarly to the numerical case, if several records are equally distant from the reference points, the alphabetical criterion can be used to induce an order within those equidistant records.

At this point, records in X can be grouped using the insensitive microaggregation algorithm, thereby yielding a set of clusters with a sensitivity of only one record per cluster. The elements in each cluster must be replaced by the cluster centroid (*i.e.*, the arithmetical mean in the numerical case) in order to obtain a k-anonymous data set. Since the mean of a sample of categorical values cannot be computed in the standard arithmetical sense, we rely again on the notion of marginality [14]: the mean of a sample of categorical values can be approximated by the least marginal value in the taxonomy, which is taken as the *centroid* of the set.

Formally, given a sample $S(A_i)$ of a nominal attribute A_i in a certain cluster, the marginality-based centroid for that cluster is defined as:

$$Centroid(S(A_i)) = \arg \min_{a_j^i \in \tau(S(A_i))} m(S(A_i), a_j^i) \tag{5}$$

where $\tau(S(A_i))$ is the minimum taxonomy extracted from $\tau(A_i)$ that includes all values in $S(A_i)$. Notice that by considering as centroid candidates all concepts in $\tau(S(A_i))$, which include all values in $S(A_i)$ and also their taxonomic generalizations, we improve the numerical accuracy of the centroid discretization inherent to categorical attributes [41].

The numerical value associated to each centroid candidate a_j^i corresponds to its marginality $m(S(A_i), a_j^i)$, which depends on the sample of values in the cluster. Given a cluster of records with a set of independent attributes $A_1, \cdots, A_m$, the cluster centroid can be obtained by composing the individual centroids of each attribute.

As in the numerical case, cluster centroids depend on input data. To fulfill differential privacy for categorical attributes, two aspects must be considered. On the one hand, the centroid computation should evaluate as centroid candidates all the values in the taxonomy associated to the *domain* of each attribute ($\tau(A_i)$), and not only the sample of values to be aggregated ($\tau(S(A_i))$), since the centroid should be insensitive to any value change of input data within the attribute's domain. On the other hand, to achieve insensitivity, uncertainty must be added to the centroid computation. Since adding Laplacian noise to centroids makes no sense for categorical values, an alternative way to obtain differentially private outputs consists in selecting centroids in a probabilistic manner. The general idea is to select centroids with a degree of uncertainty that is proportional to the suitability of each centroid and the desired level ε of differential privacy. To do so, the Exponential Mechanism proposed by McSherry and Talwar [44] can be applied. Given a function with discrete outputs t, the mechanism chooses the output that is close to the optimum according to the input data X and quality criterion $q(X,t)$, while preserving ε-differential privacy. Each output is associated with a selection probability $\Pr(t)$, which grows exponentially

Algorithm 4 Computation of ε-differentially private centroids for clusters with categorical attributes

```
let $C$ be a cluster with at least $k$ records constructed from a data set with $n$ records
for each categorical attribute $A_i$ do
    Take as quality criterion $q(\cdot,\cdot)$ for each centroid candidate $a^i_j$ in $\tau(A_i)$ the additive inverse of its marginality towards the attribute values $S(A_i)$ contained in $C$, that is, $-m(S(A_i), a^i_j)$;
    Sample the centroid from a distribution that assigns
    $$\Pr(a^i_j) \propto \exp(\frac{\varepsilon \times (-m(S(A_i), a^i_j))}{2\frac{n}{k}\Delta(m(A_i))}) \quad (6)$$
end for
```

with the quality criterion, as follows:

$$\Pr(t) \propto \exp(\frac{\varepsilon q(X,t)}{2\Delta(q)}).$$

In this manner, the optimal output or those that are close to it according to the quality criterion will be more likely to be selected. Based on the above arguments, ε-differentially private centroids can be selected as indicated in Algorithm 4.

Notice that the inversion of the marginality function has no influence on the relative probabilities of centroid candidates, since it is achieved through a *bijective linear transformation*.

With the algorithm we have the following result, which is parallel to what we saw in the numerical case: if the input data are k-anonymous, the higher k, the less the uncertainty that needs to be added to reach ε-differential privacy.

Proposition 4 *Let X be a data set with categorical attributes. Let $\overline{X}$ be a k-anonymous version of X generated using an insensitive microaggregation algorithm M with minimum cluster size k. ε-Differential privacy can be achieved by using Algorithm 4 to obtain cluster centroids in $\overline{X}$ with an amount of uncertainty that decreases as k grows.*

Proof. Without loss of generality, we can write the proof for a single attribute A_i. The argument can be composed for multi-attribute data sets.

Let $\Delta(m(A_i))$ be the sensitivity of the marginality function for attribute A_i. According to the insensitive microaggregation described earlier in Section 3, modifying one record in the data set will induce a change of at most one value in the set $S(A_i)$ of values of A_i for each cluster. Considering that marginality measures the sum of distances between a centroid candidate and all the elements in $S(A_i)$, in the worst case, in which all values in $S(A_i)$ correspond to the same boundary of $Dom(A_i)$ (defined by either Equation (3) or Equation (4)), and one of these is changed by the other boundary, the sensitivity $\Delta(m(A_i))$ of each cluster will correspond to the semantic distance between both boundaries. As stated in Section 4, since we are publishing n/k centroids when releasing the anonymized output, where n is the number of records of the data set, the global sensitivity will be $n/k \times \Delta(m(A_i))$.

We have that: i) to compute the probabilities in Expression (6), the quality criterion $-m(S(A_i), a^i_j)$ is combined with ε, which is a constant, and $n/k \times \Delta(m(A_i))$; ii) $|S(A_i)| \geq k$; iii) $m(S(A_i), a^i_j)$ is a sum of, at least, $k-1$ terms. Hence, as the cluster size k grows, the marginalities $m(S(A_i), a^i_j)$ of values a^i_j in the cluster $S(A_i)$ increase, while the global sensitivity $n/k \times \Delta(m(A_i))$ decreases. Hence, as k grows, the probability for each candidate computed with Expression (6) tends to decrease more rapidly as the marginality of the candidate increases, because k amplifies the additive inverse of the marginality within the exponent in Expression (6). Thus, the larger k, the more clearly the probability of the candidate with the smallest marginality dominates; this candidate is precisely the optimal centroid. Therefore, optimal centroids are more likely to be selected as k increases. In other words, the amount of uncertainty added to the output to fulfill differential privacy for categorical attributes decreases as the k-anonymity level of the input data increases. □

4.5 A semantic distance suitable for differential privacy

As described above, the selection of differentially private outputs for categorical attributes is based on the marginality value of centroid candidates that, in turn, is a function of the semantic distance between centroids and clustered values. Moreover, the total order used to create clusters also relies on the assessment of semantic distances between attribute values. Hence, the particular measure used to compute semantic distances directly influences the quality of anonymized outputs.

A semantic distance $\delta : o \times o \rightarrow \mathbb{R}$ is a function mapping a pair of concepts to a real number that quantifies the difference between the concept meanings. A well-suited δ to achieve semantic-preserving differentially private outputs should have the following features. First, it should capture and quantify the semantics of the categorical values precisely, so that they can be well differentiated, both when defining the total order and also when selecting cluster centroids [41]. Second, from the perspective of differential privacy, δ should have a low numerical sensitivity to outlying values, which are those that define the boundaries of the universe and, thus, the sensitivity of the quality criterion. By achieving a low numerical sensitivity, the probability of selecting optimal centroids with the exponential mechanism will increase. This will produce less noisy and, hence, more accurate differentially private outputs.

The accuracy of a semantic measure depends on the kind of techniques and knowledge bases used to perform the semantic assessments [55]. Among those relying on taxonomies, feature-based measures and measures based on intrinsic information-theoretic models usually achieve the highest accuracy with regard to human judgments of semantic distance [55]. The former measures [55,47] quantify the distance between concept pairs according to their number of common and non-common taxonomic ancestors. The latter measures [54, 52,48,53] evaluate the similarity between concept pairs according to their mutual information, which is approximated as the number of taxonomic specializations of their most specific common ancestor. Both approaches exploit more taxonomic knowledge and, hence, tend to produce more accurate results, than well-known edge-counting measures [49,60], which quantify the distance between concepts by counting the number of taxonomic edges separating them.

On the other hand, the sensitivity to outlying values depends on the way in which semantic evidences are quantified. Many classical methods [49,60] propose distance functions that are linearly proportional to the amount of semantic evidences observed in the taxonomy (*e.g.*, number of taxonomic links). As a result, distances associated to outlying concepts are significantly larger than those between other more "central" values. This leads to a centroid quality criterion with a relatively high sensitivity, which negatively affects the accuracy of the Exponential Mechanism [44]. More recent methods [55,48,13] choose to evaluate distances in a non-linear way. Non-linear functions provide more flexibility since they can implicitly weight the contribution of more specific [13,34] or more detailed [55,48,54,53] concepts. As a result, concept pairs become better differentiated and semantic assessments tend to be more accurate [55]. We can distinguish between measures that exponentially promote semantic differences [13,34] and those that aggregate semantic similarities [48,54,53] and differences [55] in a logarithmic way. Among these, the latter one is best suited for the differential privacy scenario, since the logarithmic assessment of the semantic differences helps reduce the relative numerical distances associated to outlying concepts and, hence, to minimize the sensitivity of the quality function used in the Exponential Mechanism.

Formally, this measure computes the distance $\delta : A_i \times A_i \to \mathbb{R}$ between two categorical values a_1^i and a_2^i of attribute A_i, whose domain is modeled in the taxonomy $\tau(A_i)$, as a logarithmic function of their number of non-common taxonomic ancestors divided (for normalization) by their total number of ancestors [55]:

$$\delta(a_1^i, a_2^i) = \log_2 \left(1 + \frac{|\phi(a_1^i) \cup \phi(a_2^i)| - |\phi(a_1^i) \cap \phi(a_2^i)|}{|\phi(a_1^i) \cup \phi(a_2^i)|} \right) \tag{7}$$

where $\phi(a_j^i)$ is the set of taxonomic ancestors of a_j^i in $\tau(A_i)$, including itself.

As demonstrated in [55] and [2], Expression (7) satisfies *non-negativity*, *reflexivity*, *symmetry* and *triangle inequality*, thereby being a distance measure in the mathematical sense.

Moreover, thanks to the normalizing denominator, the above distance is insensitive to the size and granularity of the background taxonomy and it yields positive normalized values in the $[0, 1]$ range. Since the distance $d : Dom(X) \times Dom(X) \to \mathbb{R}$ defined in Section 4.4 is the composition of semantic distances for individual attributes and their domains may be modeled in different taxonomies, a normalized output is desirable to coherently integrate distances computed from different sources.

Distance measures such as Expression (7) require taxonomies modeling the semantics associated to categorical values. If such taxonomies are not available, non-semantic criteria such as equality/inequality can be used to compare categorical values. Nonetheless, the omission of data semantics is likely to produce significantly less useful results [14].

4.6 Integrating heterogeneous attribute types

The above-described semantic measure provides us with a numerical assessment of the distance between categorical attributes. As a result, given a data set X with attributes of heterogeneous data types (*i.e.*, numerical and categorical), the record distance $d : Dom(X) \times Dom(X) \to \mathbb{R}$ required for microaggregation can be defined by composing numerically assessed distances for individual attributes, as follows:

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\frac{(dist(a_1^1, a_2^1))^2}{(dist(a_b^1, a_t^1))^2} + \cdots + \frac{(dist(a_1^m, a_2^m))^2}{(dist(a_b^m, a_t^m))^2}} \tag{8}$$

where a_j^i $i = 1, \ldots, m$ are the coordinates of $\mathbf{x_j}$ ($j = 1, 2$), $dist(a_1^i, a_2^i)$ is the distance (either numerical or semantic) between the values for the i-th attribute A_i in $\mathbf{x}_1$ and $\mathbf{x}_2$, and $dist(a_b^i, a_t^i)$ is the distance between the boundaries of $Dom(A_i)$, which is used to eliminate the influence of the attribute scale.

It can be noticed that Expression (8) is similar to the normalized Euclidean distance, but replacing attribute variances, which depend on input data, by distances between domain boundaries, which are insensitive to changes of input values. In this manner, the record distance function effectively defines a total order that fulfills differential privacy.

5 Empirical evaluation

In this section we give empirical results that illustrate how k-anonymous microaggregation of input data reduces the amount of noise required to fulfill differential privacy and, hence, positively influences the utility of the anonymized outputs.

5.1 Evaluation data

The above-described mechanism has been applied to three data sets consisting, respectively, of numerical attributes only, categorical attributes only and a mix of both types of attributes. These three data sets have been extracted from two reference data sets, as follows:

- "Census", which contains 1,080 records with numerical attributes [4]. This data set was used in the European project CASC and in [18,12,64,33,22,21,15]. Like in [15], we took attributes FICA (Social security retirement payroll deduction), FEDTAX (Federal income tax liability), INTVAL (Amount of interest income) and POTHVAL (Total other persons income). To fulfill differential privacy, all four attributes were masked, *i.e.*, they were considered as quasi-identifiers in all of our tests. The resulting records were all different from each other. Since all attributes represent non-negative amounts of money, we took as boundaries for the attribute domains $a_b^i = 0$ and $a_t^i = 1.5 \times max\_attr.\_value\_in\_dataset$. The domain upper bound a_t^i is a reasonable estimate if the attribute values in the data set are representative of the attribute values in the population, which in particular means that the population outliers are represented in the data set. The difference between the bounds a_b^i and a_t^i defines the sensitivity of each attribute and influences the amount of Laplace noise to be added to masked outputs, as detailed in Section 4.1. Since the Laplace distribution takes values in the range $(-\infty, +\infty)$, for consistency, we bound noise-added outputs to the $[a_b^i, a_t^i]$ range defined above.
- "Adult", a well-known data set from the UCI repository [27], which has often been used in the past to evaluate privacy-preserving methods [42,16,28, 37]. On the one hand, like in [42], we created a data set with just categorical attributes: OCCUPATION and NATIVE-COUNTRY. According to the data set description, Dom(OCCUPATION) includes 14 distinct categories, whereas the domain Dom(NATIVE-COUNTRY) covers 41. The taxonomies modeling attribute domains, τ(OCCUPATION) and τ(NATIVE-COUNTRY), were extracted from WordNet 2.1 [26], a general-purpose repository that taxonomically models more than 100,000 concepts. Mappings between attribute labels and WordNet concepts are those stated in [42]. By considering attribute categories and their taxonomic ancestors, the resulting taxonomies contain 122 concepts for OCCUPATION and 127 for NATIVE-COUNTRY. As discussed in Section 4.4, these higher figures enable a finer grained and more accurate discretization of cluster centroids in comparison with approaches based on flat lists of attribute categories. Domain boundaries for each attribute and sensitivities for centroid quality criteria were set as described in Section 4.4. For evaluation purposes, we used the training corpus from the Adult data set, which consists of 30,162 records after removing records with missing values. Due to the reduced set of attribute categories, this data set contained 388 different record tuples. We named this evaluation data set "Adult-Categorical". On the other hand, in order to consider also data with heterogeneous attribute types, we derived another evaluation data set from "Adult" that includes the two categorical attributes introduced above and two numerical attributes: AGE and (working) HOURS-PER-WEEK. Domain boundaries and sensitivities for these two numerical attributes were computed as explained for "Census". This third evaluation data set, which we named "Adult-Numerical-Categorical", also contains 30,162 records.

5.2 Evaluation measures

Differentially private algorithms and methods are usually evaluated according to the type of queries they support (*e.g.* count queries [7]) or the task to which anonymized data will be applied (*e.g.* classification [45]). As stated in the introduction, this is motivated by the fact that both interactive methods and most non-interactive ones support or preserve the utility of just a restricted type of queries. Since our proposal aims at providing differentially private outputs making as few assumptions on their uses as k-anonymity does, we used the more general evaluation criteria employed by the k-anonymity research community [17]. In such works, the quality of the anonymized output is evaluated in terms of *information loss*, which directly influences data utility, and *disclosure risk*, which measures practical privacy:

- Information loss measures the differences between original and anonymized data sets. To do so, we use the well-known Sum of Squared Errors (SSE), which is well suited to measure the practical amount of noise added to the output and is often used in the anonymization literature (*e.g.* [17]). For a given anonymized data set (*i.e.*, a k-anonymous data set $\overline{X}$ or an ε-differentially private data set X_ε), SSE is defined as the sum of squares of attribute distances between original records in X and their versions in the anonymized data set, that is

  $$SSE = \sum_{x_j \in X} \sum_{a_j^i \in x_j} (dist(a_j^i, (a_j^i)'))^2,$$

  where a_j^i is the value of the i-th attribute for the j-th original record and $(a_j^i)'$ represents its masked version. For numerical attributes, $dist(\cdot,\cdot)$ corresponds to the standard Euclidean distance, whereas for categorical ones we used the semantic distance defined in Equation (7). Notice that with a high SSE, that is, a high information loss, a lot of data uses are severely damaged, like for example subdomain analyses (analyses restricted to parts of the data set).
- The disclosure risk has been evaluated as the percentage of records of the original data that can be correctly matched from the anonymized data set, that is, the percentage of Record Linkages (RL)

  $$RL = 100 \times \frac{\sum_{x_j \in X} \Pr(x_j')}{n},$$

  where n is the number of original records and the record linkage probability for an anonymized record ($\Pr(x_j')$) is calculated as

  $$\Pr(x_j') = \begin{cases} 0 & \text{if } x_j \notin G \\ \frac{1}{|G|} & \text{if } x_j \in G \end{cases}$$

  where G is the set of original records that are at minimum distance from x_j'. The same distance functions as for SSE have been used. If the correct original record x_j is in G, then $\Pr(x_j')$ is computed as the probability of guessing x_j in G, that is, $1/|G|$. Otherwise, $\Pr(x_j') = 0$. RL measures the practical privacy from the natural perspective of a privacy attack: *e.g.* ε-differential privacy with large ε does not preclude successful record linkage. Hence, the lower RL, the lower the probability of identity disclosure and the better the privacy of the anonymized output.

We refer the reader to the appendix for additional evaluations based on specific data uses (counting queries) and comparisons with related work.

As baseline results, we have computed SSE and RL values for a standard k-anonymity scenario in which all attributes are microaggregated by means of the original MDAV algorithm [19], and also by means of its modified insensitive version with several reference points (Algorithm 2). Furthermore, we also considered the standard ε-differential privacy scenario in which Laplace noise or the Exponential Mechanism (with sensitivities corresponding to the value ranges of each attribute) are directly applied to unaggregated inputs, so that no assumptions are made on the uses of the outputs.

The ε parameter for differential privacy has been set to ε= 0.01, 0.1, 1.0, 10.0, which covers the usual range of differential privacy levels observed in the literature [24, 5, 6, 40]. Taking into consideration the cardinalities of evaluation data sets, the k-anonymity levels have been set between 2 and 100 for Census and between 2 and 500 for the two Adult data sets.

Figure 4 depicts the SSE and RL values for the different parameterizations of k and ε together with the k-anonymous and ε-differential privacy baseline methods for the Census data set; Figures 5 and 6 correspond to the two Adult data sets. Due to the broad ranges of the SSE and RL values when comparing our method with the k-anonymous baselines, the Y-axes are represented in such comparisons using a $\log_{10}$ scale. However, in comparisons between our method and ε-differential privacy baselines, ranges are narrower and we can use a linear scale for the Y-axes. The ε-differential privacy baselines are are displayed as horizontal lines, because they do not depend on the value of k. Each test involving Laplace noise shows the average results of 3 runs, for the sake of stability.

To compare our method against baseline approaches regarding the *balance* between information loss and disclosure risk, we also computed the *relative* improvement of SSE and RL values for our approach ($SSE_{k\epsilon}$, $RL_{k\epsilon}$) over the baseline values (SSE_0, RL_0) obtained with the original MDAV algorithm and with unaggregated differential privacy. First, we computed the improvement factor of SSE values as follows:

$$SSE_f = \frac{\sqrt{SSE_0}}{\sqrt{SSE_{k\epsilon}}}.$$

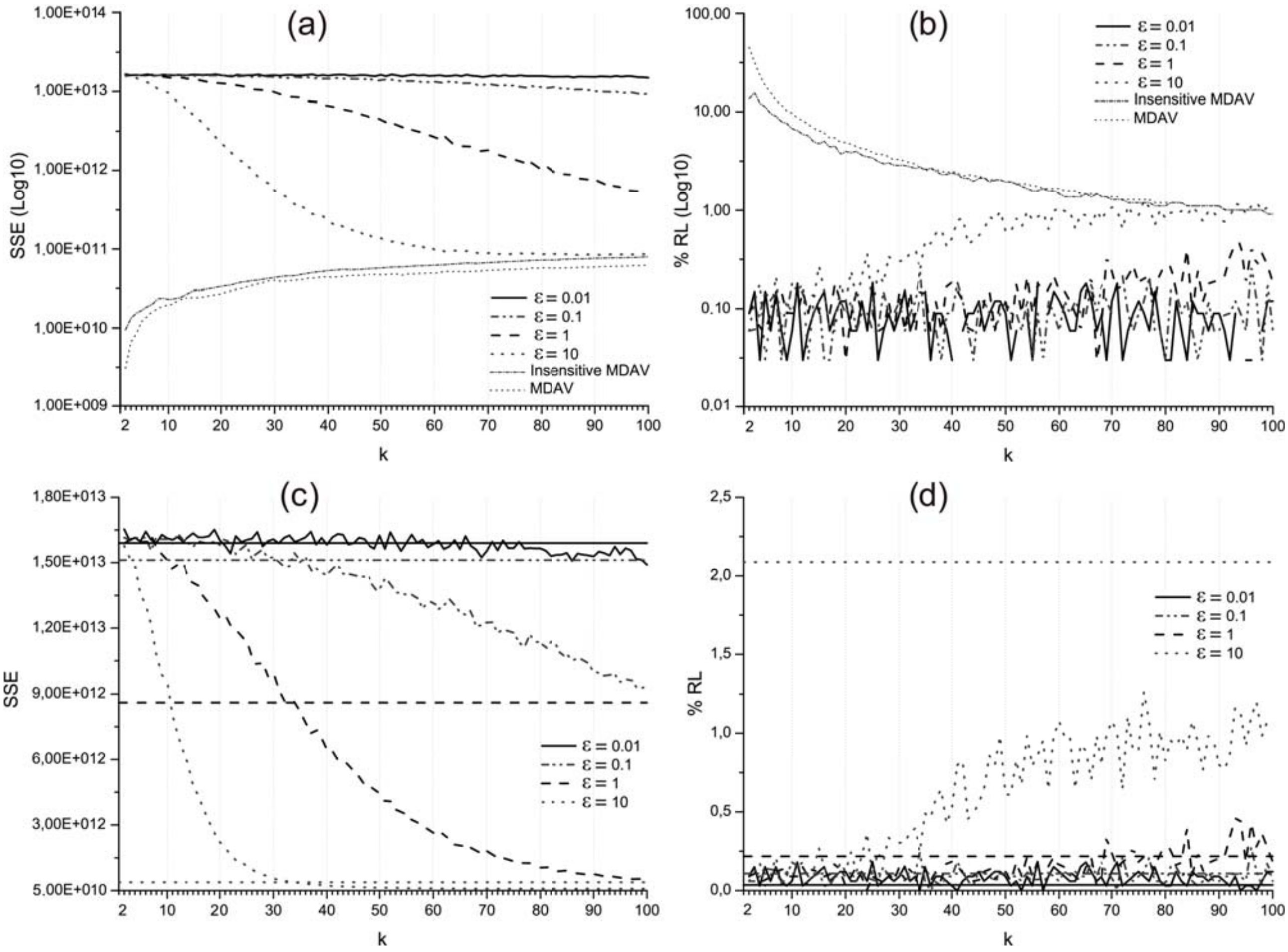

Fig. 4 Census data set. (a) and (b) show the evolution of SSE and RL in our method for different ε values vs. standard k-anonymity in its non-insensitive (MDAV) and insensitive implementations, with $\log_{10}$ scale and k varying with step 1. (c) and (d) show, for each ε value, SSE and RL in our method (non-horizontal lines, because our method depends on the microaggregation parameter k) vs. standard ε-differential privacy (horizontal lines, because this method does not depend on k), with linear scale and k varying with step 1.

Then, the improvement factor of RL values was computed as:

$$RL_f = \frac{RL_0}{RL_{k\epsilon}}.$$

The final score that balances both dimensions was the product of the two previously defined factors:

$$Score = SSE_f \times RL_f.$$

Notice that SSE values have been square rooted to provide a coherent linear integration of RL and SSE, and that $Scores$ above 1.0 show a practical improvement against baseline approaches.

Tables 1 and 2 show the SSE_f and RL_f factors and the resulting $Scores$ for different ε values and some k-anonymity degrees ($k = 2$, $k \approx \sqrt{n}$, $k \approx\ 2\sqrt{n}$, where n is the number of records of the data set) with respect to baseline approaches for the Census data set. Tables 3 and 4 and Tables 5 and 6 correspond to the two Adult data sets.

5.3 Discussion

Regarding the evolution of SSE values in Figures 4(c), 5(c) and 6(c), we observe for the three data sets that the k-anonymous microaggregation of input records reduces the loss of information compared to the standard implementation of ε-differential privacy (horizontal lines) when $k \approx \sqrt{n}$ or higher (that is $k = \sqrt{1,080} \approx 33$ for Census and $k = \sqrt{30,162} \approx 174$ for the two Adult data sets). Indeed, as stated in Section 4, for $k > \sqrt{n}$, the global sensitivity and, thus, the noise, are reduced in comparison with the standard approach of differential privacy. Even though the prior microaggregation step also produces a loss of information, this is in practice

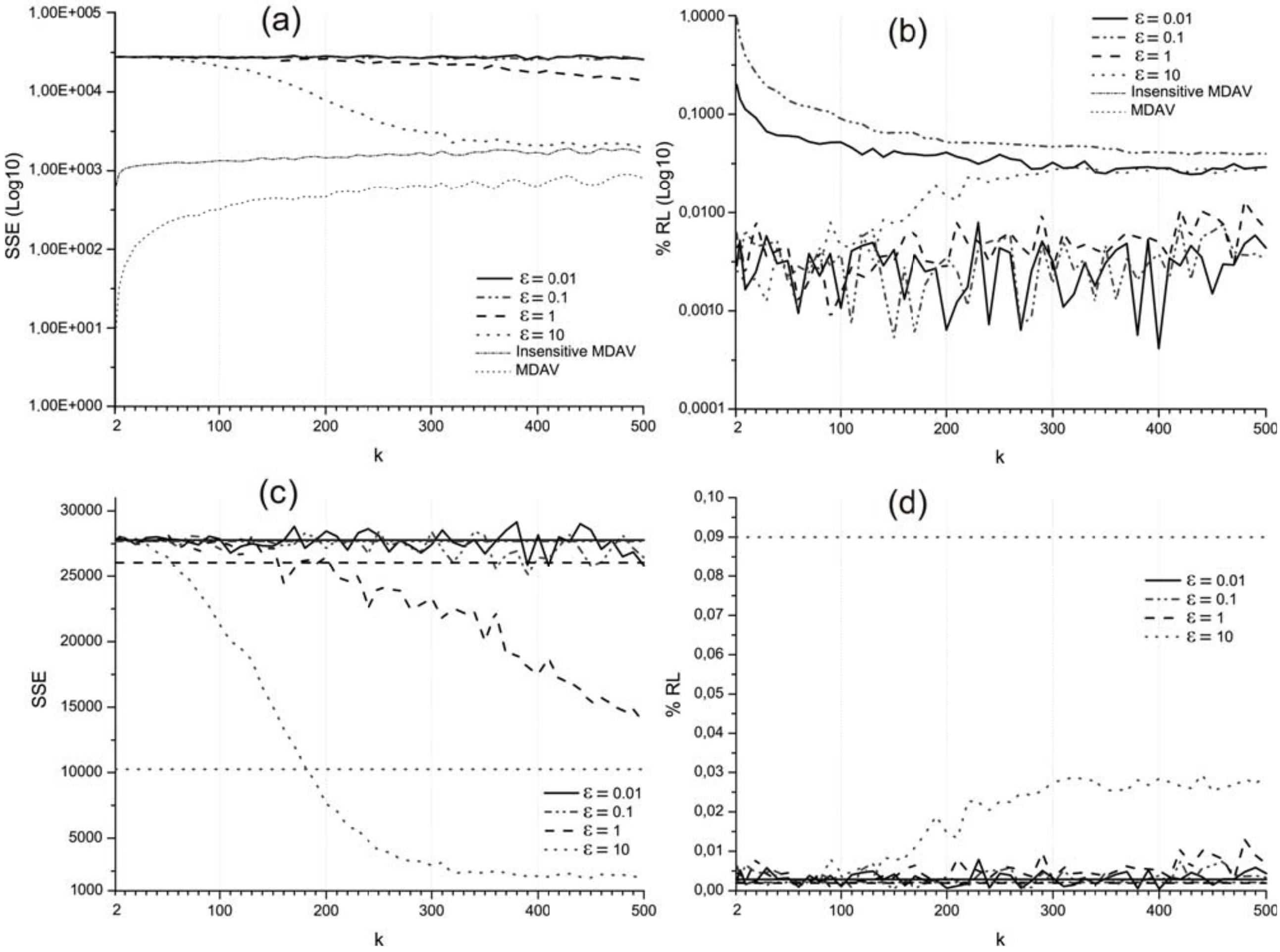

Fig. 5 Adult-Categorical data set. (a) and (b) show the evolution of SSE and RL in our method for different ε values vs. standard k-anonymity in its non-insensitive (MDAV) and insensitive implementations, with $\log_{10}$ scale and k varying with step 10. (c) and (d) show, for each ε value, SSE and RL in our method (non-horizontal lines, because our method depends on the microaggregation parameter k) vs. standard ε-differential privacy (horizontal lines, because this method does not depend on k), with linear scale and k varying with step 10.

Table 1 Census data set. SSE_f and RL_f factors, and $Scores$ for different ε values against standard MDAV microaggregation for several k-anonymity levels.

	MDAV		$\epsilon = 0.01$			$\epsilon = 0.1$			$\epsilon = 1.0$			$\epsilon = 10.0$		
	SSE_0	RL_0	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*
$k = 2$	3.07E+09	45.3	0.014	503	*6.85*	0.014	754.5	*10.39*	0.014	503.0	*6.99*	0.014	754.5	*10.52*
$k = 33$	3.95E+10	2.96	0.049	32.89	*1.61*	0.052	24.67	*1.27*	0.068	49.33	*3.36*	0.318	7.46	*2.37*
$k = 66$	5.09E+10	1.48	0.056	12.33	*0.69*	0.063	8.22	*0.52*	0.156	7.05	*1.1*	0.737	2.10	*1.55*

Table 2 Census data set. SSE_f and RL_f factors, and resulting $Scores$ for different ε values and k-anonymity levels against standard ε-differential privacy (*i.e.*, no microaggregation).

	No microggregation		$k = 2$			$k = 33$			$k = 66$		
	SSE_0	RL_0	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*
$\epsilon = 0.01$	1.59E+13	0.036	0.98	0.40	*0.39*	0.99	0.40	*0.40*	0.99	0.30	*0.30*
$\epsilon = 0.1$	1.51E+13	0.108	0.97	1.80	*1.74*	1.01	0.90	*0.91*	1.09	0.60	*0.66*
$\epsilon = 1.0$	8.61E+12	0.218	0.74	2.42	*1.78*	1.00	3.63	*3.65*	2.03	1.04	*2.10*
$\epsilon = 10.0$	3.83E+11	2.09	0.16	34.78	*5.42*	0.99	5.26	*5.21*	2.02	2.97	*6.00*

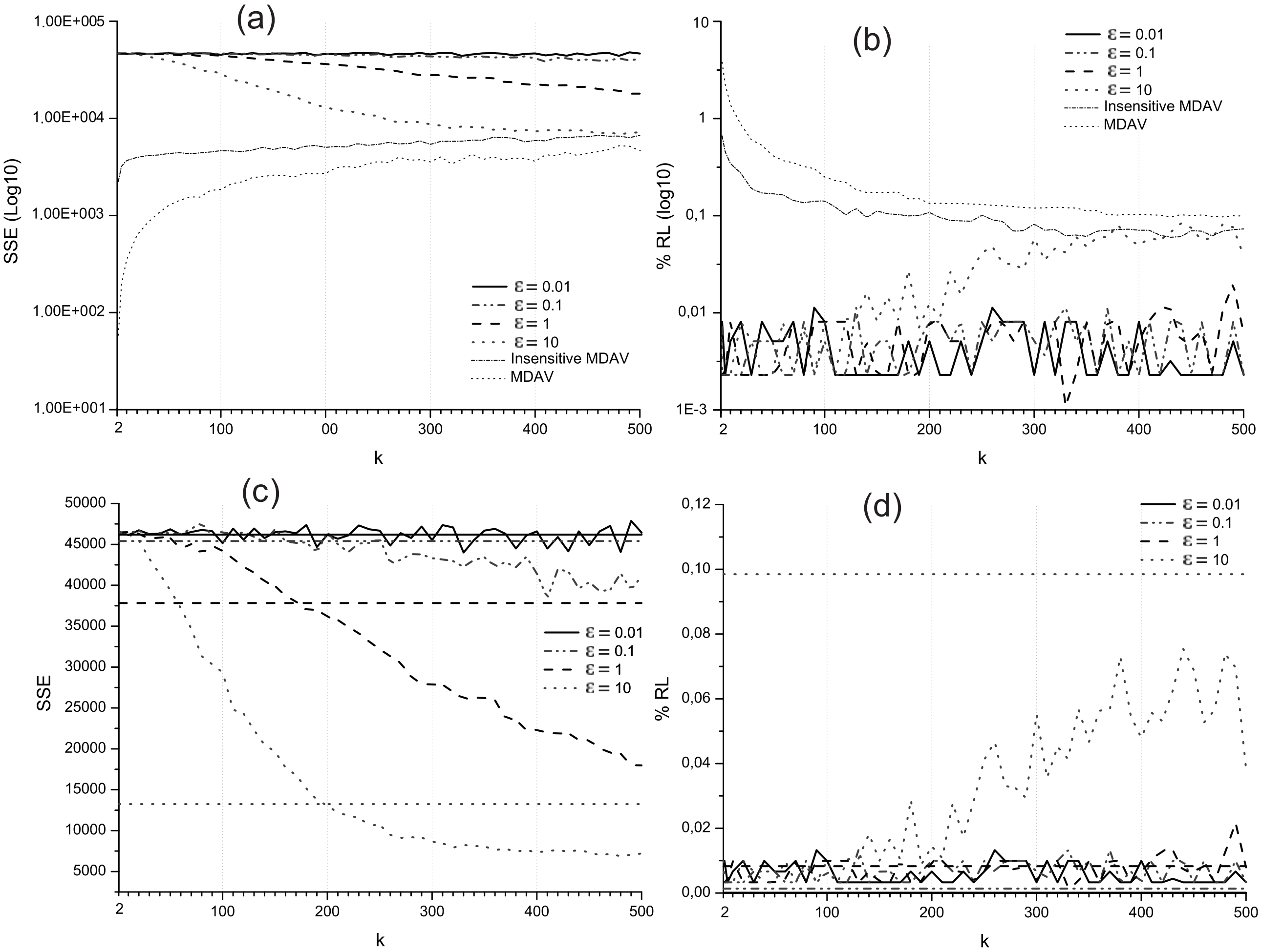

Fig. 6 Adult-Numerical-Categorical data set. (a) and (b) show the evolution of SSE and RL in our method for different ε values vs. standard k-anonymity in its non-insensitive (MDAV) and insensitive implementations, with $\log_{10}$ scale and k varying with step 10. (c) and (d) show, for each ε value, SSE and RL in our method (non-horizontal lines, because our method depends on the microaggregation parameter k) vs. standard ε-differential privacy (horizontal lines, because this method does not depend on k), with linear scale and k varying with step 10.

Table 3 Adult-Categorical data set. SSE_f and RL_f factors, and resulting *Scores* for different ε values against standard MDAV microaggregation for several k-anonymity levels.

	MDAV		$\epsilon = 0.01$			$\epsilon = 0.1$			$\epsilon = 1.0$			$\epsilon = 10.0$		
	SSE$_0$	RL$_0$	SSE$_f$	RL$_f$	*Score*	SSE$_f$	RL$_f$	*Score*	SSE$_f$	RL$_f$	*Score*	SSE$_f$	RL$_f$	*Score*
$k = 2$	10.1	0.95	0.019	328.46	*6.25*	0.019	152.98	*2.91*	0.02	213.96	*4.07*	0.019	375.46	*7.15*
$k = 174$	466	0.057	0.129	15.67	*2.03*	0.129	10.15	*1.31*	0.134	18.65	*2.50*	0.204	4.763	*0.97*
$k = 348$	650	0.044	0.150	8.83	*1.33*	0.152	10.30	*1.57*	0.184	9.97	*1.83*	0.544	1.50	*0.82*

Table 4 Adult-Categorical data set. SSE_f and RL_f factors, and resulting *Scores* for different ε values and k-anonymity levels against standard ε-differential privacy (*i.e.*, no microaggregation).

	No microaggregation		$k = 2$			$k = 174$			$k = 348$		
	SSE	RL	SSE$_f$	RL$_f$	*Score*	SSE$_f$	RL$_f$	*Score*	SSE$_f$	RL$_f$	*Score*
$\epsilon = 0.01$	2.78E+04	0.003	1.00	0.99	*0.99*	1.00	0.79	*0.79*	0.98	0.57	*0.56*
$\epsilon = 0.1$	2.77E+04	0.002	1.00	0.34	*0.34*	1.00	0.38	*0.37*	0.99	0.49	*0.49*
$\epsilon = 1.0$	2.61E+04	0.002	0.97	0.44	*0.42*	1.00	0.64	*0.64*	1.17	0.44	*0.51*
$\epsilon = 10.0$	1.02E+04	0.090	0.61	35.57	*21.60*	0.96	7.54	*7.20*	2.16	3.05	*6.58*

Table 5 Adult-Numerical-Categorical data set. SSE_f and RL_f factors, and resulting *Scores* for different ε values against standard MDAV microaggregation for several k-anonymity levels.

	MDAV		$\epsilon = 0.01$			$\epsilon = 0.1$			$\epsilon = 1.0$			$\epsilon = 10.0$		
	SSE_0	RL_0	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*
$k = 2$	58.4	2.09	0.035	210.15	*7.45*	0.035	630.46	*22.36*	0.035	630.46	*22.35*	0.035	630.46	*22.35*
$k = 174$	193	1.32	0.064	170.04	*10.93*	0.065	396.77	*25.92*	0.071	357.09	*25.19*	0.108	90.87	*9.83*
$k = 348$	631	0.58	0.117	130.85	*15.31*	0.122	523.38	*63.82*	0.154	58.15	*8.95*	0.283	11.84	*3.34*

Table 6 Adult-Numerical-Categorical data set. SSE_f and RL_f factors, and resulting *Scores* for different ε values and k-anonymity levels against standard ε-differential privacy (*i.e.*, no microaggregation).

	No microaggregation		$k = 2$			$k = 174$			$k = 348$		
	SSE	RL	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*	SSE_f	RL_f	*Score*
$\epsilon = 0.01$	4.62E+04	9.95E-05	1.00	0.01	*0.01*	1.00	0.01	*0.01*	1.00	0.02	*0.02*
$\epsilon = 0.1$	4.54E+04	0.0013	0.99	0.41	*0.41*	1.00	0.41	*0.41*	1.10	1.23	*1.27*
$\epsilon = 1.0$	3.78E+04	0.0083	0.90	2.50	*2.26*	0.99	2.25	*2.22*	1.19	0.83	*0.99*
$\epsilon = 10.0$	1.32E+04	0.098	0.53	29.70	*15.85*	0.90	6.8	*6.10*	1.30	2.02	*2.61*

so small compared to the added noise that it barely influences the threshold above which our method shows its practical benefits. Only in one case (Figure 6(c)) and just for the highest ε value ($\varepsilon = 10$), which requires very little noise addition, a higher microaggregation level ($k > 190$) was needed to observe a global reduction of information loss.

The relative improvement of SSE directly depends on the value of ε and the best results are obtained for $\varepsilon = 1.0$. For the smallest ε (that is, 0.01) the amount of noise involved is so high that even with the aforementioned noise reduction, the output data are hardly useful. For the highest ε (that is, 10.0) there is a substantial decline of SSE for low k (below $\sqrt{n}$) and, for larger k, SSE stays nearly constant and almost as low as for insensitive microaggregation (Figures 4(a), 5(a) and 6(a)). In this latter case, the noise added by prior microaggregation in larger clusters is compensated by the noise reduced at the ε-differential privacy stage due to the decreased sensitivity with larger k.

Notice also from Figures 4(a), 5(a) and 6(a) that insensitive MDAV microaggregation incurs a higher SSE than standard MDAV microaggregation. Indeed, the clusters formed by insensitive microaggregation are less homogeneous, due to the total order enforced for input records. Particularly, the Adult-Categorical data set shows a more noticeable increase of SSE figures. This is coherent with the criterion detailed in Section 4.3 to define a total order, which alternatively picks combinations of attribute domain boundaries as reference points to create clusters. Since the evaluated Adult-Categorical data set consists of two attributes, four different reference points can be defined. This contrasts with the four attributes considered for the Census and Adult-Numerical-Categorical data sets, which provide 16 different combinations of domain boundaries, giving more degrees of freedom and producing a more accurate clustering of input data. Moreover, since the Adult-Categorical data set consists of just categorical attributes with a small set of possible categories (in comparison with continuous scale numerical ranges defined by the Census attributes and the two added attribute of Adult-Numerical-Categorical), the imperfections introduced by the insensitive aggregation are amplified by the need to discretize cluster centroids. In any case, the SSE increase caused by insensitive microaggregation is around one order of magnitude smaller than the noise reduction this microaggregation enables when used as a prior step to ε-differential privacy.

RL values shown in Figures 4(b)(d), 5(b)(d) and 6(b)(d) behave the other way round compared to SSE. First, we notice that the standard MDAV algorithm results in the highest percentage of linkages. For the Census data set, a k-anonymity level $k \geq 20$ is needed to attain a percentage of linkages below 5%. For the two Adult data sets, RL is much lower because the number of distinct values of each categorical attribute is limited and because the number of records is much higher (*i.e.*, 30,162 for Adult vs. 1,080 for Census). As a result, the probability of correct record linkages is much lower (*i.e.*, below 1% from $k = 2$). Insensitive MDAV yields slightly more privacy than MDAV for the Census data set and significantly more privacy (less percentage of record linkages) for the Adult data sets. The superior RL reduction in Adult w.r.t. Census is coherent with the differences in information loss observed in SSE values, which were caused by the less homogeneous clusterization in Adult. In all data sets, the RL values of insensitive microaggregation are very similar to the ones obtained with ε-differential privacy with $\varepsilon = 10$ for large k. This shows that, for a large ε, the —theoretically— more robust privacy offered by differential privacy in comparison with k-anonymity entails no practical advantage. For ε values of 0.01, 0.1 and 1.0, the RL values hardly vary when the k-anonymity level increases, because they are already very low with the standard differential privacy (horizontal lines in Figures 4(d), 5(d) and 6(d)). Note that, for such low ε-values, the RL values stay around 0.1% for Census data, which, considering the data set size of 1,080 records, corresponds to the probability of successful random record linkage (*i.e.*, 1/1,080). For the two Adult data sets, RL behaves similarly but it shows a much lower matching probability (*i.e.*, around 0.0033%, that is, 1/30,162), because of the larger cardinality of the data set. It can also be seen that the level of privacy offered by the standard ε-differential privacy is increased (*i.e.*, lower RL) or, at least, maintained when using prior microaggregation. This is especially noticeable for $\varepsilon = 10.0$, which results in around half the RL even for the largest k values. Thus, we can conclude that the reduction in information loss achieved by using microaggregation prior to noise addition does not entail appreciable privacy penalties.

By analyzing the balance (*Score*) between the SSE and RL figures summarized in Tables 1, 2, 3, 4, 5 and 6, we can conclude that:

- *Scores* with respect to the standard MDAV algorithm (Tables 1, 3 and 5) are above 1.0 in all cases for the Census data set for $k \leq \sqrt{n}$ and for Adult when $\varepsilon = 0.01$, 0.1 or 1.0. This shows that the reduced disclosure risk brought by ε-differential privacy more than compensates the relative increase of information loss caused by noise in most configurations. *Scores* against standard microaggregation tend to decrease as the microaggregation level k increases for larger ε values (which produce less random outputs). This is explained by the sharp decrease in RL achieved by the standard microaggregation as k grows, which is more noticeable than the improvement in data utility achieved when reducing the required noise of differential privacy via prior microaggregation. This suggests that standard microaggregation, even though not offering as robust theoretical privacy guarantees as differential privacy, may achieve comparable results for large k.
- *Scores* with respect to standard ε-differential privacy (Tables 2, 4 and 6) tend to increase as ε grows. This is especially noticeable for the Census data set for $\varepsilon \geq 1.0$, thanks to the substantial information loss reduction obtained by using k-anonymous microaggregation prior to ε-differential privacy and the comparatively lower number of record linkages. For the two Adult data sets, an $\varepsilon = 10.0$ is needed to notice the significant improvement. In both cases, lower ε values tend to show no improvement because of the impossibility of reducing even more the number of record linkages that occur by chance, which define a minimum threshold achieved by the standard differential privacy approach.

5.4 Statistical analysis of anonymized results

To complement the above evaluation, in this section we provide an attribute-level analysis for some general statistical queries for the numerical data set (Census). As in [15], Θ and Θ' denote the same statistic (*e.g.*, attribute mean, attribute variance, etc.) for each attribute in the original data set and in its anonymized version (by means of k-anonymity and/or ε-differential privacy), respectively, we computed the variation of the statistic introduced by the anonymization process as:

$$\Delta(\Theta) = \frac{\mid \Theta' - \Theta \mid}{\mid \Theta \mid}.$$

Variations were computed for the *mean* of each attribute (named $\Delta(m_X)$ for FEDTAX, $\Delta(m_P)$ for POTHVAL, $\Delta(m_I)$ for INTVAL and $\Delta(m_F)$ for FICA) and also for their *variances* ($\Delta(\sigma_X)$ for FEDTAX, $\Delta(\sigma_P)$ for POTHVAL, $\Delta(\sigma_I)$ for INTVAL and $\Delta(\sigma_F)$ for FICA). In both cases, the smaller the variations, the less is the information loss and the better is the data utility. Results are reported in Table 7.

The variations of the *attribute means* directly depend on the amount of noise added to the anonymized output. Hence, for the two k-anonymous MDAV implementations, attribute means are perfectly preserved in the masked output since centroids are the exact means of clustered values. Regarding differentially privacy implementations, we observe, for most attributes, a decrease for the variations of the mean against the standard ε-differential privacy when the k-anonymity factor applied to input data is $k > \sqrt{n}$, that is $k \geq 33$. Notice that for $k > \sqrt{n}$, the use of the prior microaggregation effectively reduces the sensitivity and, thus, the amount of noise added to the output in comparison with the standard ε-differential privacy. For fixed ε, the sharpness of this decrease is similar for all attributes. However, as ε increases from 0.01 to 10.0, the decrease becomes sharper and sharper for all attributes. Indeed, for $\varepsilon = 0.01$ the decrease for the variation of the mean is barely noticeable, whereas for $\varepsilon = 10.0$ the decrease is of two orders of magnitude when $k \approx 2\sqrt{n}$. Hence, we see that $\varepsilon > 0.1$ and $k > \sqrt{n}$ are needed to significantly improve (reduce) on the baseline variations of the mean for all attributes.

The variations of the *attribute variances* increase for the two MDAV implementations as the k-anonymity level grows, since output record values tend to be more homogeneous and thereby suppress more variance as a result of the data aggregation process. The growth factor is larger for the standard MDAV algorithm in comparison with its insensitive version, because for small k the former produces significantly more homogeneous

Table 7 Census data set. Variation for several statistics between the original data set and data sets anonymized with methods using different values of k (2, $\sqrt{n}$ and $2\sqrt{n}$) and ε. Methods include standard ε-differential privacy (no prior microaggregation), ε-differential privacy with prior k-anonymous microaggregation, insensitive MDAV microaggregation and plain MDAV microaggregation.

Statistic	Microaggregation	$\varepsilon = 0.01$	$\varepsilon = 0.1$	$\varepsilon = 1.0$	$\varepsilon = 10.0$	Insensit. MDAV	MDAV
$\Delta(m_X)$	None	1.0947	1.0356	0.6925	0.0500	0.0	0.0
	k=2	1.1119	1.1058	1.0579	1.1341	0.0	0.0
	k=33	1.0762	1.0432	0.6998	0.0454	0.0	0.0
	k=66	1.1362	0.8782	0.2209	0.0045	0.0	0.0
$\Delta(m_P)$	None	14.5160	13.7959	9.0362	1.2125	0.0	0.0
	k=2	14.3173	13.5792	13.9893	13.9908	0.0	0.0
	k=33	14.6917	12.6272	8.9438	1.2064	0.0	0.0
	k=66	14.0284	11.128	3.2560	0.1802	0.0	0.0
$\Delta(m_I)$	None	25.4754	24.6380	15.9486	2.2799	0.0	0.0
	k=2	25.0788	22.8018	24.7786	24.5869	0.0	0.0
	k=33	25.1908	21.9300	15.8544	2.2675	0.0	0.0
	k=66	24.5861	18.9270	5.9530	0.4013	0.0	0.0
$\Delta(m_F)$	None	1.0126	0.9648	0.6151	0.0270	0.0	0.0
	k=2	0.9991	1.0004	1.0133	1.0035	0.0	0.0
	k=33	1.0145	0.9493	0.5708	0.0362	0.0	0.0
	k=66	1.0249	0.8142	0.1797	0.0014	0.0	0.0
$\Delta(\sigma_X)$	None	9.5299	9.2111	6.4855	0.5122	0.0	0.0
	k=2	9.5698	9.5637	9.5486	9.4422	0.0447	0.0053
	k=33	9.5308	9.2292	6.4280	0.4171	0.0804	0.0156
	k=66	9.4383	8.2329	2.0797	0.1071	0.1015	0.0398
$\Delta(\sigma_P)$	None	69.3473	67.3757	45.9873	2.1168	0.0	0.0
	k=2	69.7030	69.7434	69.5286	68.8324	0.0697	0.0247
	k=33	69.4637	67.2577	45.7902	1.8889	0.1268	0.0991
	k=66	68.7018	59.6482	11.6906	0.3972	0.2429	0.1967
$\Delta(\sigma_I)$	None	96.3225	93.3506	64.2854	3.0044	0.0	0.0
	k=2	96.5684	96.5631	96.4935	95.4336	0.0950	0.0349
	k=33	96.4524	93.1380	63.0454	2.5768	0.1358	0.1327
	k=66	95.2471	82.6625	16.1619	0.4323	0.2362	0.2614
$\Delta(\sigma_F)$	None	16.3302	15.7698	11.3811	0.9712	0.0	0.0
	k=2	16.3545	16.3723	16.3610	16.1667	0.0593	0.0067
	k=33	16.3165	15.7810	11.0145	0.8795	0.1117	0.0224
	k=66	16.1566	14.1721	3.8468	0.1436	0.1634	0.0670

clusters than the latter, but for larger k differences become less marked. Differential privacy implementations behave the other way round. For all ε values, the variations of attribute variances decrease as the k-anonymity level grows, for all attributes. In the same manner as for the variations of means, an effective improvement over the standard ε-differential privacy is observed in most cases when $k > \sqrt{n}$, especially for larger ε. This suggests that prior microaggregation helped decrease the large variance introduced by the noise added to fulfill differential privacy. Again, decrease factors for variations of variances are larger for higher ε values.

It is important to note that results with $\varepsilon = 10.0$ are quite similar to those reported for the insensitive MDAV implementation. On the other hand, results for $\varepsilon = 0.01$ are so noisy for any microaggregation level that they barely retain any utility. This suggests that we can obtain differentially private results with a level of statistical utility comparable to those of k-anonymity when the ε is relaxed enough.

Finally, the results of the above analysis of attribute-level statistics are coherent with the results based on SSE presented in previous sections. It becomes clear that, for reasonable values of ε, prior microaggregation helps ε-differentially private data to retain the utility of original data much like standard k-anonymity does.

6 Conclusions

We have presented an approach that combines k-anonymity and ε-differential privacy in order to reap the best of both models: namely, the reasonably low information loss incurred by k-anonymity and its lack of assumptions on data uses, and the robust privacy guarantees offered by ε-differential privacy. In our approach, we use a newly defined insensitive microaggregation to obtain a k-anonymous data set by considering all attributes as quasi-identifiers; then we take the k-anonymous microaggregated data set as an input to which uncertainty is added in order to reach ε-differential privacy. We have also described how our approach can be applied to numerical and categorical attributes and also to records combining heterogeneous attribute types.

In addition to a theoretical proposal, we have presented empirical results for heterogeneous data sets which show that our approach reduces the information loss of standard differential privacy, while preserving its theoretical privacy guarantee and improving the practical privacy (percentage of record linkages).

In particular, empirical results showed that, for reasonable values of ε, and thanks to the ability of the microaggregation mechanism to exploit the underlying structure of data, the loss of utility incurred by the k-anonymous microaggregation step is more than compensated by the benefits brought by the noise reduction for the ε-differential privacy stage.

Future work will involve at least the following research lines:

- Even though special care has been exerted to avoid damaging within-cluster homogeneity when making microaggregation insensitive, there is still room for improvement, especially for categorical data. New criteria to define total orders are conceivable, such as fixing sampling and sorting strategies of data spaces, so that the within-cluster homogeneity reaches levels more similar to the ones achieved by standard microaggregation.
- It would also be interesting to define a methodology that, given a data set, a target privacy level ε and fixed utility and privacy measures, determines the most suitable k for the prior k-anonymous microaggregation, in view of optimizing the data utility and/or disclosure risk.

Disclaimer and acknowledgments

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TIN2011-27076-C03-01 "CO-PRIVACY", TIN2012-32757 "ICWT", IPT2012-0603-430000 "BallotNext" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Commission under FP7 projects "DwB" and "Inter-Trust". The second author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization.

References

1. Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigraphy, R., Thomas, D., Zhu, A.: Anonymizing tables. In: Proc. of the 10th International Conference on Database Theory-ICDT 2005, pp. 246-258 (2005)
2. Batet, M., Valls, A., Gibert, K.: A distance function to assess the similarity of words using ontologies. In: XV Congreso Español sobre Tecnologías y Lógica Fuzzy, Huelva, pp. 561-566. Spain (2010)
3. Blum, A., Ligett, K., Roth, A.: A learning theory approach to non-interactive database privacy. In: Proc. of the 40th Annual Symposium on the Theory of Computing-STOC 2008, pp. 609-618 (2008)
4. Brand, R., Domingo-Ferrer, J., Mateo-Sanz, J.M.: Reference data sets to test and compare SDC methods for protection of numerical microdata. European Project IST-2000-25069 CASC. `http://neon.vb.cbs.nl/casc` (2002)
5. Charest, A.-S.: How can we analyze differentially-private synthetic data sets? J. of Priv. and Confident. 2(2), 21-33 (2010)
6. Charest, A.-S.: Empirical evaluation of statistical inference from differentially-private contingency tables. In: Proc. of Privacy in Statistical Databases-PSD 2012, LNCS 7556, pp. 257-272. Springer (2012)
7. Chen, R., Mohammed, N., Fung, B.C.M., Desai B.C., Xiong, L.: Publishing set-valued data via differential privacy. In: 37th Intl. Conference on Very Large Data Bases-VLDB 2011/Proc. of the VLDB Endowment 4(11), 1087-1098 (2011)
8. Clifton, C., Tassa, T.: On syntactic anonymity and differential privacy. Transactions on Data Privacy 6(2), 161-183 (2013)
9. Cormode, G., Procopiuc, C. M., Shen, E., Srivastava, D., Yu, T.: Differentially private spatial decompositions. In: IEEE International Conference on Data Engineering (ICDE 2012), pp. 20-31 (2012)
10. Cormode, G., Procopiuc, C. M., Shen, E., Srivastava, D., Yu, T.: Empirical privacy and empirical utility of anonymized data. In: ICDE Workshop on Privacy-Preserving Data Publication and Analysis (2013).
11. Dalenius, T.: The invasion of privacy problem and statistics production. An overview. Statistik Tidskrift 12, 213-225 (1974)
12. Dandekar, R., Domingo-Ferrer, J., Sebé, F.: LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases, LNCS 2316, pp. 153-162. Springer (2002)
13. Domingo-Ferrer, J.: Marginality: a numerical mapping for enhanced exploitation of taxonomic attributes. In: Proc. of the 9th International Conference on Modeling Attributes for Artificial Intelligence-MDAI 2012, LNCS 7647, pp. 367-381. Springer (2012)
14. Domingo-Ferrer, J., Sánchez, D., Rufian-Torrell, G.: Anonymization of nominal data based on semantic marginality. Inf. Sci. 242, 35-48 (2013)
15. Domingo-Ferrer, J., González-Nicolás, U.: Hybrid microdata using microaggregation. Inf. Sci. 180(15), 2834-2844 (2010)
16. Domingo-Ferrer, J., Martínez-Ballesté, A., Mateo-Sanz J., Sebé, F.: Efficient multivariate data-oriented microaggregation. VLDB J. 15, 355-369 (2006).
17. Domingo-Ferrer, J., Mateo-Sanz J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. on Knowl. and Data Eng. 14(1), 189-201 (2002)
18. Domingo-Ferrer, J., Mateo-Sanz, J.M., Torra, V.: Comparing SDC methods for microdata on the basis of information loss and disclosure risk. In: Pre-proceedings of ETK-NTTS'2001 (vol. 2), pp. 807-826. Eurostat (2001)
19. Domingo-Ferrer J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min. and Knowl. Discov. 11(2), 195-212 (2005)
20. Domingo-Ferrer, J.: A critique of k-anonymity and some of its enhancements. In: Proc. of ARES/PSAI 2008, pp. 990-993. IEEE Computer Society (2008)
21. Domingo-Ferrer, J., Sebé, F., Solanas, A.: A polynomial-time approximation to optimal multivariate microaggregation. Comp. & Math. with Appl. 55(4), 714-732 (2008)
22. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Min. and Knowl. Discov. 11(2), 95-212 (2005)
23. Dwork, C.: Differential privacy. In: Proc. of 33rd International Colloquium on Automata, Languages and

Programming-ICALP 2006, LNCS 4052, pp. 1-12. Springer (2006)

24. Dwork, C.: A firm foundation for private data analysis. Comm. of the ACM 54(1), 86-95 (2011)
25. Dwork, C., Naor, M., Reingold, O., Rothblum G.N., Vadhan, S.: On the complexity of differentially private data release: efficient algorithms and hardness results. In: Proc. of the 41st Annual Symposium on the Theory of Computing-STOC 2009, pp. 381-390 (2009)
26. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
27. Frank, A., Asuncion, A.: UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. `http://archive.ics.uci.edu/ml/datasets/Adult` (2010)
28. Fung, B.C.M., Wang, K., Yu., P.S.: Top-down specialization for information and privacy preservation. In: Proc. of the 21st international conference on data engineering, pp. 205-216. IEEE Computer Society (2005)
29. Goldberger, J., Tassa, T.: Efficient anonymizations with enhanced utility. Transactions on Data Privacy 3, 149-175 (2010)
30. Hardt, M., Ligett, K., McSherry, F.: A simple and practical algorithm for differentially private data release. Preprint arXiv:1012.4763v1 (2010)
31. Hay, M., Rastogi, V., Miklau, G., Suciu, D.: Boosting the accuracy of differentially private histograms through consistency. PVLDB 3(1): 1021-1032 (2010)
32. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., de Wolf, P.-P.: Statistical Disclosure Control. Wiley (2012)
33. Laszlo, M., Mukherjee, S.: Minimum spanning tree partitioning algorithm for microaggregation. IEEE Trans. on Knowl. and Data Eng. 17(7), 902-911 (2005)
34. Li, Y., Bandar, Z., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. on Knowl. and Data Eng. 15, 871-882 (2003)
35. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: privacy beyond k-anonymity and l-diversity. In: IEEE International Conference on Data Engineering (ICDE 2007), pp. 106-115 (2007).
36. Li, N., Yang, W., Qardaji, W.: Differentially private grids for geospatial data. In: IEEE International Conference on Data Engineering (ICDE 2013), pp. 757-768 (2013)
37. Lin, J.-L., Wen, T.-H., Hsieh, J.-C., Chang, P.-C.: Density-based microaggregation for statistical disclosure control. Expert Syst. with Appl. 37, 3256-3263 (2010)
38. Li, N., Qardaji, V., Su, D.: On sampling, anonymization, and differential privacy: Or, k -anonymization meets differential privacy. In: 7th ACM Symposium on Information, Computer and Communications Security (ASIACCS'2012), pages 32-33 (2012)
39. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: l-Diversity: privacy beyond k-anonymity. In: IEEE International Conference on Data Engineering (ICDE 2006), pp. 24 (2006).
40. Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., Vilhuber, L.: Privacy: theory meets practice on the map. In: IEEE International Conference on Data Engineering (ICDE 2008), pp. 277-286 (2008)
41. Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for data sets with textual attributes. Knowl.-Based Syst. 35, 160-172 (2012)
42. Martínez, S., Sánchez, D., Valls, A.: Semantic adaptive microaggregation of categorical microdata. Comp. and Secur. 31(5), 653-672 (2012)
43. McSherry, F.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, pp. 19-30. ACM (2009)
44. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: Proc. of the 48th Annual IEEE Symposium on Foundations of Computer Science-FOCS 2007, pp. 94-103 (2007)
45. Mohammed, N., Chen, R., Fung, B.C.M., Yu, P.S.: Differentially private data release for data mining. In: Proc. of the 17th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining-KDD 2011, pp. 493-501. ACM (2011)
46. Nissim, K., Raskhodnikova, S., Smith, A.: Smooth sensitivity and sampling in private data analysis. In: Proc. of the 39th ACM symposium on Theory of computing-STOC 2007, pp. 75-84. ACM (2007)
47. Petrakis, E.G.M., Varelas, G., Hliaoutakis, A., Raftopoulou, P.: X-similarity: computing semantic similarity between concepts from different ontologies. J. of Dig. Inf. Manag. 4, 233-237 (2006)
48. Pirró, G.: A semantic similarity metric combining features and intrinsic information content. Data and Knowl. Eng. 68, 1289-1308 (2009)
49. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Trans. on Syst., Man and Cybern. 19(1), 17-30 (1989)
50. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. on Knowl. and Data Eng. 13(6), 1010-1027 (2001)
51. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. SRI International Report (1998)
52. Sánchez, D., Batet, M.: Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. J. of Biomed. Inform. 44, 749-759 (2011)
53. Sánchez, D., Batet, M.: A new model to compute the information content of concepts from taxonomical knowledge. Int. J. on Semant. Web and Inf. Syst. 8, 34-50 (2012)
54. Sánchez, D., Batet, M., Isern, D.: Ontology-based information content computation. Knowl.-Based Syst. 24, 297-303 (2011)
55. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: a new feature-based approach. Expert Syst. with Appl., 39(9), 7718-7728 (2012)
56. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Martínez, S.: Improving the utility of differentially private data releases via *k*-anonymity. In: 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications-IEEE TrustCom 2013. Melbourne, Australia, July 16-18 (2013) (to appear)
57. Sweeney, L.: *k*-Anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowl.-based Syst., 10(5), 557-570 (2002)
58. Willenborg, L., De Waal, T.: Statistical Disclosure Control in Practice. Springer (1996)
59. Wong, R., Li, J., Fu, A., and Wang, K.: (α, k)-Anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), pp. 754-759 (2006).
60. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proc. of the 32nd Annual Meeting on Association for

Computational Linguistics, pp. 133-138. Las Cruces, New Mexico (1994)

61. Xiao, X., Wang, G., Gehrke, J.: Differential Privacy via Wavelet Transforms. IEEE Trans. on Knowl. and Data Eng. 23(8), pp. 1200-1214 (2010)
62. Xiao, Y., Xiong, L., Yuan, C.: Differentially private data release through multidimensional partitioning. In: Proceedings of the 7th VLDB conference on Secure data management (SDM'10), pp. 150-168 (2010)
63. Xu, J., Zhang, Z., Xiao, X., Yang, Y., Yu, G.: Differentially Private Histogram Publication. In: IEEE International Conference on Data Engineering (ICDE 2012), pp. 32-43 (2012)
64. Yancey, W.E., Winkler, W.E., Creecy, R.H.: Disclosure risk assessment in perturbative microdata protection. In: Domingo-Ferrer, J. (ed.), Inference Control in Statistical Databases, LNCS 2316, pp. 135-152. Springer (2002)