Improving the Utility of Differentially Private Data Releases via k-Anonymity

Jordi Soria-Comas, Josep Domingo-Ferrer, David Sánchez, and Sergio Martínez
Dept. of Computer Engineering and Mathematics. Universitat Rovira i Virgili.
UNESCO Chair in Data Privacy.
Tarragona, Catalonia
{*jordi.soria, josep.domingo, david.sanchez, sergio.martinezl*}*@urv.cat*

***Abstract*—A common view in some data anonymization literature is to oppose the “old” k-anonymity model to the “new” differential privacy model, which offers more robust privacy guarantees. However, the utility of the masked results provided by differential privacy is usually 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 the general-purpose anonymized data resulting from k-anonymity mechanisms, which also focus on preserving data utility. In this paper, we show that a synergy between differential privacy and k-anonymity can be found when the objective is to release anonymized data: k-anonymity can help improving the utility of the differentially private release. 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 analytical utility of the anonymized output data set is increased. The theoretical benefits of our proposal are illustrated in a practical setting with an empirical evaluation on a reference data set.**

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

I. Introduction

Publishing microdata (*e.g.* responses to polls, census information, healthcare records) is of great interest for the data analysis community. At the same time, microdata may contain sensitive information about individuals. To overcome this privacy threat, data should be anonymized before making them available [12].

In the last two decades, several models for anonymization of data have been proposed in the literature. One of the best-known and widely used is k-anonymity [16], 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 attribute generalization and local suppressions [15], [18]. An alternative procedure, especially suitable for attributes with no obvious generalization hierarchy (like the numerical ones), is microaggregation [7], [6]. 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.). 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 [8].

On the other hand, ε-differential privacy [9] 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 k-anonymity model. To do so, ε-differential privacy requires adding an amount of noise to the anonymization 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 a masked version of the data, the anonymizer returns noise-added answers to interactive queries. Compared to the general-purpose data publication offered by k-anonymity, which makes no assumptions on uses of published data, the interactive scenario of ε-differential privacy severely limits data analysis, because it only allows answering queries whose number and type are limited. Otherwise, an adversary could reconstruct some of the original data [4].

It is pointed out in [2] that the previous limitation can be circumvented by allowing an ε-differentially private data publication (*i.e.*, a *non-interactive* setting) which supports answering an unlimited number of potentially heterogeneous queries. However, since ε-differential privacy should ensure that the probability distribution of the published records is not changed by *any* modification of a single input record, the amount of noise that needs to be added to the published data in such a general setting is so large that it would severely hamper data utility [4]. This problem can be minimized in specific scenarios, but at the expense of preserving usefulness only for restricted classes of queries [2], [10], [11].

In summary, we can conclude that k-anonymity enables general-purpose data publication with reasonable utility 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 generality and/or utility of anonymized outputs.

A. Contribution and plan of this paper

We show here that a synergy between both privacy models can be found in order to achieve general-purpose ε-differentially private data publication that makes as few assumptions on the data uses as k-anonymity does. In this scenario, we show how k-anonymity can help increasing the utility of differentially private outputs. Specifically, the amount of noise required to fulfill ε-differential privacy in such a general setting can be greatly reduced if noise is applied to a k-anonymous version of the data set obtained through microaggregation of all attributes (instead of applying it to raw input data). 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, the data utility of general-purpose data publication can be improved without renouncing the strong privacy guarantee of ε-differential privacy.

Section II reviews background and related work on k-anonymity and ε-differential privacy. Section III proposes a general algorithm for generating ε-differentially private data sets. Section IV provides an empirical evaluation of the differentially private output obtained from a reference data set via k-anonymous microaggregation. Section V presents the conclusions and proposes some lines of future research.

II. Related work

A. k-Anonymity

As mentioned above, k-anonymity [16], [15], [18] 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. Several fixes/alternatives to k-anonymity also based on the idea of data set partitioning have appeared: l-diversity, t-closeness, (c, k)-safety, etc. However, none of those alternatives is free from shortcomings, see [8] for a critical survey.

In [7], it is shown how to achieve k-anonymity via microaggregation. Microaggregation [6] 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, records within each cluster are replaced by a representative of the cluster, typically the centroid record.

Clearly, when microaggregation is applied to the projection of records on their quasi-identifier attributes, the resulting data set is k-anonymous. In [7] 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.

B. Differential privacy

Differential privacy was originally proposed by [9] 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 D_1, D_2 such that one can be obtained from the other by modifying a single record, and all $S \subset Range(\kappa)$

$$P(\kappa(D_1) \in S) \leq \exp(\varepsilon) \times P(\kappa(D_2) \in S) \qquad (1)$$

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 $f(D)$ of the response to a certain user query f is computed, and then a random noise, say $Y(D)$, is added to mask $f(D)$, that is, a randomized response $\kappa(D) = f(D) + Y(D)$ is returned. To generate $Y(D)$, a common choice is to use a Laplace distribution with zero mean and $\Delta X/\varepsilon$ scale parameter, where:

- ε is the differential privacy parameter.
- Δf 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 Δf 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 Δf, 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.

Despite presented as an interactive mechanism, differential privacy has also been used in the non-interactive setting in [2], [10], [11], [4]. 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 preserving utility only for restricted classes of queries (typically count queries). This contrasts with the general-purpose data release offered by the k-anonymity model.

In [14], 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.

III. Differential privacy through k-anonymity

Assume that we have an original data set X and that we want to generate a data set X_ε —an anonymized version of X— that satisfies ε-differential privacy. 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 the query function that returns the attribute values contained in record r of X. We generate X_ε, by querying X with $I_r(X)$, for all $r \in X$. 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 X_ε also satisfies ε-differential privacy.

Algorithm 1 Generation of an ε-differentially private data set X_ε from X via microaggregation

```
let $X$ be the original data set
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 record $r$
$\overline{X} \leftarrow$ microaggregated data set $M(X)$
for each $r \in \overline{X}$ do
    $r_\varepsilon \leftarrow S_\varepsilon(I_r(\overline{X}))$
    insert $r_\varepsilon$ into $X_\varepsilon$
end for
return $X_\varepsilon$
```

The proposed approach for generating X_ε is general but naive. As each query $I_r()$ refers to a single individual, its sensitivity is large; 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 new step in the masking process: (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 all $r \in \overline{X}$.

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 allows the sensivity to be effectively reduced.

Even though the prior k-anonymous microaggregation also incurs in 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 1 details the procedure for generating the differentially private data set X_ε.

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 $I_r()$ to be effectively reduced the centroids must be stable against modifications of one record in the original data set X. This means that modification of a single record in 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 better data utility preservation. 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 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 (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 relationship.

Definition 4: A distance function $d : X \times X \to \mathbb{R}$ is said to be consistent with an order relationship $\leq_X$ 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:

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

We want to check that modifying a single record of X leads to a set of clusters that differ, at most, in 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 been added to C'_1, $C'_1 = (C_1 \cup \{x^2_1\}) \setminus \{x\}$. Clusters $C'_2, \ldots, C'_{m-1}$ contain the same elements as the respective cluster $C_2, \ldots, C_{m-1}$, except for x^i_1 that has been removed from C'_i and x^{i+1}_1 that has been added to C'_i. Cluster C'_m contains the same elements as C_m except by x^m_1 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. ∎

A. Achieving differential privacy with numerical attributes

For a data set consisting of numerical attributes, generating the ε-differentially private data set X_ε as previously described is quite straightforward. We use an insensitive MDAV as the microaggregation algorithm, and Laplace noise to mask the value of the queries $I_r(\overline{X})$.

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. To make MDAV insensitive, we need to define a total order relationship 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 relationships 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 with which the distance is consistent as follows:

Definition 5: 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 define 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.

To increase within-cluster homogeneity, microaggregation algorithms usually start 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 $(a_b^1, \ldots, a_b^m)$.

The following proposition shows that by using an intermediate k-anonymous data set the magnitude of the Laplace noise required to generate X_ε is reduced. In particular, it turns out that the scale parameter of the Laplace noise is reduced by a factor $1/k$.

Proposition 2: Let X be a data set with numerical attributes only. 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 adding to $\overline{X}$ an amount of Laplace noise that would only achieve $k\varepsilon$-differential privacy if directly added to X.

Proof: We assume that the data set X contains a single numerical attribute. For a data set with multiple attributes, the same result would be obtained by treating the attributes separately.

Let ΔI_r be the sensitivity of function I_r when applied directly to the data set X. The sensitivity of ΔI_r amounts to the difference between the maximum and the minimum possible values of the attribute. To achieve ε-differential privacy using Laplace noise, the Laplace scale parameter must be set to $\Delta I_r/\varepsilon$.

Let us now turn to the k-anonymous data set $\overline{X}$. The sensitivity of I_r when applied to $\overline{X}$ amounts to the maximum possible change in a cluster centroid due to modification of a single record in X. As M is an insensitive microaggregation, modifying a single record in X changes each cluster by at most one record; therefore, the maximum change in a centroid is $\Delta I_r/k$. To achieve ε-differential privacy using Laplace noise, the scale parameter must be set to $\Delta I_r/(k\varepsilon)$.

From the scale parameter of the Laplace noise required to attain differential privacy when evaluating I_r over X and $\overline{X}$, it is clear that 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. ■

IV. Empirical evaluation

In this section we show some empirical results that illustrate how k-anonymous microaggregation of input data reduces the amount of noise required to fulfill differential privacy. Due to space constraints, we focus on the case of numerical attributes.

The above-described mechanism has been applied to the reference data set "Census", which contains 1080 records with 13 numerical attributes [3]. This data set was used in the European project CASC. Like in [5], we take 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 will be masked, *i.e.* they will all be considered as quasi-identifiers in all our tests. Since all of them represent non-negative money amounts, we bound the attribute domains to the range $minimum = 0$ and $maximum = 1.5 \times max\_attribute\_value\_in\_the\_dataset$. The latter domain upper bound 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 *minimum* and *maximum* defines the sensitivity of each attribute and influences the amount of Laplace noise to be added to masked outputs, as detailed in Section III-A. Since the Laplace distribution takes values in the range

$(-\infty, +\infty)$, for consistency we bound noise-added outputs to the $[minimum, maximum]$ range defined above.

The quality of the masked output for different combinations of k-anonymity and ε-differential privacy levels has been evaluated from the perspectives of *information loss*, which directly influences data utility, and *disclosure risk*, which measures practical privacy:

- Information loss has been quantified by means of the Sum of Squared Errors (SSE), a measure used in a good deal of the anonymization literature (*e.g.* [6]). 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 distances between original record tuples in X and their versions in the anonymized data set, that is

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

  where x_j represents the j-th original record and x'_j is its version in the anonymized data set. Since we are dealing with numerical attributes, $dist(\cdot, \cdot)$ corresponds to the standard Euclidean distance. 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.
- On the other hand, 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)

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

  where m is the number of original records and the record linkage probability for an anonymized record ($Pr(x'_j)$) is calculated as

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

  where G is the set of original records that are at minimum distance from x'_j. For numerical attributes, the Euclidean distance can be 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: *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.

As baseline results, we have computed SSE and RL scores for a standard k-anonymity scenario in which all attributes are microaggregated by means of the original MDAV algorithm [7], and also with its modified insensitive version described in Section III. Furthermore, we also considered the straightforward ε-differential privacy scenario in which Laplace noise is directly added to unaggregated inputs; this approach is equivalent to applying our method with a k-anonymity level of $k = 1$.

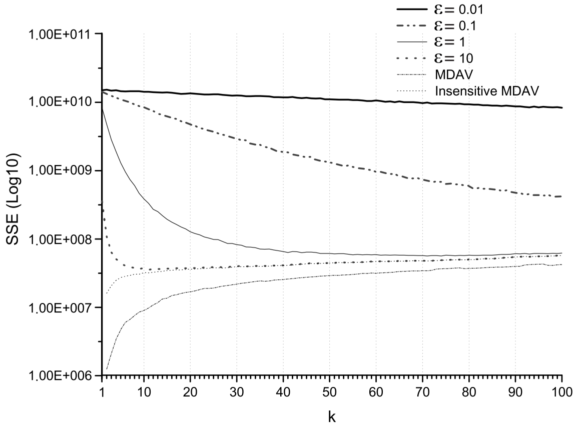

Figure 1. SSE scores for different k and ε values for the "Census" data set.

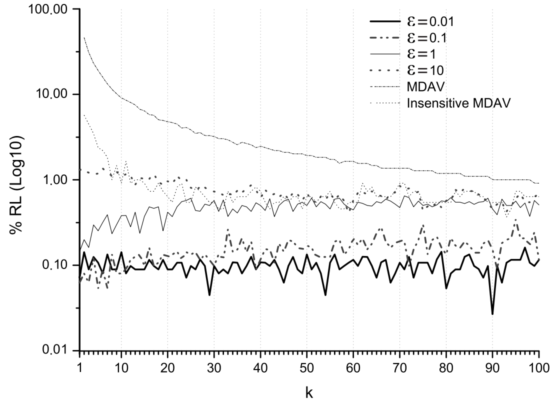

Figure 2. RL percentages for different k and ε values for the "Census" data set.

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. The k-anonymity levels have been set between 2 and 100, except for the raw sensitive and insensitive MDAV microaggregations, which start from $k = 2$, because $k = 1$ would mean that input data are not modified.

Figures 1 and 2 depict, respectively, the SSE and RL scores for the different parameterizations of k and ε. Due to the broad ranges of the SSE and RL scores, the Y-axes are represented using a $\log_{10}$ scale. Each test involving Laplace noise shows the averaged results of 10 runs, for the sake of stability.

Regarding the evolution of SSE scores, we observe in Figure 1 that k-anonymous microaggregation of input records effectively reduces the required amount of noise and hence the loss of information, compared to a straightforward implementation of ε-differential privacy (with no prior microaggregation, *i.e.* $k = 1$). Given that the figure depicts SSE in a $\log_{10}$ scale, this reduction is actually of several orders of magnitude.

When combining prior k-anonymous microaggregation ($k > 1$) with ε-differential privacy, we observe different effects depending on the value of ε:

- For small ε (that is, 0.01 or 0.1), the larger k, the smaller is SSE, because the noise reduction at the ε-differential privacy stage more than compensates the noise increase at the microaggregation stage due to greater aggregation. Anyway, the amount of noise involved for these values of ε is so high that even with the aforementioned noise reduction, the output data are hardly useful.
- For very large ε (that is, 10), there is a sharp decline of SSE for low k values (around 10); however, for larger k (above 10), there is a new and slow increase in SSE, because the noise added by ε-differential privacy being low, it is dominated by the noise added by prior microaggregation in larger clusters.
- For medium ε (that is, 1), there is a substantial decline of SSE for low k and, for larger k, SSE stays nearly constant and reasonably low. In this case, the noise added by prior microaggregation in larger clusters is compensated by the noise reduced at the ε-differential privacy stage (due to decreased sensitivity with larger k).

Notice also that insensitive MDAV microaggregation incurs a higher SSE than standard MDAV microaggregation, as anticipated in Section III. Indeed, the clusters formed by insensitive microaggregation are less homogeneous, due to the total order enforced for input records. In any case, the SSE increase caused by insensitive microaggregation is several orders of magnitude smaller than the noise reduction this microaggregation enables when used as a prior step to ε-differential privacy.

RL scores behave the other way round as SSE. First, we notice in Figure 2 that the standard MDAV algorithm results in the highest percentage of linkages; a k-anonymity level $k \geq 20$ is needed to attain a percentage of linkages below 5%. Insensitive MDAV yields noticeably more private results, in return for less homogeneous clusters and more information loss. The RL scores of insensitive microaggregation are very similar to the ones obtained with ε-differential privacy with $\varepsilon = 10$. For ε values of 0.01 and 0.1, the RL scores hardly vary when the k-anonymity level increases, because they are very low already with $k = 1$ (no prior microaggregation). Note that, for such low ε-values, the RL scores stay around 0.1% which, considering the data set size of 1080 records, corresponds to the probability of successful random record linkage (*i.e.*, 1/1080). The fact that records are almost randomly matched is reflected by the large spikes of the plot. It can also be seen that the top level of privacy offered by standard ε-differential privacy ($k = 1$) for low ε is maintained when using prior microaggregation ($k > 1$), so the reduction in information loss offered by the latter approach is achieved without privacy penalties.

For $\varepsilon = 1$, RL results are more interesting. They show an increase of the percentage of record linkages from 1.5% for $k = 1$ (no prior microaggregation) to around 5% for $k = 25$. This is the other side of the very noticeable improvement of SSE scores shown in Figure 1. Considering that SSE values were reduced by around two orders of magnitude from $k = 1$ to $k = 25$, we can conclude that, *for intermediate values of ε (around 1)*:

- The very substantial information loss reduction obtained by using k-anonymous microaggregation prior to ε-differential privacy more than compensates the small increase of record linkages with respect to standard ε-differential privacy.
- At the same time, while for $\varepsilon = 1$ and large k the information loss achieved by k-anonymous microaggregation prior to ε-differential privacy is similar to the one achieved by standard or insensitive k-anonymous microaggregation, the privacy level attained by the former approach is much higher.

The above observations suggest that, given a desired level ε of differential privacy and a specific data set, a k-anonymity degree can be determined that optimally balances data utility and privacy.

V. Conclusions

Our approach combines k-anonymity and ε-differential privacy to reap the best of each approach for anonymized data publishing: 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. 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.

In addition to a theoretical proposal, we have presented empirical results for numerical data which show that our combined approach reduces information loss by several orders of magnitude, while preserving the theoretical privacy guarantee of differential privacy and improving the practical privacy (percentage of record linkage) versus standard k-anonymity.

Future work will involve the following research lines:

- Improve our heuristics for insensitive microaggregation, so that the within-cluster homogeneity reaches levels more similar to the ones achieved by standard microaggregation. Better defining and exploiting the extreme values of the domains might help in this respect.
- Adapt the proposed procedure to work with categorical data. Unlike for numerical attributes, categorical attributes take values from a finite set of, usually, non-ordinal categories. Hence, appropriate operators to microaggregate and to add noise to the outputs should be defined [13], [17].
- Analyze information loss/data utility using measures other than SSE including statistics like means, variances and covariances (*e.g.* used for utility analysis in [5]).

Acknowledgment

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS", TIN2011-27076-C03-01 "CO-PRIVACY", TIN2012-32757 "ICWT", PT2012-0603-430000 "Ballot-Next" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission 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] C. Blake and C. Merz. Adult Data Set. UCI repository of machine learning databases, 1998. http://archive.ics.uci.edu/ml/datasets/Adult

[2] A. Blum, K. Ligett and A. Roth. 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.

[3] R. Brand, J. Domingo-Ferrer and J.M. Mateo-Sanz, Reference data sets to test and compare SDC methods for protection of numerical microdata, European Project IST-2000-25069 CASC, 2002. http://neon.vb.cbs.nl/casc

[4] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai and L. Xiong. 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.

[5] J. Domingo-Ferrer and U. González-Nicolás, Hybrid microdata using microaggregation. Information Sciences, 180(15):2834-2844, 2010.

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

[7] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, 11(2):195-212, 2005.

[8] J. Domingo-Ferrer. A critique of k-anonymity and some of its enhancements. In: Proc. of ARES/PSAI 2008, IEEE Computer Society, pp. 990-993, 2008.

[9] C. Dwork. Differential privacy. In: Proc. of 33rd International Colloquium on Automata, Languages and Programming-ICALP 2006, LNCS 4052, Springer, pp. 1-12, 2006.

[10] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In: Prof. of the 41st Annual Symposium on the Theory of Computing-STOC 2009, pp. 381-390, 2009.

[11] M. Hardt, K. Ligett and F. McSherry. A simple and practical algorithm for differentially private data release. Preprint arXiv:1012.4763v1, 21 Dec 2010.

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

[13] S. Martínez, A. Valls and D. Sánchez. Semantically-grounded construction of centroids for data sets with textual attributes. Knowledge-Based Systems, 35:160-172, 2012.

[14] N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu. Differentially private data release for data mining In: Proc. of the 17th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining-KDD 2011, ACM, pp. 493-501, 2011.

[15] P. Samarati. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010-1027, 2001.

[16] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI International Report, 1998.

[17] D. Sánchez, M. Batet, D. Isern and A. Valls. Ontology-based semantic similarity: a new feature-based approach. Expert Systems with Applications, 39(9):7718-7728, 2012.

[18] L. Sweeney. k-Anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 10(5):557-570, 2002.