Improving the utility of differential privacy via univariate microaggregation

David Sánchez, Josep Domingo-Ferrer, and Sergio Martínez

UNESCO Chair in Data Privacy,
Department of Computer Engineering and Mathematics
Universitat Rovira i Virgili,
Av. Països Catalans 26, E-43007 Tarragona, Catalonia

Abstract. Differential privacy is a privacy model for anonymization that offers more robust privacy guarantees than previous models, such as k-anonymity and its extensions. However, it is often disregarded that the utility of differentially private outputs is quite limited, either because of the amount of noise that needs to be added to obtain them or because utility is only preserved for a restricted type of queries. On the contrary, k-anonymity-like anonymization offers general purpose data releases that make no assumption on the uses of the protected data. This paper proposes a mechanism to offer general purpose differentially private data releases with a specific focus on the preservation of the utility of the protected data. Our proposal relies on univariate microaggregation to reduce the amount of noise needed to satisfy differential privacy. The theoretical benefits of the proposal are illustrated and in a practical setting.

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

1 Introduction

Data publication often faces privacy threats due to the confidentiality of the information that is released for secondary use. To tackle this problem, privacy models proposed in recent years within the computer science community [13] seek to attain a predefined notion of privacy, thus offering *a priori* privacy guarantees. Among such models, k-anonymity and the more recent ε-differential privacy have received a lot of attention.

k-Anonymity [28] seeks to make each record in the input data set indistinguishable from, at least, $k-1$ other records, so that the probability of re-identification of individuals is, at most, $1/k$. Different anonymization methods have been proposed to achieve that goal, such as removal of outlying records, generalization of values to a common abstraction [27, 30, 1, 17] or microaggregation [9, 8]. The latter method partitions a data set into groups at least k similar records and replaces the records in each group by a prototypical record (*e.g.* the centroid record, that is, the average record). Microaggregation stands out as particularly utility-preserving among the methods for k-anonymization. Indeed, microaggregation does not suffer from the loss of granularity inherent to value generalizations and can be adapted to the structure of data [10]. 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 [22, 31, 20, 7].

Unlike k-anonymity, the more recent ε-differential privacy [14] method does not make any assumption on the background knowledge available to potential attackers. ε-Differential privacy guarantees that the anonymized output is insensitive (up to a factor dependent on ε) to modification, deletion or addition of any single input record in the original data set. In this way, the privacy of any individual is not compromised by the publication of the anonymized output, which is a much more robust guarantee than the one offered by k-anonymity. The enforcement of ε-differential privacy requires adding noise to attribute values that depends on the *sensitivity* of such attributes to modification of input records. This sensitivity does not depend on the specific input values, but on the attributes domains, which satisfies the privacy guarantee but may severely distort values, thus compromising the utility of the anonymized outputs. Because of this, ε-differential privacy was originally proposed for the *interactive* scenario, in which the anonymizer returns noise-added answers to interactive queries. In this scenario, the accuracy/utility of the response to a query depends on the sensitivity of the query, which is usually lower than the sensitivity of the attribute. However, the interactive setting of ε-differential privacy limits the number and type of queries that can be performed. Most extensions of ε-differential privacy to the non-interactive setting (data set anonymization) overcome the limitation on the number of queries, but not on the type of queries for which some utility is guaranteed (see Section 2 below). In contrast, k-anonymized data sets offer more flexible utility.

1.1 Contribution and plan of this paper

In this paper, we present a procedure to improve the utility of general-purpose ε-differentially private data releases by means of a specific kind of data microaggregation. The rationale is that the microaggregation of input data helps reducing its sensitivity 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. Experiments reported on a reference data set show a significant improvement of data utility with respect to plain Laplace noise addition.

The rest of this paper is organized as follows. Section 2 details the background on ε-differential privacy and Section 3 discusses related works on ε-differentially private data releases. Section 4 proposes a new method to generate ε-differentially private data sets that uses a special type of microaggregation to reduce that amount of required noise. Section 5 reports an empirical evaluation of the proposed method, based on a reference data set. The final section gathers the conclusions and details some lines of future research.

2 Background on differential privacy

Differential privacy was originally proposed by [14] as a privacy model in the interactive setting. The assumption is that an anonymization mechanism sits between the user submitting queries and the database answering them.

Definition 1. *(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 data 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.

3 Related work on differentially private data releases

Differential privacy was also proposed for the non-interactive setting. Even though a non-interactive data release can be used to answer an arbitrarily large number of queries, in most cases, this is obtained at the cost of offering utility guarantees only for a restricted class of queries [2], typically count queries. We next review such non-interactive approaches, which are the focus of this paper.

The usual approach to releasing differentially private data sets is based on histogram queries [33, 34], 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 [23]). Several strategies have been proposed to improve the accuracy of differentially private count (histogram) queries. In [18] consistency constraints between a set of queries are exploited to increase accuracy. In [32] a wavelet transform is applied to the data and noise is added in the frequency domain. In [34, 21] the histogram bins are adjusted to the actual data. In [5], the authors consider differential privacy of attributes whose domain is ordered and has moderate to large cardinality; 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). In [26], the authors generalize similar records by using coarser categories for the classification attributes; this results in higher counts for the data partitions, which are much larger than the noise that needs to be added to reach differential privacy.

Our work differs from all previous ones in that it is not limited to histogram queries. In [29] we presented an approach that combines k-anonymity and ε-differential privacy to improve the utility of the output. In that work, we first defined the notion of insensitive microaggregation, which is a multivariate microaggregation procedure that partitions data in groups of k records with a criterion that does not depend on the input data, but just on the domain of attributes. Insensitive microaggregation ensures that, for every pair of data sets X and X' differing in a single record, the resulting clusters will differ at most in a single record. Hence, the centroids used to replace records of each cluster will have low sensitivity to changes of one input record. Specifically, when centroids are computed as the arithmetic average of the elements of the cluster, the sensitivity is as low as $\Delta(X)/k$, where $\Delta(X)$ is the distance between the most distant records of the joint domains of the input data and k is the size of the clusters. Finally, since we were finally releasing n/k centroids, each one computed on a cluster of cardinality k and having sensitivity $\Delta(X)/k$, the sensitivity of the whole data set to be released is $n/k \times \Delta(X)/k$. Thus, for numerical data sets, Laplacian noise with scale parameter $(n/k \times \Delta(X)/k)/\varepsilon$ must be added to each centroid to obtain a ε-differentially private output.

Even though this previous work effectively reduces the amount of Laplace noise to be added to achieve general-purpose ε-differentially private data releases, the fact that it requires using a microaggregation parameter k that depends on the number of records n of the input data set may be problematic for large data sets. To tackle this limitation, in this paper we present an alternative procedure that offers utility gains with respect to standard differential privacy mechanisms regardless of the number of records of the input data set.

4 Differential privacy via individual ranking

In this section we present a method to obtain differentially private data releases which can reduce noise even more than the above-mentioned prior k-anonymity approach based on multivariate microaggregation. For simplicity, we assume data sets with numerical attributes to which an amount of Laplacian noise is added to satisfy differential privacy.

In our previous approach [29] the utility gain was limited by the insensitive multivariate microaggregation used to reach k-anonymity. The sensitivity of the set of n/k centroids obtained from the multivariate microaggregation was $n/k \times \Delta(X)/k$ because, in the worst case,

- Changing a single record in the input data set can cause all n/k clusters to change by one record;
- The record changed within each cluster can alter the value of the cluster centroid by up to $\Delta(X)/k$, where $\Delta(X)$ is the maximum distance between elements in the domain of the input data (we are assuming that centroids are computed as the arithmetic average of record values in the cluster).

The above worst-case scenario overestimates the actual sensitivity of the output and, thus, the noise to be added to the centroids to achieve ε-differential privacy. Intuitively, the aggregation of the centroid variations would seem to be upper-bounded by $\Delta(X)/k$. However, this is only true if a total order for the domain of X exists for which the triangular inequality is satisfied, that is, when $d(r_1, r_2) + d(r_2, r_3) \geq d(r_1, r_3)$ holds for any records r_1, r_2 and r_3 in X. Unfortunately, this is generally not the case for multivariate data because a natural total order does not always exist. Artificial total orders defined for multivariate data (as done in [29]), do not fulfill the triangular inequality and, as discussed above, the sensitivity of individual centroids should be multiplied by the number of released centroids to satisfy differential privacy (that is, $n/k \times \Delta(X)/k$).

On the contrary, in univariate numerical data sets, a natural total order (the usual numerical order) can be easily defined with respect to the minimum or maximum value of the domain of values of the attribute so that the triangular inequality holds. In these conditions, it is shown in [8] that clusters in the optimal microaggregation partition contain consecutive values. The next lemma shows that the sensitivity of the set of centroids is indeed $\Delta(X)/k$.

Lemma 1. *Let $x_1, \cdots, x_n$ be a totally ordered set of values that has been microaggregated into $\lfloor n/k \rfloor$ clusters of k consecutive values each, except perhaps one cluster that contains up to $2k-1$ consecutive values. Let the centroids of these clusters be $\bar{x}_1, \cdots, \bar{x}_{\lfloor n/k \rfloor}$, respectively. Now if, for any single i, x_i is replaced by x'_i such that $|x'_i - x_i| \leq \Delta$ and new clusters and centroids $\bar{x'}_1, \cdots, \bar{x'}_{\lfloor n/k \rfloor}$ are computed, it holds that*

$$\sum_{j=1}^{\lfloor n/k \rfloor} |\bar{x'}_j - \bar{x}_j| \leq \Delta/k$$

Proof. Assume without loss of generality that $x'_i > x_i$ (the proof for $x'_i < x_i$ is symmetric). Assume, for the sake of simplicity, that n is a multiple of k (we will later relax this assumption). Hence, exactly n/k clusters are obtained, with cluster j containing consecutive values from $x_{(j-1)k+1}$ to x_{jk}. Let j_i be the cluster to which x_i belongs. We can distinguish two cases, namely $x'_i \leq x_{j_i k+1}$ and $x'_i > x_{j_i k+1}$.

Case 1. When $x'_i \leq x_{j_i k+1}$, x'_i stays in j_i. Thus, the centroids of all clusters other than j_i stay unchanged and the centroid of cluster j_i increases by Δ/k, because $x'_i + \Delta$. So the lemma follows in this case.

Case 2. When $x'_i > x_{j_i k+1}$, two or more clusters change as a result of replacing x_i by x'_i: cluster j_i loses x_i and another cluster j'_i (for $j'_i > j_i$) acquires x'_i. To maintain its cardinality k, after losing x_i, cluster j_i acquires $x_{j_i k+1}$. In turn, cluster $j_i + 1$ loses $x_{j_i k+1}$ and acquires $x_{(j_i+1)k+1}$, and so on, until cluster j'_i, which transfers its smallest value $x_{(j'_i-1)k+1}$ to cluster j'_i-1 and acquires x'_i. From cluster j'_i+1 upwards, nothing changes. Hence the overall impact on centroids is

$$\sum_{j=1}^{n/k} |\bar{x'}_j - \bar{x}_j| = \sum_{j=j_i}^{j'_i} |\bar{x'}_j - \bar{x}_j|$$

$$= \frac{x_{j_ik+1} - x_i}{k} + \frac{x_{(j_i+1)k+1} - x_{j_ik+1}}{k} + \cdots + \frac{x'_i - x_{(j'_i-1)k+1}}{k}$$

$$= \frac{x'_i - x_i}{k} = \frac{\Delta}{k}. \quad (2)$$

Hence, the lemma follows also in this case.

Now consider the general situation in which n is not a multiple of k. In this situation there are $\lfloor n/k \rfloor$ clusters and one of them contains between $k+1$ and $2k-1$ values. If we are in Case 1 above and this larger cluster is cluster j_i, the centroid of j_i changes by less than Δ/k, so the lemma also holds; of course if the larger cluster is one of the other clusters, it is unaffected and the lemma also holds. If we are in Case 2 above and the larger cluster is one the clusters that change, one of the fractions in the third term of Expression (2) above has denominator greater than k and hence the overall sum is less than Δ/k, so the lemma also holds; if the larger cluster is one of the unaffected ones, the lemma also holds. □

From the previous lema, it turns out that, for univariate data sets, the amount of noise needed to fulfill differential privacy after the microaggregation step is significantly lower than with the method in [29] (*i.e.* sensitivity $\Delta(X)/k$ vs. $n/k \times \Delta(X)/k$). Moreover, this noise is exactly $1/k$-th of the noise required by the standard differential privacy approach, in which the sensitivity is $\Delta(X)$ because *any* output record may change by $\Delta(X)$ following a modification of any record in the input. To benefit from such a noise reduction in the case of multivariate data sets, we rely on the following two composition properties of differential privacy.

Lemma 2 (Sequential composition [25]). *Let each sanitizing algorithm Ag_i in a set of sanitizers provide ε_i-differential privacy. Then a sequence of sanitizers Ag_i applied to a data set D provides $(\sum_i \varepsilon_i)$-differential privacy.*

Lemma 3 (Parallel composition [25]). *Let each sanitizing algorithm Ag_i in a set of sanitizers provide ε-differential privacy. Then a sequence of Ag_i each applied to a disjoint data set D_i provides ε-differential privacy.*

In the context of differentially private data publishing, we can think of a data release as the collected answers to successive queries for each attribute value in each record of the data set. Let $I_{ra}(X)$ be the query function that returns the value of attribute a (from a total of m attributes) in record r (from a total of n records) in data set X. Then if, for a fixed attribute a, we independently randomize each query $I_{ra}(X)$ for $r = 1, \cdots, n$ to attain ε-differential privacy, by parallel composition the set of n answers obtained are an ε-differentially private version of attribute a (records are disjoint from each other, so parallel composition is applicable). Now, if we publish the differentially private versions of all m attributes, by sequential composition the data set is $m\varepsilon$-differentially private (attributes are considered not disjoint from each other, since they are usually correlated; hence, sequential composition applies here).

To reduce sensitivity and hence the amount of noise needed to attain differential privacy via microaggregation, while exploiting the benefits of the natural orders available on attribute basis, we can use an univariate microaggregation: *individual ranking* [6]. Individual ranking deals with multi-attribute data sets by microaggregating one attribute at a time. Input records are sorted by the first attribute, then groups of successive k values of the first attribute are created and all values within that group are replaced by the group representative (*e.g.* centroid). The same procedure is repeated for the rest of attributes. The attribute-independent microaggregation of individual ranking fits with our goal: to independently attain differential privacy on each microaggregated attribute and then use sequential composition to obtain a differentially private data set.

As discussed in above, each numerical attribute is already equipped with a natural total order that fulfills the triangular inequality. Thus, Lemma 1 guarantees that the centroids output by individual ranking for each attribute A_i have total sensitivity $\Delta(A_i)/k$, where $\Delta(A_i)$ is the maximum distance between two values in the domain of A_i. Hence, we propose the following algorithm to obtain a differentially private version X^D of a numerical original data set X with attributes $A_1, \cdots, A_m$.

Algorithm 1

1. *Use individual ranking microaggregation independently on each attribute A_i, for $i = 1$ to m. Within each cluster,* all *attribute values are replaced by the cluster centroid value, so each microaggregated cluster consists of k repeated centroid values. Let the resulting microaggregated data set be X^M.*

2. Add Laplace noise independently to each attribute A_i^M of X^M, where the scale parameter for attribute A_i^M is

$$\Delta(A_i^M)/\varepsilon = \Delta(A_i)/(k \times \varepsilon).$$

The same *noise perturbation is applied to all repeated centroid values within each cluster.*

Now we can state:

Lemma 4. *The output of Algorithm 1 is $m\varepsilon$-differentially private.*

Proof. In Step 1 of Algorithm 1, if attribute A_i (with $i = 1, \cdots, m$) has sensitivity $\Delta(A_i)$ in X, by Lemma 1 its microaggregated version A_i^M has sensitivity $\Delta(A_i)/k$. In Step 2, an ε-differentially private version of A_i^M is obtained. By sequential composition, the noise-added data set X^D is $m\varepsilon$-differentially private. Note that sequential composition needs to be applied, because, in general, record attribute values are not independent and the correlation between the attribute centroids is preserved by the individual ranking mechanism. □

Note. In Step 2 of Algorithm 1, it is critically important to apply exactly the same noise perturbation to all repeated values within a microaggregated cluster. If we used different random perturbations for each repeated value, the resulting noise-added cluster would be equivalent to the answers to k independent queries. This would multiply times k the sensitivity of the centroid, which would cancel the sensitivity reduction brought by microaggregation in Step 1.

5 Empirical evaluation

This section details the empirical evaluation of the proposed method regarding data utility preservation.

As evaluation data we used the “Adult” data set from the UCI repository [16]. We took two numerical attributes AGE and (working) HOURS-PER-WEEK of the training corpus, which consists of 30,162 records after removing records with missing values. Since the two attributes represent non-negative numerical magnitudes, we defined their domains as $[0 \ldots (1.5 \times \mathit{max\_attr.\_value\_in\_dataset})]$, as done in [29]. The difference between the bounds of the domain of each attribute A_i determines the sensitivity of that attribute ($\Delta(A_i)$) and, as detailed above, determines the amount of Laplace noise to be added to microaggregated outputs. Since the Laplace distribution takes values in the range $(-\infty, +\infty)$, for consistency, we bound noise-added outputs to the domain ranges define above.

As done in the literature on statistical disclosure control, we evaluated the utility of the anonymized output in terms of *information loss*[19]. To do so, we used the Sum of Squared Errors (SSE), which is a well-known information loss measure. SSE is defined as the sum of squares of attribute distances between records in the original data set 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, $(a_j^i)'$ represents its anonymized version and $dist(\cdot,\cdot)$ corresponds to the standard Euclidean distance. In our experiments, the SSE value was normalized by the number of attributes considered in each test (m).

The ε parameter for differential privacy was set to $\varepsilon = \{0.1, 1.0, 10.0\}$, which covers the usual range of differential privacy levels observed in the literature $[15, 3, 4, 23]$. The two attributes in the data set were considered as not independent (which is the most usual case), so that the sequential composition should be applied. Thus, as discussed in Section 4, to obtain ε-differentially private records we need (ε/m)-differentially private attribute values. Hence, Laplace noise addition with scale parameter $\Delta(A_i)/(k \times (\varepsilon/m))$ needs to be added to each attribute A_i, where k is the level of prior microaggregation (which we set between 2 and 100) and m the number of attributes to protect (2).

As baseline methods to compare our proposal with, we considered:

- Plain Laplace noise addition for ε-differential privacy. Since attributes are considered as not independent, the sequential composition should be also applied. Thus, to obtain an ε-differentially private record we need (ε/m)-differentially private attribute values. Hence, Laplace noise addition with scale parameter $\Delta(A_i)/(\varepsilon/m) = m\Delta(A_i)/\varepsilon$ needs to be added to each attribute A_i.
- Plain individual ranking, with no subsequent Laplace noise addition. Although this method does not lead to ε-differential privacy by itself, we want to show the contribution of individual ranking to the information loss caused by our method.

Figure 1 shows the comparison between the SSE obtained with plain Laplace noise addition, plain individual ranking and our approach. Due to the broad ranges of the SSE values, a $\log_{10}$ scale is used for the Y-axes.

The plain Laplace noise addition baselines 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 5 runs, for the sake of stability.

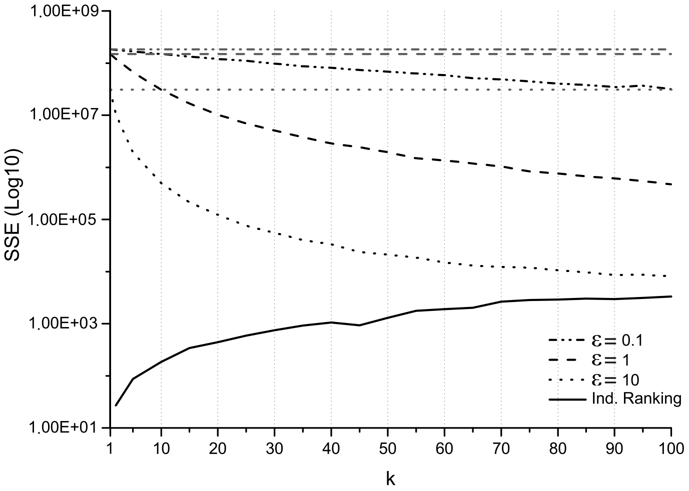

Fig. 1. SSE for the proposed method for different ε values (black non-horizontal lines, as SSE depends on the microaggregation parameter k) vs. plain Laplace noise addition with *dependent* attributes (gray horizontal lines, because SSE does not depend on k) and plain individual ranking microaggregation. Y-axes are in $\log_{10}$ scale.

Regarding plain individual ranking we observe that it causes an information loss that grows with k, but which is shown to be negligible in comparison with the information loss caused by Laplace noise addition for most values of ε (remember that the scale of Y-axes is logarithmic).

Regarding differentially private methods, we observe that, already for $k > 1$, our approach significantly reduces the noise required to attain ε-differential privacy vs. plain Laplace noise addition. The relative improvement of SSE depends on the value of ε. For the smallest ε (that is, 0.1) the amount of noise involved is so high that even with the noise reduction achieved by our method, the output's utility would be severely hampered. However, for $k = 100$ the reduction of information loss almost equals that of 1.0-differential privacy with plain Laplacian noise. This is very relevant because the more robust privacy guarantee offered by the 0.1-differentially private outputs of our method is achieved with an information loss equivalent to that of 1.0-differential privacy for the standard mechanism. For the highest ε (that is, 10.0) there is a substantial decline of SSE for low k and, for larger k, SSE stays more constant and almost as low as the SSE achieved by the individual ranking alone. In this case, the noise added by individual ranking in larger clusters dominates and limits the benefits of the noise reduction at the ε-differential privacy stage due to the decreased sensitivity with larger k. Finally, for $\varepsilon = 1.0$, a more linear decrease is achieved as the k values grow, because the information loss improvement is less limited by the lower bound of the individual ranking microaggregation.

6 Conclusions

In this paper, we have presented an anonymization method that combines the low information loss incurred by individual ranking microaggregation and its lack of assumptions on data uses and the robust privacy guarantees offered by ε-differential privacy. As a result, our method is able to effectively reduce the scale parameter of noise needed to fulfill differential privacy, and thus improve the utility of anonymized outputs. The method proposed here is easy to implement, because the individual ranking algorithm only relies on the natural order of individual attributes.

As future work, we plan to further evaluate our method with other data sets and compare the results with those of related works on differentially private data publishing, even if that means restricting the utility to specific tasks (*e.g.* counting queries). Finally, we also plan to 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 compare, sort, microaggregate and randomize the outputs should be defined [11, 24].

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". J. Domingo-Ferrer 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>: the 10th International Conference on Database Theory-ICDT 2005, pp. 246–258. (2005)
2. Blum, A., Ligett, K., Roth, A.: A learning theory approach to non-interactive database privacy. In: the 40th Annual Symposium on the Theory of Computing-STOC 2008, pp. 609–618 (2008)
3. Charest, A.-S.: How can we analyze differentially-private synthetic data sets?. Journal of Privacy and Confidentiality 2(2), 21–33 (2010)
4. Charest, A.-S.: Empirical evaluation of statistical inference from differentially-private contingency tables. In: Privacy in Statistical Databases-PSD 2012. LNCS, vol. 7556, pp. 257–272. Springer (2012)
5. 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)
6. Defays, D., Nanopoulos, P.: Panels of enterprises and confidentiality: the small aggregated method. In: the 92 Symposium on Design and Analysis of Longitudinal Surveys, pp. 195–204. (1993)
7. Domingo-Ferrer, J.: A critique of k-anonymity and some of its enhancements. In: ARES/PSAI 2008, pp. 990–993. IEEE Computer Society (2008)
8. Domingo-Ferrer, J., Mateo-Sanz J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14(1), 189–201 (2002)
9. Domingo-Ferrer J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
10. Domingo-Ferrer, J., Mateo-Sanz, J.M., Oganian, A., Torra, V., Torres, A.: On the Security of Microaggregation with Individual Ranking: Analytical Attacks. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems 18 (5), 477–492 (2002)
11. Domingo-Ferrer, J., Sánchez, D., Rufian-Torrell, G.: Anonymization of nominal data based on semantic marginality. Information Sciences 242, 35–48 (2013)
12. Domingo-Ferrer, J., Sebé, F., Solanas, A.: A polynomial-time approximation to optimal multivariate microaggregation. Computing & Mathematics with Applications 55(4), 714–732 (2008)
13. Drechsler, J.: My understanding of the differences between the CS and the statistical approach to data confidentiality. In: the 4th IAB Wokshop on confidentiality and disclosure. Institute for Employment Research (2011)
14. Dwork, C.: Differential privacy. In: the 33rd International Colloquium on Automata, Languages and Programming-ICALP 2006. LNCS, vol. 4052, pp. 1-12. Springer. (2006)
15. Dwork, C.: A firm foundation for private data analysis. Communications of the ACM 54(1), 86–95 (2011)
16. 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)
17. Goldberger, J., Tassa, T.: Efficient anonymizations with enhanced utility. Transactions on Data Privacy 3, 149–175 (2010)

18. Hay, M., Rastogi, V., Miklau, G., Suciu, D.: Boosting the accuracy of differentially private histograms through consistency. PVLDB 3(1), 1021–1032 (2010)
19. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., de Wolf, P.-P.: Statistical Disclosure Control. Wiley (2012)
20. 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)
21. 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)
22. 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)
23. 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)
24. Martínez, S., Sánchez, D., Valls, A.: A semantic framework to protect the privacy of electronic health records with non-numerical attributes. Journal of Biomedical Informatics 46(2), 294–303 (2013)
25. McSherry, F.: Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: the 2009 ACM SIGMOD International Conference on Management of Data, pp. 19–30. ACM (2009)
26. Mohammed, N., Chen, R., Fung, B.C.M., Yu, P.S.: Differentially private data release for data mining. In: the 17th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining-KDD 2011, pp. 493–501. ACM (2011)
27. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6), 1010–1027 (2001)
28. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. SRI International Report (1998)
29. Soria-Comas, J., Domingo-Ferrer, J., Sánchez, D., Martínez, S.: Enhancing Data Utility in Differential Privacy via Microaggregation-based k-Anonymity. VLDB Journal (to appear)
30. Sweeney, L.: k-Anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
31. Wong, R., Li, J., Fu, A., 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 2006), pp. 754–759 (2006)
32. Xiao, X., Wang, G., Gehrke, J.: Differential Privacy via Wavelet Transforms. IEEE Transactions on Knowledge and Data Engineering 23(8), 1200–1214 (2010)
33. Xiao, Y., Xiong, L., Yuan, C.: Differentially private data release through multidimensional partitioning. In: the 7th VLDB conference on Secure data management (SDM'10), pp. 150–168 (2010)
34. 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)