Article

Assessing the Price in Data utility of *k*-Anonymous Microaggregation

Ana Rodríguez-Hoyos [1,2], José Estrada-Jiménez[1,2], David Rebollo-Monedero[2], Jordi Forné[2], Javier Parra-Arnau[3], Luis Urquiza-Aguiar[1]

[1] Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional (EPN), Ladrón de Guevara, E11-253 Quito, Ecuador; {ana.rodriguez, jose.estrada}@epn.edu.ec
[2] Department of Telematics Engineering, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain; {david.rebollo, jforne}@entel.upc.edu
[3] Department of Computer Science and Mathematics, Universitat Rovira i Virgili (URV), CYBERCAT-Center for Cybersecurity Research of Catalonia E-43007 Tarragona, Spain; javier.parra@urv.cat
* Correspondence: jforne@entel.upc.edu

Abstract: With a data revolution underway for some time, there is an increasing demand for formal privacy protection mechanisms that are not so destructive. Hereof microaggregation is a popular high-utility approach designed to satisfy the popular *k*-anonymity criteria while applying low distortion to data. However, standard performance metrics are commonly based on mean square error, which will hardly capture the utility degradation related to a specific application domain of data. In this work, we evaluate the performance of *k*-anonymous microaggregation in terms of the loss in classification accuracy of the machine learned models built from perturbed data. Systematic experimentation is carried out on four microaggregation algorithms that are tested over four data sets. The empirical utility of the resulting microaggregated data is assessed using the learning algorithm that obtains the highest accuracy from original data. Validation tests are performed on a test set of non perturbed data. The results confirm *k*-anonymous microaggregation as a high-utility privacy mechanism in this context and distortion based on mean squared error as a poor predictor of practical utility. Finally, we corroborate the beneficial effects for empirical utility of exploiting the statistical properties of data when constructing privacy preserving algorithms.

Keywords: microaggregation; k-anonymity; privacy; data utility

1. Introduction

In the current era of information, vast amounts of data and exploitation mechanisms are more and more available; thus, more utility can be mined from data through data-analytics technologies. The potential benefits of these technologies are countless in several fields such as healthcare, advertising, and even industrial engineering, [1–3]. Said benefits entail important economic profits, so giant tech companies are leveraging data as core assets [4] that are disclosed (exploited, shared or even sold) to maximize profit.

Since this data commonly refers to individuals (personal data), the abundance of details collected about them and the growing sophistication of machine-learning analytics raise serious privacy concerns. Even if identifying attributes such as full names are suppressed, other, apparently innocuous, personal attributes, so-called quasi-identifiers, could still be used to re-identify an individual. If a sensitive attribute (gender, health status, income) were disclosed, re-identification would enable an attacker to associate an individual with such attribute, violating her privacy.

Given this privacy risk, to prevent re-identification, (user) data needs to be processed before being disclosed, which implies data distortion. In this regard, *statistical disclosure control* (SDC) offers an interesting approach to protect individual privacy while preserving some of the data utility.

SDC is usually implemented over microdata, a tabulated data representation where the attributes of several individuals are organized in records or rows, one for each individual. The purpose behind

SDC is perturbing quasi-identifier attributes such that data subjects cannot be singled out. Since re-identification is not possible, the values of sensitive attributes can be safely disclosed.

To enable SDC, different privacy models are available. Whereas a minimum protection level is guaranteed through a privacy parameter, the resulting utility of perturbed data may vary according to the implemented model. Some of such models are k-anonymity [5,6] and ε-differential privacy [7].

Since the criteria posed by privacy models are invariably met by perturbing quasi-identifiers to anonymize data, there is an impact on the data in terms of loss in utility [8]. However, said utility loss may vary according to the strategy followed by the privacy mechanism, even when the privacy parameter is already met. If the resulting data utility does not meet the requirements of the application domain (e.g., health) a different privacy parameter or mechanism should be used. Some of these mechanisms include microaggregation, suppression, generalization and noise addition.

The impact of these mechanisms on data (i.e., on the utility of data after anonymization) is commonly measured using standard, but merely syntactical, metrics, such as the mean-squared error (MSE). However, to capture the practical utility of anonymized data, other metrics related to the application domain might be more relevant, e.g., accuracy or F-measure, if data is used as input for classification in machine learning tasks. Assessing the impact of privacy mechanisms by using these metrics would help unveil the strategies that best preserve utility, but also whether or not standard metrics faithfully predict such practical utility.

1.1. Contribution and plan of the paper

In the context of SDC, we focus on k-anonymous microaggregation (aggregating numerical microdata to meet k-anonymity), which, as illustrated in Figure 1, is a high-utility privacy approach that is pertinent, e.g, for health applications. Other models such as differential privacy may pose stricter protection criteria, thus potentially implying greater costs in terms of data utility, so are beyond the scope of this work.

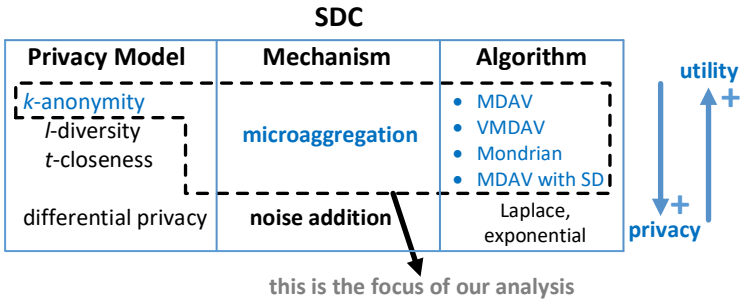


Figure 1. Since we focus on high-utility SDC, our work here is devoted to k-anonymous microaggregation algorithms.

On the other hand, also depicted in Figure 1, k-anonymous microaggregation is implemented through different mechanisms, e.g., MDAV, VMDAV, Mondrian and MDAV with preservation of statistical dependence. Accordingly, we assess several microaggregation algorithms in terms of the practical utility of anonymized data. We employ non standard, but empirical utility metrics taken from machine learning, which is currently a very common application domain of data. By systematically testing these algorithms in this context, we are able to evidence the moderate impact of microaggregation on the utility of data.

Furthermore, this evaluation enables us to compare different microaggregation algorithms in terms of their capability to preserve empirically measured utility. Since these algorithms are not usually designed to explicitly preserve the internal macrotrends within data (practical utility), we aim at identifying the parameters likely helping to this aim. In fact, we find out that efforts to preserve

the statistical dependence within quasi-identifiers and confidential attributes (such as in MDAV with statistical dependence) may effectively attenuate the impact of microaggregation on the utility of data. In this sense, we extend the study performed in [12] by assessing not only MDAV but also other microaggregation algorithms, particularly one explicitly designed to preserve the statistical information within the data.

Finally, for all these microaggregation algorithms, we assess the capability of a standard distortion metric to predict the empirical utility of anonymized data. Although these metrics are usually oriented to rather measure the syntactical distortion of quasi-identifiers, even disregarding the semantic contribution of confidential attributes, some efforts have being done to consider the statistical dependence of all the data when applying *k*-anonymous microaggregation.

The remainder of this paper is organized as follows. Section 2 reviews the background on *k*-anonymous microaggregation, presents the state of the art in algorithms for SDC, particularly those evaluated here, and explores previous work evaluating the impact on data utility caused by anonymization. Next, Section 3 describes the methodology followed to evaluate such impact. Section 4 shows the experimental results obtained for a variety of microaggregation algorithms, data sets and machine-learning algorithms. Lastly, a brief discussion is presented in 5 and conclusions are drawn in Section 6.

2. Background and state of the art on k-anonymous microaggregation

2.1. Background on microaggregation

When microdata data is to be disclosed to a not fully trusted party, suppressing *identifiers* (full names, identity numbers) is a first step to protect user privacy. But the combination of other commonly demographic attributes could still individuate data subjects; these attributes are called *quasi-identifiers*. Thus, quasi-identifiers are regularly the object of privacy protection mechanisms. Finally, *confidential attributes*, i.e., sensitive information about individuals is usually disclosed without modification since these mechanisms already protect the identity of data owners.

k-Anonymous microaggregation operates over quasi-identifiers by dividing a microdata set in cells such that every cell contains at least *k* user records (aggregation). To protect privacy, the records of each cell are replaced by a representative record (reconstruction), thus enforcing *k*-anonymity. Figure 2 depicts this process where after identifiers are suppressed from a microdata set, quasi-identifiers are microaggregated in 3-anonymous cells while confidential attributes are left untouched.

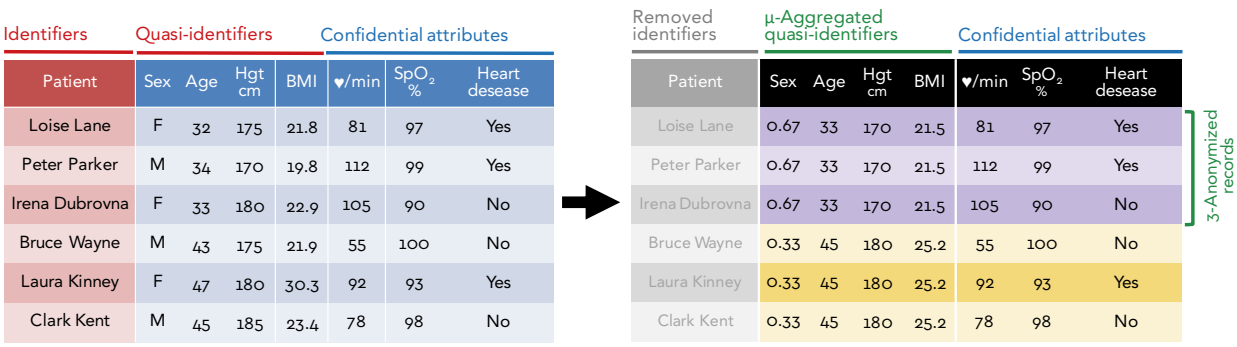


Figure 2. Toy example of *k*-anonymous micoraggregation. After suppressing identifiers, the records are clustered in groups of size *k* (microcells). Then, the quasi-identifiers in each micro cell are replaced a representative tuple (e.g., a centroid). Finally, microaggregated quasi-identifiers and original confidential attributes are published.

As illustrated in Figure 2, a representative tuple for each aggregated cell was obtained by averaging their numerical data and was used for reconstruction. However, other reconstruction

mechanisms can be used depending on the microaggregation algorithm, e.g., replacing values with intervals, directly suppressing attribute values, or even suppressing entire records from microdata.

To give some intuition, if numerical quasi-identifiers could be drawn as points in the Euclidean space, k-anonymous microaggregation could be seen as a mechanism to partition such points in cells of size of at least k. Then, each cell would be represented by a point or interval within such cell so its shape will depend on the implementation chosen. We depict such intuition in Figure 3.

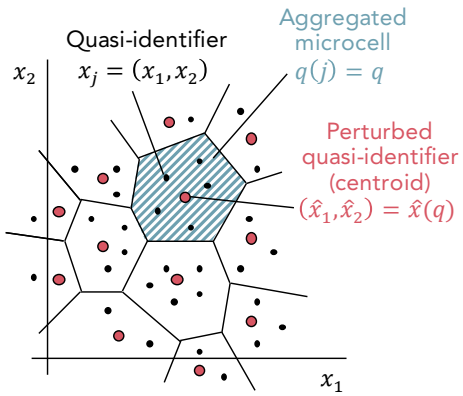


Figure 3. Intuition regarding k-anonymous microaggregation. The two-dimensional quasi-identifiers of a microdata set are depicted as points in an Euclidean space. Microaggregation partitions such points by building k-anonymous microcells to then replace each tuple with the centroid of the corresponding microcell.

2.2. *Microaggregation algorithms*

Next we briefly describe some well-known microaggregation algorithms with the aim of introducing the strategies followed to group and reconstruct microcells. This will provide with some feedback for the evaluation performed in this work that focuses on unveiling the utility preserving capabilities of k-anonymous microaggregation, but particularly on showing that some efforts to preserve the statistical dependence within data would help to increase said empirical utility.

MDAV (maximum distance to average vector algorithm) is the de facto standard for microaggregation of numerical microdata [9]. By systematically finding the furthest k-anonymous cells within the data set, MDAV replaces each record by the centroid (average) of its corresponding cell. It evolved from the multivariate fixed-size microaggregation method and was proposed in [10]. MDAV provides an excellent heuristic method for multivariate microaggregation [11] in terms of utility, measured both syntactically [11] and empirically [12], and in terms of computation complexity. Figure 4 presents a generic version of MDAV taken from [13]. Note that MDAV generates cells of fixed size k and potentially a cell with size $2k - 1$.

function MDAV
input k, $(x_j)_{j=1}^n$ ▷*Anonymity parameter k, quasi-ID portion $(x_j)_{j=1}^n$ of a data set of n records*
output q ▷*Assignment function from records to microcells $j \mapsto q(j)$*
1: **while** $2k$ points or more in the data set remain to be assigned to microcells **do**
2: find the centroid (average) C of those remaining points
3: find the furthest point P from the centroid C, and the furthest point Q from P
4: select and group the $k-1$ nearest points to P, along with P itself, into a microcell, and do the same with the $k-1$ nearest points to Q
5: remove the two microcells just formed from the data set
6: **if** there are k to $2k-1$ points left **then**
7: form a microcell with those and finish
8: **else** ▷*At most $k-1$ points left, not enough for a new microcell*
9: adjoin any remaining points to the last microcell ▷*Typically nearest microcell*

Figure 4. MDAV "generic", functionally equivalent to Algorithm 5.1 in [13].

V-MDAV [11], follows a similar strategy to MDAV but enables the aggregation process to generate variable-size cells. When k records are already aggregated, an extension step may include more records to the cell being formed (up to a total $2k-1$) if they are "close enough" to this cell. The inclusion decision is defined by a gain parameter γ that must be adjusted depending on the data set. It offers less distortion for some data sets at a computational cost comparable to that of MDAV. The V-MDAV algorithm is presente in [11].

Unlike traditional k-anonymous microaggregation (e.g., through MDAV) where only the values of quasi-identifiers X are considered when building microcells, **microaggregation with preservation of statistical dependence** (we will call it *MDAV with SD*) also includes confidential attributes [14] in the partition design. Thus, if a confidential attribute Y has to be predicted, this approach would lead to a more accurate prediction (e.g., classification) from perturbed quasi-identifiers $\hat{X}$. To involve both types of attributes, the authors propose designing a cell assignment function that minimizes a multiobjective Lagrangian distortion function

$\mathcal{D} = (1-\lambda)\mathcal{D}_X + \lambda\mathcal{D}_Y$

where $\mathcal{D}_X$ is the traditional information loss term based on MSE, $\mathcal{D}_Y$ characterizes the degradation in statistical dependence, captured through the non linear predictability of Y from X, and *lambda* controls the tradeoff between these two optimization objectives.

Finally, **Mondrian** [15] is a greedy algorithm that recursively partitions a microdata set in regions of at least k records, where a dimension (attribute) and a value about which to partition have to be heuristically chosen in each iteration. This is a microaggregation algorithm in the sense that it partitions a microdata set in variable-size cells, satisfying the k-anonymity criteria. The values of the quasi-identifiers for each cell are reconstructed as non-overlapping intervals in which such values are contained. Intuitively, such partitions are defined as hyperrectangles in the multidimensional space of quasi-identifiers.

k-Anonymous microaggregation is hardly infallible in terms of privacy, particularly because only quasi-identifiers are processed. The statistical characteristics of published confidential attributes, along with additional information an attacker might obtain, could give rise to similarity, skewness or background-knowledge attacks [16–18]. Thus, several refinements have been proposed to k-anonymity, all of them requiring a less homogeneous distribution of confidential in each k-anonymous microcell. To start, p-sensitive [19,20], requires that each microcell contains at least p different values of each confidential attribute. Going a little further, l-diversity proposes that each microcell has at least l well-represented confidential values.

In general, the implementations of microaggregation have been oriented to reduce the inherent information loss [21–23] due to perturbation, which commonly derives in more sophisticated and significantly costlier implementations in terms of computational time [24].

2.3. Utility of microaggregated data

The resulting utility of anonymized data is commonly measured inversely as the distortion applied, which is quantified through the MSE when dealing with numerical attributes. However, there are other metrics, such as accuracy, that have derived from the application domain of data, e.g., machine learning used to exploit the statistical properties of information. Evidently, the more strict the privacy criteria enforced, the less accurate the resulting (e.g., classification) models obtained from perturbed data.

Classification accuracy and other machine learning metrics have been used in previous work to assess the utility of perturbed data. However, the authors have concentrated on algorithms such as Incognito, Mondrian and DataFly. Moreover, these evaluations use to assess the performance of classifiers specifically adapted to operate on anonymized data [7,25–29], commonly using simulated data sets [30]. A lot of research has also investigated modifications ofanonymization algorithms to produce private data of "higher quality". In that context, the utility of anonymized data is evaluated in terms of classification accuracy of machine learning models [31], [32], and [33]. In [12], a systematic

evaluation is certainly done to determine the impact of *k*-anonymous microaggregation on machine learned macrotrends, but only the generic version of the MDAV algorithm is tested.

A lot of considerations should be taken into account when assessing the resulting utility of anonymized data.; for example, the variety of anonymization algorithms, that could even be proprietary [34] or non-standard [35]. Other elements that may affect the evaluation include the different application domains of data and the specific characteristics of the data tested. Finally, consider that, in practice, complying with multiple privacy criteria might render the anonymized data useless [36]. Thus, since our target application is that of data release for general statistical analysis with a focus on data utility, the restrictions imposed by differential privacy [37] and other criteria are beyond the scope of this work.

Last but not least, we would like to stress that our review of the state of the art in this section has been conducted from a strictly technological perspective. Legal and socioeconomic aspects are covered, for instance, in [38,39].

3. Methodology of evaluation

3.1. Evaluation context

Our evaluation scenario considers a microdata set whose quasi-identifiers are correlated with its corresponding confidential attribute. Moreover, this information has to be publicly released for research purposes, so *k*-anonymous microaggregation is applied over quasi-identifiers to protect the privacy of data subjects. This is the standard attack model of the SDC literature [40].

Accordingly, anonymized quasi-identifiers (here also input samples) would be published along with untouched confidential attributes (also output labels) to feed a machine learning classifier, which is the enabler of the selected application domain of data. The resulting models would allow external data analysts to build predictive models on different testing data. Intuitively, the quality of the statistical trends embedded in the resulting anonymized data would be undermined with respect to those in the original data.

Although *k*-anonymous microaggregation is known to offer interesting benefits in terms of distortion and classification accuracy [12], additional variations exist, some even incorporating utility improvements [14], have not been assessed in this context.

3.2. Privacy and utility metrics

As expected, the privacy metric we use is *k*-anonymity since microaggregation algorithms aim at guaranteeing such criteria. Thus, higher values of *k*, implying larger anonymous microcells, will offer more privacy but, at some point, less utility.

On the other hand, we assume binary classification as the application domain of data, so our utility metric is the accuracy of the classification model built from anonymized data, as performed in [31,33,35]. Basically, accuracy quantifies the rate of correctly classified samples in a test set. Besides, we also use a complementary machine learning metric, F-measure, to confirm our results in the next sections.

3.3. Scenario setup

As can be grasped from the sections above, our experimental setup builds on the algorithms for privacy protection and utility exploitation, the data sets used to assess the impact of anonymization, and the steps taken to get the results.

Being MDAV the de facto microaggregation algorithm, we extend the study performed in [12] by assessing not only MDAV but also V-MDAV [11] and MDAV with SD [14]. As explained in Section 2.2, both of them aim at increasing the data utility preserved, measured from the distortion applied by these two variants of MDAV. While V-MDAV proposes building larger microcells, when possible, to favor forming more compacted clusters, MDAV with SD builds microcells capturing the statistical

dependence between quasi-identifiers and confidential attributes. Moreover, Mondrian [15] is also considered in our setup to corroborate the performance of microaggregation algorithms, no matter the strategy used to build k-anonymous microcells. Some of the implementation details of these algorithms and further references are included in Section 2.

To measure the utility of microaggregated data, we use the machine learning algorithms that obtain the best performance, in terms of classification accuracy, from each of our data sets. Since the intrinsic nature of the data sets might vary, we experimentally determine the best performer by testing a series of algorithms such as boosted trees, logistic regression, Support Vector Machine, and k-nearest neighbor on the original data. This way we more rigorously adapt our evaluation to the specific utility context.

The data where microaggregation algorithms are assessed includes both real and synthetic data sets. Essentially, we look for data sets meeting two main requirements: include demographic attributes and evidence a correlation between the quasi-identifiers and a confidential attribute. We briefly describe their characteristics in Table 1. The first is the "Adult" data set [41], which is a standard when assessing microaggregation algorithms. We also tested the "Breast Cancer Wisconsin" data set [42] and the "Heart disease" data set [43] that contain medical data extensively used to evaluate binary classification tasks. Finally, we created an elementary synthetic data set with three attributes mimicking two quasi-identifiers and a binary confidential attribute; to do it, two groups of two-dimensional quasi-identifiers are generated following two different, but overlapping, normal distributions.

We employ Matlab 2018B to implement the aforementioned microaggregation algorithms [11,13, 14], except for Mondrian, as well as to deploy the evaluation of perturbed data sets, and to process and plot results. Said evaluation implies loading data, building machine learning models over it, and applying such models over new data to measure classification accuracy, F-measure, and distortion. The implementation of Mondrian is written in Python and was taken from [44]. Since the reconstruction method applied by Mondrian returns intervals instead of single values for each microaggregated attribute, we adapt this reconstruction such that the multidimensional hyperrectangles (microcells) are replaced by their corresponding centroids. The exploratory analysis to define the best suitable classification algorithm for each data set is performed with the Classification Learner application included in Matlab 2018B and then the model training and evaluation are automatized using specific embedded functions for each algorithm.

3.4. Methodology

Next we describe the experimental methodology we use to assess the performance of microaggregation algorithms in terms of the resulting empirical utility of perturbed data. Figure 5 synthesizes the main elements of such procedure.

First, the original data set is *preprocessed* through three steps. To start, since MDAV based algorithms only works with numerical data, any categorical values for quasi-identifiers are represented numerically (e.g., the values female and male for sex are replaced with 1 and 0). Moreover, for validation purposes explained in the next paragraphs, we split each data set in two sets: a training set and a test set such that the former's size is $3/4$ of the data set. Afterwards, each column of the training set, involving only quasi-identifiers, are normalized such that each column has zero mean and unit variance. Note that normalization is useful to avoid the harmful impact on microaggregation resulting from attributes having different ranges.

Once normalized, the *microaggregation* algorithm is fed with the training set for data perturbation. Only in the case of MDAV with SD confidential attributes are also considered since this algorithm exploits the statistical dependence between quasi-identifiers and confidential attributes. We use progressively increasing values of k to then measure the utility degradation of data due to k-anonymous microaggregation. Besides de generic privacy criteria k, other parameters are configured for some algorithms.

Table 1. Description of the Data sets Used to Evaluate the Impact of *k*-Anonymous microaggregation

Data set	# of records	# of attributes used as quasi-identifiers	list of quasi-identifiers used	confidential attribute (output of the data set in ML terms)
Adult [41]	45,222	15	Age, education-num, marital-status, sex, capital-gain, hours-per-week	Salary (>50K?)
Breast Cancer Wisconsin [42]	699	9	clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses	class (benign/malignant)
Heart Disease [43]	303	13	age, sex, chest pain type, trestbps, serum cholestoral, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal	diagnosis of heart disease
Synthetic	1000	2	x_1, x_2	y

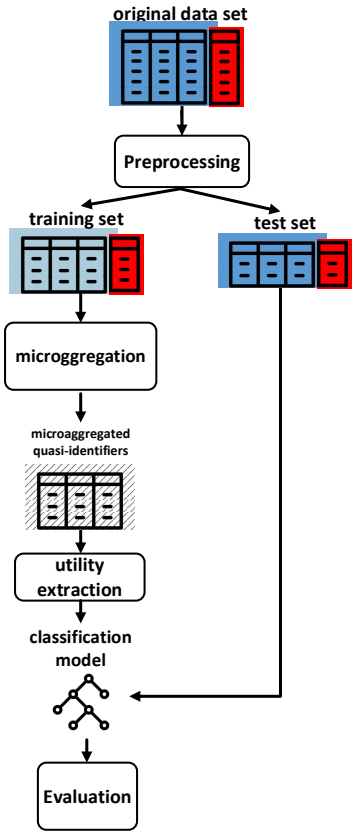
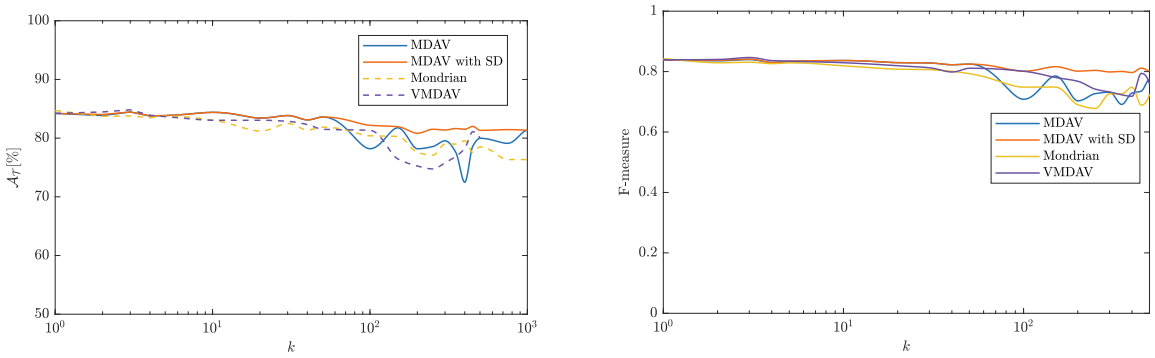


Figure 5. Experimental methodology followed to assess *k*-anonymous microaggregation algorithms in terms of the empirical utility preserved.



(a) Accuracy degradation

(b) F-measure degradation

Figure 6. Degradation of the empirical utility of the microaggregated "Adult" data set.

V-MDAV requires a gain parameter γ that we set in 0.9 as set in [11]. Additionally, MDAV with SD can be tuned by a λ parameter that regulates the tradeoff between distortion of quasi-identifiers and distortion of confidential attributes; we test different values of λ from 0 to 1 in order to get those showing the highest utility (maximum utility trace).

Once quasi-identifiers are perturbed, we implement the *utility extraction* phase. For this, we build a classification model using the microaggregated version of each data set as input. The algorithms showing the best performance in terms of utility are *boosting trees* and *logistic regression*, and the specific functions implemented in Matlab 2018b are used for training using 5-fold cross validation. Finally, each resulting classification model is evaluated over the test set originally extracted during the preprocessing phase; then accuracy and F-Measure are obtained. Namely, the machine-learned model built from microaggregated data is tested on a different portion of original data. This scenario mimics the (e.g., medical) context in which a researcher would use a machine learning model obtained from anonymized shared data to predict a given condition from their own (non perturbed) data.

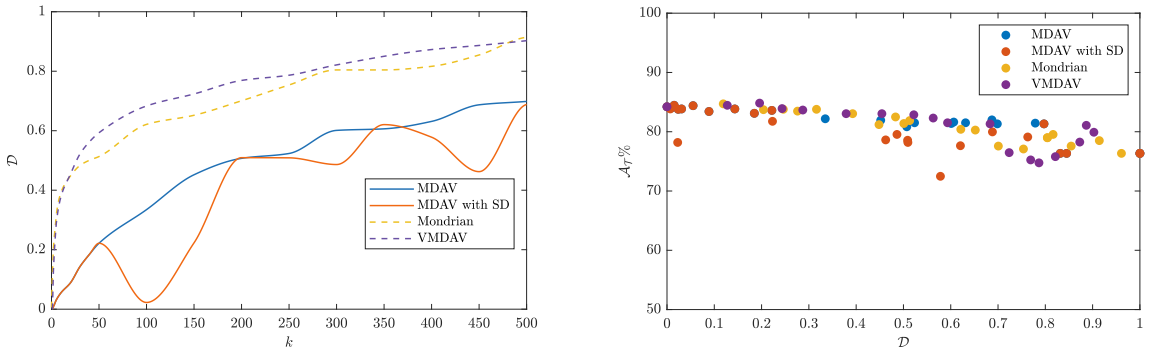
4. Experimental results

In this section, we present the results obtained from measuring the degradation of empirical utility of microdata due to k-anonymous microaggregation. This implies assessing the accuracy of machine learned models when trained over data microaggregated using an increasing value of k. Also distortion as MSE is measured in these terms to validate its capability to estimate the practical utility of data.

Said two main results are depicted in two groups of figures for each data set: one where accuracy and F-measure are shown and another where distortion is drawn against accuracy to unveil their potential correlation.

Our first experiment builds on the *UCI Adult data set*. In this particular case, we do not use the entire data set of more than 45 thousand records, but only 10% of them, i.e., a random sample that preserves the prevalence of the output (confidential) attribute. Suppressing potentially valuable data might reduce even more the data utility after microaggregation, an effect that we are interested in studying.

Accordingly, we illustrate in Figure 6 how empirical utility is affected when microaggregation is applied over the UCI Adult data set. As expected, data perturbation eventually renders data useless, as shown by the decreasing trend in accuracy as k gets higher values. Note that the lowest value in accuracy does not reach zero since, in the worst case, when the data input (quasi-identifiers) is completely perturbed, machine learned models predict based only on the prevalence of classes in the output data.



(a) Distortion measured for different values of k-anonymity.

(b) Representation of accuracy vs distortion

Figure 7. Distortion of the microaggregated "Adult" data set. The distortion corresponding to MDAV with SD is measured according to the hybrid metric proposed in [14].

Despite this inevitable degradation in the long term, as stated in [12], microaggregated data shows high levels of utility even up to $k = 50$. Namely, for such values of k, accuracy easily keeps greater than 80% for any of the four microaggregation algorithms evaluated. Interestingly enough, in the case of the UCI Adult data set, this means that said utility in terms of machine learning accuracy might be kept even when vast amounts of data are suppressed.

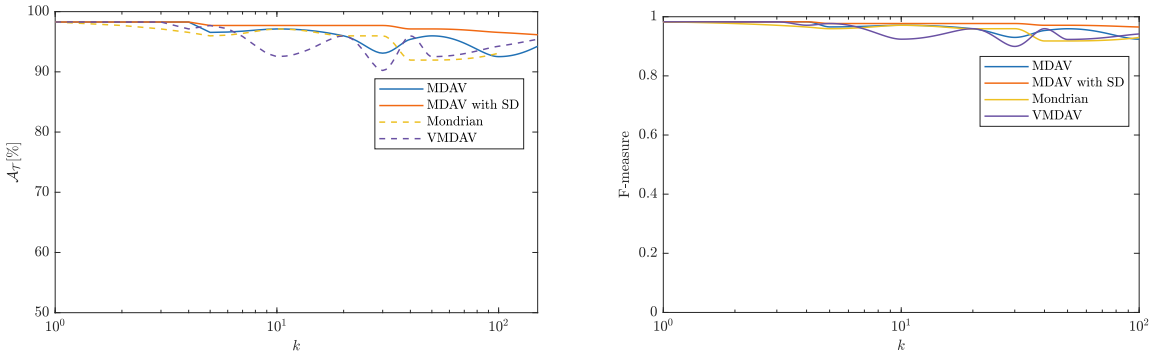
Furthermore, from Figure 6a, utility is remarkably preserved by MDAV with SD. In fact, accuracy do not drop below 80%, even for $k = 1000$. Similar encouraging results are obtained when measuring F-measure 6b. Besides, we can see that the original MDAV is the second best performer in regards to practical utility, at least up to $k = 60$. On the other hand, V-MDAV and Mondrian are the worst performers, although for very few small values of k, V-MDAV gets the best results.

When plotting the evolution of distortion as k is progressively increased, while microaggregating the Adult data set, Figure 7a confirms that MDAV with SD applies less distortion (as measured through the combined metric proposed in [14]) than the other algorithms. Original MDAV repeats as the second best performer, now in terms of MSE, but Mondrian and V-MDAV seem to introduce more perturbation. In any case, distortion grows exponentially so, according to this metric, data would render useless very quickly. In fact, when $k = 50$, MDAV and MDAV with SD would have injected more than 20% of distortion while Mondrian and V-MDAV more than 40%.

The utility metric obtained empirically may not go hand in hand with a more syntactical measure based on MSE. This is confirmed in Figure 7b where we plot accuracy vs data distortion. The scatter plot shows that, although the distortion increases, e.g. up to 0.5, the corresponding accuracy keeps more or less stable in 80% for all the microaggregation algorithms. This implies that distortion is not a good predictor of the practical utility of microaggregated data, at least in the application domain here studied.

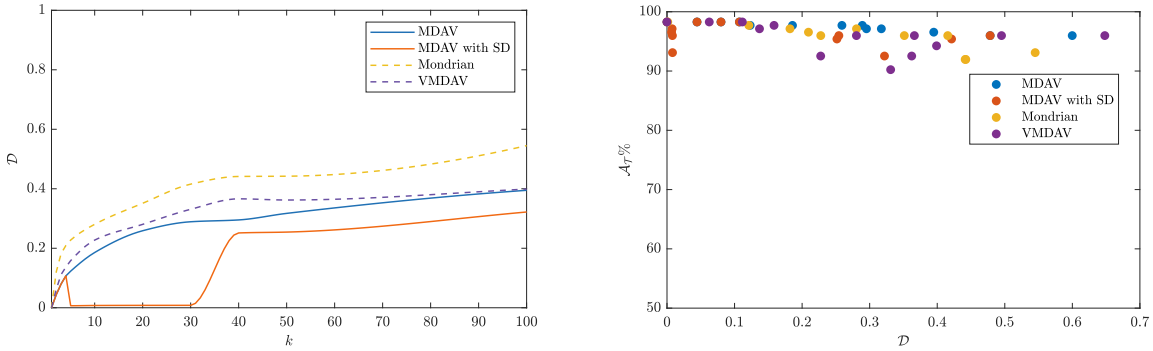
As described in 3, the results aforementioned are corroborated in experiments with three more data sets. When testing the *Breast Cancer Wisconsin* data set, the resilience of empirical data utility manifests again when k-anonymous microaggregation is enforced. Once again, the benefits of MDAV with SD are evident when outperforming the accuracy obtained by the rest of algorithms, as can be seen in Figure 8. Beyond the clear superiority of MDAV with SD, it is not clear for this data set which of the other algorithms performs the best in terms of accuracy.

Regarding the standard metric, note in Figure 9a that MDAV with SD also has the least distortion, that Mondrian performs the worst, and that both MDAV and V-MDAV show a similar distortion trend. As with the previous data set, the results of distortion hardly explain the practical utility of microaggregated data because it can be seen in 9a that accuracy does not vary as significantly as MSE when measuring the impact of microaggregation algorithms.



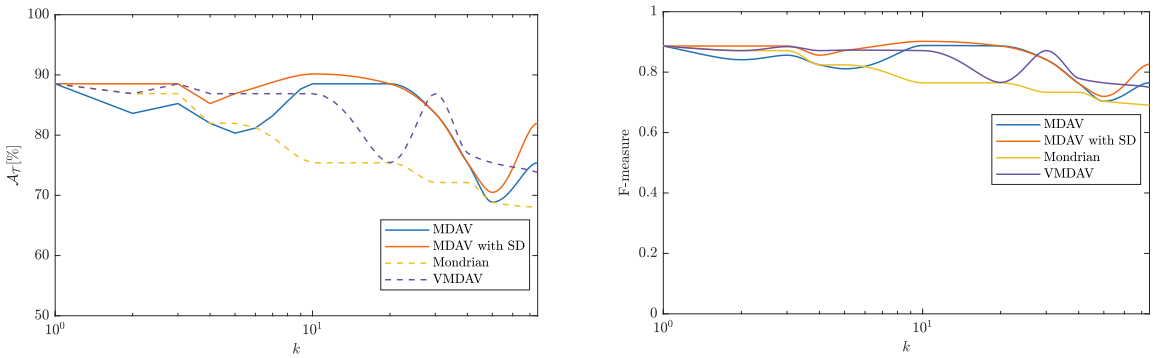
(a) Accuracy degradation **(b)** F-measure degradation

Figure 8. Degradation of the empirical utility of the microaggregated "Breast Cancer Wisconsin" data set.



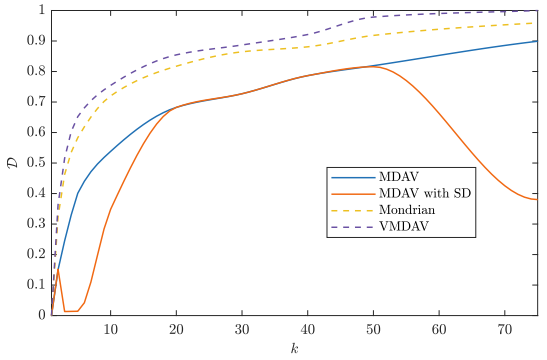
(a) Distortion measured for different values of k-anonymity. **(b)** Representation of accuracy vs distortion

Figure 9. Distortion of the microaggregated "Breast Cancer Wisconsin" data set. The distortion corresponding to MDAV with SD is measured according to the hybrid metric proposed in [14].

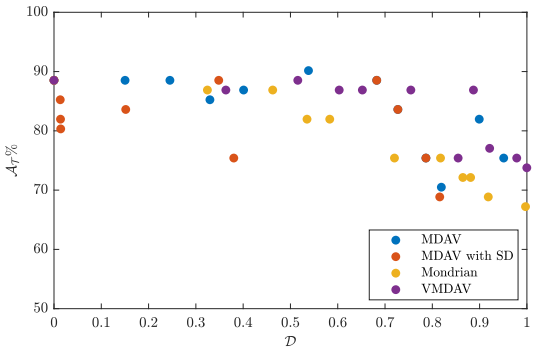


(a) Accuracy degradation **(b)** F-measure degradation

Figure 10. Degradation of the empirical utility of the microaggregated "Heart Disease" data set.

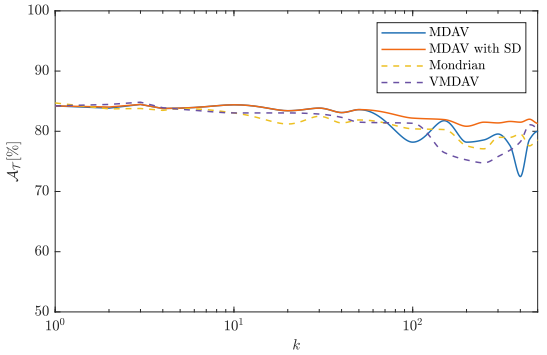


(a) Distortion measured for different values of k-anonymity.

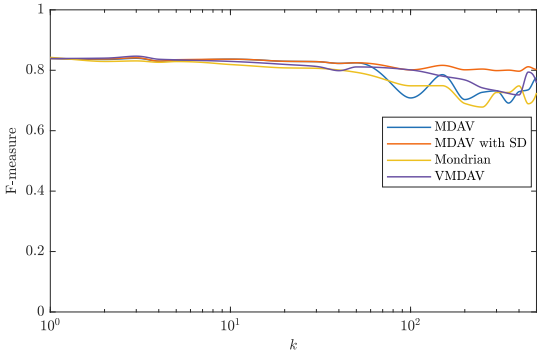


(b) Representation of accuracy vs distortion

Figure 11. Distortion of the microaggregated "Heart Disease" data set. The distortion corresponding to MDAV with SD is measured according to the hybrid metric proposed in [14].

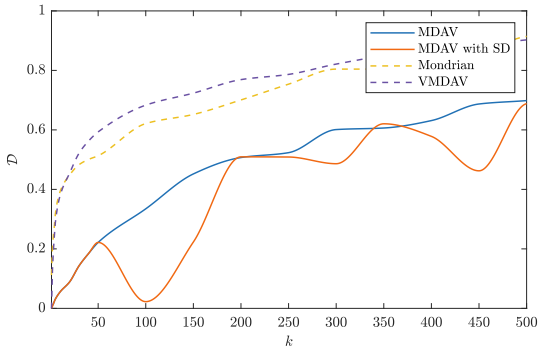


(a) Accuracy degradation

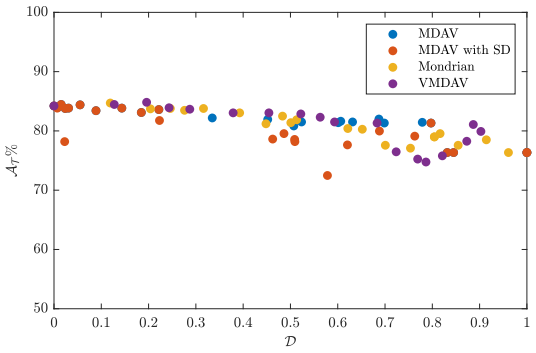


(b) F-measure degradation

Figure 12. Degradation of the empirical utility of the microaggregated synthetic data set.



(a) Distortion measured for different values of k-anonymity.



(b) Representation of accuracy vs distortion

Figure 13. Distortion of the microaggregated synthetic data set. The distortion corresponding to MDAV with SD is measured according to the hybrid metric proposed in [14].

Figures 10, 11, 12, and 13 illustrates the results of assessing microaggregation algorithms over the Heart Disease and synthetic data sets. For both of them, microaggregation, in general, performs quite well in terms of practical utility (see Figures 10a and 12a) while distortion grows much faster (see Figures 11a and 13a). In any case, the original MDAV exhibits anonymized data with lower distortion and stable accuracy, only improved by its statistically dependent variant, MDAV with SD). Finally, V-MDAV and Mondrian show interesting results on these data: while the former would spawn more distortion than the latter, V-MDAV apparently preserves better the data utility when measured as accuracy of the resulting machine learning model.

5. Discussion

Our systematic experimentation shows that k-Anonymous microaggregation has a benevolent, still destructive, effect on microdata in terms of its empirical utility, which is measured as the accuracy of learning models built from such data. Namely, while meeting a k-anonymity criteria, microaggregation preserves data utility even for high values of k, which is consistent with the results obtained in [12], where only MDAV is assessed.

This positive effect is attributed to the averaging operations to find a centroid that would be denoising the data, making it more resistant to perturbation.

In addition, although said averaging, inherent to microaggregation, might be even convenient, the distortion metric based on MSE would measure it as utility degradation. In this sense, MSE is a pessimistic metric that, in general, is not able to predict the practical utility of microaggregated data in this domain. As a matter of fact, not even the combined distortion metric proposed in [14] for MDAV with SD is capable of estimating such practical utility, despite its great performance in terms of accuracy.

The results obtained by MDAV with SD confirms that adapting privacy protection mechanisms to the intrinsic statistical properties of microdata and to the specific application domain might open the door to interesting improvements in utility preservation. This approach has not been addressed for microaggregation algorithms and particularly for MDAV-based approaches, so there is an appealing avenue for future work.

The "positive" impact of anonymization algorithms is indirectly reported by previous work that accounts for, e.g., the reduced degradation of obfuscated data under certain conditions [33,45], and the beneficial contribution to utility of some anonymization techniques [31] that may act as feature selection mechanisms, particularly when the protection strategy is selectively tailored to the application domain [35].

V-MDAV and Mondrian show, in general, a lower performance than the ones of MDAV and MDAV with SD in terms of both distortion and accuracy. However, since the strategies of V-MDAV and Mondrian operate on the internal distribution of the microdata set, such results could vary according to the data set being microaggregated.

Beyond the promising results, it is worth noting that our approach has inevitably some limitations that arise, essentially, from the bounded evaluation context we have defined . For instance, the application domain, where utility is empirically measured, is binary classification. However, many other domains may exist where utility is extracted differently.

Furthermore, a statistical dependence should exist between quasi-identifiers and confidential attributes such that something can be learned and preserved when microaggregating. Evidently, if this is not the case, another utility metric should be assessed.

Some avenues of future work could be addressed to complement this work. First, further fixes could be designed to adapt existing microaggregation algorithms to better preserve data utility. Some inspiration could certainly be taken from machine learning strategies. Additionally, this analysis could be extended considering alternate contexts (or attacker scenarios), e.g., where more confidential attributes are disclosed, or where a multi-class classification problem (more than two classes in the confidential attribute) is involved.

6. Conclusions

k-Anonymous microaggregation algorithms are able to preserve much of the data utility while protecting the privacy of each subject in groups of k individuals. Their clustering and averaging operations may contribute to filter, normalise, or consolidate the statistical information within microdata, e.g., when exploiting data through machine learning applications. This is confirmed in this paper through systematic experimentation with several microaggregation algorithms, data sets and machine learning mechanisms.

Interestingly enough, further catching and processing such statistical properties of microdata (e.g., the statistical dependence between quasi-identifiers and confidential attributes) when building microaggregation algorithms cause an additional slowdown in the degradation of empirical utility. This is clearly evidenced by MDAV with SD through our extensive tests.

Although Mondrian and V-MDAV consistently perform worse than MDAV and MDAV with SD, the two former algorithms behave differently between each other in terms of accuracy and MSE-based distortion. This would evidence the dependence of their performance on the internal distribution of the data set, as claimed by their creators. Such dependency calls again our attention to the need of considering the application domain of data (size, exploitation mechanisms, distribution of tuples) when designing or adapting privacy protection.

These considerations pave the way for future work on improving the performance of microaggregation algorithms. For instance, other anonymization algorithms could be assessed under these conditions to test their behaviour when empirical utility is measured. Though, some of their reconstruction techniques, e.g., using other than numerical representations for microaggregated data, could complicate the measurement of utility when the application domain is machine learning, so further assumptions or preprocessing should be done. Additionally, it is worth exploring adaptations or novel contributions for privacy protection that exploit to the maximum the statistical properties of all the information available within microdata. Intuitively, it seems that some of the strategies available for machine learning could be used to preserve the utility of microaggregated data.

Author Contributions: "Conceptualization, A.R.-H., D.R.-M. and J.E.-J.; methodology, A.R.-H., D.R.-M. and J.E.-J.; software, A.R.-H.; validation, A.R.-H. and J.E.-J.; formal analysis, A.R.-H. and D.R.-M.; investigation, A.R.-H.; resources, A.R.-H. and J.P.-A; data curation, A. R.-H. and J.E.-J.; writing–original draft preparation, A. R.-H. and J.E.-J.; writing–review and editing, J. P.-A ; visualization, L.U-A.; supervision, J. F. and L.U-A.; project administration, J. F. and L.U-A.; funding acquisition, J. F., D.R-M and L.U-A.".

Funding: This work is partly supported by the Spanish Ministry of Industry, Energy and Tourism (MINETUR) through the "Acción Estratégica Economía y Sociedad Digital (AEESD)" funding plan, through the project, "Data-Distortion Framework (DDF)", ref. TSI-100202-2013-23. Additional funding supporting this work has been granted to UPC by the Spanish Ministry of Economy and Competitiveness (MINECO) through the "Anonymized Demographic Surveys (ADS)" project, ref. TIN2014-58259-JIN, under the funding program "Proyectos de I+D+i para Jóvenes Investigadores", and through the project "MAGOS", ref. TEC2017-84197-C4-3-R, as well as by the Government of Catalonia, under grant 2014 SGR 1504.

J. Parra-Arnau was supported by the Spanish government under grant TIN2016-80250-R and by the Catalan government under grant 2017 SGR 00705. J. Parra-Arnau is also the recipient of a Juan de la Cierva postdoctoral fellowship, IJCI-2016-28239, from the Spanish Ministry of Economy and Competitiveness.

Acknowledgments: The authors gratefully acknowledge the support provided by the Escuela Politécnica Nacional, for the development of the project PII-DETRI-2019-01 - "Privacidad Sintáctica Funcional: Análisis y adaptación de mecanismos de anonimato con enfoque en la preservación de utilidad de los datos". We gratefully acknowledge the invaluable assistance of Irene Carrión-Barberà, M.D., in the preparation of the medical example in Fig. 2.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDAV	Maximum distance to average vector
MDAV with SD	MDAV with preservation of statistical dependence
V-MDAV	Variable MDAV
MSE	Mean squared error

References

1. Mehta, N.; Pandit, A. Concurrence of big data analytics and healthcare: A systematic review. *Int. J. Med. Inf.* **2018**, *114*, 57–65.
2. Tiwari, S.; Wee, H.M.; Daryanto, Y. Big data analytics in supply chain management between 2010 and 2016: Insights to industries. *Comput., Ind. Eng.* **2018**, *115*, 121–135.
3. Tiwari, S.; Wee, H.M.; Daryanto, Y. The strategic value of data resources in emergent industries. *Int. J. Inf. Mgmt.* **2018**, *39*, 146–155.
4. Bean, R. Every Company Is a Data Company. *Forbes* **2018**.
5. Samarati, P. Protecting respondents' identities in microdata release. *IEEE Trans. Knowl., Data Eng.* **2001**, *13*, 1010–1027.
6. Sweeney, L. Uniqueness of simple demographics in the U.S. population. Tech. Rep. LIDAP-WP4, Carnegie Mellon Univ., Sch. Comput. Sci., Data Priv. Lab., Pittsburgh, PA, 2000.
7. Dwork, C. Differential privacy. Proc. Int. Colloq. Automata, Lang., Program. (ICALP); , 2006; Vol. 4052, *Lect. Notes Comput. Sci. (LNCS)*, pp. 1–12.
8. Sankar, L.; Rajagopalan, S.R.; Poor, H.V. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Trans. Inform. Forensics, Secur.* **2013**, *8*, 838–852.
9. Hundepool, A.; de Wetering, A.V.; Ramaswamy, R.; Franconi, L.; Capobianchi, A.; de Wolf, P.P.; Domingo-Ferrer, J.; Torra, V.; Brand, R.; Giessing, S. *μ-ARGUS version 3.2 software and user's manual*. Stat. Neth., Voorburg, Netherlands, 2003.
10. Domingo-Ferrer, J.; Mateo-Sanz, J.M. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. Knowl., Data Eng.* **2002**, *14*, 189–201.
11. Solanas, A.; Martínez-Ballesté, A. V-MDAV: Multivariate microaggregation with variable group size. Proc. Int. Conf. Comput. Stat. (CompStat); , 2006; pp. 917–925.
12. Rodríguez-Hoyos, A.; Estrada-Jiménez, J.; Rebollo-Monedero, D.; Parra-Arnau, J.; Forné, J. Does *k*-anonymous microaggregation affect machine-learned macrotrends? *IEEE Access* **2018**, *6*, 28258–28277.
13. Domingo-Ferrer, J.; Torra, V. Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation. *Data Min., Knowl. Discov.* **2005**, *11*, 195–212.
14. Rebollo-Monedero, D.; Forné, J.; Soriano, M.; Puiggalí Allepuz, J. *k*-Anonymous microaggregation with preservation of statistical dependence. *Inform. Sci.* **2016**, *342*, 1–23.
15. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Mondrian multidimensional *k*-anonymity. Proc. IEEE Int. Conf. Data Eng. (ICDE); , 2006; pp. 25–35.
16. Domingo-Ferrer, J.; Torra, V. A critique of *k*-anonymity and some of its enhancements. Proc. Workshop Priv., Secur., Artif. Intell. (PSAI); , 2008; pp. 990–993.
17. Rebollo-Monedero, D.; Forné, J.; Domingo-Ferrer, J. From *t*-closeness-like privacy to postrandomization via information theory. *IEEE Trans. Knowl., Data Eng.* **2010**, *22*, 1623–1636.
18. Rebollo-Monedero, D.; Parra-Arnau, J.; Díaz, C.; Forné, J. On the measurement of privacy as an attacker's estimation error. *Int. J. Inform. Secur.* **2013**, *12*, 129–149.
19. Truta, T.M.; Vinay, B. Privacy protection: *p*-Sensitive *k*-anonymity property. Proc. Int. Workshop Priv. Data Mgmt. (PDM); , 2006; p. 94.
20. Sun, X.; Wang, H.; Li, J.; Truta, T.M. Enhanced *p*-sensitive *k*-anonymity models for privacy preserving data publishing. *Trans. Data Priv.* **2008**, *1*, 53–66.
21. Lin, J.L.; Wen, T.H.; Hsieh, J.C.; Chang, P.C. Density-based microaggregation for statistical disclosure control. *Expert Syst., Appl.* **2010**, *37*, 3256–3263.
22. Matatov, N.; Rokach, L.; Maimon, O. Privacy-preserving data mining: A feature set partitioning approach. *Inform. Sci.* **2010**, *180*, 2696–2720.
23. Domingo-Ferrer, J.; González-Nicolás, Ú. Hybrid microdata using microaggregation. *Inform. Sci.* **2010**, *180*, 2834–2844.

24. Rebollo-Monedero, D.; Forné, J.; Soriano, M. An algorithm for *k*-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers. *Data, Knowl. Eng.* **2011**, *70*, 892–921.
25. Lin, K.P.; Chen, M.S. Privacy-preserving outsourcing support vector machines with random transformation. Proc. ACM SIGKDD Int. Conf. Knowl. Discov., Data Min. (KDD); , 2010; pp. 363–372.
26. Lin, K.P.; Chen, M.S. On the design and analysis of the privacy-preserving SVM classifier. *IEEE Trans. Knowl., Data Eng.* **2010**, *23*, 1704–1717.
27. Inan, A.; Kantarcioglu, M.; Bertino, E. Using anonymized data for classification. Proc. IEEE Int. Conf. Data Eng. (ICDE); , 2009; pp. 429–440.
28. Mancuhan, K.; Clifton, C. Decision tree classification on outsourced data. *arXiv Preprint* **2016**.
29. Zaman, A.N.K.; Obimbo, C.; Dara, R.A. A novel differential privacy approach that enhances classification accuracy. Proc. Int. C* Conf. Comput. Sci., Softw. Eng. (C3S2E); , 2016; pp. 79–84.
30. Schmid, M.; Schneeweiss, H. The effect of microaggregation procedures on the estimation of linear models: A simulation study. *J. Econ., Stat.* **2005**, *225*, 529–543.
31. LeFevre, K.; DeWitt, D.J.; Ramakrishnan, R. Workload-aware anonymization. Proc. ACM SIGKDD Int. Conf. Knowl. Discov., Data Min. (KDD); , 2006; pp. 277–286.
32. Jafer, Y.; Matwin, S.; Sokolova, M. Task oriented privacy preserving data publishing using feature selection. Proc. Can. Conf. Artif. Intell.; , 2014; pp. 143–154.
33. Kisilevich, S.; Rokach, L.; Elovici, Y.; Shapira, B. Efficient multidimensional suppression for *k*-anonymity. *IEEE Trans. Knowl., Data Eng.* **2010**, *22*, 334–347.
34. Gursoy, M.E.; Inan, A.; Nergiz, M.E.; Saygin, Y. Privacy-preserving learning analytics: Challenges and techniques. *IEEE Trans. Learn. Technol.* **2017**, *10*, 68–81.
35. Malle, B.; Kieseberg, P.; Weippl, E.; Holzinger, A. The right to be forgotten: Towards machine learning on perturbed knowledge bases. Proc. Int. Conf. Availab., Rel., Secur. (ARES); , 2016; Vol. 9817, *Lect. Notes Comput. Sci. (LNCS)*, pp. 251–266.
36. Brickell, J.; Shmatikov, V. The cost of privacy: Destruction of data-mining utility in anonymized data publishing. Proc. ACM SIGKDD Int. Conf. Knowl. Discov., Data Min. (KDD); , 2008.
37. Matwin, S.; Nin, J.; Sehatkar, M.; Szapiro, T. A review of attribute disclosure control. In *Advanced research in data privacy*; Navarro-Arribas, G.; Torra, V., Eds.; Springer Int. Publ.: Switzerland, 2015; Vol. 567, *Stud. Comput. Intell.*, pp. 41–61.
38. D'Acquisto, G.; Domingo-Ferrer, J.; Kikiras, P.; Torra, V.; de Montjoye, Y.; Bourka, A. Privacy by design in big data. Tech. Rep. TP-04-15-941-EN-N, EU Agency for Netw., Inform. Secur. (ENISA), 2015.
39. Anonymisation techniques. Opinion 05/2014, Art. 29 Data Prot. Work. Party (indep. EU advis. cmte., Dir. 95/46/EC, Art. 29), 2014.
40. Machanavajjhala, A.; Gehrke, J.; Kiefer, D.; Venkitasubramanian, M. *l*-Diversity: Privacy beyond *k*-anonymity. Proc. IEEE Int. Conf. Data Eng. (ICDE); , 2006; p. 24.
41. UCI machine learning repository: Adult dataset.
42. UCI machine learning repository: Breast cancer Wisconsin dataset.
43. UCI machine learning repository: Heart disease dataset.
44. Gong, Q.; Kun, L. Mondrian. https://github.com/qiyuangong/Mondrian, 2018.
45. Cormode, G.; Shen, E.; Gong, X.; Yu, T.; Procopiuc, C.M.; Srivastava, D. UMicS: From anonymized data to usable microdata. Proc. ACM Int. Conf. Inform., Knowl. Mgmt. (CIKM); , 2013; pp. 2255–2260.

Sample Availability: Samples of the compounds ..... are available from the authors.