Preserving Empirical Data Utility in k-Anonymous Microaggregation via Linear Discriminant Analysis

Ana Rodríguez-Hoyos[a,b], David Rebollo-Monedero[b], José Estrada-Jiménez[a,b,*], Jordi Forné[b], Luis Urquiza-Aguiar[a]

[a] *Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional (EPN), Ladrón de Guevara, E11-253 Quito, Ecuador*

[b] *Department of Telematic Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain*

Abstract

Today's countless benefits of exploiting data come with a hefty price in terms of privacy. k-Anonymous microaggregation is a powerful technique devoted to revealing useful demographic information of microgroups of people, whilst protecting the privacy of individuals therein. Evidently, the inherent distortion of data results in the degradation of its utility. This work proposes and analyzes an anonymization method that draws upon the technique of linear discriminant analysis (LDA), with the aim of preserving the empirical utility of data. Further, this utility is measured as the accuracy of a machine learning model trained on the microaggregated data. By transforming the original data records to a different data space, LDA enables k-anonymous microaggregation to build microcells more tailored to an intrinsic classification threshold. To do this, first, data is rotated (projected) towards the direction of maximum discrimination and, second, scaled in this direction by a factor α that penalizes distortion across the classification threshold. The upshot is that thinner cells are built along the threshold, which ends up preserving data utility in terms of the accuracy of machine learned models for a number of standardized data sets.

Keywords: data privacy, statistical disclosure control, LDA, microaggregation, data utility

1. Introduction

Modern technologies and massive access to them by billions of people have enabled the generation of vast amounts of data. Also, more powerful and sophisticated information systems are developed to exploit such data with the aim of getting unprecedented intelligence and personalization. The potential benefits of these technologies are countless in several fields such as healthcare, advertising, and even industrial engineering [33, 59, 28]. For most of such fields, more utility can be mined from data to unveil qualitatively superior insight into challenges and opportunities that may otherwise remain undiscovered [15, 44].

A compelling example of application where data utility is absolutely critical is, undoubtedly, health and, particularly, precision or personalized medicine. In this domain, a large data sample could reveal otherwise subtle patterns. To illustrate this point, we recall a well-known medical experiment conducted in 1989, in which a large number of participants in a study allowed practitioners to find out a slight but clinically relevant effect of aspirin tablets in participants who had a myocardial infarction [44]. From a sample of $22,071$ individuals, the study found that heart attacks were 0.77% less frequent when participants took an

*Corresponding author.

Email addresses: ana.rodriguez@epn.edu.ec (Ana Rodríguez-Hoyos), david.rebollo@entel.upc.edu (David Rebollo-Monedero), jose.estrada@epn.edu.ec (José Estrada-Jiménez), jforne@entel.upc.edu (Jordi Forné), luis.urquiza@epn.edu.ec (Luis Urquiza-Aguiar)

aspirin table every other day, a phenomenon that would have been much harder to observe without such a large sample.

But exploitation of data encompasses serious privacy risks when information is associated with individuals. Since abundant details are usually collected about them, even after suppressing identifier attributes such as full names, other, apparently innocuous, personal attributes (quasi-identifiers), could still be used to re-identify an individual [56]. Thus, if a sensitive attribute (gender, health status, income) were disclosed along with other information, re-identification would enable an attacker to associate an individual with such attribute, thus violating her privacy. But this risk is exacerbated by the fact that data has become a core asset for companies [2], so there is a great incentive to exploit, share and even sell data to maximize profit.

Statistical disclosure control (SDC) is commonly used to tackle these privacy risks when disclosing microdata files (individual user data tabulated in records). Such SDC techniques build on perturbing quasi-identifier attributes to de-identify records, a process also called *anonymization*. The privacy models enforced through user data perturbation, e.g., k-anonymity [55, 46] or ϵ-differential privacy [11], are usually conditioned by a privacy parameter that defines an upper bound on the re-identification risk.

Differential privacy and other privacy criteria such as multi-party computation [65, 5] and integral privacy [60] are out of the scope of this work, since our target application is that of data release for general statistical analysis with a focus on data utility. Recall that differential privacy is conceived for online querying on predefined computations, and that in general it imposes stringent restrictions, both in terms of usability and data utility. Those restrictions, explained also in [31], render it rather inadequate for our purposes.

On the other hand, *k-anonymous microaggregation* is a high-utility mechanism to protect privacy in microdata by obfuscating demographic attributes. Carefully aggregating these attributes, a minimum level of distortion must be applied to original data. In fact, k-anonymous microaggregation is an excellent approach to applications requiring the preservation of data utility [42].

Obfuscating data to protect privacy naturally affects its resulting utility [49]. Consequently, there is a trade-off that must be addressed so that data exploitation keeps feasible and usable. In this line, the role of SDC mechanisms is guaranteeing a given level of privacy while preserving (some of) the utility of anonymized data.

The impact of these mechanisms on the utility of data has been commonly measured using standard, but merely syntactical, metrics, such as mean-squared error (MSE). However, to capture the practical utility of anonymized data, other metrics related to its application domain might be more relevant. For example, since a very common domain of application is building machine learning models, accuracy or F-measure of these models could be reasonable metrics of empirical utility.

Aiming to find a balance among privacy and empirical utility, some research is devoted, not only to design new less-"destructive" protection algorithms, but also to "adapt" already existing algorithms that increase the resulting utility of anonymized data. In this line, recent work is increasingly oriented to propose semantic (more empirical) approaches to the preservation of data utility when protecting privacy [48, 41, 1].

Although utility is certainly the *raison d'etre* of our effort, another parameter key to privacy protection usability is computational complexity. If protection mechanisms cannot cope with the (sometimes real-time) requirements of modern applications, they render unusable no matter how much utility is preserved. A few works have been proposed recently in this direction [35, 42].

In this work, we present and assess a strategy to preserve (empirical) utility of data after a k-anonymous microaggregation algorithm is applied. By representing original data in a new rotated and scaled domain, we adjust the implementation of the microaggregation algorithm to the specific application domain of data, which in this case is binary classification. As a result, the error of the machine learning model, when evaluated over new testing data, is reduced, at no cost, even for high anonymity levels.

1.1. Contribution and plan of the paper

The anonymization method addressed in this work is computationally and functionally efficient since the utility of data is preserved while the privacy level offered by an underlying microaggregation algorithm is left intact, at no additional cost in terms of running time.

Interestingly, data utility preservation at no (computational) cost could be a great incentive to adopt privacy protection technologies. In fact, some big tech companies are turning their privacy stance into a

huge competitive advantage. Thus, the companies that best adapt their operation to privacy requirements (preserving data utility and algorithm usability) will be in better position to exploit such advantage. In this context, these parameters could become a powerful value generator.

Below we briefly describe the main contributions of this work. For the sake of illustration, the highlights of our contributions are summarized in Fig. 1.

- Firstly, we propose the novel application of a powerful algebraic-statistical method to the problem of preserving empirical data utility when microaggregating microdata for privacy protection. Secondly, unlike the standard syntactical utility metrics, we verify the soundness of our contribution using an empirical metric from machine learning methods.
- Precisely, our method is based on applying Linear Discriminant Analysis to find the direction of maximum discrimination within the data space. This enables the microaggregation mechanism to intelligently tailor its anonymization strategy to the specific application domain of such data (binary classification in this case).
- This approach also involves weighting, i.e., adequately scaling, said discriminating direction in such a way that distances in this direction are penalized when creating k-anonymous microcells. The upshot is that microcells are grouped much less coarsely along the classification threshold.
- As this is, to the best of our knowledge, the first application of LDA to the field of SDC, to give some intuition regarding our approach, we include in this work a running example that allows us to graphically illustrate the transformation applied to data for preserving utility. This is presented along with an introduction of the theoretical foundations of LDA.
- Finally, we systematically evaluate this method on several data sets, both real and synthetic, using different machine learning algorithms and increasing anonymity levels and scaling factors.

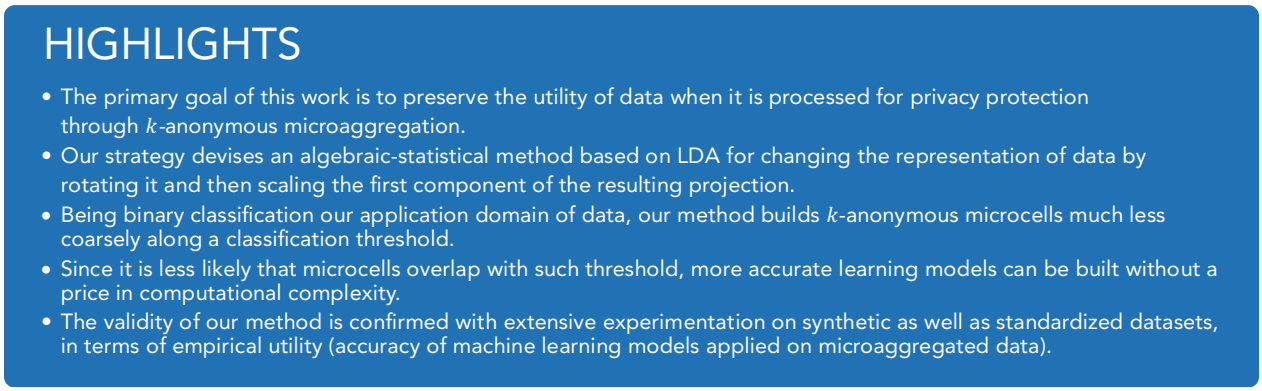


Figure 1: Highlights of our contribution.

The rest of this paper is organized as follows. §2 briefly reviews the current state of the art in k-anonymous microaggregation metrics and algorithms in the SDC literature. Some related works are presented in §2. §3 formally presents the proposed formulation of our privacy preserving approach, while §4 presents the experimental analysis and outcomes of this strategy. Finally, conclusions are drawn in §5.

2. State of the art on k-anonymous microaggregation

2.1. Background on microaggregation

When microdata 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*. These quasi-identifiers are regularly object of privacy protection mechanisms. Finally, *confidential attributes*, i.e., sensitive information about individuals, are usually disclosed without modification since de-identification of data owners is assumed when quasi-identifiers are anonymized.

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.

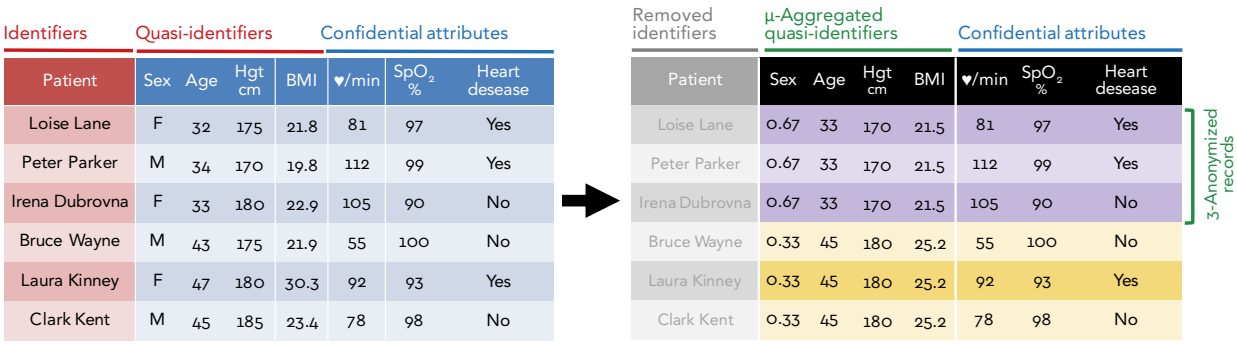


Figure 2: Toy example of k-anonymous micoraggregation. After suppressing identifiers, the records are clustered in groups of size k (microcells). Then, the quasi-identifiers in each micro cell are replaced a representative tuple (e.g., a centroid). Finally, microaggregated quasi-identifiers and original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

As illustrated in Fig. 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 employed 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 Fig. 3.

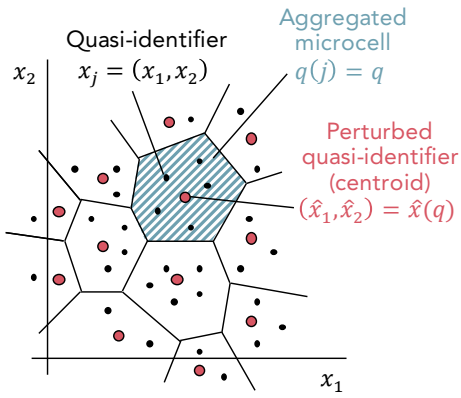


Figure 3: Intuition regarding k-anonymous microaggregation. The two-dimensional quasi-identifiers of a microdata set are depicted as points in an Euclidean space. Microaggregation partitions such points by building k-anonymous microcells to then replace each tuple with the centroid of the corresponding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

2.2. Algorithms for k-anonymous microaggregation

Getting groups of exactly k records from a microdataset is a strong restriction. In fact, multivariate microaggregation is an NP-hard problem. Thus, several heuristic algorithms have been proposed to cope with such complexity. First, the maximum distance (MD) [8] and its variation, maximum distance to average vector

(MDAV) ([8, 9]) are cataloged as fixed-size algorithms because all aggregated groups but one have exactly k elements. Variable-size algorithms include, on the other hand, μ-Approx by [14], minimum spanning tree (MST) by [16], variable MDAV (V-MDAV) by [18] and two-fixed reference points algorithms (TFRP) [4].

The *de facto* standard for numerical microaggregation is the MDAV algorithm. It was proposed by [16] as a practical evolution of a multivariate fixed-size microaggregation method and conceived by [7]. Since we use MDAV to illustrate our method in this work, for the sake of reproducibility, we provide in Fig. 4 a simplified version of that given by [9] and termed "MDAV generic".

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 [9].

In general, the implementations of microaggregation have been oriented to preserve the utility of data [22, 30, 6], which is evidently affected due to perturbation. Although the usual metric to measure such utility is MSE, other semantic-oriented metrics could be considered, aiming to conceive realistic implementations.

We use MDAV since it is a well-known microaggregation algorithm for numerical data in the literature of database anonymization. In fact, many of these works refer to MDAV not only as a standard method (or the most widely used) for microaggregation [57, 26] and use it as a baseline for comparison purposes ([54, 34]), but also recommend it due to its efficiency and performance [58] in terms of the resulting data utility. Even in recent years, MDAV is used as the baseline to find new and improved microaggregation approaches [17, 45, 24, 12, 66].

Talking about its impact on data utility, MDAV is even being actively used to enhance the utility of differentially private data sets via record masking [36, 47, 52]. Interestingly, its averaging operations to find a representative centroid turn to be a mechanism to reduce the amount of noise required to meet a differential privacy criteria. Due to this de-noising effect, it is not surprising that microaggregation is commonly used to face the privacy/utility trade-off together with machine learning techniques [3].

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 [10, 37, 40]. 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 [61, 53], 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 [22, 30, 6] due to perturbation, which commonly derives in more sophisticated and significantly costlier implementations in terms of computational time [39].

2.3. Utility of microaggregated data

Resulting utility of anonymized data is commonly measured inversely as 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. Moreover, these evaluations use to assess the performance of classifiers specifically adapted to operate on anonymized data [23, 18, 11, 29, 67], commonly using simulated data sets [50]. A lot of research has also investigated modifications of anonymization 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 ([21], [19], and [20]). In [42], a systematic evaluation is performed to determine the impact of k-anonymous microaggregation on the extraction of machine learned macrotrends from data.

3. Application of LDA to k-anonymous microaggregation

To explain the concept of LDA and then illustrate its application to preserving data utility while implementing k-anonymous microaggregation, we next introduce some principles and notation that are explained later through a running example. This example builds on a synthetic data set, generated according to the scenario and parameters described below.

3.1. Introduction to the preservation of the utility of microaggregated data through LDA

In order to assess data utility, a metric is required; it commonly derives from the application domain of data. For our approach, we use binary classification as the application domain since machine learning is increasingly used to exploit data. Namely, we assume that data requiring anonymization through microaggregation will be further processed to extract a binary classification model.

However, k-anonymous microaggregation groups records (building cells) without considering any application domain, so both privacy protection and data exploitation might be naturally incompatible in terms of utility preservation. Thus, or aim is modify this aggregation process such that it adjusts to the binary classification algorithm while privacy is still protected.

Binary classification, in general, obtains a threshold that enables classifying the elements of a given set that, in our scenario, consists of multidimensional numeric points. Since k-anonymous microaggregation groups such points in cells without any particular shape or direction, it is likely that said threshold will split some of the cells, implying that their corresponding centroids misrepresent their aggregated points when obtaining a classification model. In order to address this issue that would affect the resulting utility of data, we resort to LDA.

LDA ([32, 13]) is a method commonly used as a preprocessing step before implementing machine learning classification. It aims at modeling the difference between classes of data by projecting a data set onto a lower-dimensional space. To do this, loosely speaking, LDA looks for maximizing the distance (separability) among the data of different classes (their means) while minimizing the variation within each class. Such projection enables good class separability and even a reduction of computational costs on classification tasks ([51]) .

LDA and Fisher's linear discriminant technique ([13]) are often used interchangeably, but there is a subtle difference. On the one hand, with Fisher's linear discriminant, we seek to maximize the ratio between the determinants of the between-class covariance and the within-class covariance. On the other hand, LDA fits a Gaussian homoscedastic mixture to the generative model via maximum likelihood estimation. The original linear discriminant was described for a 2-class problem, and it was generalized later for multiple classes. Both methods result in the same direction of best discrimination for the corresponding class from the multivariate observation.

Interestingly, such direction of best discrimination can be used to tailor the microaggregation process such that microcells are built aligned to such direction; by, basically, a *rotation*. In addition, we propose a *weighing* step of the records. Both of this building blocks (rotation and scaling/weighting) aim at increasing the separability of the two classes embedded in data to facilitate the construction of utility-preserving microcells. Namely, our approach would be implemented before applying the original microaggregation process, as depicted in the scheme of Fig. 5 .

In the next subsections we try to depict by example how this direction of best discrimination is obtained.

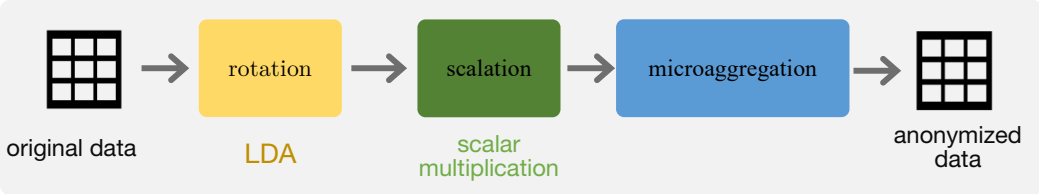


Figure 5: Main building blocks of our proposal to preserve utility from microaggregated data.

3.2. Integration of LDA into k-anonymous microaggregation

In this section we explain our proposed method in detail. We include a description on the scope considered –in particular for data utility exploitation– and a step-by-step illustration of the integration of LDA into k-anonymous microaggregation.

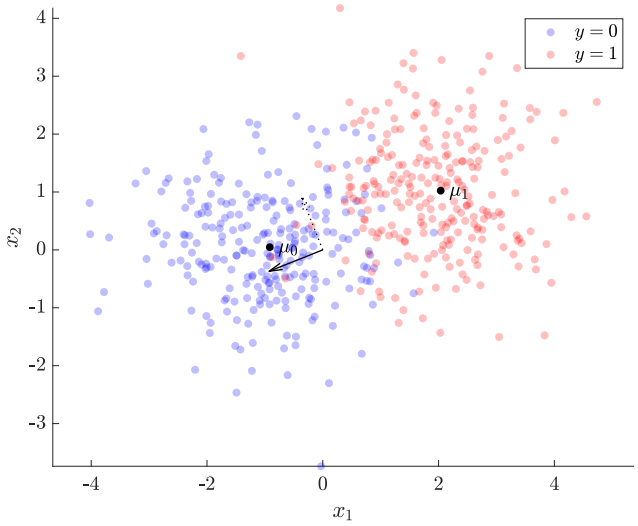


Figure 6: Depiction of the quasi-identifiers (x_2 vs x_1) of our toy synthetic data set. Samples are colored according to their class, y; blue for $y = 0$ and red for $y = 1$. The direction defined by mean points of both classes is the direction of maximum discrimination on which data will be projected to maximize its separability.

3.2.1. Scope and preliminar notation

As stated in §3.1, the scope of our work, in terms of data utility extraction (and application domain of data), is binary classification. Thus, we next make a brief description of the main elements of this scenario, the math connecting them, and the notation that will be used along the rest of this section.

First, consider a population of patients whose attributes (e.g., height/weight) and diabetes status are studied to build a model capable of detecting diabetes in new individuals, based on said attributes, i.e., a binary classification problem.

Then, let x be a numeric random variable (r. v.) in $\mathbb{R}^n$, i.e., an n-dimensional vector representing these attributes for an individual. Also, let Y be a binary random variable representing whether a patient has a diabetes condition ($Y = 1$) or not ($Y = 0$), i.e., a label. Let μ_1 and μ_0 be the mean vectors of the diabetic and non-diabetic subpopulations, respectively, considering only their attributes. Accordingly, let Σ_1 and Σ_2 be the corresponding covariance matrices, and p the prevalence of diabetics in this example. Finally, let

$$\Sigma_W = (1-p)\Sigma_0 + p\Sigma_1$$

be the within-class covariance matrix associated to the two-class data mentioned above. For single-class Fisher's discriminant, there is no need to compute the between-class matrix.

$$\Sigma_B = (1-p)\, p(\mu_1 - \mu_0)\,(\mu_1 - \mu_0)^T.$$

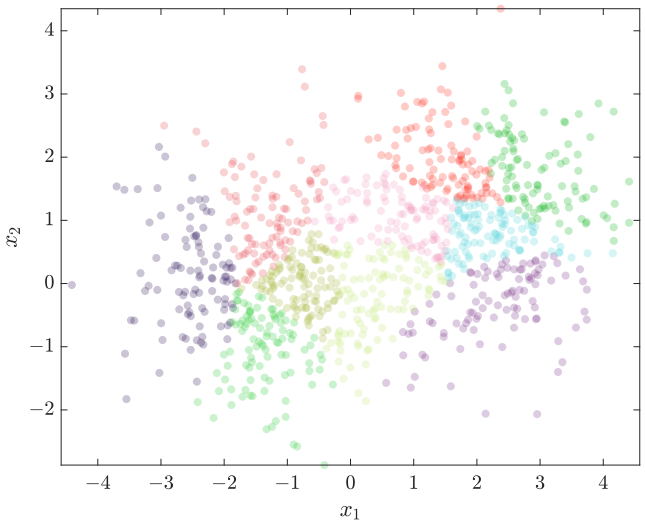


Figure 7: Microcells of samples obtained by applying k-anonymous microaggregation with MDAV on our toy synthetic data set ($k = 100$). Note how the single criteria to group points in clusters is their relative closeness.

Based on the previous definitions of x and Y, suppose a data set with n numerical attributes, resembling n quasi-identifiers, and a binary label as the confidential attribute. Besides, assume Y is to some extent predictable from the quasi-identifiers represented by x so the data set is useful in the realm of machine learning classification. Accordingly, consider a *generative model* defined by

$$\begin{aligned} x|Y &\sim \mathcal{N}(\mu_1, \Sigma) \\ x|\bar{Y} &\sim \mathcal{N}(\mu_0, \Sigma), \text{and} \\ p & \end{aligned}$$

that builds a Gaussian homoscedastic mixture fit via machine learning estimation. After characterizing a generic representation of the data on which our approach would be applicable, below we describe the method for preserving its utility when microaggregated.

3.2.2. Data rotation and scaling

Our strategy for preserving data utility when microaggregating consists of building microcells shaped in parallel to a discriminative direction and scaling data; all this with the aim to increase the separability of numeric records when a learning model is built. Accordingly, the following paragraphs describe the steps for finding such direction and implementing scaling of data.

To discern between Y and $\bar{Y}$, we use a *discriminative model* defined by $\mathrm{P}(Y|x)$, i.e., the probability *a posteriori* of the event Y. Recall that the corresponding Bayes factor (BF)

$$\frac{\mathrm{P}(x|Y)}{\mathrm{P}(x|\bar{Y})}$$

can be perfectly used as the discrimination function since it is a minimal sufficient statistic for Y from x under this homoscedastic and multivariate Gaussian model.

If we obtain the natural logarithm of the BF (which can be seen as a unit change), it can be finally expressed as a simple scalar product, i.e.,

$$\ln \mathrm{BF} = \left\langle \mu_1 - \mu_0,\, x - \frac{\mu_0 + \mu_1}{2} \right\rangle_{\Sigma_W^{-1}} = (\mu_1 - \mu_0)^T \, \Sigma_W^{-1} \left(x - \frac{\mu_0 + \mu_1}{2} \right).$$

We obtain a linear discriminant function whose *direction of maximum discrimination* (given that Σ_W is symmetric and applying some properties of the matrix multiplication) can be expressed as

$$U = \Sigma_W^{-1}(\mu_1 - \mu_0).$$

In general, for the multi-class Fisher's discriminant, the compression matrix U contains the orthonormal eigenvectors associated with the $L-1$ largest eigenvalues of $\Sigma_W^{-1}\,\Sigma_B$ (regarded as the solution to a generalized eigenvalue problem), where L denotes the number of classes. The optimization criterion is

$$\max_U \frac{\det(U^T \Sigma_B\, U)}{\det(U^T \Sigma_W U)}.$$

Rotation. LDA projects the data set (the part defined by x) on U, which defines the direction on which the distance among the different classes of the data is maximized while their variance is minimized. As a note, this direction can be more efficiently calculated, e.g., in MATLAB, without resorting to the calculation of an inverse matrix but by solving a system of linear equations.

Then, with full QR decomposition, we find an orthonormal base extension of U, V (an orthonormal base where one of the axes is U). This contains the normalized Fisher's discriminant direction. Next, the original attributes of the data set, which are points in the Euclidean space, are represented in terms of the new axes defined by V. Thus, we get the projection

$$x' = V^T\, x,$$

where x' is a transformed version of the original attributes represented by x. The first component of x' is the linear combination of the original attributes that best discriminate between the classes, while the rest can be considered less relevant.

Scaling. In line with the spirit of increasing the separability of two-class data, we complement the application of LDA with another strategy. We propose weighting the *first* transformed component, that is, first component of the LDA projection, by a factor $\alpha \geq 1$. In this manner, distance and distortion calculations will penalize the discrimination direction. Namely, we increase the distance among points in this direction so that they can be more easily grouped into microcells that do not overlap with the classification threshold. This scaling operation turns the new representation of data into the product $S\,V^T\,x$, for $S = \mathrm{diag}(\alpha, 1, \ldots, 1)$. Note that the scaling affects the first rotated component only, and this scaling can be regarded as a multiplication by a diagonal matrix. This product can be equally computed as $(S\,V^T)\,x$ or $S\,(V^T\,x)$, but if the data set to be transformed is very large, the former is much faster. Namely, this scaling by S can be regarded as matrix multiplication and the rotation by V can be associatively lumped into a transformation by a linear operator incorporating both scaling and rotation, for efficiency.

In Fig. 9 we summarize the main building blocks of the theoretical analysis of our proposal.

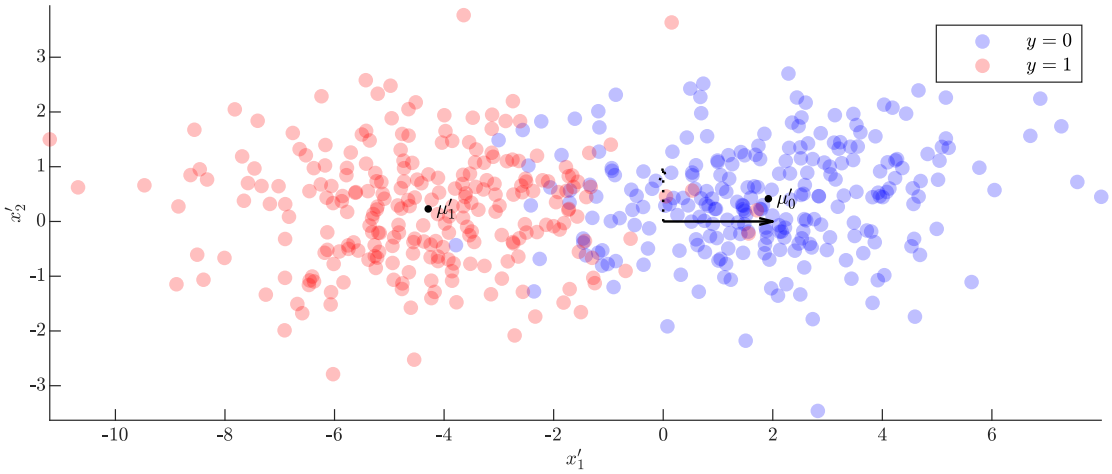


Figure 8: LDA projection of our toy synthetic data set on the direction of maximum discrimination x'_1. Scaling is also applied with $\alpha = 2$.

To graphically illustrate the wellness of our utility-preserving methods, we next depict their application in a simple scenario. In §4.6 we assess them experimentally using real data sets.

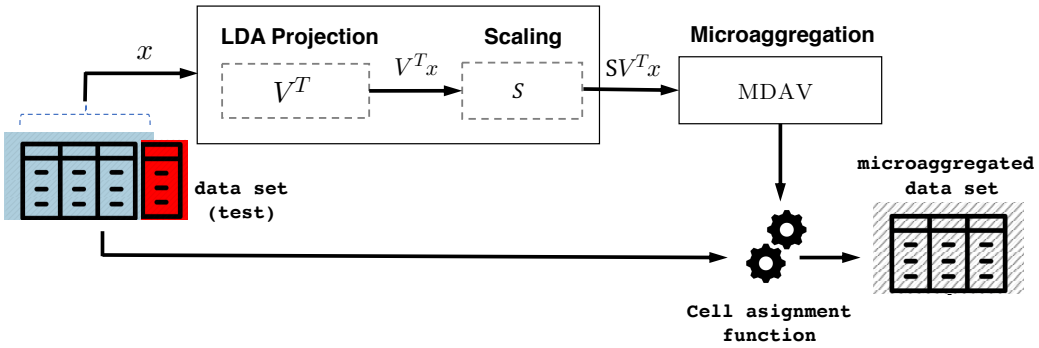


Figure 9: Main building blocks and theoretical operations involved in our proposal for preserving data utility. This can also be read as the particular experimentally methodology followed for its implementation.

From the scenario and generative model proposed in §3.2, assume a toy synthetic dataset of 1000 records, with two numerical quasi-identifiers (say, e.g., weight and height) x_1, x_2, and a corresponding binary confidential attribute y for each individual (e.g., diabetes status). For the sake of clarity, let us illustrate the distribution of these quasi-identifiers in Fig. 6, where x_1 and x_2 are plotted as points in two dimensions in the Cartesian plane. Evidently, the confidential attribute y is somewhat dependent on the contribution of the quasi-identifiers x_1, x_2, so a model can be learned to predict the former one from the latter ones.

If k-anonymous microaggregation through MDAV is employed to protect the identity of data owners, these points are grouped in cells of size k as graphically depicted in Fig. 7. As can be seen in this figure, microcells are built considering only relative closeness among points, so they tend to be grouped more or less equidistantly from a centroid. This produces "thick" groups with no particular orientation in any direction. Such thickness, and the omnidirectional distribution of cells, however, makes them more prone to fall over the classification threshold; thus, their corresponding centroids will likely misrepresent such points when a classification model is built. This evidently may contribute to reducing data utility.

Finding a maximally discriminative direction over which this data can be represented, LDA seems to be a convenient technique for k-anonymous microaggregation in terms of resulting empirical utility of anonymized data. In practice, LDA will maximize separation of data of the two classes and the inherent distortion would be weighted using an empirical parameter α. While in Fig. 6 we draw such direction, defined by the mean points of both classes of data, in Fig. 8 we can see the LDA projection of the data set on this direction. Said otherwise, data is rotated and scaled in this direction.

3.2.3. Brief discussion

Within this new representation of data, MDAV builds "thinner" microcells in the direction of maximum discrimination. Namely, increasing the separability between classes will enable MDAV to tailor the shape of resulting microcells to the intrinsic classification threshold estimated by LDA. This new distribution of cells is illustrated in Fig. 10 for our toy example. There we plot the microcells built from the original data set, following the microcell assignment obtained from microaggregating the LDA projection of the data set.

Since the resulting cells are clearly distributed in parallel to the intrinsic classification threshold gotten by LDA (Fig. 10), it is much less likely that such threshold falls over multiple cells. Thus, very few centroids would misrepresent data when a machine learning model is built from microaggregated data, preserving, in this way, its utility.

Besides preserving data utility, our method does not involve any additional computational complexity since the microaggregation process is not essentially changed but the representation of data before being anonymized. Fortunately, rotating and scaling data to change its representation are tasks performed once and does not entail significant complexity with respect to that of the iterative and complex process of microaggregation.

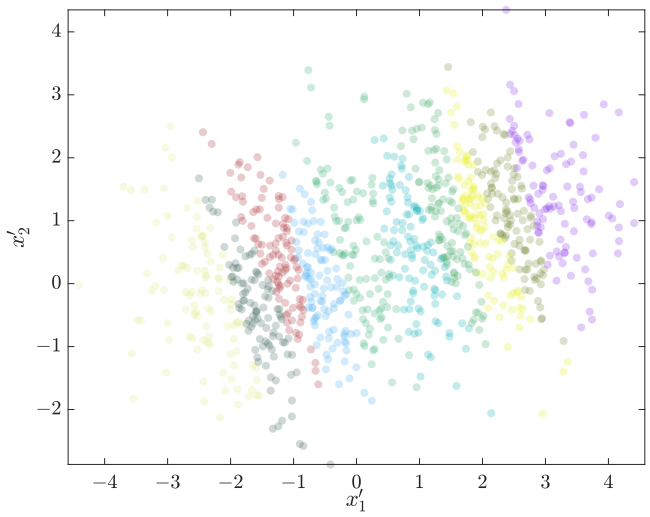


Figure 10: Microcells built in our original toy example by using the microcell assignment obtained from microaggregating the LDA projection of the data set ($k = 100$). Note how microcells are thinner in the direction of maximum discrimination, favoring the separation of the two classes by a classification task.

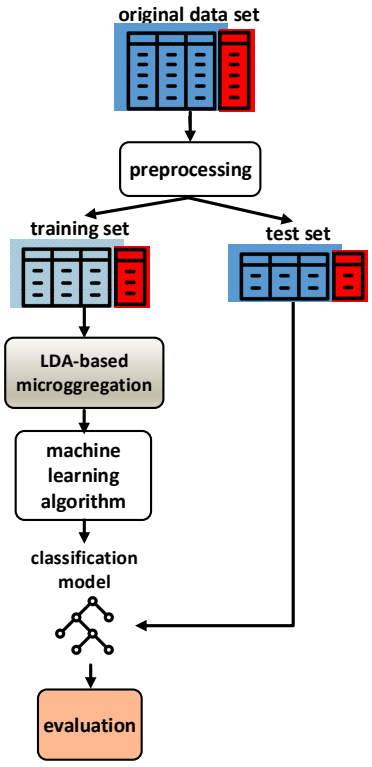


Figure 11: Main experimental methodology followed to implement our utility-preserving privacy protection approach on top of MDAV-based k-anonymous microaggregation.

4. Experimental evaluation

In this section we aim at describing the general context of the evaluation of our proposal on preserving the utility of anonymized data. For this, we describe the scenario assumed, the evaluation criteria (privacy and utility metrics), the tools used, and the phases implemented.

4.1. Evaluation scenario

Our evaluation revolves around the standard attack model in the SDC literature ([25]). To start, we assume a microdata set that needs to be released for research purposes. This microdata set has quasi-identifiers and a single confidential attribute. In this case, the utility of data lies in the statistical dependence among quasi-identifiers and a confidential attribute (such a diagnosis). In particular, such dependence would derive in a learning model to classify other individuals, e.g., as sick or healthy. In this data mining context, quasi-identifier records used to build the model are *input samples*, while the confidential records are *output labels*.

Besides, due to evident privacy concerns in this context, k-anonymous microaggregation is applied over quasi-identifiers to protect the privacy of data subjects. Thus, instead of original data, anonymized quasi-identifiers along with untouched confidential attributes are released. However, the utility of anonymized data would be undermined since obfuscating quasi-identifier records will most likely affect the quality of statistical trends embedded.

As mentioned in previous sections, to preserve such utility, we propose using LDA and scaling on the data as part of the microaggregation process. To assess this approach, we test it on several data sets and compare the resulting utility with that of data anonymized only with MDAV.

4.2. Data sets

With respect to the data, we use real and synthetic data sets. Furthermore, given the scenario proposed in this work, two main conditions are met when selecting data, in particular for real data sets. First, we look for microdata sets, i.e., data containing demographic information about actual individuals, such that a privacy concern might be involved. Second, we require data whose confidential attribute evidence a clear statistical dependence on its quasi-identifiers, since data utility is measured in terms of the capability of a machine learning algorithm to exploit such dependence. Given the last condition, standardized data sets that do now show such statistical characteristic were excluded.

We use four data sets: three real and one synthetic. The first one is "UCI Adult" data set [62], standardized in the evaluation of microaggregation algorithms but, conveniently, also employed to assess machine learning algorithms. The other two real data sets are "Breast Cancer Wisconsin" data set [63] and "Heart disease" data set [64], both containing 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, in the same way as the toy example illustrated in §3.2.2. In table 1 we include greater details of these data sets.

4.3. Evaluation criteria

To assess the performance of our utility-preserving method we need two metrics: a privacy metric and a utility metric. Both enable us to measure how data utility is preserved as privacy protection is increased. The privacy metric we use is k-anonymity since microaggregation algorithms aim at guaranteeing such criteria. Higher values of k imply larger anonymous microcells, so will offer more privacy to the subjects involved. Naturally, less utility is expected from data anonymized with higher values of k.

As described in §4.1, our evaluation scenario assumes that binary classification is the application domain of data. Thus, the corresponding utility metric here employed is classification accuracy, i.e., the accuracy of the classification model built from data, whether anonymized or not. Basically, accuracy quantifies the rate of correctly classified samples in a test set. Previous work has used accuracy as a utility metric, an empirical alternative to distortion measured as MSE ([20, 21, 27]). For the sake of confirmation, we also use F-measure as another machine-learning-based utility metric.

4.4. *Algorithms and tools*

In order to assess the effectiveness of our approach, we use some inputs and tools that we put together and describe next. We refer to the type of data and to the algorithms used for privacy protection and utility exploitation.

As expected, the privacy protection mechanism we use is MDAV, the de facto microaggregation algorithm. Besides its benefits in terms of time complexity, it has demonstrated to offer interesting results in terms of distortion and classification accuracy [42].

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. For the aforementioned data sets, the machine learning algorithms that provide the best results are boosting trees (Adult) and logistic linear regression (for the rest).

Finally, all the tests whose results are here presented were implemented with MATLAB 2018B. This includes loading and preprocessing data, the implementation of MDAV [9], as well as the evaluation of the resulting utility of perturbed data sets. This evaluation implies building machine learning models over data and applying such models over new data to measure classification accuracy and F-measure; all of this automatized using specific embedded functions for each algorithm. Greater detail is given in the next subsection.

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 (input)**	**confidential attribute (output label of the data set in ML terms)**
Adult [62]	45,222	15	Age, education-num, marital-status, sex, capital-gain, hours-per-week	Salary (>50K?)
Breast Cancer Wisconsin [63]	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 [64]	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

4.5. *Methodology*

Next we describe the experimental methodology we use to assess the effectiveness of our (empirical) utility-preserving approach for k-anonymous microaggregation. Figure 11 synthesizes the flow of the evaluation procedure, while Fig. 9 illustrates the specific methodology implemented for our utility-preserving strategy.

In general, our evaluation builds on determining whether the utility of microaggregated data is preserved better when LDA is considered as part of the anonymization process. In this scenario, two main steps are carried out: anonymization through k-anonymous microaggregation, and utility extraction through the application of a machine learning algorithm over anonymized data. Figure 11 illustrates the flow of these steps. To assess the benefits of our LDA-based approach, then, we measure the performance of such algorithm when LDA is used and when not.

But some preprocessing involving three steps is required. First, we transform the categorical values of the data set to numeric since MDAV only deals with numerical data. 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, we implement zero-mean, unit-variance normalization to each column of the training set, involving only quasi-identifiers.

Once normalized, the *microaggregation* algorithm is fed with the training set for data perturbation. We test progressively increasing values of k to then measure the utility degradation of data due to k-anonymous microaggregation. Figure 9 shows the specific process followed to obtain the anonymized data set from our approach proposed here. To start, the quasi-identifier values of the training set are transformed by projecting them through LDA and scaling them by a factor α. Then, the resulting transformed data is microaggregated using MDAV. Finally, the microcell assignment (a vector indicating the cell to which each record belongs) from the last step is applied on the original data to obtain the microaggregated data set, as depicted in Fig. 9.

With respect to the scaling, we made several tests varying the factor α from 1 (no scaling) to 64. Then, when presenting the results, we drew the corresponding maximum trace, i.e., the highest accuracy and F-measure values reached for each value of k.

After the anonymization phase, we implement the *utility extraction* phase. For this, we build a classification model using the microaggregated version of each data set (the training subset) as input. The algorithms showing 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 look for predicting a patient's condition (based on his data) by employing a machine learning model trained from anonymized shared data about other patients.

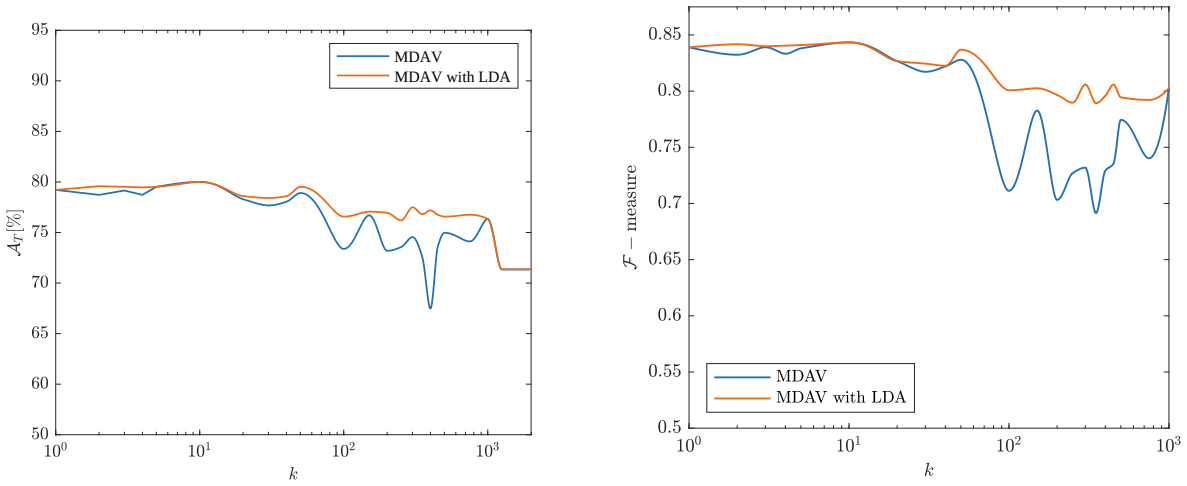
4.6. Experimental results

In this section, we describe the results of assessing the performance of our LDA-based k-anonymous microaggregation in terms of utility preservation. To this end, we present a series of figures where such performance is compared with that of MDAV. As previously explained, since we address the empirical utility of data, the metrics used are accuracy and F-measure of machine learned models when trained over data microaggregated, using an increasing value of k.

To start, we assess our approach on *UCI Adult data set*. In this case, we do not use all the records but a sample of 10% of them, looking for reducing even more the data utility after microaggregation. To keep the structure of the original data set, we take a random sample that preserves the prevalence of the output (confidential) attribute. By reducing the baseline utility, we think we can better visualize the effects of data utility preservation.

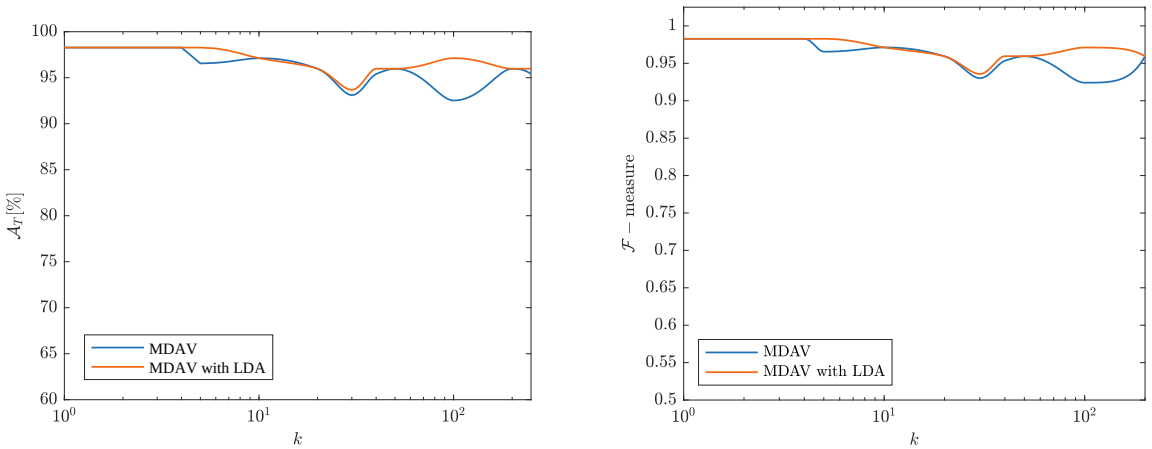
In Fig. 12 we depict the results of empirical utility extracted from the *UCI Adult* data set after applying k-anonymous microaggregation. Note that, as expected, the values of both metrics show a decreasing trend as the value of k increases: the impact of anonymization eventually renders data useless.

However, as depicted in Fig. 12, despite the inevitable degradation, the improvement, both in terms of accuracy and F-measure, is not only clear but significant in some cases when using MDAV with LDA. For example, when $k = 50$, the accuracy of the machine learning model goes from 81.8% to 83.9%, i.e., the error is reduced from 16.1% to 13.2%, which is a relative reduction of 18%. In the general, curves of utility look more stable when LDA and scaling are introduced, which implies that utility gets preserved even with relatively high values of k.



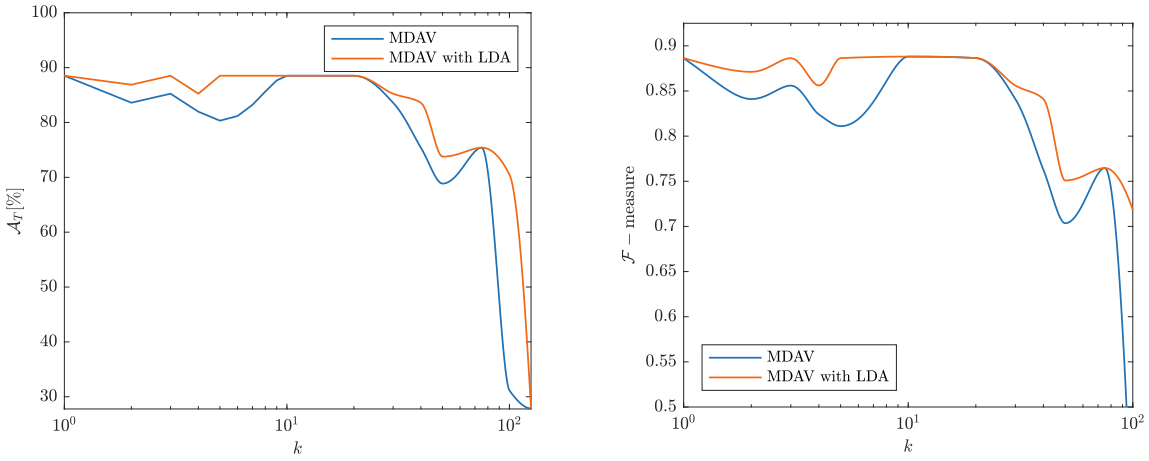
(a) Accuracy (b) F-measure

Figure 12: Empirical utility extracted from the UCI Adult dataset, microaggregated with original MDAV (blue) and with LDA-based MDAV (orange). Both in terms of accuracy and F-measure, LDA-based MDAV preserves better the utility of anonymized data.



(a) Accuracy (b) F-measure

Figure 13: Empirical utility extracted from the Breast Cancer Wisconsin dataset, microaggregated with original MDAV (blue) and with LDA-based MDAV (orange). Both in terms of accuracy and F-measure, LDA-based MDAV seems to preserve better the utility of anonymized data.



(a) Accuracy

(b) F-measure

Figure 14: Empirical utility extracted from the Heart disease dataset, microaggregated with original MDAV (blue) and with LDA-based MDAV (orange). Both in terms of accuracy and F-measure, LDA-based MDAV preserves better the utility of anonymized data.

As described in §3, the results aforementioned are corroborated in experiments with three more data sets. When testing *Breast Cancer Wisconsin* data set, the benefits of MDAV with LDA are again evident. Also in this case, for some values of k, the reduction is significant. Fig. 13 illustrates this in terms of accuracy and F-measure. Although the results of our method are better than those of "plain" MDAV, they do not seem as good as those obtained with the UCI Adult data set. There are several reasons that justify this behavior. Different data sets might naturally involve different macrotrends whose quality, in terms of utility, could also vary depending even on the amount of data. In addition, learning models built from the Breast Cancer Wisconsin data set show a maximum reachable accuracy of about 97% (i.e., very high), while it is about 80% for UCI Adult. Thus, we suspect that, when the room for improvement is greater, it is more likely that higher increases in accuracy can be reached.

Figures 14 and 15 illustrate the results of assessing microaggregation algorithms over Heart Disease and synthetic data sets, respectively. For this two data sets, we confirm that MDAV with LDA achieves its goal of preserving utility of microaggregated data sets better than with MDAV. Once more we verify the benefits of our proposed mechanism but also the difficulty to do so given that MDAV already offer a privacy preserving approach.

Even though experimenting over real data sets might be enough for validation purposes, we use a synthetic data set with the aim to validate the results obtained over real data.

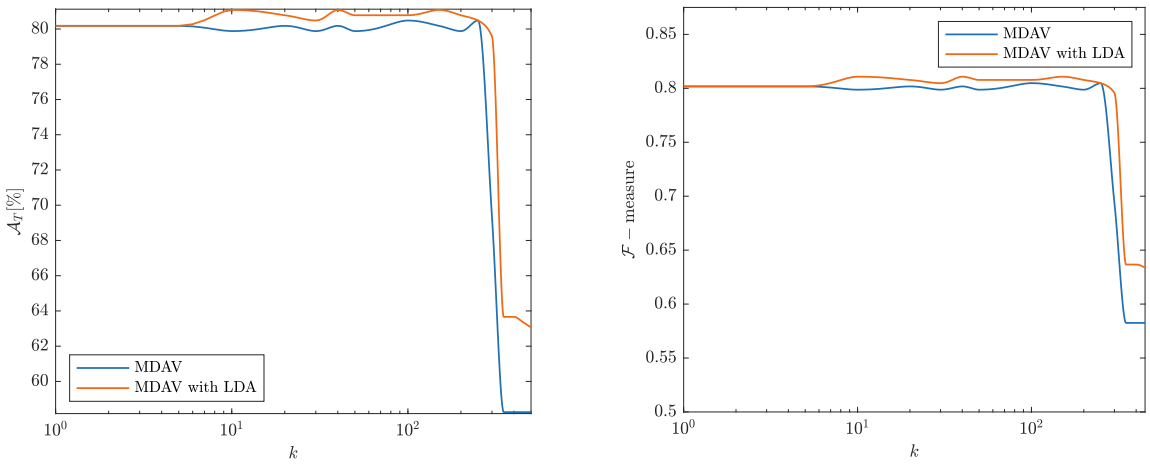
As a last note, classical distortion metrics based on MSE does not make sense in this study since the transformation based on LDA does not modify distances among points. In the case of scaling points are indeed separated in the direction of maximum discrimination, so it is even possible that the resulting distortion in this context is even greater than 1 although the empirical results are improved.

4.6.1. Discussion on results

The results obtained by our method are encouraging in that they show a consistent and, in some cases, significant preservation of data utility for microaggregated data. We would like to make some points below about this matter.

First, although MDAV with LDA behaves consistently better, in terms of data utility, than classical MDAV, the increase in utility may depend on the data set at hand, particularly on the information it can contribute to a learning model to improve its performance. Little could be done if machine learning algorithms cannot obtain practical accurate models from data even before applying privacy protection methods.

Second, in practice, our proposal does not imply any modification of the iterative process performed by MDAV. Given that our method modifies the representation of data before being microaggregated, the



(a) Accuracy

(b) F-measure

Figure 15: Degradation of the empirical utility for the synthetic data set.

resulting computation complexity remains invariable. This detail is important because, in times when the world revolves around big data, processing time quickly becomes a bottleneck with respect to the potential applications of large-scale databases. Moreover, domains as critical as health, vehicular traffic, or network intrusion detection are currently using tons of data to help computational systems make real-time, and even life-or-death decisions. Due to such demanding requirements, privacy issues related to data processing are commonly overshadowed. Thus, from the perspective of privacy, we feel that any improvement in preserving data utility *without a price in (computing) efficiency* is not negligible and some works are currently being purposed in this direction [43, 35].

Finally, we would like to point out that, since our approach resorts to changing the representation of data –although not necessarily its semantics–, conventional, syntactic, utility metrics such as distortion (measured as MSE) would be hardly applicable in this context. This fact gradually characterized syntactic metrics as less meaningful in practical, real-word applications.

5. Conclusion

Our method successfully preserves the empirical utility of data when microaggregated through MDAV. This is done by transforming quasi-identifier values in such a way that, after microaggregated, the resulting k-anonymous cells enable the construction of a more effective machine learning classifier.

Graphically illustrated, our proposal gets "thinner" microcells in the direction of maximum discrimination, obtaining a distribution of cells and reconstruction that better preserve the statistical properties on microaggregated data. Linear Discriminant Analysis and scaling are applied to find this direction and to weight the inherent distortion by an empirical parameter α.

In terms of accuracy and F-measure of resulting machine learning models, LDA applied to MDAV outperforms the classical implementation of MDAV. Although MDAV is by default benign when affecting the statistics within data, our approach successfully preserves the utility of data after microaggregation. This is confirmed trough systematic experimentation over synthetic and real data sets.

Conveniently, this benefit comes at no cost, e.g., in terms of running time, as other utility preserving proposals do ([38]). Thus, our approach is both functionally and computationally effective. Furthermore, ours is the first application of LDA to the domain of statistical disclosure control, applying a substantial and non trivial modification of any microaggregation algorithm, although here is assessed with MDAV.

Further research in this direction could involve the generalization of this method to address multi-class classification and not only binary classification scenarios. More generally, it might be interesting to study

other machine-learning-based models as mechanisms to represent and microaggregate data to reduce the distortion introduced to variables, combination of variables. or directions that contribute to a more accurate classification.

As other research in this field, our proposal paves the way for future work on improving the performance of microaggregation algorithms for specific application domains of data. This mainly implies exploring adaptations or novel contributions for privacy protection that exploit to the maximum the statistical properties of all the information available within microdata. This work confirms the intuition that some of the strategies already available for machine learning could be used to preserve the utility of microaggregated data.

Acknowledgment

We gratefully acknowledge the invaluable assistance of Irene Carrión-Barberà, M.D., in the preparation of the medical example in Figure 2.

This work is supported by the Escuela Politécnica Nacional through the project 'Privacidad Sintáctica Funcional: Análisis y adaptación de mecanismos de anonimato con enfoque en la preservación de utilidad de los datos", ref. PII-DETRI-2019-01.

This work is partly supported by the Spanish Ministry of Economy and Competitiveness (MINECO) through the project "MAGOS", ref. TEC2017-84197-C4-3-R.

Ana Rodríguez-Hoyos and José Estrada-Jiménez acknowledge the support from Escuela Politécnica Nacional (EPN) for their doctoral studies at Universitat Politècnica de Catalunya (UPC).

References

[1] M. Batet and D. Sánchez, "Semantic disclosure control: semantics meets data privacy," *Semantic Disclosure Control: semantics meets data privacy*, vol. 42, pp. 290–303, Jan. 2018.

[2] R. Bean, "Every company is a data company," *Forbes*, Sep. 2018. [Online]. Available: https://www.forbes.com/sites/ciocentral/2018/09/26/every-company-is-a-data-company/%2523d9840d45cfc5

[3] A. Blanco-Justicia, J. Domingo-Ferrer, S. Martínez, and D. Sánchez, "Machine learning explainability via microaggregation and shallow decision trees," *Knowledge-Based Systems*, Jan. 2020, in press.

[4] C.-C. Chang, Y.-C. Li, and W.-H. Huang, "TFRP: An efficient microaggregation algorithm for statistical disclosure control," *Journal of Systems and Software*, vol. 80, no. 11, pp. 1866–1878, Nov. 2007.

[5] F. K. Dankar, R. Brien, C. Adams, and S. Matwin, "Secure multi-party linear regression," in *Proceedings of the International Joint Conference on Extending Database Technology, and Database Theory (EDBT/ICDT)*, Athens, Greece, Mar. 2014, pp. 406–414.

[6] J. Domingo-Ferrer and Ú. González-Nicolás, "Hybrid microdata using microaggregation," *Information Sciences*, vol. 180, no. 15, pp. 2834–2844, 2010.

[7] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 189–201, 2002.

[8] J. Domingo-Ferrer, A. Solanas, and J. Castellà-Roca, "$h(k)$-private information retrieval from privacy-uncooperative queryable databases," *Online Information Review*, vol. 33, no. 4, pp. 720–744, 2009.

[9] J. Domingo-Ferrer and V. Torra, "Ordinal, continuous and heterogeneous k-anonymity through microaggregation," *Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 195–212, 2005.

[10] ——, "A critique of k-anonymity and some of its enhancements," in *Proceedings of the Workshop on Privacy and Security by means of Artificial Intelligence (PSAI)*, Barcelona, Spain, Mar. 2008, pp. 990–993.

[11] C. Dwork, "Differential privacy," in *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*, ser. Lecture Notes in Computer Science (LNCS), vol. 4052, Venice, Italy, Jul. 2006, pp. 1–12.

[12] E. Fayyoumi and O. Nofal, "Applying genetic algorithms on multi-level micro-aggregation techniques for secure statistical databases," in *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, Aqaba, Jordan, Oct. 2018, pp. 1–6.

[13] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[14] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, "Privacy-preserving learning analytics: Challenges and techniques," *IEEE Transactions on Learning Technologies*, vol. 10, no. 1, pp. 68–81, Sep. 2017.

[15] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Journal on Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[16] A. Hundepool, A. V. de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. de Wolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing, *μ-ARGUS version 3.2 software and user's manual*, Statistics Netherlands, Voorburg, Netherlands, 2003. [Online]. Available: http://neon.vb.cbs.nl/casc

[17] M. Iftikhar, Q. Wang, and Y. Lin, "Publishing differentially private datasets via stable microaggregation," in *Proceedings of the International Joint Conference on Extending Database Technology, and Database Theory (EDBT/ICDT)*, Lisbon, Portugal, Mar. 2019, pp. 662–665.
[18] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, Shanghai, China, Apr. 2009, pp. 429–440.
[19] Y. Jafer, S. Matwin, and M. Sokolova, "Task oriented privacy preserving data publishing using feature selection," in *Proceedings of the Canadian Conference on Artificial Intelligence*, Montréal, Canada, May 2014, pp. 143–154.
[20] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira, "Efficient multidimensional suppression for *k*-anonymity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 3, pp. 334–347, Apr. 2010.
[21] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Philadelphia, PA, Aug. 2006, pp. 277–286.
[22] J. L. Lin, T. H. Wen, J. C. Hsieh, and P. C. Chang, "Density-based microaggregation for statistical disclosure control," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3256–3263, Apr. 2010.
[23] K.-P. Lin and M.-S. Chen, "On the design and analysis of the privacy-preserving SVM classifier," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 11, pp. 1704–1717, Oct. 2010.
[24] H. Liu, Q. Zhang, K. Guo, and Y. Wu, "Grey maximum distance to average vector based on quasi-identifier attribute," *Journal of Grey System*, vol. 30, no. 1, 2018.
[25] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian, "*l*-Diversity: Privacy beyond *k*-anonymity," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, Atlanta, GA, Apr. 2006, p. 24.
[26] A. N. Mahmood, M. E. Kabir, and A. K. Mustafa, "New multi-dimensional sorting based *k*-anonymity microaggregation for statistical disclosure control," in *Proceedings of the EAI International Conference on Security and Privacy in Communication Networks (SecureComm)*, Padua, Italy, Sep. 2012, pp. 256–272.
[27] B. Malle, P. Kieseberg, E. Weippl, and A. Holzinger, "The right to be forgotten: Towards machine learning on perturbed knowledge bases," in *Proceedings of the International Conference on Availability, Reliability, and Security (ARES)*, ser. Lecture Notes in Computer Science (LNCS), vol. 9817, Salzburg, Austria, Aug. 2016, pp. 251–266.
[28] S. Mamonov and T. M. Triantoro, "The strategic value of data resources in emergent industries," *International Journal of Information Management*, vol. 39, pp. 146–155, Apr. 2018.
[29] K. Mancuhan and C. Clifton, "Decision tree classification on outsourced data," *arXiv Preprint*, no. 1610.05796, Oct. 2016. [Online]. Available: http://arxiv.org/abs/1610.05796
[30] N. Matatov, L. Rokach, and O. Maimon, "Privacy-preserving data mining: A feature set partitioning approach," *Information Sciences*, vol. 180, no. 14, pp. 2696–2720, 2010.
[31] S. Matwin, J. Nin, M. Sehatkar, and T. Szapiro, "A review of attribute disclosure control," in *Advanced research in data privacy*, ser. Studies in Computational Intelligence, G. Navarro-Arribas and V. Torra, Eds. Switzerland: Springer International Publishing, 2015, vol. 567, pp. 41–61.
[32] G. McLachlan, *Discriminant analysis and statistical pattern recognition*. New York, NY: John Wiley & Sons, 2004, vol. 544.
[33] N. Mehta and A. Pandit, "Concurrence of big data analytics and healthcare: A systematic review," *International Journal of Medical Informatics*, vol. 114, pp. 57–65, Jun. 2018.
[34] R. Mortazavi and S. Jalili, "Fine granular proximity breach prevention during numerical data anonymization," *Transactions on Data Privacy*, vol. 10, no. 2, pp. 117–144, Aug. 2017.
[35] E. Pallarès, D. Rebollo-Monedero, A. Rodríguez-Hoyos, J. Estrada-Jiménez, A. Mohamad Mezher, and J. Forné, "Mathematically optimized, recursive prepartitioning strategies for *k*-anonymous microaggregation of large-scale datasets," *Expert Systems with Applications*, no. ESwA-D-19-02372, 2019.
[36] J. Parra-Arnau, J. Domingo-Ferrer, and J. Soria-Comas, "Differentially private data publishing via cross-moment microaggregation," *Information Fusion*, 2019, in press.
[37] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From *t*-closeness-like privacy to postrandomization via information theory," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010. [Online]. Available: http://doi.org/10.1109/TKDE.2009.190
[38] D. Rebollo-Monedero, J. Forné, E. Pallarès, and J. Parra-Arnau, "A modification of the Lloyd algorithm for *k*-anonymous quantization," *Information Sciences*, vol. 222, pp. 185–202, Feb. 2013. [Online]. Available: http://doi.org/10.1016/j.ins.2012.08.022
[39] D. Rebollo-Monedero, J. Forné, and M. Soriano, "An algorithm for *k*-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers," *Data and Knowledge Engineering*, vol. 70, no. 10, pp. 892–921, Oct. 2011. [Online]. Available: http://doi.org/10.1016/j.datak.2011.06.005
[40] D. Rebollo-Monedero, J. Parra-Arnau, C. Díaz, and J. Forné, "On the measurement of privacy as an attacker's estimation error," *International Journal of Information Security*, vol. 12, no. 2, pp. 129–149, Apr. 2013. [Online]. Available: http://doi.org/10.1007/s10207-012-0182-5
[41] M. Rodriguez-Garcia, M. Batet, and D. Sánchez, "Utility-preserving privacy protection of nominal data sets via semantic rank swapping," *Information Fusion*, vol. 45, pp. 282–295, Jan. 2019.
[42] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "Does *k*-anonymous microaggregation affect machine-learned macrotrends?" *IEEE Access*, vol. 6, pp. 28 258–28 277, May 2018. [Online]. Available: http://doi.org/10.1109/ACCESS.2018.2834858
[43] ——, "The fast MDAV (F-MDAV) algorithm: An algorithm for *k*-anonymous microaggregation in big data," *Engineering Applications of Artificial Intelligence*, 2020.

[44] R. L. Rosnow and R. Rosenthal, "Statistical procedures and the justification of knowledge in psychological science," *American Psychologist*, vol. 44, no. 10, pp. 1276–1284, Oct. 1989.
[45] J. Salas and V. Torra, "A general algorithm for k-anonymity on dynamic databases," in *Data privacy management, cryptocurrencies and blockchain technology*, ser. Lecture Notes in Computer Science (LNCS), J. García-Alfaro, J. Herrera-Joancomartí, G. Livraga, and R. Ríos, Eds. Switzerland: Springer, 2018, vol. 11025, pp. 407–414.
[46] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.
[47] D. Sánchez, J. Domingo-Ferrer, S. Martínez, and J. Soria-Comas, "Utility-preserving differentially private data releases via individual ranking microaggregation," *Information Fusion*, vol. 30, pp. 1–14, 2016.
[48] D. Sánchez, S. Martínez, J. Domingo-Ferrer, J. Soria-Comas, and M. Batet, "μ-ANT: semantic microaggregation-based anonymization tool," *Bioinformatics*, vol. 36, no. 5, pp. 1652–1653, Mar. 2020.
[49] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, Jun. 2013.
[50] M. Schmid and H. Schneeweiss, "The effect of microaggregation procedures on the estimation of linear models: A simulation study," *Journal of Economics and Statistics*, vol. 225, no. 5, pp. 529–543, Sep. 2005.
[51] O. Siohan, "On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, MI, May 1995, pp. 125–128.
[52] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, , and S. Martínez, "Enhancing data utility in differential privacy via microaggregation-based k-anonymity," *Very Large Database (VLDB) Journal*, vol. 23, no. 5, pp. 771–794, 2014.
[53] X. Sun, H. Wang, J. Li, and T. M. Truta, "Enhanced p-sensitive k-anonymity models for privacy preserving data publishing," *Transactions on Data Privacy*, vol. 1, no. 2, pp. 53–66, 2008.
[54] X. Sun, H. Wang, J. Li, and Y. Zhang, "An approximate microaggregation approach for microdata protection," *Expert Systems with Applications*, vol. 39, no. 2, pp. 2211–2219, Feb. 2012.
[55] L. Sweeney, "Simple demographics often identify people uniquely," Carnegie Mellon University, Working Paper 3, 2000.
[56] ——, "Uniqueness of simple demographics in the U.S. population," Carnegie Mellon University, School of Computer Science, Data Privacy Lab, Pittsburgh, PA, Technical Report LIDAP-WP4, 2000.
[57] M. Templ, *Statistical disclosure control for microdata: Methods and applications in R.* Cham, Switzerland: Springer International Publishing, 2017.
[58] M. Templ, B. Meindl, A. Kowarik, and S. Chen, "Introduction to statistical disclosure control (SDC)," International Household Survey Network (IHSN), Working Paper 7, Aug. 2014. [Online]. Available: www.ihsn.org/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf
[59] S. Tiwari, H. M. Wee, and Y. Daryanto, "Big data analytics in supply chain management between 2010 and 2016: Insights to industries," *Computers and Industrial Engineering*, vol. 115, pp. 319–330, Jan. 2018.
[60] V. Torra and G. Navarro-Arribas, "Probabilistic metric spaces for privacy by design machine learning algorithms: Modeling database changes," in *Data privacy management, cryptocurrencies and blockchain technology*, ser. Lecture Notes in Computer Science (LNCS). Switzerland: Springer, 2018, vol. 11025, pp. 422–430.
[61] T. M. Truta and B. Vinay, "Privacy protection: p-Sensitive k-anonymity property," in *Proceedings of the International Workshop on Privacy Data Management (PDM)*, Atlanta, GA, Apr. 2006, p. 94.
[62] "UCI machine learning repository: Adult dataset." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Adult
[63] "UCI machine learning repository: Breast cancer Wisconsin (original) dataset," 1992. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)
[64] "UCI machine learning repository: Heart disease dataset," 1988. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease
[65] J. Vaidya, C. W. Clifton, and Y. M. Zhu, *Privacy preserving data mining.* New York, NY: Springer, 2006.
[66] R. Wei, H. Tian, and H. Shen, "Improving k-anonymity based privacy preservation for collaborative filtering," *Computers and Electrical Engineering*, vol. 67, pp. 509–519, Apr. 2018.
[67] A. N. K. Zaman, C. Obimbo, and R. A. Dara, "A novel differential privacy approach that enhances classification accuracy," in *Proceedings of the International C* Conference on Computer Science and Software Engineering (C3S2E)*, Porto, Portugal, Jul. 2016, pp. 79–84.