IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 5, 2018, accepted April 25, 2018, date of publication May 16, 2018, date of current version June 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2834858

Does k-Anonymous Microaggregation Affect Machine-Learned Macrotrends?

ANA RODRÍGUEZ-HOYOS[1,2], JOSÉ ESTRADA-JIMÉNEZ[1,2], DAVID REBOLLO-MONEDERO[2], JAVIER PARRA-ARNAU[3], AND JORDI FORNÉ[2]

[1]Departamento de Electrónica, Telecomunicaciones y Redes de Información, Escuela Politécnica Nacional, Ladrón de Guevara, E11-253 Quito, Ecuador
[2]Department of Telematics Engineering, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain
[3]CYBERCAT-Center for Cybersecurity Research of Catalonia, Department of Computer Science and Mathematics, Universitat Rovira i Virgili, E-43007 Tarragona, Spain

Corresponding author: David Rebollo-Monedero (david.rebollo@entel.upc.edu)

This work was supported in part by the Spanish Ministry of Economy and Competitiveness (MINECO) through the Anonymized Demographic Surveys Project under Grant TIN2014-58259-JIN, in part by the Proyectos de I + D + i para Jóvenes Investigadores Funding Program through the MAGOS Project under Grant TEC2017-84197-C4-3-R, the INRISCO Project under Grant TEC2014-54335-C4-1-R, and the Sec-MCloud Project under Grant TIN2016-80250-R, and in part by the European Commission through the H2020 CLARUS Project and the CIPSEC Project under Grant 700378.

ABSTRACT In the era of big data, the availability of massive amounts of information makes privacy protection more necessary than ever. Among a variety of anonymization mechanisms, microaggregation is a common approach to satisfy the popular requirement of k-anonymity in statistical databases. In essence, k-anonymous microaggregation aggregates quasi-identifiers to hide the identity of each data subject within a group of other $k - 1$ subjects. As any perturbative mechanism, however, anonymization comes at the cost of some information loss that may hinder the ulterior purpose of the released data, which very often is building machine-learning models for macrotrends analysis. To assess the impact of microaggregation on the utility of the anonymized data, it is necessary to evaluate the resulting accuracy of said models. In this paper, we address the problem of measuring the effect of k-anonymous microaggregation on the empirical utility of microdata. We quantify utility accordingly as the accuracy of classification models learned from microaggregated data, and evaluated over original test data. Our experiments indicate, with some consistency, that the impact of the de facto microaggregation standard (maximum distance to average vector) on the performance of machine-learning algorithms is often minor to negligible for a wide range of k for a variety of classification algorithms and data sets. Furthermore, experimental evidences suggest that the traditional measure of distortion in the community of microdata anonymization may be inappropriate for evaluating the utility of microaggregated data.

INDEX TERMS k-anonymity, microaggregation, machine learning, privacy, large-scale databases.

I. INTRODUCTION

With the advent of modern data-analytics technologies, the availability of massive amounts of information –the so-called big data era– has changed the landscape in a big manner: more data now means more useful data. The myriad of benefits these technologies can bring to private companies, public institutions and, in general, our society are innumerable. Healthcare, transportation, banking and marketing are just a few fields in which big-data analytics is leading a profound transformation of the traditional models [26], [37], [38].

However, the availability of sheer volumes of personal data and the growing sophistication of machine-learning analytics pose serious risks to individual privacy. Clearly, more data and better analytic capabilities increase the risk of reidentification of the individuals to whom a database records refer. The increasing demand for data-sharing among different data collectors, besides, only exacerbates this risk.

To cope with this, when a data set is to be shared, the privacy of the individuals therein must be taken into account very seriously. The purpose of *statistical disclosure control* (SDC) is precisely to ensure that only useful macrotrends are learned by the recipients of such data and individual privacy is therefore protected.

Microdata are database tables whose records carry data concerning individual subjects. The typical scenario in

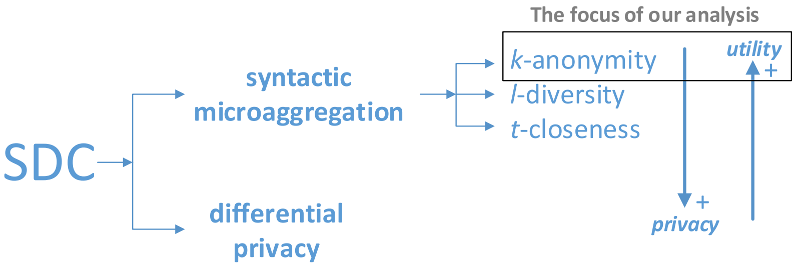


FIGURE 1. Our work focuses on high-utility SDC, involving k-anonymous microaggregation, which has a direct application, e.g., in the health domain.

microdata SDC is a data curator holding the original data set and perturbing the so-called quasi-identifier attributes (i.e., attributes that, in combination, may be linked with external information to reidentify individuals in the data set), so that disclosure risk is kept as low as possible. One of the most common strategies to keep this risk under control is the ''privacy first'' approach. Here, the data curator enforces a privacy model, which usually depends on a privacy parameter, to ensure an upper bound on the re-identification risk. Some of the best-known privacy models comprise k-anonymity [43], [48] and ε-differential privacy [18].

Although anonymization methods for microdata rely on a variety of mechanisms, the common denominator binding all them is *data perturbation*. Essentially, all such mechanisms modify the original data set to guarantee the chosen privacy model, inevitably at the cost of some loss in data utility [44]. If the resulting utility of the anonymized microdata does not satisfy the data curator's requirements, then a less stringent privacy parameter is applied or the privacy model is replaced. Examples of these mechanisms include microaggregation, suppression, generalization and noise addition.

When it comes to striking a balance between privacy and utility, appropriate and effective measures of the latter aspect are as important as the privacy model. In real practice, however, the standard measures of data utility employed by the SDC community may not capture the performance of a given data-analytics task, and therefore may not be useful in evaluating the anonymization mechanism in question. In SDC, the most widely used utility metric is the mean-squared error (MSE), while accuracy is the most popular one in machine learning.

A. CONTRIBUTION AND PLAN OF THIS PAPER

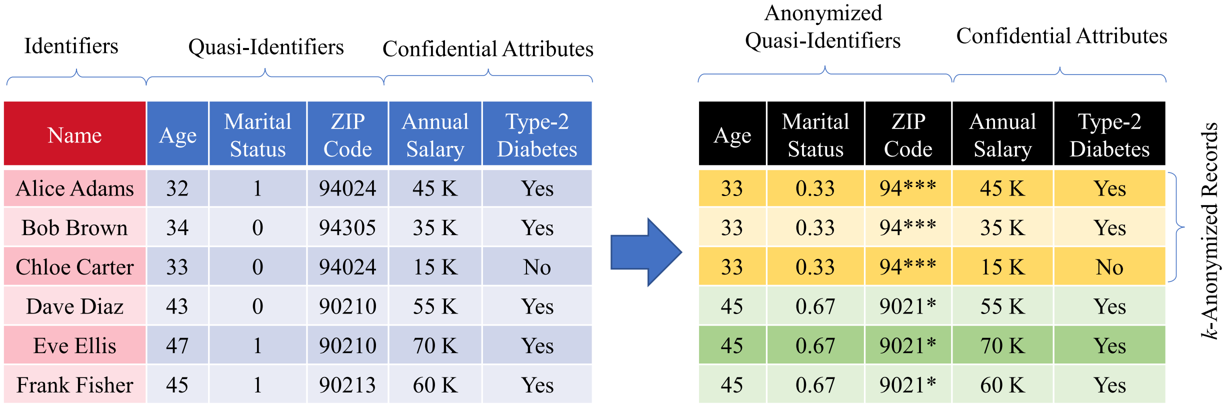
In the context of microdata anonymization, and particularly k-anonymous microaggregation, strong privacy protection requires masking the original data significantly, thus reducing their utility notably. However, the impact on utility caused by microaggregation is typically measured in terms of the syntactic distortion introduced to data, and not in terms of the performance of the intended data-analytic task. Besides, traditional utility metrics of k-anonymized data often neglect the statistical dependence of quasi-identifier attributes, and microaggregation algorithms tend to disregard the potential of confidential attributes as quasi-identifiers. All this makes it difficult to determine the effect of microaggregation on the practical utility of protected data.

The leading object of this paper is to investigate the impact on the performance of machine-learning tasks caused by data perturbation in the k-anonymous microaggregation process. To the best of our knowledge, the effect of standard k-anonymous microaggregation on the macrotrends learned from anonymized data has not been systematically studied, reported and discussed.

We apply a rigorous methodology for evaluating the specific impact of microaggregated data on machine-learning tasks. Our methodology uses accuracy and F-measure as utility metrics. The two are standard measures of performance in machine learning and allow for the statistical dependence among quasi-identifiers. The impact of microaggregation on the utility of anonymized data is quantified, accordingly, as the resulting accuracy (or F-measure) of a machine-learning model trained on a portion of microaggregated data and evaluated on a different portion of original data.

Since the utility extracted from data could depend on the learning algorithm used, we conduct an extensive, thorough evaluation of a wide range of machine-learning algorithms amply used in classification tasks. The results of utility we present correspond to the algorithms that obtain the greatest accuracy from each anonymized data set. Among others, our experiments investigate naïve Bayes, logistic regression, SVM, bagging and C4.5. As for microaggregation algorithms, we focus on MDAV, the SDC de facto standard for k-anonymous microaggregation. The evaluation of MDAV and all those machine-learning algorithms is conducted in four data sets, three real and well-known data sets and one synthetic. We would like to stress that our analysis focus on high-utility SDC, which involves plain k-anonymous microaggregation using *numerical* microdata. Although more strict privacy criteria exist, e.g., in the domain of syntactic microaggregation (such as t-closeness or l-diversity), or in the domain of semantic privacy (such as differential privacy), we examine those privacy mechanisms offering greater utility guarantees for anonymized data, which may be highly desirable in domains like health. We show the context of our analysis in Fig. 1.

The remainder of this paper is organized as follows. Section II reviews the background on k-anonymous microaggregation, reviews the state of the art in microaggregation


Identifiers
Quasi-Identifiers
Confidential Attributes
Anonymized Quasi-Identifiers
Confidential Attributes
k-Anonymized Records

Name	Age	Marital Status	ZIP Code	Annual Salary	Type-2 Diabetes
Alice Adams	32	1	94024	45 K	Yes
Bob Brown	34	0	94305	35 K	Yes
Chloe Carter	33	0	94024	15 K	No
Dave Diaz	43	0	90210	55 K	Yes
Eve Ellis	47	1	90210	70 K	Yes
Frank Fisher	45	1	90213	60 K	Yes

Age	Marital Status	ZIP Code	Annual Salary	Type-2 Diabetes
33	0.33	94***	45 K	Yes
33	0.33	94***	35 K	Yes
33	0.33	94***	15 K	No
45	0.67	9021*	55 K	Yes
45	0.67	9021*	70 K	Yes
45	0.67	9021*	60 K	Yes

FIGURE 2. Example of *k*-anonymous microaggregation of published data with $k = 3$. Quasi-identifiers in the left table are anonymized on the right.

algorithms for SDC, and explores previous work evaluating the impact on data utility caused by anonymization. Next, Section III describes our experimental methodology. Section IV shows the experimental results obtained for a variety of data sets and machine-learning algorithms. Lastly, conclusions are drawn in Section V.

II. BACKGROUND AND STATE OF THE ART

The old polarized landscape of data transmission with only intended and non-intended receivers has changed due to big data. Nowadays, data is more prone to be shared with external parties or even openly, e.g., for research purposes, in order to better exploit its utility. Thus, malicious observers may illegitimately take advantage, since sensitive information might still be encoded within released data. Unfortunately, conventional privacy services against unintended observers and based on cryptography (such as confidentiality), fail to address the practical dilemma when the intended recipient of the information is not fully trusted.

As a first approach to protect the anonymity of individuals, it is common to just eliminate their identifiers. However, this practice was proved to be insufficient in [48], where it was shown that 87% of the population in the United States could be unequivocally identified solely on the basis of the triple consisting of their date of birth, gender and 5-digit ZIP code, according to 1990 census data. Due to the discriminative potential of a few combined demographic attributes, more sophisticated approaches have been proposed to obscure the identity of the respondents appearing in the released data set.

A. BACKGROUND ON *k*-ANONYMOUS MICROAGGREGATION

Privacy protection techniques usually focus on databases carrying information concerning individual respondents (from a survey or a census). Said databases (also known as microdata sets) contain a set of attributes that may be classified into identifiers, quasi-identifiers and confidential attributes. Firstly, *identifiers*, such as full names or medical record numbers, can single out individuals from a data set and are commonly removed in order to preserve the anonymity of respondents. Secondly, *quasi-identifiers* or *key attributes* may include age, gender, address, or physical features, which combined and linked with other external information can be used to reidentify respondents. Finally, a data set may contain confidential attributes with sensitive information on the respondents, such as salary, health condition, and religion.

In Fig. 2, we illustrate how a perturbed, and thus more private, version of a data set is obtained to be published instead of the original one. In the figure, the original data set combines attributes common in census and medical surveys. It has three quasi-identifiers, age, marital status and ZIP code, and two confidential attributes, annual salary and type-2 diabetes condition. The figure at hand shows how, in order to preserve the privacy of respondents, perturbation is applied to quasi-identifiers. This technique, called *microaggregation*, is applied to enforce *k*-anonymity [48], a privacy model that guarantees that each tuple of key-attribute values is identically shared by at least *k* records in a data set. Rather than making the original table available, a perturbed version is published where aggregated records of quasi-identifying values are replaced by a common representative tuple. The result is a microaggregated data set that may prevent reidentification attacks.

As illustrated in Fig. 3, if tuples of key attributes in a data set could be represented as points in the Euclidean space, *k*-anonymous microaggregation would consist in partitioning these points in cells of size *k*, and quantizing each cell and its elements with a representative point. Perturbed key attributes would be characterized by the set of representative points. The de facto standard for numerical microaggregation is the maximum distance to average vector algorithm (MDAV). It was proposed in [20] as a practical evolution of a multivariate fixed-size microaggregation method and conceived in [12]. We provide, in Algorithm 1, a simplified version of that given in [15] and termed ''MDAV generic''.

B. STATE OF THE ART OF *k*-ANONYMITY AND *k*-ANONYMOUS MICROAGGREGATION

Microaggregation is a technique aimed to protect the privacy of those individuals whose personal records are included in

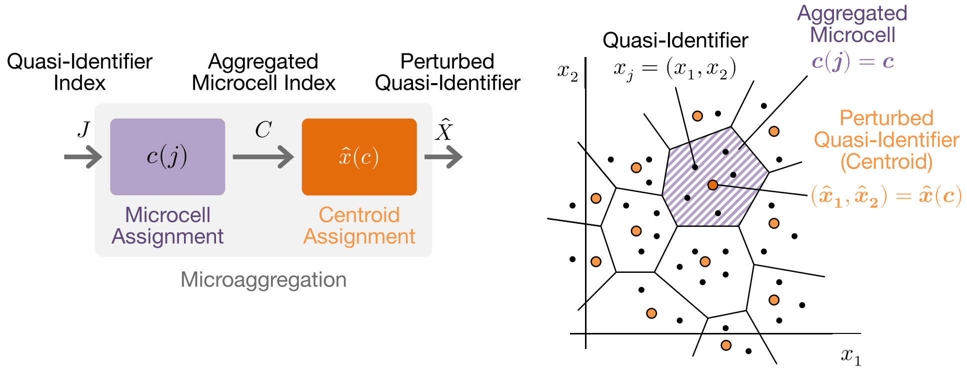


FIGURE 3. *k*-Anonymous microaggregation as a minimum-distortion quantizer design problem with a constraint on the size of the quantizer cells [41].

Algorithm 1 MDAV "Generic", Functionally Equivalent to [15, Algorithm 5.1]

```
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 $j \mapsto q(j)$
while $2k$ points or more in the data set remain to be assigned to microcells do
    find the centroid (average) $C$ of those remaining points
    find the furthest point $P$ from the centroid $C$, and the furthest point $Q$ from $P$
    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 $Q$
    remove the two microcells just formed from the data set
if there are $k$ to $2k-1$ points left then
    form a microcell with those and finish
else                               ▷At most $k-1$ points left, not enough for a new microcell
    adjoin any remaining points to the last microcell        ▷Typically nearest microcell
```

a released microdata set. With microaggregation, as with generalization and suppression, the distortion is applied to the key attributes to satisfy the *k*-anonymity privacy model [43], [48]. This model guarantees that each individual's information contained in a released data set cannot be distinguished from that of at least $k-1$ individuals whose information also appears in the data set. The original formulation of *k*-anonymity as a privacy criterion was modified into the microaggregation-based approach in [9], [12], [13], and [15].

Although *k*-anonymity is a very popular privacy criterion, it is not flawless. Since the criterion strictly operates with the key attributes, the statistical properties of confidential attributes (and thus their disclosure potential), both in the data set and in the entire population, are neglected. In general, *k*-anonymity overlooks the knowledge a potential attacker may already have or obtain about the data set, giving rise to similarity, skewness or background-knowledge attacks [16], [40], [42]. In spite of its shortcomings, the application of *k*-anonymous microaggregation does not only concern the publication of databases but also some variants thereof like search engine querying, online data collection and data streaming [5], [14], [54].

Additional criteria have been proposed that refine *k*-anonymity and prevent some of the above-mentioned attacks. The former, *p*-sensitive [47], [49], requires that each group of *k*-anonymized records contains at least *p* different values of each confidential attribute. In the same but broader spirit, *l*-diversity proposes that each group have at least *l* well-represented confidential values. None of these criteria assures complete protection against skewness attacks, nor against similarity attacks when confidential attributes within a group are semantically similar.

Other privacy criteria dealing with similarity and skewness attacks pose requirements in the distribution of confidential attributes within groups. The aim is that confidential attributes in each group of anonymized records are stratified according to their distribution in the original data set. Depending on the discrepancy allowed between the within-cluster and overall distributions, these privacy criteria yield *t*-closeness [27], δ-disclosure [4], and average privacy risk [39], [40].

To cope with the NP-hardness of multivariate microaggregation, several heuristic algorithms have been proposed. These algorithms can be classified as fixed-size and variable-size. Among the former ones, we find the maximum distance [12] (MD) and its variation, maximum distance to average vector [12], [15] (MDAV). Variable-size algorithms include, on the other hand, the μ-Approx [13], the minimum spanning tree [24] (MST), the variable MDAV [46] (V-MDAV) and the two-fixed reference points algorithms (TFRP).

In general, the implementations of microaggregation have been oriented to reduce the inherent information loss [10], [29], [35] due to perturbation, which commonly derives in more sophisticated and significantly costlier implementations in terms of computational time [41].

C. MEASURING THE IMPACT OF MICROAGGREGATION ON MACHINE-LEARNED MACROTRENDS

Usually, the impact of microaggregation is measured in terms of the distortion introduced in the data. MSE is commonly used to quantify such *distortion* in the case of numerical attributes. Equivalently, the *utility* of microaggregated data is measured inversely as the distortion resulting from data perturbation.

Within a broader scope, in an effort to tailor anonymization mechanisms to the application domain of data (e.g., building classifiers to predict someone's health condition), some previous research work has also used other utility metrics. One of such metrics is the accuracy of machine-learned macrotrends built using anonymized data. A model built with perturbed data would be less accurate than another built with original data. Accordingly, a higher degree of anonymization would result in less accurate models. Surprisingly, to the best of our knowledge, this metric has not been used to systematically evaluate microaggregation-based anonymization algorithms, but other anonymization algorithms based on generalization and suppression of records, such as Incognito, Mondrian and DataFly.

In previous work, classification accuracy has been used to evaluate the utility of (or, equivalently, the distortion introduced to) anonymized data, just to compare the performance of adapted classifiers or anonymization mechanisms. One of these works is [45], where the effects of four microaggregation algorithms on the estimation of a linear regression is compared, when solely applied to simulated data sets. Other works propose improvements on machine learning algorithms and methodologies, to obtain higher utility (classification accuracy) from anonymized data. This is the case of [21], where the authors develop a method to increase the level of utility obtained from support vector machine (SVM) and k-nearest neighbor (kNN) machine learning

algorithms, when data are anonymized with the DataFly algorithm. By feeding these algorithms with statistics from original data, in addition to anonymized data, greater utility ensues from the latter. In the same line, [6] describes an adjustment to logistic regression that provides differential privacy [18]. Furthermore, decision tree learning methods are developed in [34] and [53] that enforce l-diversity and differential privacy, respectively, as privacy criteria and whose accuracy levels approach those of a non-private decision tree. Using a different focus, [30] and [31] address the privacy risk resulting from the release of SVM and the anonymized data. Privacy preserving versions of SVM are proposed and their classification accuracies are used to compare them with the original SVM.

A great deal of research has also investigated adaptations of anonymization algorithms that generate private data of "higher quality". In that context, the utility of anonymized data is evaluated in terms of classification accuracy of machine learning models [22], [23], [25]. The cited works rely on generalization and suppression as perturbation techniques and include preprocessing steps such as selective anonymization of attributes, to adapt the released data to machine learning applications, and hence preserve their utility. On the other hand, [7] proposes publishing synthetic microdata generated from differentially private models applied on original data. For that, machine learning techniques are integrated to improve utility.

Ironically, although enhancements in the utility of anonymized data are reported, it is not clear what the overall impact of original anonymizing mechanisms in the first place is. Some approaches do attempt to evaluate the tradeoff between privacy gain and information loss (measured as accuracy reduction) due to anonymization. However, various considerations should be done for such evaluation. To start, there is a variety of anonymization algorithms. For example, [19] focuses on a proprietary anonymization algorithm whereas [33] examines a non-standard one.

Other caveat is the variable application domain of the data. While classification is the most popular workload for anonymized data, machine learning algorithms would perform differently depending on the particular data set used, so the utility would vary accordingly. This also applies to the number of records, or the size of the data set, which may affect the performance of anonymization algorithms, e.g., when k-anonymity is applied, a given value of k shall affect the utility of small data sets more than the utility of bigger ones.

A last limitation has to do with the baselines to measure privacy gain and utility loss. Utility, measured as the accuracy of machine learning models, reaches its lower bound when all the key attributes are discarded; or, for k-anonymity, when k equals the number of records of the data set. Utility's upper bound is attained when no anonymization is applied[1]

Even in this variable scenario, one thing is certain about how machine-learned trends are affected by anonymization: simultaneously satisfying various privacy criteria, e.g., k-anonymity, l-diversity, and t-closeness, may make the data completely useless, as reported by [4], a study where not only syntactic but also semantic requirements of privacy are evaluated. Those privacy criteria, together with differential privacy, are out of the scope of this work, since our

[1] Further considerations regarding baseline performance can be found in [28].

TABLE 1. Summary of related contributions.

Reference	Anonymization algorithm	Type of attributes used	Application domain	Max size of data sets	Max value of k	Main focus
Inan et al, 2009 [21]	DataFly	Hybrid	Classification	5,000	128	Comparing classifiers on anonymized data
LeFevre et al, 2006 [25]	Mondrian, TDS	Hybrid	Classification	49,657	1,000	Algorithms to anonymize data while preserving utility
Chaudhuri and Monteleoni, 2008 [6]	Differential Privacy	Numeric	Classification	N/A	N/A	Improving ML algorithm to work with anonymized data
Lin and Chen, 2010 [31]	DataFly	Numeric	Classification	270-49,990	128	Improving ML algorithm to work with anonymized data
Kisilevich et al, 2010 [23]	kACTUS, TDS, TDR, Mondrian, kADET	Hybrid	Classification	42,244	1,000	Building an algorithm to protect privacy in classification tasks (comparing accuracy with others)
Jaffer et al, 2014 [22]	Mondrian	Hybrid	Classification	1,000	50	Building an algorithm to protect privacy in classification tasks (comparing accuracy with others)
Malle et al, 2016 [33]	SaNGreeA	Hybrid	Classification	42,244	19	Showing the destructive effect of an anonymization algorithm on classification tasks
Gursoy et al, 2017 [19]	k-Map	Hybrid	Classification	42,244	5	Evaluating an anonymization algorithm based on differential privacy
Brickell and V Shmatikov, 2008 [4]	Mondrian	Hybrid	Classification	42,244	1,000	A methodology to measure the tradeoff between loss of privacy and gain of utility

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, introductorily explained also in [36], render it useless for our purposes.

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 [8] and [17]. Table 1 summarizes the main conclusions of this section.

III. METHODOLOGY OF EVALUATION

Attack and Usability Model In this work, we assume the standard attack model of the SDC literature [32]. When a microdata set is released, it is assumed to be available to any privacy attacker. For research and statistical purposes, the released microdata contains key attributes (basically demographic data) that are correlated with another, probably confidential, attribute. In the k-anonymity model, besides, the attacker knows a target individual's record –although microaggregated– is in the released data set.

To protect that individual's privacy, an anonymized version of the microdata set is released. To keep the information usable, i.e., "truthful" [43], microaggregation is applied to the key attributes, while the confidential attribute is unperturbed. Researchers may leverage the key attributes by building classifiers on the microaggregated data, for example to predict a given condition. Recall that classification is a machine learning task that aims to predict the class, or label, of a tuple of information. To do so, it requires learning a model from a group of labeled input samples. In our case, we can assume a large anonymized data set of patients that is publicly released so that researchers can build classifiers.

As another example of this model, suppose that the taxation authority publishes a microaggregated data set with 3 key attributes: gender, age, and marital status; additionally, a confidential binary attribute is published without being modified, specifying whether a respondent has paid taxes or not. Both perturbed key attributes and the confidential attribute could be used by researchers to develop algorithms that predict the propensity of other people to pay taxes. At the same time, the privacy of a specific individual would be preserved as a result of microaggregation. However, as commented in previous sections, the macrotrends embedded in the original data, which are necessary to get more accurate classifiers, might be affected by the perturbation of the key attributes values caused by microaggregation.

A. MEASURING PRIVACY AND UTILITY

To evaluate the impact of anonymization on the utility of a released microdata set, we need quantifiable metrics of privacy and utility. Since our experiments focus on microaggregation as anonymization mechanism, we shall assume k-anonymity as privacy criterion. In this manner, the identity of a respondent will be protected in a group of k tuples sharing the same key attribute values. Higher values of k will imply

TABLE 2. Description of the data sets used to evaluate the impact of *k*-anonymous microaggregation.

Data set	# of instances	# of attributes	Selected key attributes	Confidential (label) attribute
Synthetic	30,000	2	x_1, x_2	y
Adult [50]	45,222	15	Age, education-num, marital-status, sex, capital-gain, hours-per-week	Salary (>50K?)
Pima Indians Diabetes [51]	768	9	Number of pregnancies, glucose concentration, blood pressure, triceps skin fold thickness, serum insulin, body mass index, diabetes pedigree function, age	Health condition (diabetes?)
Irish Census [3]	100,000	10	Gender, age, marital status, highest education completed	Economic status (employed?)

TABLE 3. Machine learning algorithms used in our experimental evaluation.

Data set	ML algorithm used		ML algorithm description
	Type	Name	
Synthetic	Classification tree	C4.5	It builds decision trees from training data, where attribute nodes are selected based on their information gain (mutual information).
Adult [50]	Ensemble	Bagging	Bootstrap aggregation is an ensemble of decision trees that improves classification tasks by combining the classification results of randomly (bootstrap) generated training data sets obtained from the original data set.
Pima Indians [51]	Regression	Logistic Regression	It is a regression model that probabilistically estimates a binary response (binary classification) based on a set of predictors. It is based on the logistic or sigmoid function.
Irish Census [3]	Classification tree	C4.5	It builds decision trees from training data, where attribute nodes are selected based on their information gain (mutual information).

more anonymity and then more privacy, although, eventually, less utility.

To measure the utility of anonymized data, we must decide the application domain of such data. We choose *binary classification* since it is a very popular workload for released microdata sets. Accordingly, we measure utility through the performance of a binary classifier, when executed on anonymized data. Several metrics exist that measure the performance of binary classification tests. Next, we elaborate on them with a medical example.

Let D be a binary random variable (r. v.) representing whether a patient has a given condition ($D = 1$) or not ($D = 0$). Let T be a binary r. v. representing the outcome of a medical test, being $T = 1$ a positive detection, and $T = 0$ a negative detection. By the law of total probability,

$$\mathrm{P}\{T = D\} = \mathrm{P}\{T = D \mid D = 0\}\, \mathrm{P}\{D = 0\} + \mathrm{P}\{T = D \mid D = 1\}\, \mathrm{P}\{D = 1\},$$

and thus,

$$\mathrm{P}\{T = D\} = \mathrm{P}\{T = 0 \mid D = 0\}\, \mathrm{P}\{D = 0\} + \mathrm{P}\{T = 1 \mid D = 1\}\, \mathrm{P}\{D = 1\}.$$

Specificity (true negative rate) and sensitivity (true positive rate) are two metrics of the performance of a binary classifier and can be defined as $\mathrm{P}\{T = 0|D = 0\}$ and $\mathrm{P}\{T = 1\,|D = 1\}$, respectively. In our evaluation, we follow the same approach as [23], [25], and [33] and measure utility as the *accuracy* of a binary classifier. In our example, accuracy can be defined as the probability that the test and disease coincide, that is $\mathcal{A} = \mathrm{P}\{T = D\}$. Accuracy can also be expressed in terms of specificity and sensitivity as the convex combination

$$\mathcal{A} = (1 - \text{prevalence}) \times \text{specificity} + \text{prevalence} \times \text{sensitivity}$$

weighted by the prevalence, that is, the a priori probability of having a disease.

Although accuracy is a very popular metric, when the class of the data is significantly unbalanced this metric might incorrectly measure the goodness of a classifier. Fortunately, other stricter indicators are available such as F-measure, ROC curve and area under the ROC curve (AuC).

Accuracy quantifies how well a binary classifier performs, in terms of the rate of correctly classified (as positive or negative) samples in a test set. For example, a binary classifier constructed to predict diabetes would be 100% accurate if, when applied on a test set of 600 samples, it correctly identifies the class of the 500 samples labeled with ''no diabetes'' and the class of the 100 samples labeled as ''diabetes''.

F-Measure (or F_1 score) is a machine learning metric that combines other metrics, particularly recall and precision.

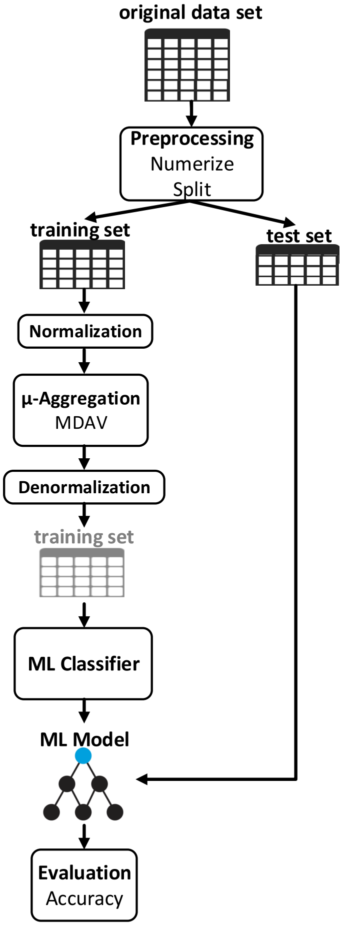


FIGURE 4. Experimental methodology followed to evaluate the impact of MDAV-based *k*-anonymous microaggregation on the empirical utility of microdata.

In fact, F-Measure is defined as the harmonic mean of precision and recall. Furthermore, another composed metric is the ROC curve, which measures the performance of a classifier based on the graphical representation of the sensitivity in function of the specificity.

For our application domain (binary classification), we first measure the utility of a microdata set before being microaggregated. Since no perturbation is applied to the data, the classifier built from that data set would yield the highest accuracy. The data would therefore give the best achievable utility, but the worst privacy.

In our experiments, we shall generate several microaggregated versions of a data set, by varying the value of the privacy parameter k incrementally for a wide range. For each of these versions, we shall compute the corresponding classification performance to observe the progressive degradation of data utility due to microaggregation. We use accuracy and F-measure to assess the performance of classifiers built with microaggregated data. Naturally, as k increases, we expect a lower data utility, but obviously in exchange for higher privacy. Note that, for binary classifiers computed over a set of data samples and their corresponding labels, the lowest possible accuracy is not zero. To see this, suppose that "positive" is the majority class (more than 50% of the training samples are labeled as "positive"). Accordingly, the simplest classifier would classify any new input as "positive". Then, interestingly, a binary classifier should not have accuracy values lower than 50%.

B. EXPERIMENTAL SETUP

Next, we describe the algorithms, tools and data we use to quantify the impact of k-anonymous microaggregation on the performance of machine-learned classifiers.

1) ALGORITHMS

With regard to microaggregation, our experiments employ MDAV [11], the de facto standard protocol described in Section I. The specification of the MDAV algorithm used here can be found in [15]. The algorithm in question is referred in the cited work as "MDAV-generic".

With the aim of constructing classifiers from microdata, we use the *Weka toolkit* [52], a collection of algorithms extensively employed by the machine learning community. In the interest of fairness when measuring the impact of microaggregation, we assign each data set the machine learning algorithm that extracts the greatest utility from it. Accordingly, we measure said impact with respect to the highest achievable utility. In order to find the corresponding algorithm for a data set, we tried on it a range of machine learning algorithms, including naïve Bayes, logistic regression, SVM, bagging, and C4.5. The reasons for choosing this set is manifold. First, we include different algorithms to observe whether the effects of microaggregation are consistent along different utility extraction techniques. Moreover, we select naïve Bayes and SVM since in several previous works [7], [23], [25], [33] they have been adapted to obtain more utility from anonymized data. Additionally, logistic regression, C4.5 and bagging were considered to represent the main families of machine learning classifiers, i.e., regression, decision tree, and ensemble algorithms, respectively. For each data set, we choose the algorithm showing the best performance in the classification task, i.e., the highest accuracy. This way, we test the impact of microaggregation in the different utility contexts or domains defined by a variety of data sets and machine learning algorithms.

2) DATA

For the purpose of illustration, we shall evaluate the impact of microaggregation first on a synthetic data set. The effect of microaggregation on real scenarios will be assessed afterwards in data sets satisfying these four properties. First, we require data sets containing demographic attributes so that they reflect the typical characteristics of microdata. Secondly, we consider only data sets whose potential key attributes are correlated with a given sensitive (label) attribute, so the latter could be effectively predicted (classified). Thirdly, we need a relatively large number of records (e.g., more than 500) to have a better view of the overall effect of

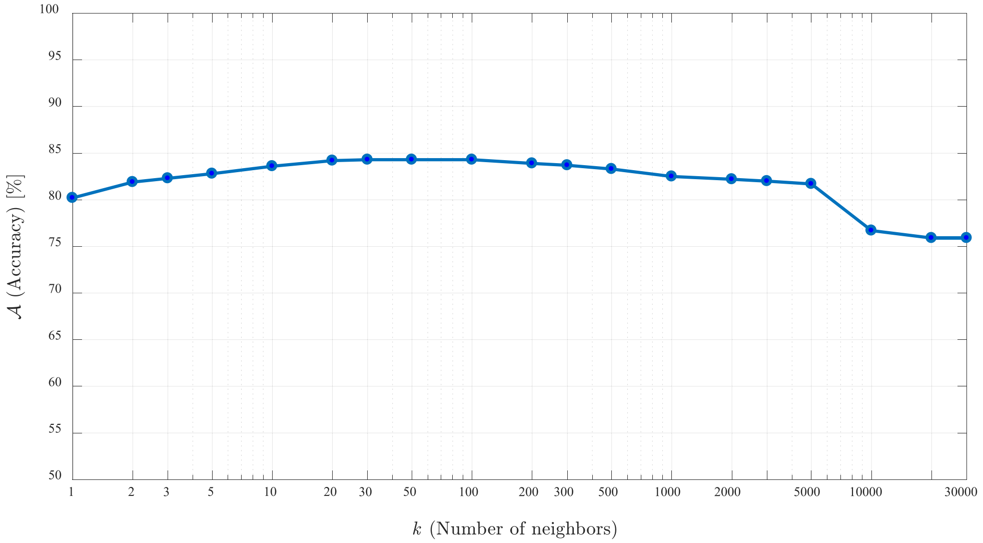


FIGURE 5. Accuracy of the *k*NN machine learning algorithm applied on the UCI Adult data set, for different values of *k* (here, *k* is not related with *k*-anonymity).

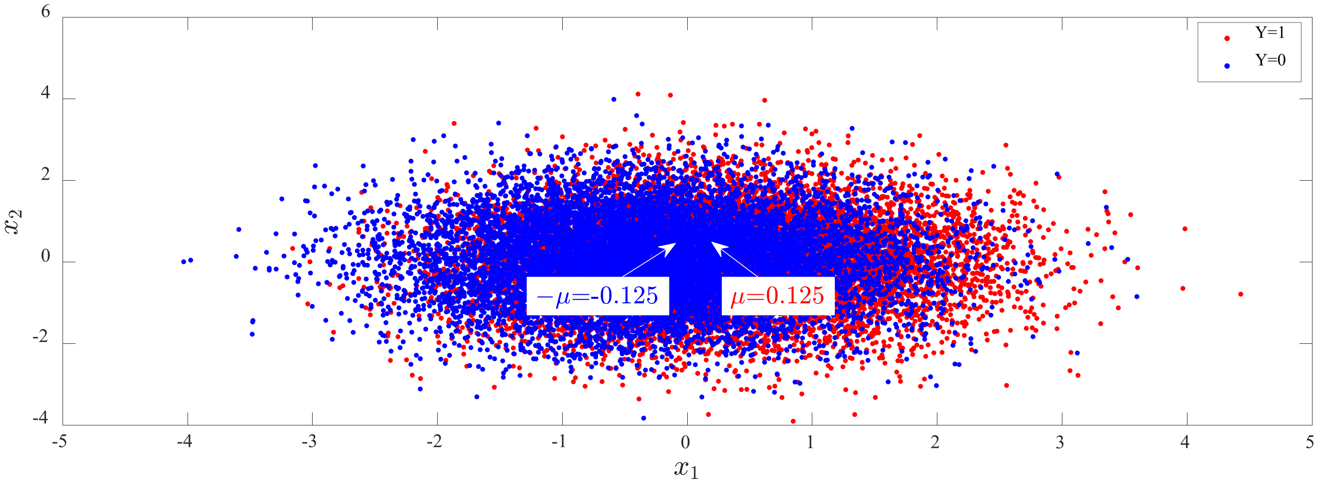


FIGURE 6. Depiction of the quasi-identifiers (x_2 *vs* x_1) of our synthetic data set. Samples are colored according to their class, *y*; blue for $y = 0$ and red for $y = 1$.

microaggregation, using an incremental value of the privacy parameter k. Finally, we use standardized or already tested data sets so that our results can be easily reproduced. It is worth noting that predictive demographic data turned out to be a very restrictive condition when we searched for data sets to carry out the tests.

For the sake of simplicity and ease in its graphic representation, we build the synthetic data set with only two numerical attributes (x_1, x_2) resembling quasi-identifiers, and a binary attribute (y) as the confidential attribute. The data set is generated so that y is to some extent predictable from x_1 and x_2 and has 30,000 records. In Section IV-B, we describe in greater detail the process by which the synthetic data set was generated and show a preliminary experiment to illustrate the effects of microaggregation.

Regarding the experiments on real data sets, we first employ the standardized "Adult" data set from the UCI Machine Learning Repository [50], described in Table 2. The data set in question has been widely used to evaluate binary classifiers and privacy preserving mechanisms. Its 45,222 records are already split into two parts, for training (2/3) and testing (1/3) purposes. The data set contains 15 input demographic attributes and a binary label attribute, the salary, which is the attribute the machine learning algorithm will try to predict. In particular, the attribute specifies whether a person makes over 50K a year or not. The attributes we use as quasi-identifiers are age, education-num, marital-status, sex, capital-gain, and hours-per-week.

The second standardized data set is "Pima Indians Diabetes" [51] which contains 768 records and 9 demographic attributes. Available at the UCI Machine Learning Repository, this data set has been used in [22], [23], and [31]. The 8 key attributes we selected allow predicting whether an individual will be diagnosed with diabetes or not. The

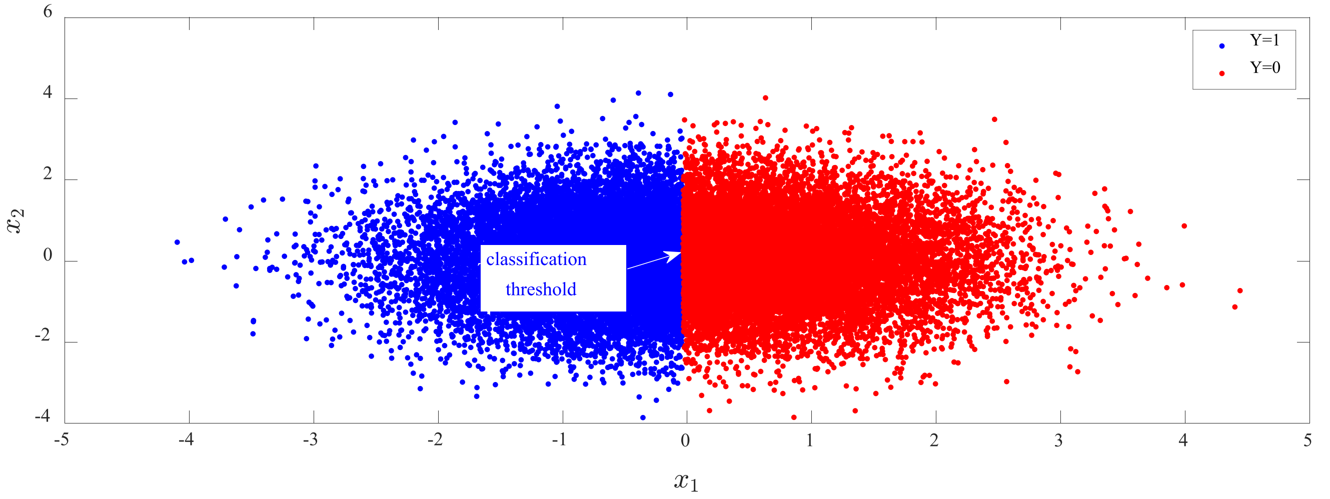


FIGURE 7. Samples of our our synthetic data set, colored according to their *predicted* class, $\bar{y}$; blue for $\bar{y} = 0$ and red for $\bar{y} = 1$.

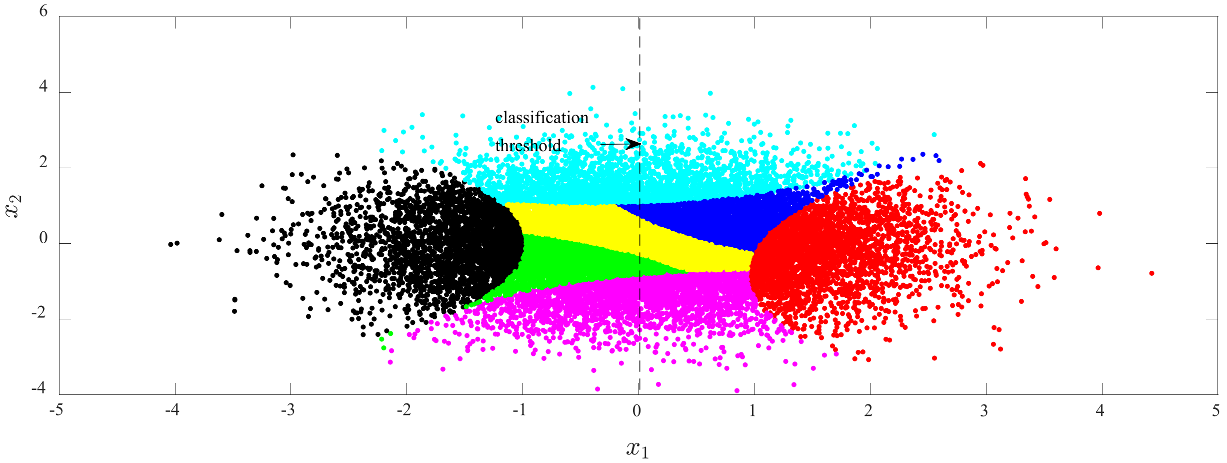


FIGURE 8. Cells of samples obtained after k-anonymous microaggregation with MDAV on the quasi-identifiers of our synthetic data set ($k = 3000$).

third real data set we consider in our experiments is the "Irish Census" [3], a synthetic version of the data from the 2011 Irish Census, which has been used in [1] and [2] to evaluate and compare k-anonymization algorithms. It contains 100,000 records and 10 demographic attributes. Originally, it was not built with a predictive task in mind, but 5 of its attributes could be used to predict an individual's economic status (employed or unemployed).

Table 2 describes the main characteristics of the data sets tested in our experiments, and Table 3 shows the machine learning algorithms employed for each data set.

3) ADDITIONAL TASKS

Since our implementation of MDAV only operates with numerical attributes, we conducted some preprocessing tasks on the data sets described in the previous subsection. Specifically, we converted some useful categorical attributes to numeric, where possible, and binarized the sensitive attribute, where necessary, so that the application domain of data was binary classification.

C. EXPERIMENTAL METHODOLOGY

The steps we follow to evaluate the impact of microaggregation on the utility of microdata are in line with the attack and utility models described at the beginning of Section III and are illustrated in Fig. 4. As a first data preprocessing step, we extract the key attribute information of our interest from each data set, according to the guidelines described in the previous subsection. Moreover, from the selected key attributes, we "numerize" the categorical data so that they are compatible with MDAV. Finally, we identify the key attributes that are then used as input samples and the sensitive label attribute that will serve as the class to be predicted by the classification model.

The next step splits each microdata set into training and test sets. As is common in the evaluation of machine learning

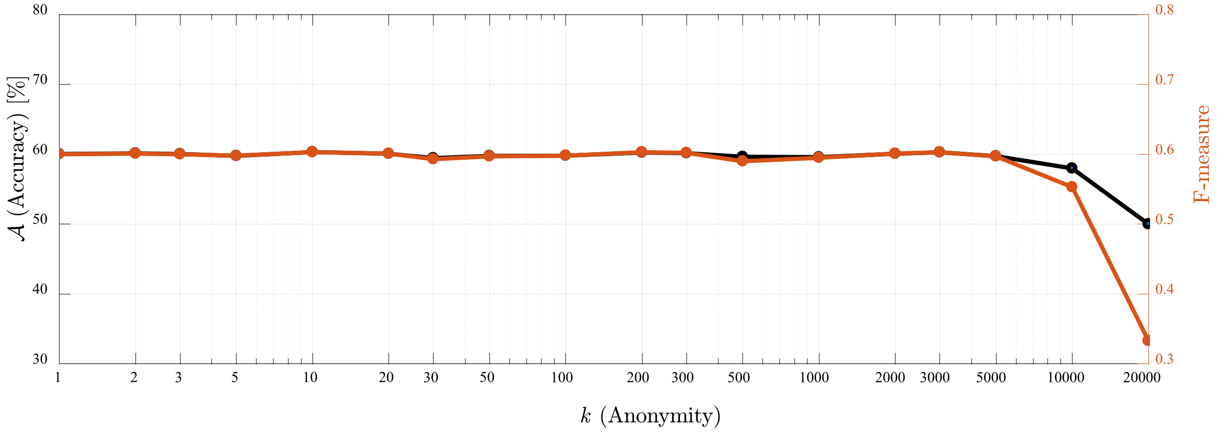


FIGURE 9. Degradation of the empirical utility (accuracy and F-measure) of our synthetic data set when microaggregated (using MDAV) for a wide range of k.

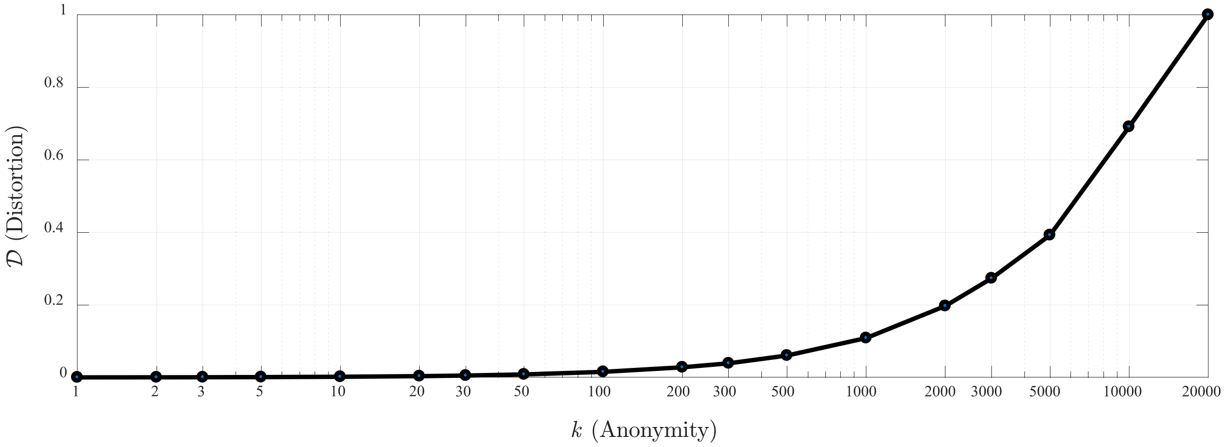


FIGURE 10. Distortion, measured as MSE, introduced by MDAV k-anonymous microaggregation to our synthetic data set, considering a wide range of k.

algorithms, a model is constructed from a training subset of the data and is evaluated on the test subset. Following such methodology, we use two-thirds of the data for training and one-third for testing. The splitting is done in such a way that the class attribute is stratified in each subset, according to its original distribution in all the data set.

After splitting the data into training and test sets, the microaggregation process is performed using MDAV over the latter set. To this end, previously we follow the common practice of normalizing each column of the data to have zero mean and unit variance.

With the microaggregated versions of each (training) data set, we then construct a classification model over each of those versions using Weka and 10-fold cross validation. The learning algorithms we use for each data set are listed in Table 3. Finally, we evaluate the accuracy of the resulting classification models over the non-anonymized test subset, reproducing the application scenario where a database user would use the classification model to classify their original samples of data.

IV. EXPERIMENTAL RESULTS

Preliminary Experiment To get some intuition about the impact of microaggregation and its clustering capability on the empirical utility of anonymized data, we next make an analogy with the operation of some machine learning algorithms. Consider the k-nearest neighbors algorithm (kNN), a simple classifier, and assume a data set with n training tuples, each one assigned to a binary class label. kNN classifies a new tuple according to a majority vote of its k closest ‘‘neighboring training tuples’’ in the feature space. Note that, in this context, k has nothing to do with anonymity. A small k implies considering few neighboring samples for classification, which would be the most representative ones, being the closest, but would not be so reliable in terms of predictability. On the other hand, a large k implies taking more (and not so close) neighboring samples, being demographically less representative, but predictably more reliable. This tradeoff is illustrated in Fig. 5, where we measure the accuracy of kNN on the original UCI Adult data set for several values of k. As depicted in Fig. 5, the classification accuracy of

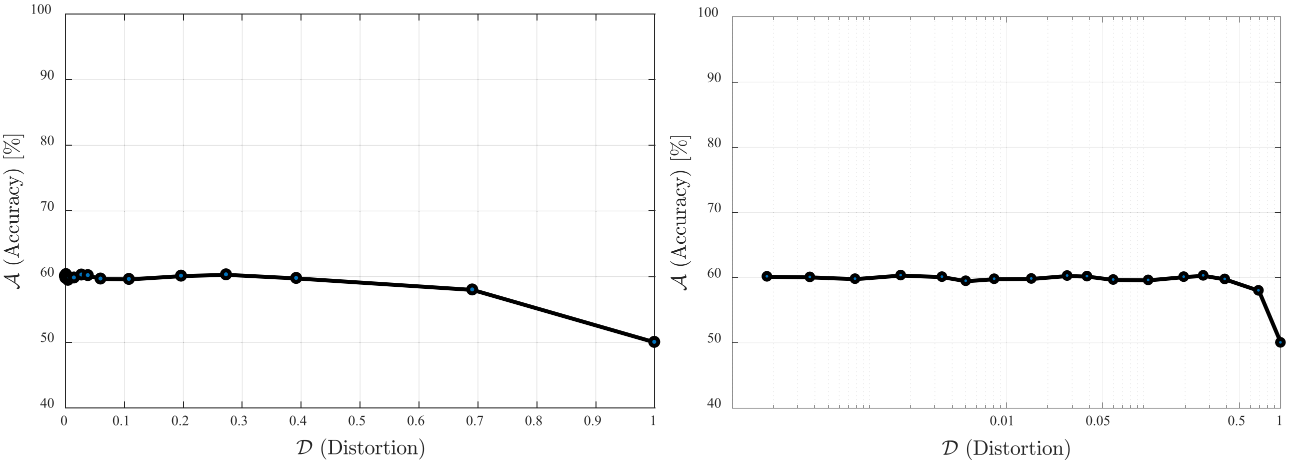


FIGURE 11. Accuracy of the *bagging* machine learning model trained on our microaggregated synthetic data set, against the distortion due to MDAV.

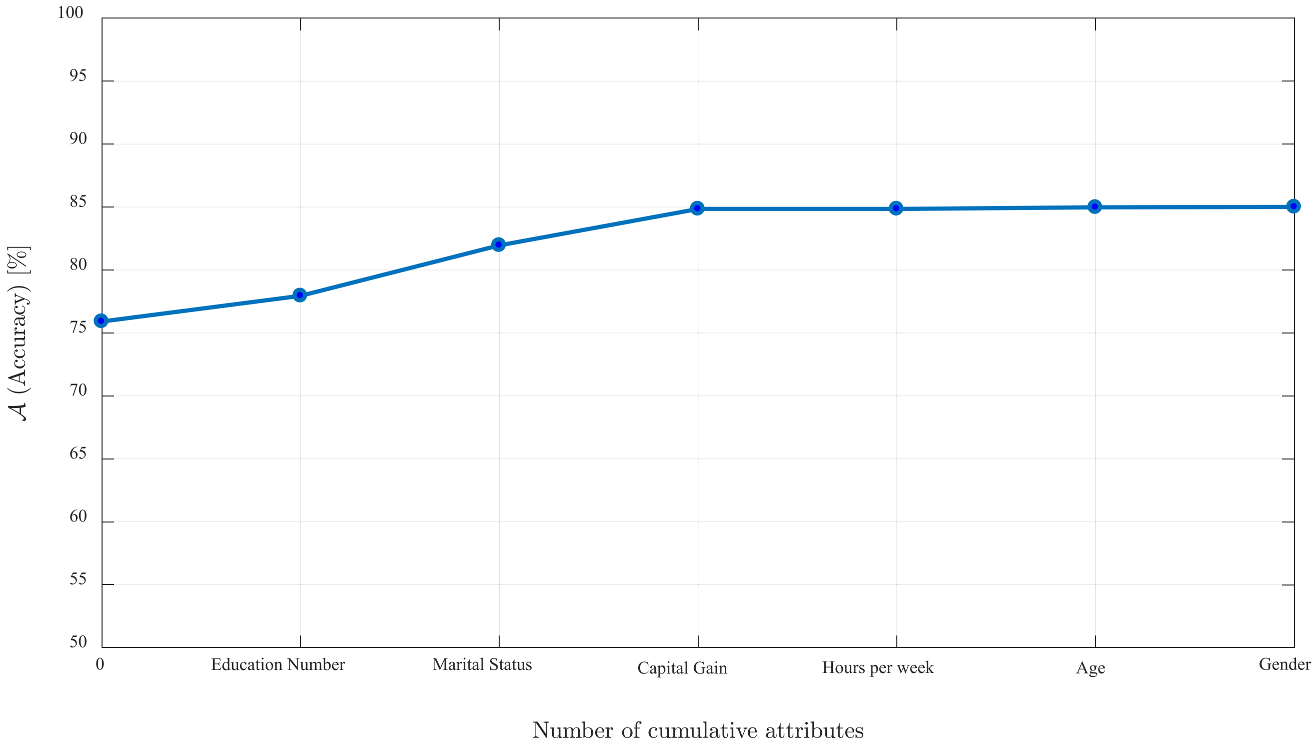


FIGURE 12. Relevance of the cumulative number of selected attributes from the UCI Adult data set as predictors of the class attribute (Annual Salary).

kNN improves as groups rather than individual samples are considered to robustly infer what would effectively constitute a macrotrend.

We argue that microaggregation would be acting analogously to kNN when aggregating neighboring data points to construct cells, and computing averages to get representative centroids for each cluster. Such clustering could be regarded as a denoising process. In fact, the benefit of preprocessing data with unsupervised techniques based on clustering, prior to supervised learning, is known in the machine-learning literature. Therefore, it seems reasonable to expect k-anonymous microaggregation to have a minor (and sometimes even positive) impact on the empirical utility of data, measured as the accuracy of machine learning models when deriving macrotrends.

A. MEASURING THE IMPACT OF MICROAGGREGATION ON A SYNTHETIC DATA SET

We begin our experiments by analyzing the effect of microaggregation on synthetic data. To this end, we generate 30,000 samples of 3-dimensional Gaussian data. The first two dimensions are assumed to be quasi-identifiers, and the

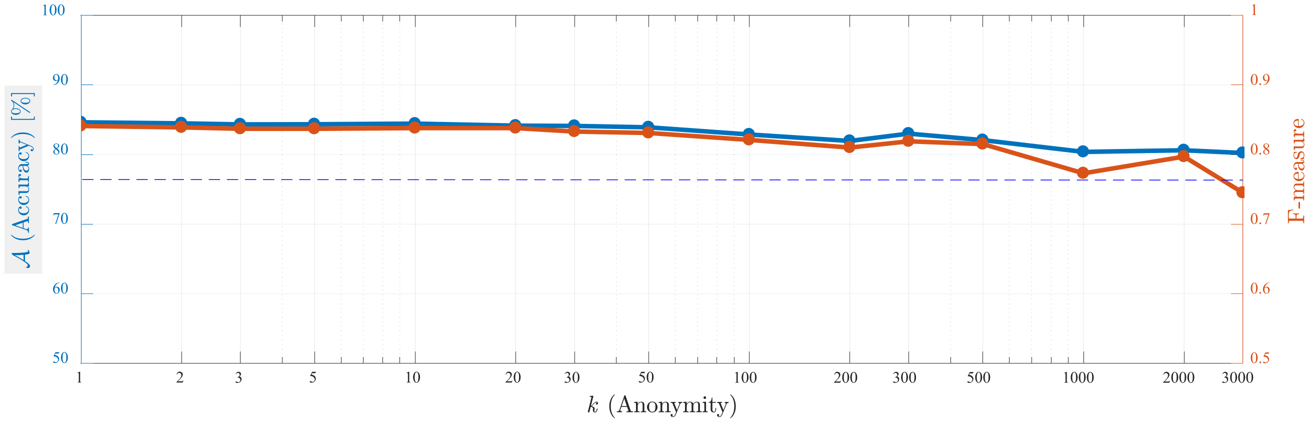


FIGURE 13. Degradation of the empirical utility (accuracy and F-measure) of the UCI Adult data set when microaggregated (using MDAV) for a wide range of ***k***.

TABLE 4. Different utility metrics for the UCI Adult data set when microaggregated for a wide range of ***k***.

k	Accuracy	F-Measure	AuC
1	84.63	0.841	0.902
2	84.48	0.839	0.898
3	84.33	0.837	0.897
5	84.35	0.837	0.897
10	84.44	0.838	0.898
20	84.15	0.838	0.891
30	84.11	0.833	0.887
50	83.91	0.831	0.883
100	82.88	0.821	0.875
200	81.95	0.810	0.861
300	83.00	0.819	0.848
500	82.07	0.815	0.827
1000	80.38	0.773	0.794
2000	80.61	0.797	0.693
3000	80.22	0.745	0.585

third dimension represents a binary confidential attribute. Since we require that the quasi-identifiers be predictors of the confidential attribute (as would be, e.g., the weight and height predictors of the existence or not of a disease in an individual), we introduce in the data a learnable macrotrend or dependence among the quasi-identifiers and the confidential attribute.

Next, we describe how we generate this synthetic data set. Let X be a bidimensional continuous r.v. representing the two quasi-identifiers $(x_1,\ x_2)$, and let Y be a binary r.v. indicating whether an individual has a disease ($Y = 1$) or not ($Y = 0$). The data set is generated in two parts, each matched to a different value of Y. Accordingly, X is distributed as a unit-variance Gaussian distribution with mean μ for $Y = 1$, and with mean $-\mu$, for $Y = 0$. In Fig. 6, we represent this data set by plotting the values of X for each record as coordinates of a point in a plane, coloring each point according to the class to which it belongs. As expected, two clouds of points are obtained (the red one, for $Y = 1$, slightly on the right; and the blue one, for $Y = 0$, on the left) where we can guess the optimal threshold to estimate the class $\hat{Y}$ of each point.

Let $\mathrm{P}\{Y = 1|x\}$ be the discriminative model of this problem. The prevalence p of a disease in this data set is the proportion of records matched to the class $Y = 1$. It is routine to represent this model, using logarithmic odds, as

$$\mathrm{L}\{Y = 1 \mid X = x\} = 2\mu x + \ln\frac{p}{1-p}.$$

We denote the cumulative distribution function (CDF) of the zero-mean, unit-variance Gaussian distribution as Φ. The accuracy $\mathcal{A}$ of our model to find the estimated class $\hat{Y}$ can be expressed as

$$\mathcal{A} = \mathrm{P}\{Y = \hat{Y}\} = (1-p)\Phi(\theta+\mu) + p\,\Phi(\mu-\theta),$$

for a given threshold $x = \theta$. It is straightforward to derive the optimal threshold θ^* for maximum accuracy of our discriminative model, which is

$$\theta^* = -\frac{1}{2\mu}\ln\frac{p}{1-p}.$$

In order to have a balanced data set, we use $p = 0.5$, thus half of the samples are matched to each class. Consequently, the optimal threshold to classify both parts of the data set is $\theta^* = 0$. Additionally, we choose $\mu = 0.125$ so that the distribution of both groups of samples are close; evidently, the more overlapped the two groups are, the more difficult the classification task.

Next, we train a machine learning model over a stratified part of the synthetic data, using the C4.5 algorithm. Since μ is low, the accuracy obtained from the classifier is 60%. Based on this model, we predict the class using the quasi-identifiers. Then, in Fig. 7, we plot the same clouds of samples of Fig. 6, but now we color them according to the *predicted* class. Accordingly, the classification threshold is evident.

To analyze the impact of microaggregation, we apply MDAV to the training set of this data set with $k = 3000$,

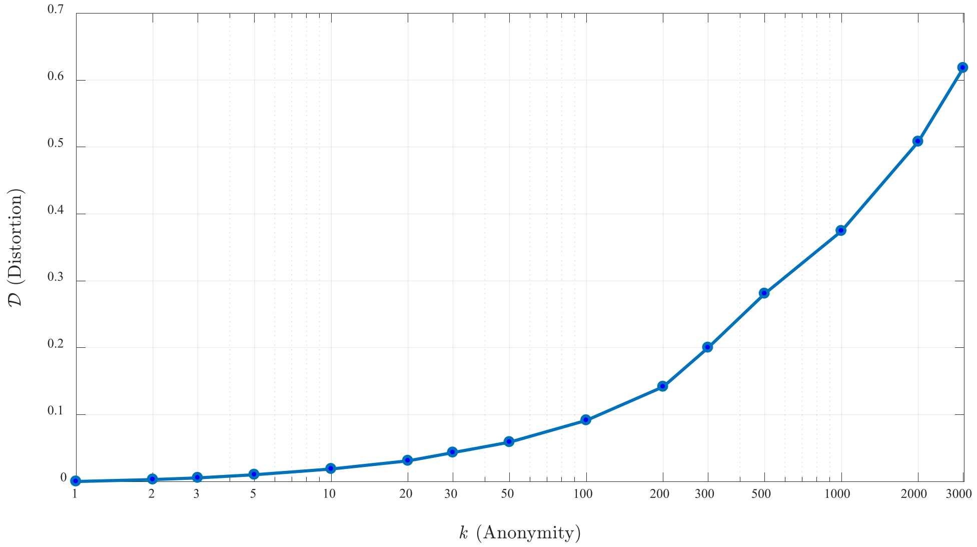


FIGURE 14. Distortion introduced by MDAV k-anonymous microaggregation to the UCI Adult data set, when microaggregated for a wide range of k.

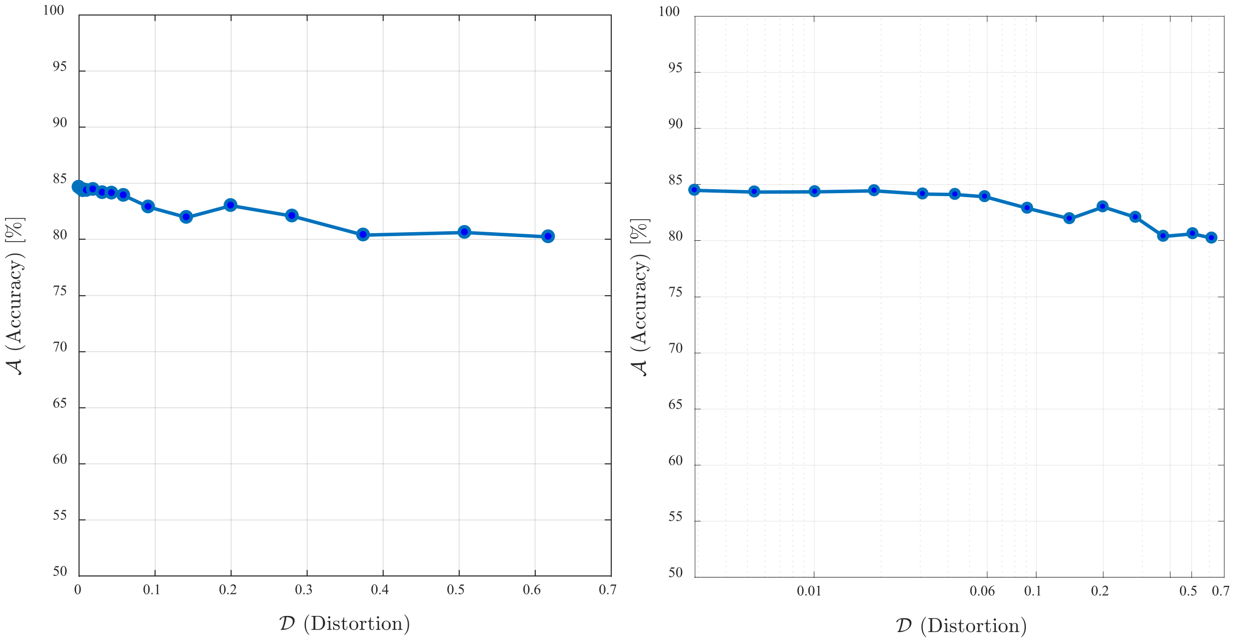


FIGURE 15. Accuracy of the *bagging* machine learning model trained on the microaggregated UCI Adult data set, against the distortion due to MDAV.

which is a very large value of cluster size. Accordingly, we get 7 cells that we plot in Fig. 8 with distinct colors; the classification threshold is also plotted. Notice in the figure that, after the clustering applied by MDAV, the samples of 3 out of 7 cells might be misclassified with a higher probability since such samples are distributed on both sides of the classification threshold. However, the remaining 4 cells, which account for about 57% of the data, are clearly defined on one side of the classification threshold, so they would be correctly classified. Hence, even after microaggregation, machine-learned macrotrends might not suffer a significant impact, i.e., the accuracy obtained from original data is not harshly reduced, even for high values of k.

To illustrate more systematically this effect on data utility, we plot in Fig. 9 the accuracy and F-measure of the learning model obtained from our synthetic data, after anonymizing it with different values of k. Consistently with the previous experiment, none of these utility metrics is drastically affected by the influence of microaggregation, for practical values of k. Another metric of the impact of microaggregation (not necessarily in terms of utility degradation) is also depicted in Fig. 10. Here, we observe that distortion,

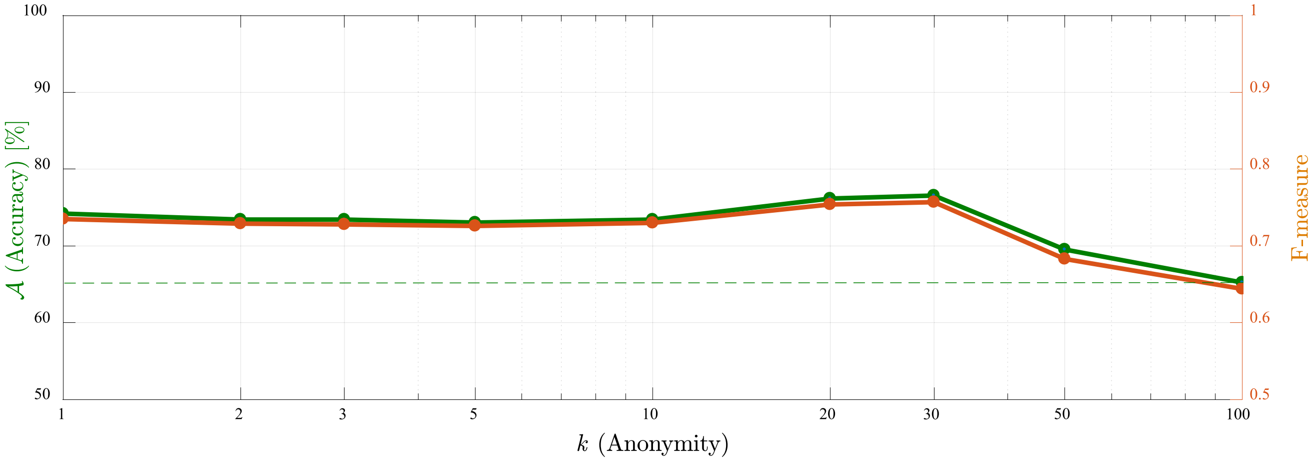


FIGURE 16. Degradation of the empirical utility of the UCI Pima Indians Diabetes data set when microaggregated (using MDAV) for a wide range of *k*.

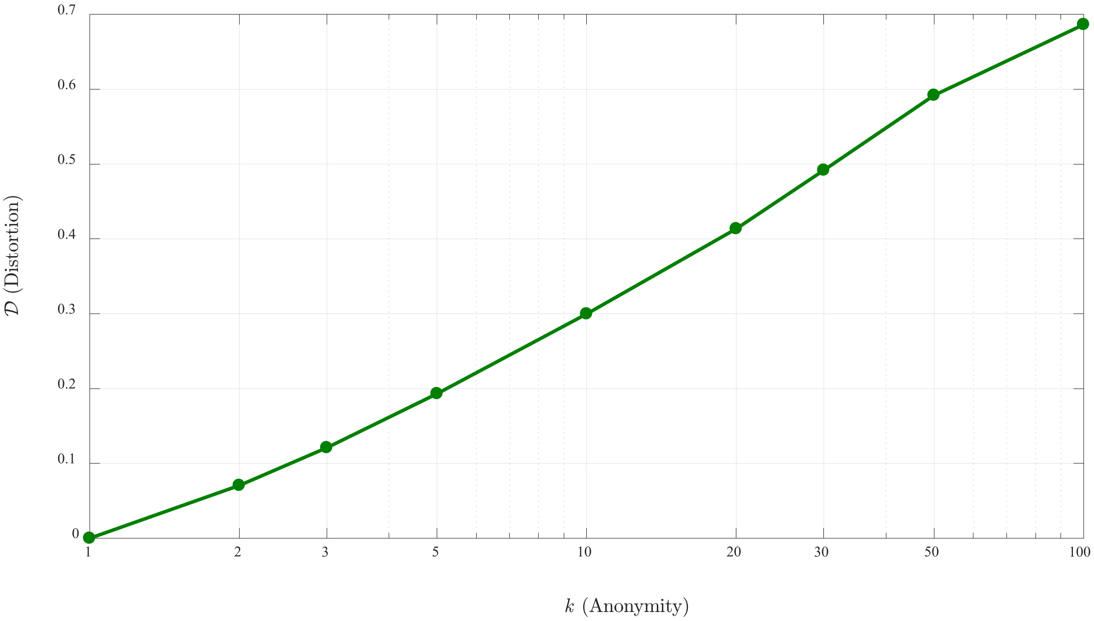


FIGURE 17. Distortion introduced by MDAV *k*-anonymous microaggregation to the UCI Pima Indians data set, for a wide range of *k*.

measured in terms of MSE, increases with k. However, distortion starts soon to increase significantly from $k = 100$. This divergence between accuracy and distortion is evidenced in Fig. 11, where the connection between both seems non-specific and nonlinear. A more detailed discussion regarding these results is presented in the next section, where real data is considered.

B. GENERAL RESULTS FROM REAL DATA SETS

We begin our first series of experiments by computing the relevance of the number of predictive attributes in each data set. The aim is to analyze how the accuracy of the classification task varies with the number of predictive attributes. To determine the order of the attributes employed, we used sequential forward selection, which consists in sequentially adding attributes to an empty set until the addition of further attributes does not decrease the accuracy of the classification task. Fig. 12 illustrates the variation of accuracy with the number of predictive attributes for UCI Adult.

Although intuition could suggest that even small levels of data perturbation might yield important reductions in utility, riveting results are found in our experiments when using microaggregation. First, Fig. 13 shows how the accuracy and F-measure of the classifier degrades as the privacy parameter k increases, when anonymizing the UCI Adult data set. As expected, accuracy attains its highest value (about 85%) when no anonymization is applied ($k = 1$). For $k = 200$, which is a relatively large value of cluster size, accuracy only

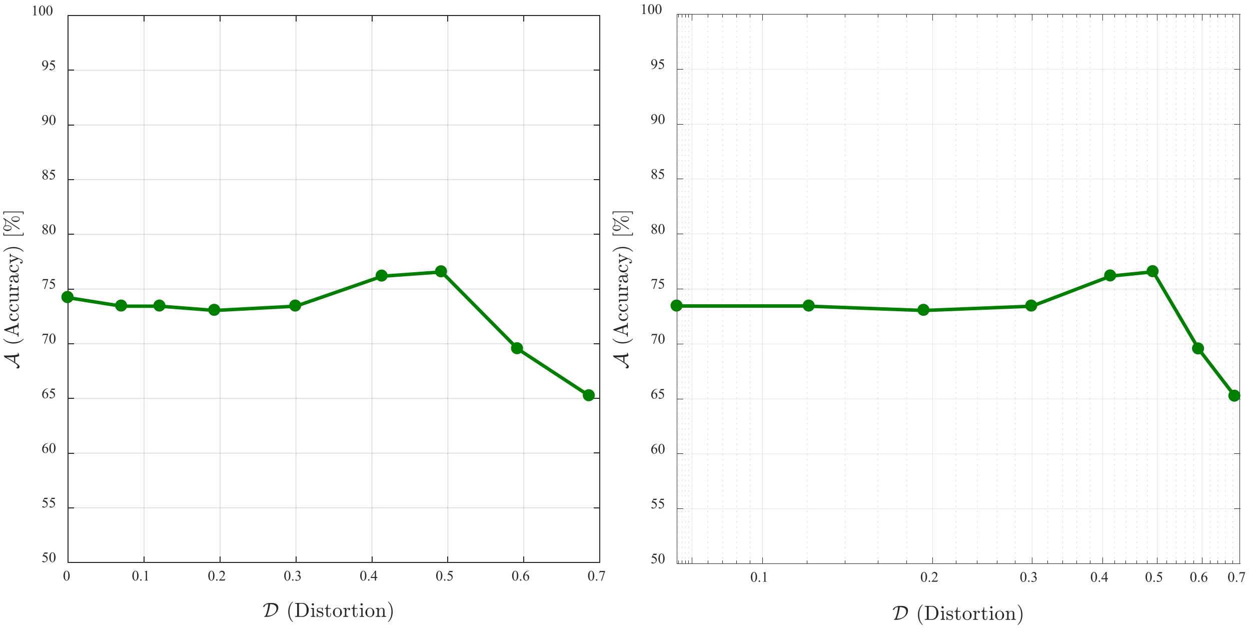


FIGURE 18. Accuracy of the *logistic regression* model trained on the microaggregated UCI Pima Indians Diabetes data set, against the distortion due to MDAV.

decreases up to 82%. We also note that, even for a value of k of 3,000, which implies a strong level of anonymity, accuracy is approximately 80%.

Fig. 13 also depicts a dotted line to represent the lowest accuracy achieved by the machine learning algorithm (76.37%) when no predictor attributes are used (suppression of all key attributes); this provides the highest level of privacy protection. Note that, when all key attributes are suppressed, the machine learning model always classifies a new instance depending on the majority value of the class attribute.

From the figure, we observe that a reduction in accuracy from 85% to 82% (attained for $k = 200$) when the key attribute (important predictor) "Capital Gain" is eliminated. Similarly, even when $k = 3,000$, we obtain a smaller impact on utility (accuracy of 80%) than when all predictors–except "Education Number"– are suppressed. This are good news for microaggregation, since it suggests that we can still get useful microdata after applying more than reasonable levels of privacy. The reported values of accuracy and other metrics (F-measure and AuC) are shown, in more detail, in Table 4.

The impact of MDAV on the UCI Adult data set is also measured in terms of the distortion introduced to quasi-identifiers. We use MSE to quantify such distortion. In Fig. 14, we can see how distortion increases from 0 (when $k = 1$) to 0.62 (for $k = 3,000$). Specifically, we observe a pronounced growth from $k = 100$, although for values of k smaller than 100, distortion does not seem significant.

In Fig. 15, we plot accuracy vs distortion. The most relevant conclusion that can be drawn from this figure is that accuracy stays relatively stable (greater than 80%) up to distortions of 0.7. Precisely, although MSE is conventionally used in SDC to compare the utility of microaggregation algorithms, we observe that this distortion metric says little about the impact on the performance of a machine-learning classifier. In other words, the data yielded by this figure seems to provide convincing evidence that MSE is not a suitable measure of utility for classification tasks.

TABLE 5. Different utility metrics for the UCI Pima Indians data set when microaggregated for a wide range of k.

k	Accuracy	F-Measure	AuC
1	74.21	0.735	0.813
2	73.43	0.729	0.810
3	73.43	0.728	0.808
5	73.04	0.726	0.804
10	73.43	0.730	0.806
20	76.17	0.754	0.807
30	76.56	0.757	0.789
50	69.53	0.683	0.758
100	65.23	0.644	0.716

In our evaluation of the UCI Pima Indian Diabetes data set in Fig. 16, we note that the degradation margin of utility goes from 74.2% (when $k = 1$, thus without perturbation) to 65.23% (from $k = 100$). Microaggregation shows a similar behavior to that observed in the UCI Adult data set but, being 50 times smaller, it evidently degrades more quickly as k increases. However, a noticeable stability is appreciated in accuracy up to $k = 30$ and, in fact, this performance metric remains close to the upper baseline at around 74%. For values of k between 10 and 30, accuracy is even improved, which could be explained by the denoising effect of averaging through clever clustering, that may positively contribute to

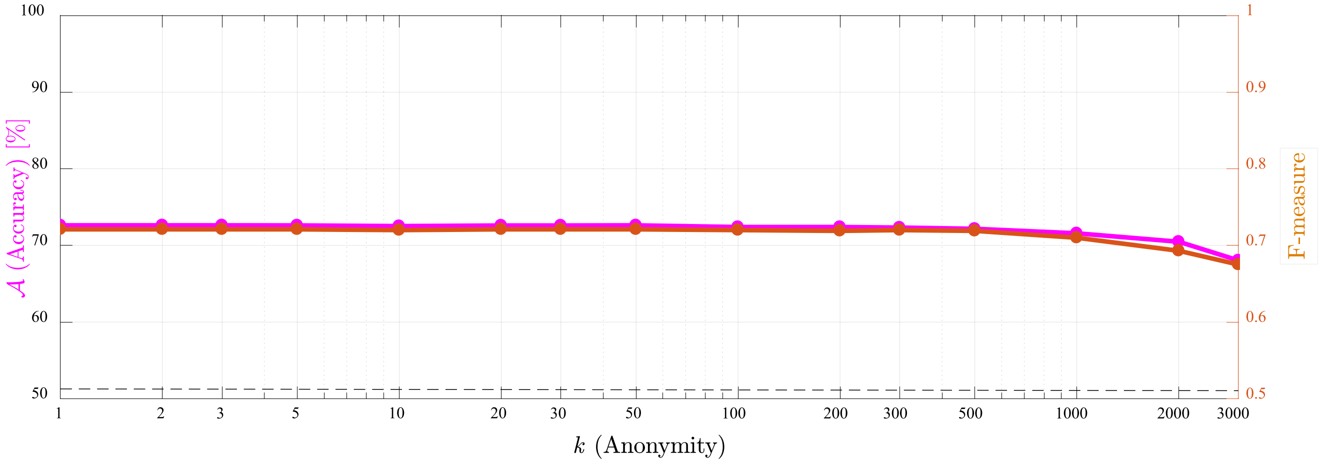


FIGURE 19. Degradation of the empirical utility (accuracy) of the Irish Census data set when microaggregated for a wide range of k.

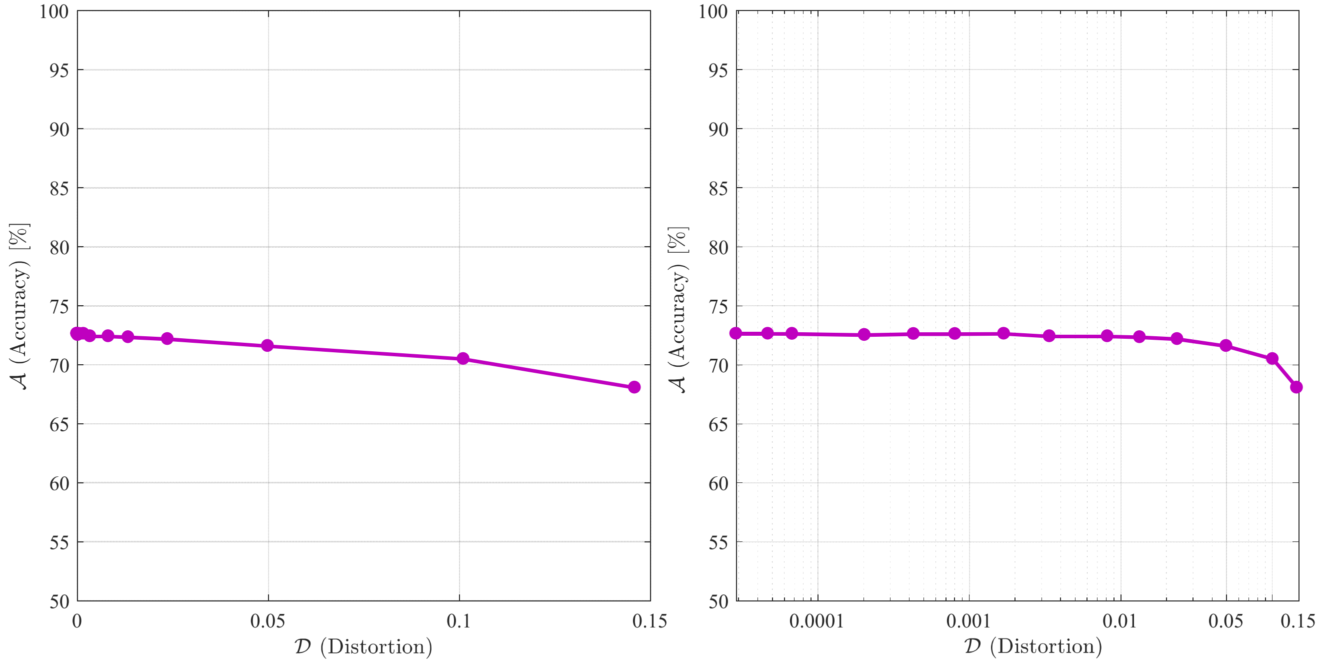


FIGURE 20. Accuracy of the *C4.5* machine learning model trained over the microaggregated Irish Census data set, against the distortion due to MDAV.

a more robust inference. This effect is illustrated in §IV-A and §V. Interestingly, Fig. 17 shows a sustained increase in distortion as k becomes larger. To gain insight into this relative stability in accuracy, we also plot accuracy vs distortion in Fig. 18 and confirm that, up to distortions of 50%, utility remains close to the upper baseline. The values of accuracy and other metrics (F-measure and AuC) obtained for this data set are also shown in Table 5.

Finally, we examine the Irish data set in Fig. 19. Here, we observe a wide degradation margin since its label attribute has balanced classes. Specifically, accuracy goes from 72.62% to about 68.04% when the privacy parameter k equals 3,000. Also, we can see, once again, that accuracy remains quite high (more than 70%) and stable up to $k = 2,000$. A similar behavior is observed for F-measure. Although the size of the data set at hand is relatively large (100K instances), the available evidence suggests that the reduction of empirical utility of the data due to microaggregation is not significant for a wide range of values of k. Such effect is also noticeable in Fig. 20, where we plot accuracy vs distortion. Table 6 shows the reported values of accuracy, as well as other metrics (F-measure and AuC), in greater detail.

Our experimental findings confirm that MDAV introduces sufficiently small levels of perturbation in the quasi-identifiers, so that the statistical properties of the published data

TABLE 6. Different utility metrics for the Irish Census data set when microaggregated for a wide range of k.

k	Accuracy	F-Measure	AuC
1	72.62	0.721	0.736
2	72.62	0.721	0.733
3	72.62	0.721	0.733
5	72.61	0.721	0.733
10	72.52	0.720	0.733
20	72.60	0.721	0.733
30	72.60	0.721	0.733
50	72.62	0.721	0.734
100	72.40	0.720	0.731
200	72.40	0.719	0.735
300	72.33	0.720	0.729
500	72.17	0.719	0.718
1000	71.58	0.710	0.729
2000	70.48	0.693	0.739
3000	68.04	0.675	0.703

can be preserved to a large extent, while satisfying a given k-anonymity constraint. The upshot is that much of the empirical utility is retained within the microaggregated data. In fact, the results of our experiments suggest that such impact is often minor, since microaggregation preserves machine-learned macrotrends. We believe that the average operations performed by MDAV to find a centroid representative of k tuples are working as a noising removal filter that prevents the classifier algorithm from adjusting to the existing noise in the data.

Interestingly, although not explicitly reported in these terms, previous work surveyed in Section I appears to be consistent with our findings. For example, in [23], where different algorithms based on generalization and suppression are compared, the degradation in accuracy is certainly small in many cases. Other works in the literature give some clues about a potential "constructive effect" of anonymization mechanisms. In that sense, [25] mentions that anonymization might sometimes behave as a form of feature selection or construction. Moreover, Malle *et al.* [33] conclude that a selective anonymization may not be so destructive. Finally, using a less conclusive argument, [7] states that, while making no changes to existing tools and systems, significant utility can be extracted from anonymized data.

Testing a wide range of values of the privacy parameter helps to make visible the overall effect of anonymization on data utility. Doing so also assists in noticing the influence of other critical criteria such as the size of the data set and the absolute upper and lower bounds of utility. As shown in our experimental results, the utility of anonymized microdata, measured as classification accuracy, may not take values strictly from 0 to 100%. The intrinsic statistical properties of released data would already limit the capabilities of machine learning algorithms and, thus, the improvements they get over baseline methods (e.g., always predicting the most frequent class in the training set). Evidently, very little utility can be maintained after anonymization if machine learning (classification) algorithms perform poorly, by default, with respect to the baseline. Unfortunately, these considerations are not always made when evaluating the performance of k-anonymous microaggregation or, in general, of anonymization mechanisms.

V. CONCLUSION

With the advent of the Internet and the development of sophisticated data analytics, the availability of massive amounts of information has increased the demand for data sharing. In the context of structured data, microdata are an invaluable source of information for their potential to reveal patterns or macrotrends about the population there represented.

Before these data can be made public or shared with external entities, data holders must ensure individual privacy is safeguarded. Perturbing quasi-identifiers attributes is the usual approach to prevent identity disclosure in microdata. Nonetheless, while perturbation may prevent reidentification attacks, it may have a large impact on data utility, particularly on the performance of machine-learning tasks. To cope with it, several works have proposed adapting data-anonymization or machine-learning algorithms to get more utility from anonymized data. We claim in this work, however, that the default operation of some anonymization mechanisms may not affect data utility significantly.

In this paper, we have investigated the high-utility SDC spectrum, implemented by syntactic k-anonymous microaggregation, which has a direct application on the health domain where utility is critical. Our experiments have shown, with some consistency, that k-anonymous microaggregation implemented through MDAV does not have a significant impact on machine-learned macrotrends for multiple data sets and a wide range of machine-learning algorithms. Trying to consider the domain of data in or evaluation, we not only tested different data sets but also multiple learning algorithms to extract the maximum utility from the data. Then, these algorithms were selected to get the highest accuracy from each data set.

These excellent results on learning performance from microaggregated data deserve careful attention. As the lack of substantial degradation in classification accuracy for a generous range of microcell sizes k may be somewhat counterintuitive, we conducted further verification on such remarkable finding. Specifically, we applied the k-nearest neighbor algorithm (kNN) to the original, unperturbed data, in order to verify our working hypothesis that clustering effectively acts as a form of averaging and thus denoising. In our verification, k is the usual name for the parameter governing the size of the cluster of the kNN algorithm, analogous to some extent to the anonymity parameter. Fig. 5 illustrates that the classification accuracy of kNN improves as groups rather than individual samples are considered to robustly infer what would effectively constitute a macrotrend. We stress that kNN was applied in Fig. 5 to the original data, with no alteration or protection, to shed light on this matter.

We contend that a similar denoising effect, akin to averaging through clustering, is the underlying cause of the striking utility of k-anonymous microaggregation. Conceivably, for reasonable values of the anonymity parameter k, microaggregation should not substantially devalue the process of inference of macrotrends carried out by the machine learning algorithm. Moreover, high-utility microaggregation algorithms such as MDAV may, in some cases, positively contribute to a more robust inference by denoising through clever clustering of demographically similar individuals. The benefit of preprocessing data with unsupervised techniques based on clustering, prior to supervised learning, is known in the machine-learning literature. The lack of substantial degradation in classification performance due to k-anonymous microaggregation, and the occasional slight improvement in utility, is a novel result of strategic importance in the privacy arena.

Finally, our results provide confirmatory evidence that, while keeping a monotonicity relationship with accuracy, the traditional utility metric of SDC (i.e., MSE) is not an ideal metric to determine the impact on the utility of microaggregated data, since there exists a non-specific non-linear dependence.

ACKNOWLEDGMENT

The authors are thankful to V. Ayala-Rivera for sharing the Irish Census synthetic data set. J. Parra-Arnau is the recipient of a Juan de la Cierva postdoctoral fellowship, FJCI-2014-19703, from the MINECO. Ana Rodríguez and José Estrada acknowledge the support from Escuela Politécnica Nacional (EPN).

REFERENCES

[1] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy. (Nov. 2013). "Synthetic data generation using Benerator tool." [Online]. Available: https://arxiv.org/abs/1311.3312

[2] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners," *Trans. Data Privacy*, vol. 7, no. 3, pp. 337–370, 2014.

[3] V. Ayala-Rivera, A. O. Portillo-Dominguez, L. Murphy, and C. Thorpe, "COCOA: A synthetic data generator for testing anonymization techniques," in *Proc. Int. Conf. Privacy Stat. Databases (PSD)*, Dubrovnik, Croatia, Sep. 2016, pp. 163–177.

[4] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Las Vegas, NV, USA, Aug. 2008, pp. 70–78.

[5] J. Cao, B. Carminati, E. Ferrari, and K.-L. Tan, "CASTLE: Continuously anonymizing data streams," *IEEE Trans. Depend. Sec. Comput.*, vol. 8, no. 3, pp. 337–352, May/Jun. 2011.

[6] K. Chaudhuri and C. Monteleoni, "Privacy-preserving logistic regression," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, Dec. 2008, pp. 289–296.

[7] G. Cormode, E. Shen, X. Gong, T. Yu, C. M. Procopiuc, and D. Srivastava, "UMicS: From anonymized data to usable microdata," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, San Francisco, CA, USA, Oct. 2013, pp. 2255–2260.

[8] G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y. de Montjoye, and A. Bourka, "Privacy by design in big data," EU Agency Netw., Inf. Secur., Heraklion, Greece, Tech. Rep. TP-04-15-941-EN-N, Dec. 2015. [Online]. Available: http://doi.org/10.2824/641480

[9] D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: The small aggregates method," in *Proc. Symp., Design Anal. Longitudinal Surv.*, Ottawa, ON, Canada, Nov. 1993, pp. 195–204.

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

[11] J. Domingo-Ferrer, A. Martínez-Ballesté, J. M. Mateo-Sanz, and F. Sebé, "Efficient multivariate data-oriented microaggregation," *VLDB J.*, vol. 15, no. 4, pp. 355–369, 2006.

[12] J. Domingo-Ferrer and J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 1, pp. 189–201, Jan./Feb. 2002.

[13] J. Domingo-Ferrer, F. Sebé, and A. Solanas, "A polynomial-time approximation to optimal multivariate microaggregation," *Comput., Math., Appl.*, vol. 55, no. 4, pp. 714–732, Feb. 2008.

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

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

[16] J. Domingo-Ferrer and V. Torra, "A critique of k-anonymity and some of its enhancements," in *Proc. Workshop Privacy, Secur., Artif. Intell. (PSAI)*, Barcelona, Spain, Mar. 2008, pp. 990–993.

[17] "Opinion 05/2014 on anonymisation techniques," Article 29 Data Protection Working Party, Apr. 2014.

[18] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloq. Autom., Lang. Programm.*, vol. 4052. Jul. 2006, pp. 1–12.

[19] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, "Privacy-preserving learning analytics: Challenges and techniques," *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 68–81, Jan./Mar. 2017.

[20] A. Hundepool *et al.* (2003). μ-ARGUS version 3.2 software and user's manual, Statistics Netherlands, Voorburg, Netherlands. [Online]. Available: http://neon.vb.cbs.nl/casc

[21] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Shanghai, China, Mar./pr. 2009, pp. 429–440.

[22] Y. Jafer, S. Matwin, and M. Sokolova, "Task oriented privacy preserving data publishing using feature selection," in *Proc. Can. Conf. Artif. Intell.*, Montréal, BC, Canada, May 2014, pp. 143–154.

[23] S. Kisilevich, L. Rokach, Y. Elovici, and B. Shapira, "Efficient multidimensional suppression for k-anonymity," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 3, pp. 334–347, Mar. 2010.

[24] M. Laszlo and S. Mukherjee, "Minimum spanning tree partitioning algorithm for microaggregation," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 7, pp. 902–911, Jul. 2005.

[25] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Workload-aware anonymization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Philadelphia, PA, USA, Aug. 2006, pp. 277–286.

[26] D. Li, X. Shen, and L. Wang, "Connected geomatics in the big data era," *Int. J. Digit. Earth*, vol. 11, no. 2, pp. 139–153, Apr. 2017.

[27] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.

[28] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Paris, France, Jun. 2009, pp. 517–526.

[29] J.-L. Lin, T.-H. Wen, J.-C. Hsieh, and P.-C. Chang, "Density-based microaggregation for statistical disclosure control," *Expert Syst., Appl.*, vol. 37, no. 4, pp. 3256–3263, Apr. 2010.

[30] K.-P. Lin and M.-S. Chen, "On the design and analysis of the privacy-preserving SVM classifier," *IEEE Trans. Knowl., Data Eng.*, vol. 23, no. 11, pp. 1704–1717, Nov. 2010.

[31] K.-P. Lin and M.-S. Chen, "Privacy-preserving outsourcing support vector machines with random transformation," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Washington, DC, USA, Jul. 2010, pp. 363–372.

[32] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k-anonymity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Atlanta, GA, USA, Apr. 2006, p. 24.

[33] B. Malle, P. Kieseberg, E. Weippl, and A. Holzinger, "The right to be forgotten: Towards machine learning on perturbed knowledge bases," in *Proc. Int. Conf. Availability, Rel., Secur. (ARES)*, vol. 9817. Salzburg, Austria, Aug. 2016, pp. 251–266.

[34] K. Mancuhan and C. Clifton. (Oct. 2016). "Decision tree classification on outsourced data." [Online]. Available: https://arxiv.org/abs/1610.05796

[35] N. Matatov, L. Rokach, and O. Maimon, "Privacy-preserving data mining: A feature set partitioning approach," *Inf. Sci.*, vol. 180, no. 14, pp. 2696–2720, 2010.
[36] S. Matwin, J. Nin, M. Sehatkar, and T. Szapiro, "A review of attribute disclosure control," in *Advanced Research in Data Privacy* (Studies in Computational Intelligence), vol. 567, G. Navarro-Arribas and V. Torra, Eds. Cham, Switzerland: Springer, 2015, pp. 41–61.
[37] A. Monaco. (Oct. 2016). *Big Data: The Measure of Humankind*. Times Higher Education. [Online]. Available: http://www.timeshighereducation.com/comment/big-data-measure-humankind#survey-answer
[38] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health Inf. Sci., Syst.*, vol. 2, no. 3, pp. 1–10, Feb. 2014.
[39] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From *t*-closeness to PRAM and noise addition via information theory," in *Proc. Int. Conf. Privacy Stat. Databases (PSD)*, Istanbul, Turkey, Sep. 2008, pp. 100–112.
[40] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From *t*-closeness-like privacy to postrandomization via information theory," *IEEE Trans. Knowl., Data Eng.*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190
[41] 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, Knowl. Eng.*, vol. 70, no. 10, pp. 892–921, Oct. 2011. [Online]. Available: http://doi.org/10.1016/j.datak.2011.06.005
[42] D. Rebollo-Monedero, J. Parra-Arnau, and C. Díaz, and J. Forné, "On the measurement of privacy as an attacker's estimation error," *Int. J. Inf. Secur.*, vol. 12, no. 2, pp. 129–149, Apr. 2013. [Online]. Available: http://doi.org/10.1007/s10207-012-0182-5
[43] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Trans. Knowl., Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov./Dec. 2001.
[44] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 6, pp. 838–852, Jun. 2013.
[45] M. Schmid and H. Schneeweiss, "The effect of microaggregation procedures on the estimation of linear models: A simulation study," *J. Econ., Stat.*, vol. 225, no. 5, pp. 529–543, Sep. 2005.
[46] A. Solanas and A. Martínez-Ballesté, "V-MDAV: A multivariate microaggregation with variable group size," in *Proc. Int. Conf. Comput. Stat. (CompStat)*, Rome, Italy, 2006, pp. 1–8.
[47] X. Sun, H. Wang, J. Li, and T. M. Truta, "Enhanced *p*-sensitive *k*-anonymity models for privacy preserving data publishing," *Trans. Data Privacy*, vol. 1, no. 2, pp. 53–66, 2008.
[48] L. Sweeney, "Uniqueness of simple demographics in the U.S. population," School Comput. Sci., Data Privacy Lab, Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. LIDAP-WP4, 2000.
[49] T. M. Truta and B. Vinay, "Privacy protection: *p*-sensitive *k*-anonymity property," in *Proc. 22nd Int. Conf. Data Eng. Workshops*, Apr. 2006, p. 94.
[50] *UCI Machine Learning Repository: Adult Dataset*. Accessed: Jun. 15, 2017. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Adult
[51] *UCI Machine Learning Repository: Pima Indians Dataset*. Accessed: Jun. 15, 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes
[52] I. Witten, E. Frank, M. Hall, and C. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington, VT, USA: Morgan Kaufmann, 2016.
[53] A. N. K. Zaman, C. Obimbo, and R. A. Dara, "A novel differential privacy approach that enhances classification accuracy," in *Proc. Int. C Conf. Comput. Sci., Softw. Eng. (C3S2E)*, Porto, Portugal, Jul. 2016, pp. 79–84.
[54] S. Zhong, Z. Yang, and T. Chen, "*k*-Anonymous data collection," *Inf. Sci.*, vol. 179, no. 17, pp. 2948–2963, Aug. 2009.

ANA RODRÍGUEZ-HOYOS received the bachelor's degree from Escuela Politécnica Nacional (EPN), Quito, Ecuador, in 2010, and the M.S. degree from the Universitat Politècnica de Catalunya, Barcelona, Spain, in 2013, where she is currently pursuing the Ph.D. degree in telematics engineering with the Information Security Group. She is a Lecturer at EPN. Her research interests encompass data privacy and machine learning.

JOSÉ ESTRADA-JIMÉNEZ received the bachelor's degree from Escuela Politécnica Nacional (EPN), Quito, Ecuador, in 2007, and the M.S. degree from the Universitat Politècnica de Catalunya, Barcelona, Spain, in 2013, where he is currently pursuing the Ph.D. degree in telematics engineering with the Information Security Group. He is a Lecturer at EPN. His research interests encompass data privacy, online advertising, and machine learning.

DAVID REBOLLO-MONEDERO received the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, USA, in 2003 and 2007, respectively. From 1997 to 2000, he was an Information Technology Consultant with PricewaterhouseCoopers, Barcelona. He is currently a Senior Researcher with the Information Security Group, Department of Telematic Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain, where he investigates the application of information theoretic and operational data compression formalisms to privacy in information systems. His current research interests encompass data privacy, information theory, data compression, and machine learning.

JAVIER PARRA-ARNAU received the M.S degree in telecommunications engineering and the M.S. and Ph.D. degrees in telematics engineering from the Universitat Politècnica de Catalunya in 2004, 2009, and 2013, respectively. He was a Post-Doctoral Fellow with INRIA Grenoble and Paris-Saclay, France, and a Visiting Researcher with NEC Laboratories Europe, Germany. He is currently a Juan-de-la-Cierva Post-Doctoral Researcher with the Department of Computer Science and Mathematics, Universitat Rovira i Virgili, Tarragona, Spain. Among other honors, he received the prize to the Best Ph.D. thesis on information and communication technologies in banking from the Official College of Telecommunication Engineers and Banco Sabadell and the prize Data Protection by Design 2016 from the Catalan Data Protection Authority.

JORDI FORNÉ received the M.S. and Ph.D. degrees in telecommunications engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 1992 and 1997, respectively. From 2007 to 2012, he was a Coordinator of the Ph.D. Program in telematics engineering and the Director of the master's research program in telematics engineering. He is currently an Associate Professor with the Telecommunications Engineering School of Barcelona, UPC. His research interests span a number of subfields within information security and privacy.