Preprint Submitted to
Information Sciences
Jan. 31st, 2019

Elsevier
Journal

Efficient k-Anonymous Microaggregation of Multivariate Numerical Data via Principal Component Analysis

David Rebollo-Monedero*[,1], Ahmad Mohamad Mezher[1], Xavier Casanova Colomé[1], Jordi Forné[1], and Miguel Soriano[1,2]

[1] Department of Telematic Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain
[2] Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), E-08860 Castelldefels, Barcelona, Spain

Article Info

Article History:
Revised Jan. 31st, 2019

Keywords:
- data privacy
- statistical disclosure control
- k-anonymity
- microaggregation
- principal component analysis
- large-scale datasets

Abstract

k-Anonymous microaggregation is a widespread technique to address the problem of protecting the privacy of the respondents involved beyond the mere suppression of their identifiers, in applications where preserving the utility of the information disclosed is critical. Unfortunately, microaggregation methods with high data utility may impose stringent computational demands when dealing with datasets containing a large number of records and attributes.

This work proposes and analyzes various anonymization methods which draw upon the algebraic-statistical technique of principal component analysis (PCA), in order to effective reduce the number of attributes processed, that is, the dimension of the multivariate microaggregation problem at hand. By preserving to a high degree the energy of the numerical dataset and carefully choosing the number of dominant components to process, we manage to achieve remarkable reductions in running time and memory usage with negligible impact in information utility. Our methods are readily applicable to high-utility SDC of large-scale datasets with numerical demographic attributes.

I. Introduction

Over recent years, big-data technologies have acquired extreme relevance, becoming commonplace for all kinds of companies and research organizations to focus their efforts on the development of methodologies to process vast volumes of data efficiently. Feeding this data to powerful data-analysis systems and machine-learning algorithms has set in motion a virtuous circle of digitalization in a wide variety of arenas, running the whole gamut from targeted advertising to precision medicine, an effect that can only continue to accelerate in coming years. But all too often, a substantial portion of this data consists in personal information posing significant privacy risks, particularly when releasing sensitive data to untrusted parties or openly publishing it for any number of statistical studies.

* Corresponding author, author@domain.edu.
ORCID 0000-0000-0000-0000, sites.google.com/site/author.

http://dx.doi.org/<DOI>

The field of *statistical disclosure control* (SDC) emerged to address this conundrum in the release of personal data. Specifically, k-anonymous microaggregation permits protecting the privacy of the respondents involved, beyond the mere suppression of their identifiers, by carefully aggregating demographic attributes. This reduces the risk of reidentification in applications where preserving the utility of the information disclosed is also critical. Unfortunately, microaggregation methods with high data utility may impose stringent computational demands when dealing with datasets containing a large number of records and attributes.

In this work, we propose and analyze various anonymization methods which draw upon the algebraic-statistical technique of *principal component analysis* (PCA), in order to effective reduce the number of attributes processed, that is, the dimension of the multivariate microaggregation problem at hand. By preserving to a high degree the energy of the numerical dataset and carefully choosing the number of dominant components to process, we manage to achieve remarkable reductions in running time and memory usage with negligible impact in information utility. Our methods are readily applicable to high-utility SDC of large-scale datasets with numerical demographic attributes.

A. Fundamentals of Statistical Disclosure Control and Microaggregation

As famously shown in [41, 42], 87% of the population in the United States has reported characteristics that likely made them unique based only on the tuple consisting of 5-digit ZIP, gender, and date of birth. The findings in [41, 42] mean that the mere elimination of identifiers such as first and last name, or social security number, is grossly insufficient to effectively protect the anonymity of the participants of published statistical studies containing confidential data linked to demographic information.

In the field of SDC, introduced earlier, a *microdata set* is a database table whose records carry information concerning identifiable individuals or organizations. Each of these records contains attributes that may be divided into identifiers, quasi-identifiers, and confidential attributes.

- *Identifiers* unequivocally identify respondents in the microdata set. Examples of identifiers are full name or SSNs. Certainly, they must be removed before publishing the microdata set, in order to guarantee anonymity.
- *Quasi-identifiers*, typically demographic attributes, may still pose a risk of *reidentification* when considered jointly, by cross-referencing them with external, usually publicly available information. Examples of quasi-identifiers are age, height, weight, gender, and job.
- *Confidential attributes* contain sensitive information on the respondents, such as salary, political affiliation, and health condition.

A conceptual example of microdata set is shown in Fig. 1, where respondents are identified by their full names, educational stage or years of schooling, age, and ZIP code play the role of quasi-identifiers, and where confidential attributes consist of family income and a consumer profile representing purchasing preferences along predefined categories.

Identifiers	Quasi-Identifiers (Demographic Attributes)			Confidential Attributes	
Name	Edu Yrs	Age	ZIP Code	Family Income	Consumer Profile
Alice Adams	14	32	94024	$39 250	
Bob Brown	10	34	94305	$21 700	
Chloe Carter	12	33	94024	$32 150	
Dave Diaz	17	43	90210	$57 400	
Eve Ellis	16	47	90210	$56 300	
Frank Fisher	15	45	90213	$54 100	

Fig. 1. Synthetic example of microdata set containing demographic attributes (educational stage or years of schooling, age, and ZIP code) along with confidential information (family income, consumer profile representing interests along predefined categories).

As already mentioned, the mere suppression of identifiers is not sufficient to guarantee anonymity, although it is certainly necessary. In order to address the risk of reidentification, the quasi-identifiers in the microdata set must be subjected to some form of perturbation, according to the process outlined next and conceptually depicted in Fig. 2. Precisely, in k-anonymous microaggregation, groups of k demographically similar respondents are formed, and their corresponding tuples of quasi-identifiers, replaced by a common representative tuple. This prevents reidentification via cross-referencing of quasi-identifiers. Because respondents within each group of k records are indistinguishable from each other on the basis of their quasi-identifiers, this form of protection is naturally known as k-anonymity. The procedure described is shown in Fig. 3 for our synthetic example. Recall that full name is an identifier, years of education, age and ZIP code constitute quasi-identifiers, and family income and consumer profile play the role of confidential attributes. As illustrated, identifiers are removed before publishing the table. Further, the published table contains groups of k

records with a common value for their quasi-identifiers. The published table is a k-anonymous version of the original, where demographic attributes have been adequately aggregated.

As we have seen, in k-anonymous microaggregation, quasi-identifiers are perturbed in order to preserve *privacy*, at the cost of losing some of the *data utility*, the latter characterized as the accuracy or absence of discrepancy with respect to the original dataset. The design and operation of this type of algorithms must take into consideration the *trade-off* between these two contrasting aspects. The careful assignment of individual records to microcells of size at least k in order to guarantee this form of anonymity, while preserving the utility of the data released, is no trivial matter.

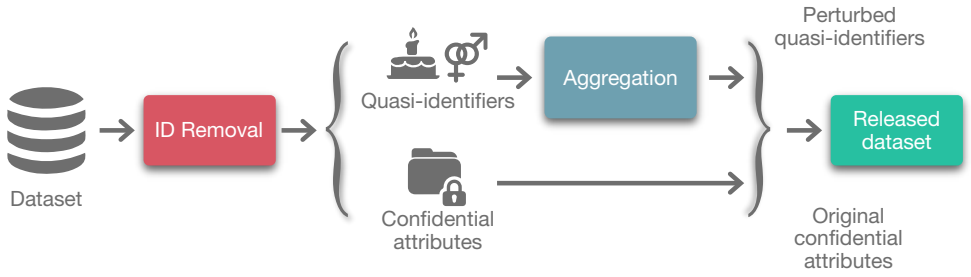


Fig. 2. Block diagram of the k-anonymous microaggregation process. The removal of identifiers is necessary yet insufficient. Quasi-identifiers, typically demographic attributes that may be cross-referenced to infer the identity of the respondent, are aggregated by similarity. Aggregated tuples of quasi-identifiers are replaced by a common representative tuple for each group. The published dataset contains the original confidential attributes, but they are linked to perturbed quasi-identifiers.

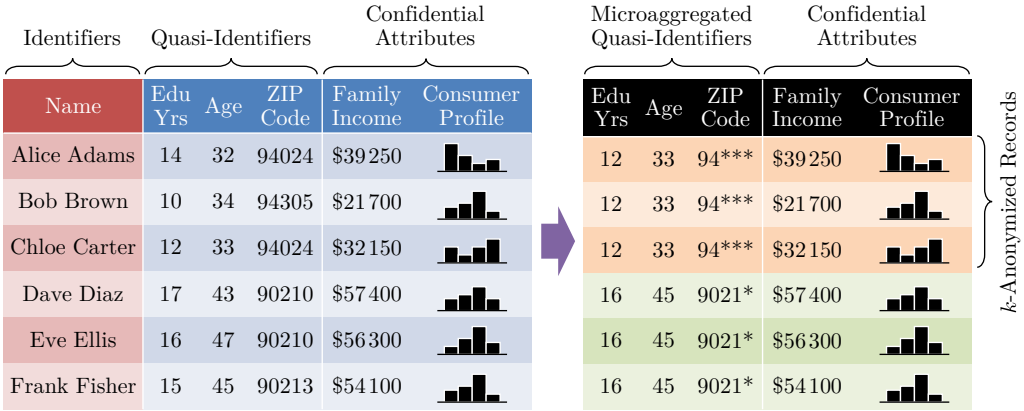


Fig. 3. k-Anonymous microaggregation applied to our synthetic example. Prior to its release or publication, identifiers are removed, and k-anonymous microaggregation applied with $k = 3$. By grouping quasi-identifiers, individuals remain demographically indistinguishable among a group of k uncertain possibilities.

B. Contribution and Organization

This work deals with the specific problem of computational complexity of k-anonymous microaggregation for large datasets with a substantial amount of numerical quasi-identifiers and records. Although our work is illustrated with the special case of the widely used algorithm known as maximum distance to average vector (MDAV), the methods outlined would be readily applicable to other microaggregation techniques. Certainly, the focus is not on the particular algorithm employed, as we operate merely under the mild assumption that the running time significantly increases with the number of numerical quasi-identifiers in the microdata set.

In order to accelerate the anonymization process, we capitalize on PCA as an algebraic-statistical technique for dimensionality reduction, in the form of two novel methods. Here, dimension refers to the number of numerical quasi-identifiers in the dataset. Both of our methods are able to microaggregate large databases significantly faster than MDAV, and they do maintain the statistical quality of the released information, which makes them suitable for a great number of applications across different fields. We also report extensive experimentation on standardized datasets. As an added practical advantage, the use of dimensionality-reduction techniques is also the source of substantial memory savings, and thus, the new algorithms not only greatly improve the time required to execute them, but they should also increase memory efficiency, particularly through memory paging and cache-memory storage. The highlights of our contribution are summarized in Fig. 4.

More concretely, our contributions are the following.

- We introduce two dimensionality-reduction methods in the microaggregation field, by means of PCA, which greatly reduce the running time of k-anonymous microaggregation of multivariate data, maintaining the quality of the

released information. The first method is more conservative in terms of distortion and corresponds to a direct application of PCA to the entire dataset.

- The second method attains more aggressive speed-ups by splitting the dataset into a proximal portion of data points closer to each other, for which PCA is directly applied as in the first method, and a distal portion with higher distortion that is not subject to PCA at all. The speed-up achieved through this second method is the resulting synergy of dimensionality reduction and the superadditivity of the running time of microaggregation algorithms with the number of records. Recall that *superadditive complexity* in the number of records means that the running time on $a+b$ records satisfies $t(a+b) \geqslant t(a)+t(b)$, making it conducive to the celebrated algorithmic approach of "divide and conquer". This is a typical characteristic of high-utility k-anonymous microaggregation algorithms, as they analyze distances between pairs of data points. In fact, we shall see that our proximal-distal prepartition technique, combined or not with PCA, represents a valuable contribution by itself.

HIGHLIGHTS

- The primary goal of this work is to reduce the running time of k-anonymous microaggregation algorithms operating on datasets with a large quantity of numerical demographic attributes, acting as quasi-identifiers. Principal component analysis (PCA), an algebraic-statistical procedure that constructs an orthogonal projection onto a lower-dimensional subspace, permits the effective reduction of the number of attributes of the original dataset. The optimality principles of multivariate PCA strive to preserve Euclidean distances between the projected data points.
- The compressed data is fed to the microaggregation algorithm, but the k-anonymous microcells or groups obtained are directly applied to the original data. The distance-preservation properties of multivariate PCA help construct a micropartition of the set of respondents similar to that obtained when the original data is microaggregated in the conventional fashion, but in fewer dimensions.
- This means that we are able to achieve significant time gains ($\approx$ 14–31%) with very little impact on information utility (<2%, with respect to the total variance) with respect to the traditional procedure on the original data.
- Additional variants of the above method are devised and analyzed with extensive experimentation on standardized datasets, in terms of running time and information loss, pushing the already substantial speed-up even further ($\approx$ 48–64%), with mild distortion impact (<3%, with respect to the total variance).

Fig. 4. Highlights of our contribution.

- Although our work is illustrated with the special case of the widely used algorithm known as maximum distance to average vector (MDAV), we must hasten to point out that the computational improvements proposed in our work are compatible with any other multivariate microaggregation algorithm. The focus of this work is not on the particular k-anonymous microalgorithm algorithm employed, as we operate merely under two mild assumptions. First, that the running time significantly increases with the number of numerical quasi-identifiers in the microdata set, and secondly, although this is only assumed in our second method, that running times are superadditive with the number of records. We would like to remark that in the case of MDAV, running time is approximately an affine function of the amount of numerical quasi-identifiers, that is, the dimension of our dataset, but any form of increasing monotonicity would be sufficient for our methods to prove valuable. In terms of the number of records, MDAV is known to have approximately quadratic complexity, and therefore, superadditive.
- Further, these improvements do not require any modification of the internal code of said microaggregation algorithm, as they simply resort to modifying the representation of the data fed to it. In this manner, most conceivable computational improvements to the microaggregation process should be compatible with the dimensionality-reduction techniques put forth in this manuscript, finally resulting in a synergically multiplicative speed-up.
- We verify the expected performance of the two methods devised, by carrying out extensive experimentation on standardized datasets, illustrated with the popular algorithm MDAV. One of the datasets, "Forest" was intentionally chosen as a challenge to PCA, owing to the strong linear independence of its attributes. For the other main dataset, "Large Census", a widely popular choice in the SDC literature, we discovered high quasi-identifier redundancy, leading to substantial dimensionality reduction with negligible impact on distortion.

The potential applicability of this work encompasses information systems designed for the collection, analysis or dissemination of large amounts of anonymized data with a significant number of numerical quasi-identifiers, typically demographic attributes. The ulterior purpose is permitting the swift release of large amounts of data between organizations, departments, or to the public, for statistical study, in contexts including, but not limited to, socioeconomics, healthcare, targeted advertising, personalized content recommendation, social networks, and politics. A conceptually summarized list of assumptions and applicability of this work is provided in Fig. 5.

It is important to stress that our method widely applies to any anonymization method with time or memory complexity increasing with the number of numerical attributes. The choice of k-anonymous microaggregation is chiefly motivated for its simplicity, as we keep our focus on efficiency rather than on the specific privacy criteria. Certainly, k-anonymity as a privacy criterion is not without flaws, as we explain in our review of the state of the art, in §II. On

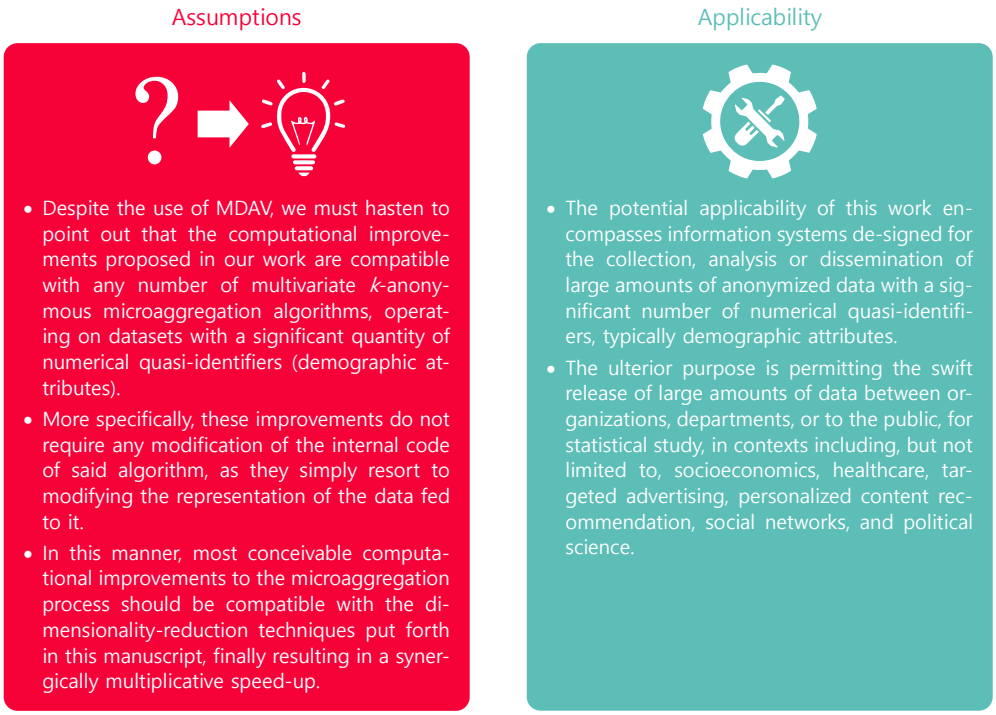


Fig. 5. Conceptually summarized list of assumptions and applicability of this work.

the flip side, stronger privacy guarantees have a price in information loss. For practical implementations beyond the scope of the research conducted here, depending on the privacy and the utility requirements of the application at hand, and whether data is to be released or accessible via online querying, one anonymization approach may be preferred over another. Specifically, high-utility approaches such as k-anonymous microaggregation may be preferred when data accuracy is critical, for offline data release. But they may be discarded in favor of alternatives with stricter privacy guarantees, such as l-diversity, t-closeness for data release, or *differential privacy* for online querying, at the expense of significant utility loss. This point, marginal to the intended focus of our contribution, is nevertheless discussed further in §II.

The paper is structured as follows. §II gives an overview of the state of the art on k-anonymous microaggregation. §III outlines the theoretical foundation of the application of principal component analysis to k-anonymous microaggregation. The algorithms developed are presented in §IV, while §V reports our experimental results. Finally, conclusions are drawn in §VI.

II. State of the Art on k-Anonymous Microaggregation

Next, we proceed to give a brief review of the state of the art on k-anonymous microaggregation more pertinent to this work, focusing on the methods and algorithms used to perform k-anonymous microaggregation while mitigating data utility loss. In addition, a critical view of k-anonymity and of its variants is provided. Later, in §III, we shall present a review of the theoretical foundations of multivariate numerical k-anonymous microaggregation, followed by the specifics of the microaggregation algorithm employed.

A. Methods and Algorithms for k-Anonymous Microaggregation

A number of algorithms for microaggregation have been developed, with the goal of minimizing the perturbation of the key attributes with accordance to a variety of distortion measures, while meeting a given k-anonymity constraint.

As multivariate microaggregation is known to be NP-hard [27], several heuristic methods have been proposed, which can be categorized into fixed-size and variable-size methods, according to whether all aggregated groups but one have exactly k elements. The maximum distance (MD) algorithm [9] and its less computationally demanding variation, the maximum distance to average vector (MDAV) algorithm [8, 11, 13, 43], are fixed-size algorithms that perform particularly well in terms of the distortion they introduce, for many data distributions. Popular variable-size algorithms include the μ-Approx [10], the minimum spanning tree (MST) [16], the variable MDAV (VMDAV) [35] and the two fixed reference points (TFRP) [4] algorithms. Efforts to circumvent the complexity of multivariate microaggregation exploit projections onto one dimension, but are reported to yield a much higher disclosure risk [26].

Research on microaggregation algorithms has continued recently. In particular, an approach recommends creating clusters of k records according to their densities [18]. Still in the case of perturbative algorithms, [21] contemplates the partition of the original dataset into several projections such that each projection satisfies the k-anonymity requirement, with the help of genetic algorithms. A well-known alternative to perturbative algorithms is the generation of synthetic

data that preserves some pre-established statistics of the original dataset. The combination of perturbed and synthetic data is exactly the approach followed by [7], which proposes a method for the generation of hybrid data through microaggregation.

More recently, an analysis of theoretical optimality in k-anonymous microaggregation [29] extends the necessary (not sufficient) optimality conditions that gave rise to the Lloyd-Max algorithm [19, 23], a celebrated quantization method for lossy data compression, also known as the k-means method in the areas of statistics and computer science. The properties of theoretical optimality and the excellent behavior of the Lloyd-Max algorithm in practice motivated the conception of the probability-constrained Lloyd (PCL) algorithm [29–31], which additionally incorporates a variation of the Levenberg-Marquardt algorithm [25], in order to adjust cell sizes. PCL is capable of outperforming even the popular MDAV in terms of distortion, typically by a reduction in MSE of roughly 10–30%, under the same exact k-anonymity constraint, for a wide variety of synthetic and standardized datasets [31]. Unfortunately, the distortion improvement offered by PCL comes at the expense of increased mathematical sophistication, which translates into a significantly costlier implementation and a substantially longer running time.

Due to its excellent performance in numerical microaggregation within reasonable computational demands, the most widely used fixed-size microaggregation algorithm for numerical data is the aforementioned MDAV, employed in this work to illustrate our novel methods. For reproducibility, we give the precise specification of MDAV used here in §III-B, formalized as Algorithm A, which is a functionally equivalent simplification of Algorithm 5.1 in [11], referred to as "MDAV-generic". We also comment later in that section on the computational complexity of the algorithm chosen, in terms of both the number of attributes and the number of records.

Certainly, other flavors of microaggregation problem exist, with their corresponding algorithms. In that regard, we would like to mention two intriguing extensions on the traditional k-anonymous microaggregation setup. First, the k-anonymity criterion as a measure of privacy can be given a probabilistic twist in order to encompass the more general case of uncertain respondent participation [33]. Secondly, the usual mean squared error as a measure of utility can be extended to include a Lagrangian term accounting for the degradation of statistical dependence between quasi-identifiers and confidential attributes [32].

B. Computational improvements for k-Anonymous Microaggregation

Two studies concerning computational improvements for k-Anonymous microaggregation have been identified. Firstly, in [24], authors improved the efficiency of MDAV by developing algebraic modifications that enables the use of the basic linear algebraic subprograms (BLAS), for its efficient parallel computation on CPU, where no additional distortion is incurred at all. Secondly, in [34], authors tackle the need of running k-Anonymous microaggregation efficiently with soft distortion loss, dealing with the fact that the data may arrive over an extended period of time. Additionally, in [34], authors have presented a detailed mathematical formulation which gives them the ability to compute the optimal time for the fastest optimization as well as for minimum distortion under a given deadline. As we can observe, [34] attacks different kind of problem as the one treated in this paper, concerning computational improvements for k-Anonymous microaggregation. However, regarding [24], we can stress on the fact that is totally compatible with the work done in this manuscript (i.e., introducing the dimensionality-reduction methods in the microaggregation field by means of PCA), and even more, they could be joined together to achieve even higher remarkable reductions in running time and memory usage than the one obtained separately, with negligible impact in information utility.

C. Previous Uses of Principal Component Analysis in the Context of Statistical Disclosure Control

We have identified two studies bearing some relation to our approach, with essential differences which we proceed to describe next. In the first study [22, 28], a single principal component was used for scalar microaggregation, resulting in faster execution, but at the cost of severe loss in data utility. In the cited work, the data is simply projected onto a single axis, the principal component with dominant eigenvalue, thereby representing the best univariate approximation to the data. We shall see in our own investigation that the multivariate approach involving several components rather than one is far superior in performance, turning an impractical application of PCA into an excellent method for SDC. Readers familiar with principal component analysis may quickly refer to Fig. 9 in §III, which represents the normalized energy (information) for the Large Census dataset, stored across dimensions. We shall see that we need to take into account 6 to 8 dimensions out of the original 13 dimensions in order to keep 92 to 98% of the total normalized energy. In contrast, the rather simplistic approach in the cited work is limited to the first component, barely containing 56% of the total energy.

In the second loosely related work [3], the author proposes a method for limiting disclosure in continuous microdata based on principal components analysis. Strictly speaking, the goal and investigation of the cited work differ dramatically from the approach presented here, but we opted to mention it in order to prevent confusion:

- The author's focus is not on computational efficiency but on limiting the risk of statistical disclosure.
- Rank swapping is used in lieu of microaggregation, a completely different strategy.
- The author's strategy uses one or a few principal components at a time.

- Finally, authors do not take into account the normalized energy per dimension, an approach that we did and was useful to choose a subset of principal components while preserving a high percentage of the total energy.

D. A Critical View of k-Anonymity and of its Variants

Despite the popularity of k-anonymity as a privacy measurement criterion in the SDC community, this criterion is based entirely on processing the quasi-identifiers and it is important to stress that it does not always prevent the disclosure of confidential attributes.

In some cases, confidential attributes may be repeated or too similar. Revisiting the example presented in Fig. 3, an attacker who may know the educational stage, age, and ZIP code of one of the three individuals belonging to the second cluster knows that his or her family income is in the range from \$54,100 to \$57,400, fairly similar values. This inference is known as *homogeneity* or *similarity attack*. The attack is often formulated in qualitative terms as a privacy deficiency of k-anonymity.

Observe however that in practice, the severity of a homogeneity attack depends on the prevalence of the sensitive values of the confidential attributes, and the microcell size k. For example, the prevalence of type-2 diabetes in the general population in the U.S. is close to 9%, but for senior citizens 65 years and older that figure may rise to more than 25%. For $k = 10$, the risk of homogeneity attack in a microcell corresponding to aged individuals, with $p = 1/4$, can be coarsely estimated as $p^k = 1/1,048,576$, less than once in a million. For microcells representative of the average population, with p=9%, the risk is even lower. And even for high prevalence nearing $p = 1/2$ in symmetric studies, $p^k = 1/1024$.

In order to mitigate this kind of attack, certain countermeasures, such as *p-sensitive* k-anonymity [39, 44], have been proposed This stronger requirement advocates for at least p different values for each confidential attribute within each microcell. Although privacy is improved, it comes at the price of data utility. A slight generalization of this concept was introduced in [14, 20], and termed *l-diversity*. It requires at least l "well-represented" confidential attributes. Depending on the definition of well represented, l-diversity can be reduced to p-sensitive or be more restrictive, again at the expense of higher information loss.

Other attacks against k-anonymity, of a more probabilistic nature, known as *skewness attacks*, exploit the discrepancy between the distribution of confidential attributes of the entire table, or the population, and the distribution within a given k-anonymous cell. In the hypothetical example in Fig. 3, suppose that it is widely known in the country of reference that 33% of the entire population has a family outcome above \$33,000. A privacy attacker looks for a female individual aged 32 and resident in the area with ZIP code 94024. The attacker notes that there is a 66% probability that this individual has a family income above \$33,000, which is well above the population's average.

One of the best-known palliative measures against this probabilistic risk is the *t-closeness* criterion [17], which requires that the distribution of a confidential attribute inside a given cluster be similar to the distribution of the overall dataset. More recently, *differential privacy* [5, 12] emerged as a proposal with strong privacy guarantees, but at considerable cost in data utility and conceived for online querying rather than microdata release. Although this work deals exclusively with microdata release, for offline use, the differential privacy may be implemented as a form of t-closeness, as described in [37].

Strongly restrictive privacy criteria such as t-closeness, or differential privacy under the representation in [37], require that the within-cell probability be similar to that of the table or the general population. However, unveiling the absence or low prevalence of a sensitive condition below the population's average may pose no privacy risk. In the above diabetes example, a cell comprising only healthy individuals may be acceptable from a privacy perspective. In general, privacy criteria are the object of ongoing investigation, and while k-anonymity may produce excellent utility with limited privcy guarantees, t-closeness and differential privacy may be stricter than required in some applications, deteriorating utility unnecessarily. Another critique on the overprotection of differential privacy can be found in [38].

It is essential to bear in mind the general principle that stronger privacy criteria come at the expense of a higher price on data utility. Hence, these restrictive flavors of k-anonymity must be employed with caution in applications where data utility is critical, as in certain medical studies directed toward the diagnosis and treatment of serious ailments, or might simply be rendered inapplicable.

To complete our basic description of privacy attacks, we would like to remark that an attacker can gain further insight if he is equipped with certain side information. In the synthetic example of Fig. 3, imagine that the attacker knows that the individual is an African-American male aged 34 who lives in the area with ZIP code 94305. Suppose that external demographic studies pointed out that African-Americans of this age were having a family income less than \$30,000. The attacker could discard 2 out of 3 records and guess the individual's consumer profile. This form of statistical inference is known as *background knowledge attack*. These kind of attacks are studied in [40], where the authors propose strategies based on graph theory and inference paths.

Although the methodology proposed in this work is illustrated with k-anonymous microaggregation, it is readily extensible to most of the variants aforementioned.

III. Theoretical Foundation of the Application of Principal Component Analysis to k-Anonymous Microaggregation

This section introduces the basic notation employed and the fundamentals of multivariate numerical k-anonymous microaggregation. Subsequently, it describes our main assumptions and the specifics on the k-anonymous microaggregation algorithm MDAV, chosen to illustrate this work. Next, it reviews the basic principles of PCA, explaining its application to the problem at hand. The following section, §IV, proceeds to present the two methods for efficient anonymization of large-scale datasets put forth in this work.

A. Basic Notation and Fundamentals of Multivariate Numerical k-Anonymous Microaggregation

The scope of our study encompasses the important case when the n records of the dataset contain tuples of m numerical quasi-identifiers, representable as n points $(x_j)_{j=1}^n$ in the m-dimensional Euclidean space $\mathbb{R}^m$. We shall employ two convenient representations of the dataset: first, as a matrix, and secondly, as a *random vector* (r.v.). Under the former representation, we take the *data matrix* $X \in \mathbb{R}^{m\times n}$ as the collection of n column vectors $(x_1, \ldots, x_n)$, each with m real-valued entries. Under the latter statistical representation, X is an r.v. taking on values on the set $\{x_1, \ldots, x_n\}$ of m-dimensional points uniformly at random. The slight albeit intended abuse of notation by reusing the same symbol X to denote both a matrix and an r.v. should pose no ambiguity when considered in its context.

Either representation shall be used interchangeably according to its convenience in the context at hand. For example, we may write the arithmetic average $\frac{1}{n}\sum_{j=1}^n x_j$ as a matrix product $X\frac{1}{n}1$, where $\frac{1}{n}1$ represents the uniform probability (column) vector, or more compactly as a probabilistic expectation $\mathrm{E}\, X$. Under the assumption of zero-mean normalization, the *covariance matrix* $\Sigma_X \in \mathbb{R}^{m\times m}$ can be written indistinctly as

$$\Sigma_X = \frac{1}{n}\sum_{j=1}^{n} x_j\, x_j^T = \frac{1}{n} XX^{\mathrm{T}} = \mathrm{E}\, XX^{\mathrm{T}},$$

where the matrix form $\frac{1}{n}XX^{\mathrm{T}}$ follows immediately from the outer product interpretation of matrix multiplication.

We mentioned in the introductory review of k-anonymous microaggregation that practical algorithms are designed to perturb quasi-identifiers in a way such that the statistical quality of the published data is guaranteed. Technically speaking, microaggregation is similar to a quantization problem: the algorithms find a partition of the set of quasi-identifying tuples in cells of k elements, and try at the same time to reduce the distortion incurred when replacing each element in a cell by its representative within this cell. Fig. 6 conceptualizes k-anonymous microaggregation as minimum-distortion vector quantization, with the added restriction that cells be at least of size k. The function $c(j)$ assigns the quasi-identifier tuple x_j to microcell c, which will contain at least $k-1$ other points, and whose value will be replaced by the common reconstruction tuple or centroid $\hat{x}(c)$.

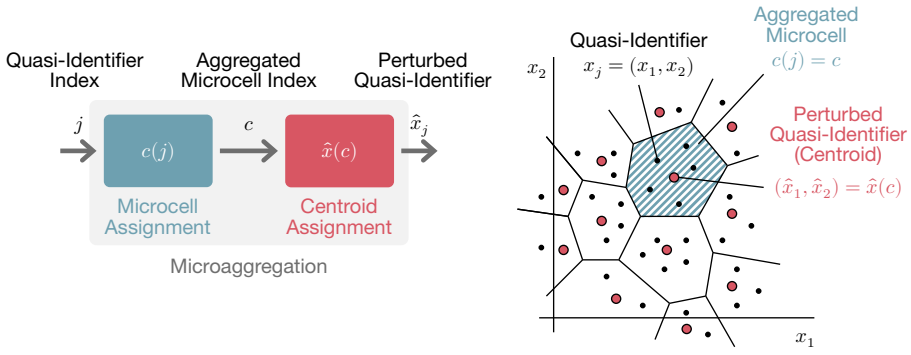


Fig. 6. Traditional microaggregation interpreted as a quantization problem on the record indices j, represented by a microcell assignment function $c(j)$, together with a centroid assignment function $\hat{x}(c)$ that reconstructs the perturbed version $\hat{x}_j$ of the original quasi-identifier x_j. The figure also shows an example of microaggregation of 2-dimensional quasi-identifiers with anonymity parameter $k = 5$. Each microcell of five points is assigned a representative centroid. In this example, the two-dimensional quasi-identifiers could correspond to a pair of demographic attributes such as age and number of school years. The centroids are the value of the published, perturbed quasi-identifiers within each cell.

Recall the common practice in SDC of normalizing each quasi-identifier for zero mean and unit variance. *Zero-mean normalization* is merely a convenience that facilitates the computation of variances. *Unit-variance columnwise normalization* is essential for perturbation errors inherent in the microaggregation process to remain invariant with respect to arbitrary choices of units, say pounds or kilograms for weights, and inches or meters for heights. Normalization also confers equal importance to all quasi-identifiers. Of course, reweighting is possible in applications where certain quasi-identifiers are deemed of greater importance than others in the quantification of data utility. This normalization also means that the usual measure of *distortion*, precisely, the ratio between the sum of squared errors

$$\mathrm{SSE} \stackrel{\mathrm{def}}{=} \sum_{j=1}^{m} \|x_j - \hat{x}_j\|^2$$

and the sum of squares total

$$\mathrm{SST} \stackrel{\mathrm{def}}{=} \sum_{j=1}^{n} \|x_j\|^2 = mn,$$

matches the usual definition of distortion in the field of vector quantization, as *mean squared error* (MSE) normalized by dimension:

$$\mathcal{D} \stackrel{\mathrm{def}}{=} \frac{\mathrm{SSE}}{\mathrm{SST}} = \frac{1}{mn} \sum_{j=1}^{n} \|x_j - \hat{x}_j\|^2 = \frac{1}{m} \mathrm{E}\, \|X - \hat{X}\|^2.$$

A discussion of the optimality conditions of k-anonymous microaggregation can be found in [29]. Let $n(c) \geqslant k$ denote the size of microcell c. While it is well known that the centroid or conditional expectation

$$\hat{x}(c) = \frac{1}{n(c)} \sum_{j \,|\, c(j)=c} x_j = \mathrm{E}\,[X|c]$$

minimizes the MSE within each microcell, and thus it constitutes the optimal reconstruction for a given microcell assignment function $c(j)$, the problem of constructing such microcell assignment, under the restriction that it contain at least k points, may prove difficult. In practice, the k-anonymous microaggregation algorithm MDAV, introduced in our review of the state of the art in §II, is an excellent heuristic in terms of MSE, with manageable computational complexity in small datasets. In fact, our choice of MDAV is motivated by its performance with respect to the sophisticatedly optimized method known as *probability-constrained Lloyd* (PCL) algorithm [29, 31], being MDAV slightly above in terms of distortion, albeit with far less computation.

B. Main Assumptions and Specifics on the k-Anonymous Microaggregation Algorithm Employed

As we mentioned in §I-B, although this work is illustrated with the specific k-anonymous microaggregation algorithm MDAV, the methods devised are readily applicable to other anonymization algorithms under two mild assumptions on their time complexity. First, that their running time monotonically increases with the number m of quasi-identifiers, and secondly, albeit only for the second method, that their running time is superadditive in the number of records n.

We also recalled that *superadditive complexity* in the number of records means that the running time on $a+b$ records satisfies

$$t(a+b) \geqslant t(a) + t(b),$$

making it conducive to the celebrated algorithmic approach of "divide and conquer". We mentioned that superadditivity is a typical characteristic of high-utility k-anonymous microaggregation algorithms, as they carefully analyze Euclidean distances between pairs of data points. We would like to illustrate the elegant potential of this property, somewhat counterintuitive at first, with a very simple example. Suppose that an anonymization algorithm requires an amount of time $t(n) = n^2$ in order to process n records (in time units relative to $t(1)$). Suppose further that we split the dataset into two portions of $n/2$ records each, process each part separately, and reassemble the result. Provided that the cost of splitting and recombination were negligible, the total time required by this strategy would be $2(n/2)^2 = n^2/2$, in words, half the time required to process the entire recordset in a single pass. As well shall see, the second method described in this paper exploits superadditivity in a single division of the dataset, without further recursion, in addition to PCA, in a synergic manner. Future work may certainly consider further, progressive recursion[(a)].

For reproducibility, we give our specification of MDAV, formalized as Algorithm A, which is a functionally equivalent simplification of Algorithm 5.1 in [11], referred to as "MDAV-generic". Simple inspection of its pseudocode leads to the conclusion that its running-time complexity in terms of the number m of quasi-identifiers is approximately affine, that is, the running time t of the algorithm is of the form $t(m) = m_0 + m$ (in appropriate time units), for some constant m_0. In terms now of the number n of records, it is also straightforward to show that the running time t of the algorithm grows asymptotically as $t(n) = n^2/k$ (in time units relative to the case $n^2/k = 1$) for $n \gg k$, that is, it has quadratic complexity and therefore, superadditive running time.

C. Basic Principles of PCA and Application to k-Anonymous Microaggregation

We offer a brief review the basic principles behind PCA [15] and describe the main ideas in its application to k-anonymous microaggregation. Consider a set of n points x_j in the m-dimensional Euclidean space, representing the tuples of numerical quasi-identifiers of the microdata set with n records. The goal of PCA is to find a linear subspace of dimension $\tilde{m} < m$ and a set of points $\hat{x}_j$ within said subspace that approximate the original data x_j.

Consider any orthonormal basis of the subspace, rearranged as a sequence of column vectors in the form of a matrix $\tilde{U} \in \mathbb{R}^{m \times \tilde{m}}$. Since $\tilde{U}$ is columnwise orthonormal, $\tilde{U}^{\mathrm{T}}\tilde{U} = I$. Each approximation $\hat{x}_j$ can be written as a unique linear

[(a)] Even though further recursion does not immediately apply to our method, it is still interesting to observe its effect on the idealized algorithm used in this example. Progressively recursive application of the "divide and conquer" strategy for such an idealized algorithm, satisfying $t(n) = 2\,t(n/2) + O(1)$, according to the master theorem [6], would yield linear complexity $t(n) = \Theta(n)$. If the cost of splitting and recombination were not asymptotically negligible, but linear instead, that is, $t(n) = 2\,t(n/2) + \Theta(n)$, then a recursive implementation would have complexity $t(n) = \Theta(n \log n) = o\,(n^{1+\epsilon})$ for any $\epsilon > 0$.

Algorithm A: MDAV "generic", functionally equivalent to Algorithm 5.1 in [11].

function MDAV
input k, $(x_j)_{j=1}^n$ ▷*Anonymity parameter k, quasi-ID portion $x_1, \ldots, x_n \in \mathbb{R}^m$ of a dataset of n records*
output q ▷*Assignment function from records to microcells $j \mapsto q(j)$*

while $2k$ points or more in the dataset 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 dataset
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*

combination of this base $\hat{x}_j = \tilde{U}\tilde{x}_j$, thus determining what we shall call the compressed vector $\tilde{x}_j \in \mathbb{R}^{\tilde{m}}$, on account of its reduced number of components. Each compressed point is not a selection of dimensions, but a linear combination of dimensions. This means that a principal component may represent a combination of several quasi-identifiers, for example, height and weight, with multiplicative weights given by the corresponding basis, real valued and possibly negative.

The precise optimization objective of PCA is the MSE

$$\frac{1}{n}\sum_{j=1}^{n} \|x_j - \hat{x}_j\|^2 = \mathrm{E}\,\|X - \hat{X}\|^2.$$

The minimization of this objective implies that the approximations $\hat{x}_j$ must be the orthogonal projections of x_j onto the subspace determined by the basis $\tilde{U}$. In this manner, the problem of minimizing the MSE is now reduced to finding the appropriate columnwise orthonormal matrix $\tilde{U}$, for a desired reduced dimension $\tilde{m}$. It has been long established that the optimal choice for $\tilde{U}$ is the set of $\tilde{m}$ dominant eigenvectors of the covariance matrix Σ_X, associated with the dominant eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_{\tilde{m}}$ in its spectral decomposition $\Sigma_X = U\Lambda U^{\mathrm{T}}$, with U orthonormal and $\Lambda = \operatorname{diag}(\lambda_i)_{i=1}^m$. Furthermore, the principal components thus obtained are uncorrelated.

The cost of computing the covariance matrix scales linearly with the number of records n. The cost of the solution to the PCA problem given this matrix only depends on the number of dimensions m, typically much smaller than n for most microdata sets, solution which is swiftly computed by extremely efficient algebraic algorithms. Since microaggregation is usually superlinear in the number of records n, the overall cost of PCA, including the computation of Σ_X, is utterly negligible. For a perfectly fair comparison with traditional microaggregation, the experiments in this manuscript most certainly take into consideration this additional time, however insignificant.

Recall that an orthogonal projection for a given basis orthonormal can be computed in two steps. First, we obtain the compressed components $\tilde{x}_j = \tilde{U}^{\mathrm{T}}x_j$, and then we reconstruct the projections via $\hat{x}_j = \tilde{U}\tilde{x}_j = \tilde{U}\tilde{U}^{\mathrm{T}}x_j$, where $P = \tilde{U}\tilde{U}^{\mathrm{T}}$ is the associated projection matrix. Often, and this will also be our case, it suffices to work with the compressed versions, and the reconstruction is unnecessary. A compact way to write the compression formula in terms of the associated data matrix $X \in \mathbb{R}^{m\times n}$ and its compressed analogue $\tilde{X} \in \mathbb{R}^{\tilde{m}\times n}$ is $\tilde{X} = \tilde{U}^{\mathrm{T}}X$.

Since PCA is an orthogonal projection, the projection error $x_j - \hat{x}_j$ is orthogonal to the projection $\hat{x}_j$, and the Pythagorean identity

$$\|x_j\|^2 = \|\hat{x}_j\|^2 + \|x_j - \hat{x}_j\|^2$$

holds. If the error is small, as it is customarily the case, the norm of the original vector is preserved, that is, $\|x_j\|^2 \approx \|\hat{x}_j\|^2$. Since $\tilde{U}$ has orthonormal columns by construction and $\hat{x}_j = \tilde{U}\tilde{x}_j$, it follows immediately that

$$\|\hat{x}\|^2 = \tilde{x}_j^{\mathrm{T}}\tilde{U}^{\mathrm{T}}\tilde{U}\tilde{x}_j = \|\tilde{x}\|^2.$$

More generally, orthonormal reconstructions preserve inner products and norms, a fact intuitively consistent with their interpretation as combinations of rotations and reflections. This means that the distances between compressed samples match distances between projected points, which in turn approximate the corresponding distances between original vectors. As a consequence, any geometric computation can be carried out approximately in the compressed space, in less dimensions. This is, in fact, the key to our proposal, as the k-anonymous microaggregation algorithm will work with distances between compressed points, instead of distances between the original points. The computation of those distances will require a reduced number of operations, based on $\tilde{m} < m$.

The representation of the microdata set as the data matrix X and the compression matrix $\tilde{U} \in \mathbb{R}^{m \times \tilde{m}}$ are shown in Fig. 7. The compressed version $\tilde{X}$ of the dataset as a matrix is represented in Fig. 8, with emphasis on the orthogonality of the projections $\hat{x}_j$ and their errors $x_j - \hat{x}_j$.

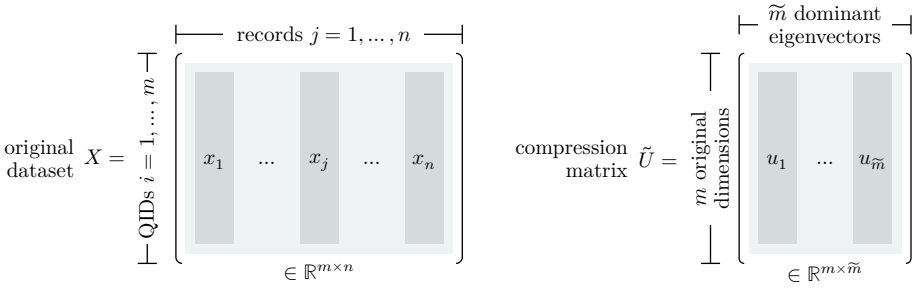


Fig. 7. Dataset interpreted as a series of n vectors $x_1, \ldots, x_n$ in the m-dimensional Euclidean space $\mathbb{R}^m$, then expressed as a matrix $X \in \mathbb{R}^{m \times n}$, where m is the number of numerical quasi-identifiers, and n the number of records. The compression matrix $\tilde{U} \in \mathbb{R}^{m \times \tilde{m}}$ retrieves the compressed version $\tilde{X}$ of the dataset, also arranged as a matrix, but consisting of vectors of a lower dimension $\tilde{m} < m$. Each of these compressed dimensions are linear combinations of the m original quasi-identifiers, typically linear combinations of demographic attributes.

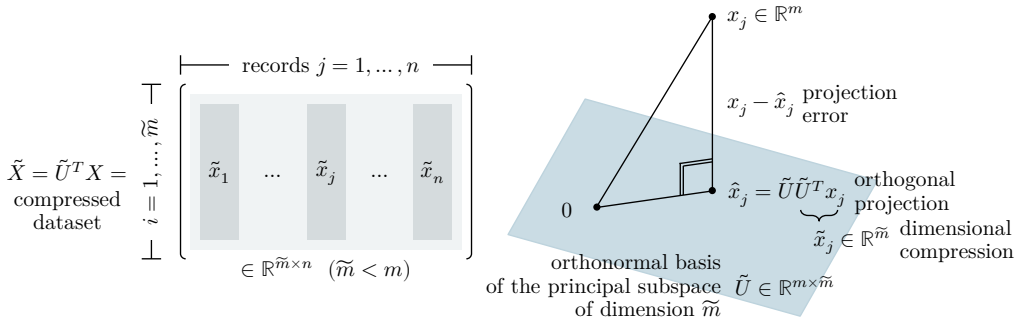


Fig. 8. Compressed version $\tilde{X}$ of the original dataset X, with data vectors in the lower-dimensional Euclidean space $\mathbb{R}^{\tilde{m}}$, with $\tilde{m} < m$. Since PCA is an orthogonal projection with projection matrix $P = \tilde{U}\tilde{U}^{\mathrm{T}}$, the projection error $x_j - \hat{x}_j$ is orthogonal to the projection $\hat{x}_j$, and the Pythagorean identity $\|x_j\|^2 = \|\hat{x}_j\|^2 + \|x_j - \hat{x}_j\|^2$ holds. If the error is small, as it is customarily the case, the norm of the original vector is preserved, that is, $\|x_j\|^2 \approx \|\hat{x}_j\|^2$. Since $\tilde{U}$ has orthonormal columns by construction and $\hat{x} = \tilde{U}\tilde{x}$, it follows immediately that $\|\hat{x}_j\|^2 = \|\tilde{x}_j\|^2$. This means that the distances between compressed samples approximate the corresponding distances between the original vectors, and any geometric computation can be carried out approximately in the compressed space, in less dimensions.

Regarding the choice of the number $\tilde{m}$ of principal components, one may equivalently impose a condition on the quality of the approximation instead, in the sense that additional components will yield a lower MSE in the PCA problem. More precisely, the MSE in the approximation by $\tilde{m}$ principal components is simply the sum of residual eigenvalues

$$\mathrm{E}\,\|X - \hat{X}\|^2 = \sum_{i=\tilde{m}+1}^{m} \lambda_i,$$

and the individual unit-variance normalization of each original dimension implies that the total energy of the dataset, which in general is the sum of all eigenvalues, becomes the number m of original dimensions:

$$\mathrm{E}\,\|X\|^2 = m = \operatorname{tr}\Sigma_X = \operatorname{tr}\Lambda = \sum_{i=1}^{m} \lambda_i.$$

Therefore, the relative error ϵ^2 in the PCA approximation is

$$\epsilon^2 \stackrel{\text{def}}{=} \frac{\mathrm{E}\,\|X - \hat{X}\|^2}{\mathrm{E}\,\|X\|^2} = 1 - \frac{\mathrm{E}\,\|\hat{X}\|^2}{\mathrm{E}\,\|X\|^2} = \frac{\sum_{i=\tilde{m}+1}^{m} \lambda_i}{\sum_{i=1}^{m} \lambda_i} = 1 - \frac{\sum_{i=1}^{\tilde{m}} \lambda_i}{\sum_{i=1}^{m} \lambda_i} = 1 - \frac{1}{m}\sum_{i=1}^{\tilde{m}} \lambda_i \ \in [0, 1].$$

Notwithstanding the theoretical appeal of PCA, it remains to demonstrate that practical microdata sets may offer a substantial reduction $\tilde{m} < m$ in the number of quasi-identifiers, and at the same time, that the distances between compressed samples approximate the corresponding distances between the original vectors, which we may quantify through a small relative energy loss ϵ^2. In this manner, we will ensure that MDAV can be carried out on the compressed space and yield similar micropartitions as the conventional anonymization procedure on the original data. Intuitively, the use of this technique will have greater impact in datasets where most of the information is stored in just a few quasi-identifiers; in other words, whenever there exists significant redundancy (linear dependence) between demographic attributes. Although the goal of the experimental section §V is precisely the demonstration that such redundancy indeed

exists and that our methods are able to exploit it efficiently, we would like to illustrate the point made here with an example of a standardized dataset.

Specifically, Fig. 9 shows a histogram of normalized eigenvalues $(\lambda_i)_{i=1}^m/m$ in order of decreasing dominance, along with the normalized cumulative energy function $\frac{1}{m}\sum_{i=1}^{j}\lambda_i$ for the Large Census dataset. The horizontal axis shows the number $\tilde{m}$ of cumulative principal components, ranging from the use of a single principal component $\tilde{m}=1$ with significant energy loss $\epsilon^2=1-\lambda_1$, all the way to $\tilde{m}=m$, for which all energy is preserved and thus $\epsilon^2=0$, but no effective dimensionality reduction occurs. Observe that the extreme case $\tilde{m}=1$ corresponds to the preliminary study in [2], where univariate PCA was hastily discarded as a practical approach due to its distortion overhead. However, quite remarkably, just 5 principal components out of the total of 13 dimensions suffice to capture $1-\epsilon^2\approx 89\%$ of the energy of the dataset. This means that multivariate PCA should achieve significant time reduction with a fraction of the dimensions and mild distortion overhead.

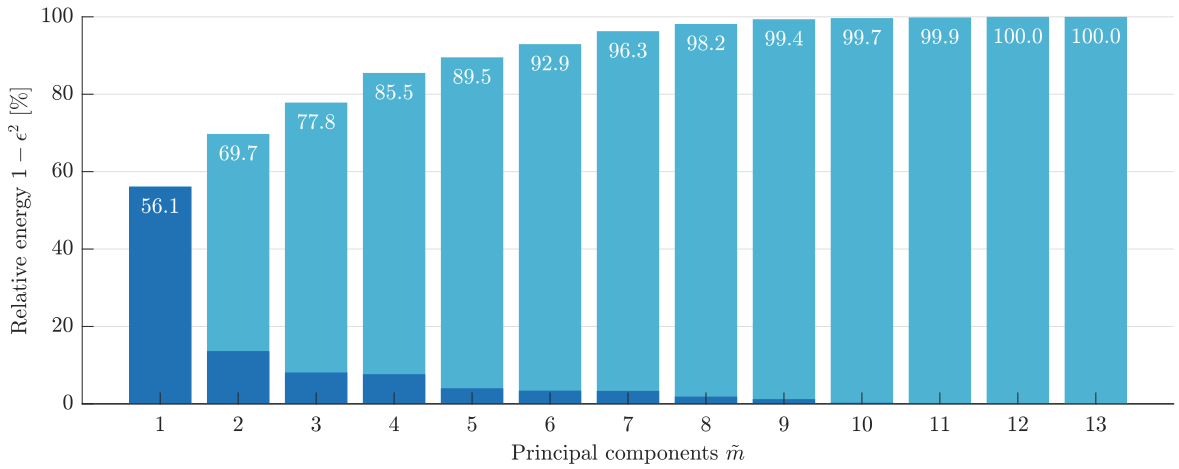


Fig. 9. The graphic shows the normalized energy per dimension and the cumulated energy per dimension for the Large Census dataset. The horizontal axis shows the number $\tilde{m}$ of cumulative principal components, ranging from the use of a single principal component $\tilde{m}=1$ with significant energy loss $\epsilon^2=1-\lambda_1$, all the way to $\tilde{m}=m$, for which all energy is preserved and thus $\epsilon^2=0$, but no effective dimensionality reduction occurs. Observe that 92.9% of the energy is retained by keeping only 6 principal components out of the total of 13 dimensions, and 99.4% with 9 out of 13.

IV. Two Novel Methods for Efficient k-Anonymous Microaggregation via Principal Component Analysis

Dimensionality reduction via PCA is a widely used technique in machine learning, precisely, as an unsupervised means to reduce the number of features involved in subsequent supervised logic, thereby reducing the complexity of the learning problem. From that perspective, reducing the total number of dimensions of the dataset comes naturally when trying to find efficient methods for data anonymization. Our goal is to use PCA for the reduction of dimensions of quasi-identifiers in large datasets, in order to attain a significant reduction in the running time required by k-anonymous microaggregation, at the expense of a slight degradation in data utility. This section builds on the theoretical foundation of the application of PCA to k-anonymous microaggregation exposed in the previous one, employing the notation, assumptions, and principles described there. Having reviewed said foundation, we may now turn to the description of the two novel methods devised in this work.

A. *MDAV with PCA*

Our first method consists in the direct application of PCA to MDAV, method which we naturally term *MDAV with PCA*, and which proceeds as follows.

1. We perform PCA on the m numerical quasi-identifiers of our microdata set, where the number $\tilde{m}<m$ of principal components may be given directly, or indirectly through a constraint on the maximum energy loss ϵ^2 allowed.
2. We carry out the k-anonymous microaggregation procedure with the algorithm of choice, on the compressed $\tilde{m}$-dimensional data rather than on the original m-dimensional vectors. Absolutely no change is required inside the code of the microaggregation algorithm. In our experiments, we run MDAV on the compressed vector space. The algorithm will simply act as if the data consisted of $\tilde{m}$ quasi-identifiers in lieu of of m, and simply run faster.
3. We argued in the previous subsection that the distances between the original data vectors will approximate the distances between the compressed versions. This due to the small magnitude of the relative error ϵ^2 assumed, and the orthonormality of the compression/reconstruction matrix $\tilde{U}$.

4. Typically, the running time of a microaggregation algorithm scales according to an affine function $t \approx a + b\,m$ of the number of quasi-identifiers, which is, incidentally, subadditive. For large values of the number of quasi-identifiers, this dependence may be approximately linear. The time gain will correspond then to the reduction in dimensions, that is, to the proportion between $\tilde{m}$ and m. In practice, whenever subadditivity applies, the time reduction should be somewhat smaller.
5. The k-anonymous microcells created for the compressed data are immediately applied to the original data. The computation of centroids and distortion could also be estimated in the compressed domain, but being a relatively swift process, it is preferred to compute both centroids and distortion more accurately in the original domain. The data released should of course be represented in the original domain.

Because the compressed data to be anonymized will not be a perfectly accurate version of the original data, the groups created by MDAV operating on the compressed space may differ slightly from those when MDAV is conventionally applied to the original data, without PCA. In practice, we shall expect a slight increment in distortion, fact that will be confirmed in the experimental section. On the other hand, for datasets where the energy is concentrated on a few components, we should attain a significant reduction in running time with negligible distortion degradation. This would be the case for the dataset with the spectral profile depicted in Fig. 9. The essential steps of our first method are succinctly outlined as Algorithm B. A summary in greater detail is offered in Fig. 10, which should serve as a convenient recapitulation of our proposal.

Algorithm B: MDAV with PCA.

1: Perform principal component analysis on the dataset X, obtaining $\tilde{X} = \tilde{U}^{\mathrm{T}} X$.
2: Execute MDAV on the dataset of reduced dimension $\tilde{X}$ in lieu of X, according to the specification Algorithm A, with absolutely no internal modification of the microaggregation algorithm.
3: Apply the microcell assignment function $j \mapsto q(j)$ found on the compressed dataset $\tilde{X}$ directly to the original dataset X, maintaining the same exact record indexing $j = 1, \ldots, n$, in order to obtain the reconstruction centroids for publication.

B. Proximal-Distal Prepartitioning and MDAV with PCA on Proximal Data

The key to our second method consists in splitting the original dataset into two parts. One part, which we shall call *proximal*, contains points closer to each other, contributing less to the overall distortion after the microaggregation process. The other part, called *distal*, contains the rest of points, further from each other. This partitioning offers a twofold advantage. First, PCA will only be applied to the proximal portion, less sensitive in terms of distortion, and the distal portion will remain dimensionally unchanged. As a second advantage, by force of the "divide and conquer principle", the superadditive running time of microaggregation will be significantly reduced. The object of this proximal-distal method is a synergy between PCA and prepartitioning, in hopes of attaining significant computation gains with negligible distortion cost.

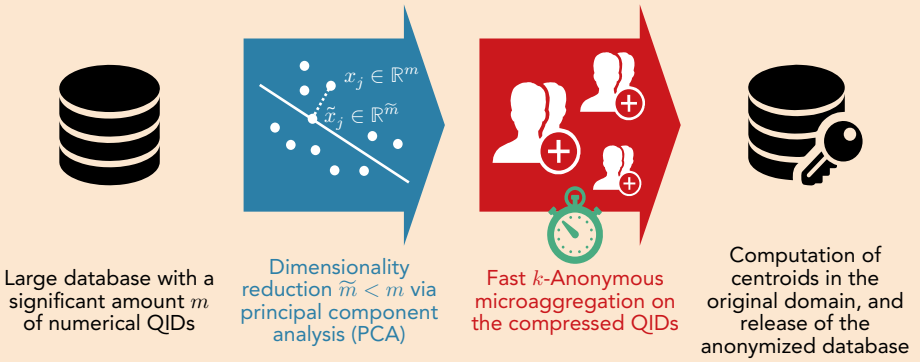
The excellent performance of MDAV for k-anonymous microaggregation makes it a suitable contender for the proximal-distal prepartitioning process introduced, which is as follows.

1. For a given macroaggregation parameter $K > k$, we first apply MDAV on the dataset of n records, to divide it into $\lfloor n/K \rfloor$ large macrocells of size K. For example, for a dataset of $n = 150,000$ records and for an anonymity parameter $k = 10$, choosing a macrocell size $K = 1000$ gives 150 macrocells. For large K in comparison with the anonymity parameter k, the computational cost n^2/K of this prepartition should be a negligible fraction of the cost n^2/k of applying MDAV to the entire dataset. In our example, macroaggregation will be $K/k = 100$ times faster.
2. Next, we compute the centroid and the mean squared error for each of the resulting macrocells, the latter constituting a reasonable indicator of the proximity of the data points contained.
3. Finally, we subdivide the points into two portions according to said macrocell distortion, a proximal part including points in macrocells with lower distortion, and a distal part with the remainder, either according to a distortion threshold, or equivalently, according to a predefined fraction ν of the number $\lfloor n/K \rfloor$ of macrocells to be categorized as distal.

We stress that the initial macroaggregation into macrocells of size K is only an (efficient) means to obtain this final prepartition into two portions. Certainly, alternatives such as the Lloyd algorithm or k-means method could be explored. Additionally, it remains to consider the impact on the choice of the parameters involved in this prepartition, namely the macroaggregation parameter K, and the fraction ν of distal points.

Equipped with a candidate implementation for the proximal-distal prepartitioning, we may turn to the specification of our second method, which is essentially the application of PCA only to the proximal portion, as detailed next.

Brief Recapitulation of k-Anonymous Microaggregation Accelerated through Principal Component Analysis



- In the following procedure we view the numerical quasi-identifiers as a series of m-dimensional data vectors x_j, for each record $j = 1, \dots, n$. The $m \times n$ data matrix X contains the data vectors arranged as columns.
- Zero-mean normalization of each of the m scalar quasi-identifiers. Additionally, as it is customary in the field of SDC, for scale-independent measurement of distances, normalize for unit-variance.
- Compute the $m \times m$ covariance matrix Σ, and its spectral decomposition $U \Lambda U^T$. (In practice, only a partial decomposition for the largest eigenvalues may be required.)
- Sort the eigenvalues in decreasing order, maintaining their correspondence with the eigenvectors. (A partial selection with Quickselect could prove faster for very large m.)
- Due to the unit-variance normalization carried out previously, $\operatorname{tr} \Lambda = \sum_{i=1}^{m} \lambda_i = \operatorname{tr} \Sigma = m$. Select the desired amount $\widetilde{m} \leqslant m$ of dominant eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_{\widetilde{m}}$ to satisfy a predefined constraint on the preserved normalized energy $1 - \epsilon^2 = \frac{1}{m} \sum_{i=1}^{\widetilde{m}} \lambda_i$, or on the approximation error $\epsilon^2 = \frac{1}{m} \sum_{i=\widetilde{m}+1}^{m} \lambda_i$ incurred. In practice, some datasets will allow a value of $\widetilde{m}$ significantly smaller than m, with very small error ϵ^2.
- Zero eigenvalues indicate that some of the original quasi-identifiers are exact linear combinations of others, and therefore effectively redundant.
- Compose the $m \times \widetilde{m}$ compression matrix $\tilde{U}$ consisting of the $\widetilde{m}$ dominant eigenvectors, arranged as columns, and compress the data vectors according to $\tilde{X} = \tilde{U}^T X$, where X is the data matrix, with data points also arranged as columns. The compressed vectors will contain $\widetilde{m}$ components instead of m, each an independent linear combination of the original data component. The new $\widetilde{m}$ quasi-identifiers are each a different mixture of the original m.
- Carry out the k-anonymous microaggregation procedure with the algorithm of choice, on the compressed $\widetilde{m}$-dimensional data rather than on the original m-dimensional vectors. Absolutely no change is required inside the code of the microaggregation algorithm. The algorithm will simply act as if the data consisted of $\widetilde{m}$ quasi-identifiers in lieu of of m, and run faster.
- It can be shown that the distances between the original data vectors will approximate the distances between the compressed versions. This due to the small magnitude of ϵ^2 assumed, and the orthonormality of the compression/reconstruction matrix $\tilde{U}$.
- Typically, the running time of a microaggregation algorithm scales according to an affine function $t \approx a + bm$ of the number of quasi-identifiers, which is in fact subadditive. For large values of the number of quasi-identifiers, this dependence may be approximately linear. The time gain will correspond then to the reduction in dimensions, that is, to the proportion between $\widetilde{m}$ and m. In practice, whenever subadditivity applies, the time reduction should be somewhat smaller.
- The k-anonymous microcells created for the compressed data are immediately applied to the original data. The computation of centroids and distortion could also be estimated in the compressed domain, but being a relatively swift process, it is preferred to compute them accurately in the original domain. The data released should of course be represented in the original domain.

Fig. 10. Brief recapitulation of our proposal to use principal component analysis (PCA) for the reduction of the dimensionality of the quasi-identifiers in k-anonymous microaggregation, which usually translates in a similar reduction in the running time of the microaggregation algorithm.

1. We split the original dataset into two portions, a proximal portion of data points closer to each other and a distal portion of data points far from each other, according to the proximal-distal prepartitioning method just outlined.
2. We perform PCA on the m numerical quasi-identifiers of our proximal portion microdata set, where the number $\tilde{m} < m$ of principal components may be given directly, or indirectly through a constraint on the maximum energy loss ϵ^2 allowed. PCA is *not* applied to the distal portion, which remains completely unchanged prior to microaggregation.
3. We carry out the k-anonymous microaggregation procedure with the algorithm of choice, on the compressed $\tilde{m}$-dimensional proximal dataset rather than on the original m-dimensional vectors as well as on the original distal portion of data points without having previously applied PCA on it, as mentioned in Step 2. Again, absolutely no change is required inside the code of the microaggregation algorithm. In our experiments, we run MDAV separately on both, the compressed vector space for the proximal dataset and the original distal dataset without compression.
4. It is worth remembering that the distances between the original proximal dataset vectors will approximate the distances between the compressed versions of it. This due to the small magnitude of the relative error ϵ^2 assumed, and the orthonormality of the compression/reconstruction matrix $\tilde{U}$. However, for the distal dataset, and as mentioned before, no PCA will be applied at all. The reason is that the data points of the distal dataset are far from each other and the result of applying PCA would not lead to a good approximation between the original distal dataset vectors and the distances between the compressed versions of it even if ϵ^2 is assumed to be small.

5. The k-anonymous microcells created for both, the compressed data of the proximal dataset and the uncompressed data of the distal dataset, are immediately applied to the original data. The computation of centroids and distortion is done in the original domain. The data released should of course be represented in the original domain.

We elaborate further on the prepartition into a proximal fraction $1-\nu$ of macrocells, and a distal fraction ν. Recall that PCA is only applied to the proximal portion, contributing less to the overall distortion, whereas the distal portion is microaggregated without dimensionality reduction. A further goal in our methodology is to find an adequate value for ν, or equivalently an adequate number of distal macrocells C_{D}, taking into consideration both the speed up and the distortion incurred. In our experiments, we shall simply report the results of scanning C_{D} from zero to the maximum number of possible cells $\lfloor n/K \rfloor$. For $C_{\mathrm{D}} = 0$ or $\nu = 0$, we reproduce our first method, MDAV with PCA, since the proximal dataset, where PCA will be applied, will be identical to the original dataset and the distal dataset will be empty. By contrast, for $C_{\mathrm{D}} = \lfloor n/K \rfloor$ or $\nu = 1$, we revert to the traditional use MDAV, since the distal dataset, where no PCA will be applied is defined to be the entire dataset, and the proximal dataset is empty. We may scan through all possible values of C_{D}, to explore the best value that fits with our abovementioned objective.

In conclusion, as the compressed data of the proximal dataset, to be anonymized, will be almost a perfect accurate version of the original proximal dataset thanks to that fact that the chosen points are close to each other, the cells that are created by a MDAV operating on the compressed space of the proximal dataset may probably not differ from those that could be obtained when a MDAV is operating on the original proximal dataset. Nevertheless, for the distal dataset, as the points can be far from each other, no PCA will be applied at all so no additional distortion due to compression will be incurred, then, only MDAV will be applied. In practice, by applying this method, we are expecting even a better performance result in terms of speed up and probably a slight decrement in terms of distortion, in comparison to the MDAV with PCA method, while the enhancements of the second method over the first one strictly depend on the used dataset, fact that will be confirmed in the experimental section. Remember that the speed-up achieved through *MDAV with PCA on proximal data* is the resulting synergy of dimensionality reduction and the superadditivity of the running time of microaggregation algorithms with the number of records.

The main steps of the algorithm are outlined as Algorithm C.

Algorithm C: MDAV with PCA on proximal data.

1: Split the original dataset into two parts: a proximal dataset containing a fraction $1-\nu$ of the records, and a distal dataset containing the remaining fraction ν.
2: Perform principal component analysis (PCA) only on the proximal dataset X_{P}, obtaining $\tilde{X}_{\mathrm{P}} = \tilde{U}_{\mathrm{P}}^{\mathrm{T}} X_{\mathrm{P}}$.
3: Apply traditional MDAV, that is, Algorithm A, to the proximal portion $\tilde{X}_{\mathrm{P}}$ of the data.
4: Separately apply traditional MDAV to the distal portion of the data X_{D}.
5: Combine both micropartitions, obtained separately, into a single cell-assignment function. Publish the overall microaggregation with the corresponding centroids.

V. Experimental Results

In this experimental section, we aim to confirm the algorithmic efficiency of our two dimensionality-reduction methods for k-anonymous microaggregation, specifically by means of PCA, in terms of time gain, along with their performance, in terms of additional distortion incurred. We employ one of the best-known and most widely used fixed-size microaggregation algorithms for numerical data, namely MDAV [8, 11, 13, 43]. Bear in mind that although our work is illustrated with the special case of MDAV, the two methods outlined as Algorithm B and Algorithm C would apply to other microaggregation algorithms and variations of the privacy criteria. However, this versatility is not really the focus of our work, and we shall content ourselves with a standard application of MDAV for k-anonymity.

We should stress, once more, that the speed-up achieved through MDAV with PCA on proximal data is the resulting synergy of dimensionality reduction and the superadditivity of the running time of microaggregation algorithms with the number of records. Now, for a perfectly fair comparison with traditional microaggregation MDAV, the experiments in this manuscript most certainly take into consideration this additional time of applying PCA technique, however insignificant. Furthermore, all experiments in their entirety were implemented and executed in Matlab R2017b, Intel® Core™ i7-6820HQ @2.70 GHz, Windows 10 64-bit, explicitly disabling any form of parallelization for fair and clear comparison.

Additionally, two standardized datasets are considered to evaluate the performance of our novel methods versus MDAV. The standardized datasets are "Large Census" and "Forest", previously used in [36]. The "Large Census" dataset contains 149,642 records with 13 numerical attributes that, in this contribution, will be considered as quasi-identifiers; the "Forest" dataset contains 581,012 records with 10 numerical attributes, all considered as quasi-identifiers as well. We adhere to the common practice of normalizing each attribute of the dataset for unit variance.

The choice of Large Census was motivated by its widespread use in the SDC literature. We shall see that its considerable dimensional redundancy offers excellent results, constituting a perfect illustration of the enormous potential

behind our methods. The Forest dataset was selected as a representation of the opposite case, in which low dimensional redundancy may hinder the effective applicability of PCA.

The reason for introducing PCA in k-anonymous in microaggregation on multivariate data is to substantially reduce the running time of the algorithm employed, while maintaining the quality of the information released. In order to demonstrate how both methods described in §IV achieve our goal, we shall first direct our attention to the histogram of normalized eigenvalues. Precisely, we shall look at $(\lambda_i)_{i=1}^m/m$ in order of decreasing dominance, and at the normalized cumulative energy function $\frac{1}{m}\sum_{i=1}^m \lambda_i$ for each of our two datasets, Large Census and Forest. This will offer a general idea of the extent to which dimensional redundancy can be exploited. Then, we shall report relative time with respect to the conventional use of MDAV, as well as distortion increments.

We shall measure the relative performance gain $\tau \stackrel{\text{def}}{=} t'/t$, defined as the execution time t' of the novel method considered (either MDAV with PCA or MDAV with PCA on proximal data) with respect to the time t of the traditional microaggregation procedure MDAV. In this manner, relative running times τ will potentially range from 0% to 100%, where 100% indicates a running time identical to that of MDAV. Similarly, but not quite identically, we shall report the incurred distortion increment $\Delta\mathcal{D} \stackrel{\text{def}}{=} \mathcal{D}' - \mathcal{D}$, where $\mathcal{D}'$ is the distortion corresponding to the novel method, and $\mathcal{D}$ is the distortion corresponding to conventional MDAV. Therefore, a distortion increment $\Delta\mathcal{D}$ of 0% representing a distortion equal to conventional MDAV, but we should expect small positive values, denoting a cost in distortion that we wish to keep to a minimum. For convenience, τ is relative, but $\Delta\mathcal{D}$ constitutes an increment. Distortion are normalized as it is customary in the literature, with respect to the total variance of the dataset. Times are also normalized, with respect to the running time of conventional use of MDAV, without dimensionality reduction.

A. Large Census dataset

The histogram of normalized eigenvalues in order of decreasing dominance, along with the normalized cumulative energy function for the Large Census dataset was previously shown in Fig. 9 where we clearly observed that 89% of the energy was retained by keeping only 5 principal components out of the total of 13 dimensions.

Fig. 11 and Fig. 12 report the relative time and the distortion increment, respectively. Both results are given as a function of the number $\tilde{m} = 1, \ldots, m$ of compressed dimensions, after applying MDAV with PCA to a subset of 75,000 samples randomly selected from the abovementioned dataset, with a representative anonymity parameter, $k = 10$. Obviously, when the projected dataset contains all $m = 13$ dimensions, that is, when the original information stays intact, $\tau = 100\%$. As we increase the number $\tilde{m}$ of compressed dimensions of the projected data, running time also increases roughly linearly. Regarding the extreme case when the dimension of the projected dataset is $\tilde{m} = 1$, the relative time is approximately $\tau \approx 32\%$. However, Fig. 12 clearly shows that the distortion increment due to this extreme reduction ($\tilde{m} = 1$) is $\Delta\mathcal{D} \approx 28\%$ with respect to conventional MDAV, an unacceptable increment in high-utility applications, the focus of our work. Because our main goal is to preserve data quality, the trade-off between relative time τ and distortion increment $\Delta\mathcal{D}$ must be considered. In practice, one may set a tolerance threshold for the distortion increment, for the energy lost in the projection, or for the number of compressed dimensions.

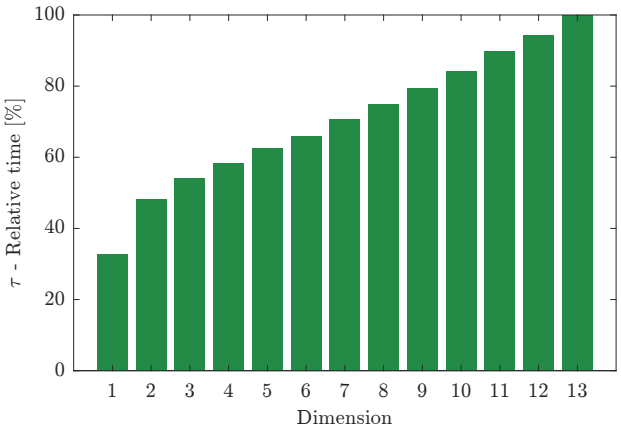


Fig. 11. Relative time after applying MDAV with PCA on a subset of 75,000 samples randomly selected from the Large Census dataset. Observe that by reducing the dimension of the dataset only from 13 to 6, a gain of 34% in time is obtained with respect to classical MDAV.

Furthermore, we can clearly see from Fig. 11 and Fig. 12 that by reducing the dimension of the dataset to $\tilde{m} = 6$, we obtain a relative time of 66% with an increase of only 1.13% in the incurred distortion increment. The possibility of reducing the execution time while almost maintaining the quality of the released information comes from the statistical structure of the Large Census dataset, which keeps a great part of its relevant information in a few dimensions.

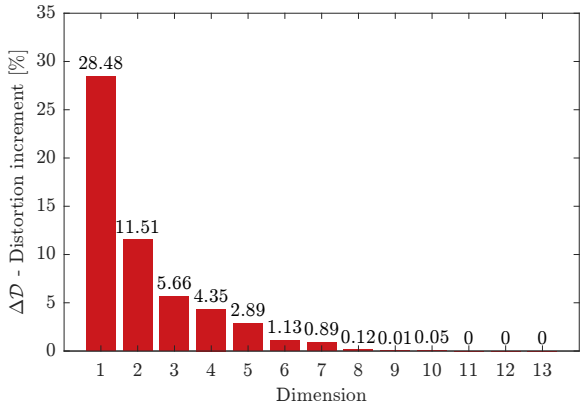


Fig. 12. Distortion increment incurred after applying *MDAV with PCA* on a subset of 75,000 samples randomly selected from the Large Census dataset. Observe that by reducing the dimension of the dataset only from 13 to 6, a distortion increment of approximately 1.13% is incurred with respect to the classical MDAV. Note that, the smaller the dimension of the projected dataset, the higher the distortion increment, since the loss in information is greater.

Now, as explained in §IV-B, we proceed to analyze the adequacy of the choice of the fraction ν of distal points and the remaining fraction $1-\nu$ of the records for the proximal part, both relative to n, of the split original Large Census dataset. To such end, we must first create the macrocells so we can later compute a suitable number of distal macrocells C_{D} taking into consideration both relative time and distortion increment. Macroaggregation will be carried out with MDAV on the entire Large Census dataset with a macrocell size $K = 1000$, much greater than the anonymity parameters k employed later. Since the Large Census dataset has 149,642 records, 149 macro cells will be created with 1000 records each, except for the last macrocell, which will contain 1642 records.

Fig. 13 shows the distortion value $\mathcal{D}$ for each one of the macrocell created, in linear and algorithmic scale. We can clearly observe that the first macrocell incurs by itself 11.37%, with respect to the total variance of the dataset, which corresponds to more than 48% of the whole distortion incurred by the remaining macrocells. In general, Fig. 13 illustrates that there is a dramatic difference in terms of intracell distortion between macrocells, where we can find macrocells with a very small incurred distortion, especially those macrocells created at the end by MDAV (from 100 to 149), and macrocells with high incurred distortion especially the first macrocells created by MDAV. This behavior often happens in real datasets and this special behavior is basically the reason why we decided to split the original dataset into proximal and distal datasets and only apply PCA to the proximal one.

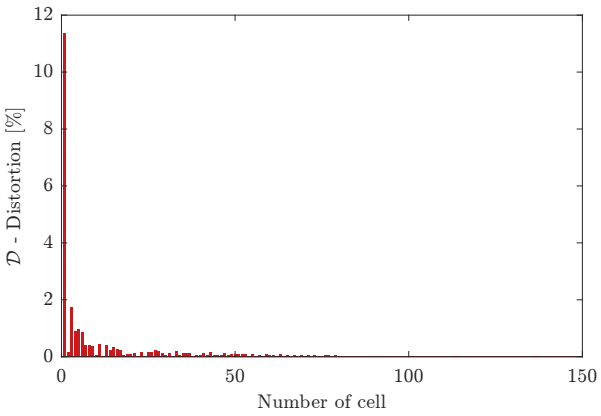


Fig. 13. Distortion $\mathcal{D}$ for each one of the macrocells created by applying MDAV on the whole Large Census dataset with a macrocell size $K = 1000$. Observe that the first macrocell incurs by itself more than 48% of the whole distortion incurred by the remaining macrocells.

As a result, Fig. 14 illustrates relative time τ vs. distortion increment $\Delta\mathcal{D}$, by varying the number of distal macrocells C_{D} from zero (same case as if we are only applying *MDAV with PCA* to the whole dataset) till the maximum number of possible cells $\lfloor n/k \rfloor = 149$ (same case as if we are only applying *MDAV* to the whole dataset), and in each one of the cases *MDAV with PCA on proximal Data* algorithm is applied, with a k-anonymity parameter considered, $k = 10$. From Fig. 14, we define an adequate number of distal macrocells $C_{\mathrm{D}} = 59$, that is, $\nu \approx 40\%$. In this specific case, a reduction of time to almost 65% is obtained with a distortion increment of less than 1%.

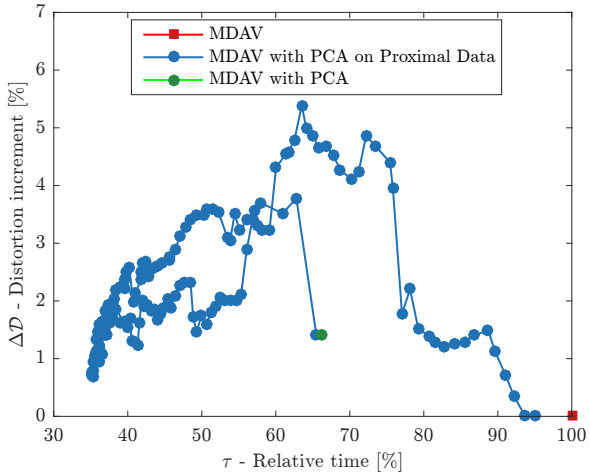


Fig. 14. Relative time vs. distortion increment for each one of the 150 cases (from zero till 149) after applying *MDAV with PCA on proximal Data* algorithm. We have defined an adequate number of distal macrocells $C_{\mathrm{D}} = 59$, that is, $\nu \approx 40\%$. With $\nu \approx 40\%$, a reduction of time to almost 65% is obtained with a distortion increment of less than 1%. The values of compressed dimensions $\tilde{m}$ and the relative energy loss ϵ^2 employed in the PCA experiments for which the relative time and the distortion increment are reported, are equal to 6 and 0.1, respectively.

As the ν fraction has been chosen ($\nu \approx 40\%$), the proximal dataset as well as the distal dataset are well created. Fig. 15 shows the relative time for the two novel methods, *MDAV with PCA* and *MDAV with PCA on proximal Data*, with respect to MDAV only, for various values of k-anonymity parameter. Additionally, two energy loss values ϵ^2 have been used, 0.1 and 0.2, that is, 90% and 80% of the total energy will be preserved, respectively.

However, Fig. 16 illustrates the distortion increment of the two novel methods MDAV with PCA and MDAV with PCA on proximal data for various values of k-anonymity parameter, with respect to MDAV. Observe that by applying MDAV with PCA on the whole Large Census dataset, with energy loss of 10% and 20%, an average time gain of 33% and 41% approximately has been obtained, respectively, with an average distortion increment of 1% and 4%. Also, we see that by applying MDAV with PCA on the proximal dataset with an energy loss of 10% and 20%, where the distal dataset is not subject to PCA at all, only MDAV is applied on it, an average time gain of 64 % and 67% has been obtained, respectively, with an average distortion increment of approximately 1% and 3%. Besides, we have also evaluated the case of applying MDAV without PCA on the proximal dataset instead of the MDAV with PCA on proximal data algorithm, that is, prepartitioning the whole dataset into distal and proximal dataset and later apply MDAV on both of them, and effectively, an average time gain of 45% has been obtained, far from the 64% and 67% obtained by MDAV with PCA on the proximal dataset, without any additional cost in distortion.

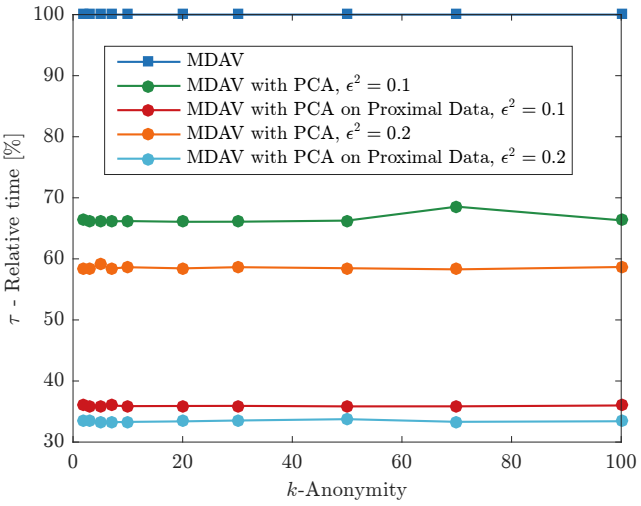


Fig. 15. Relative time of the two novel methods MDAV with PCA and MDAV with PCA on proximal data for many common values of k-anonymity parameter, with respect to MDAV. Observe that by applying MDAV with PCA on the whole Large Census dataset, with energy loss of 10% and 20%, an average time gain of 33% and 41 % approximately has been obtained, respectively. Also, we see that by only applying MDAV with PCA on the proximal dataset with an energy loss of 10% and 20%, an average time gain of 64% and 67% has been obtained, respectively.

As we have observed, the prepartitioning method into distal and proximal dataset works very well and this is because it looks for the very distorted zone as well as for the less distorted one, and this is basically what MDAV does. Nonetheless, one of the great advantages, less evident more significant of prepartitioning into distal and proximal dataset, is that the prepartitioning method allows posteriori methods to be applied, and the price in distortion that those posteriori methods will pay to obtain even more average time gain will be smaller since they will be applied only on the proximal part.

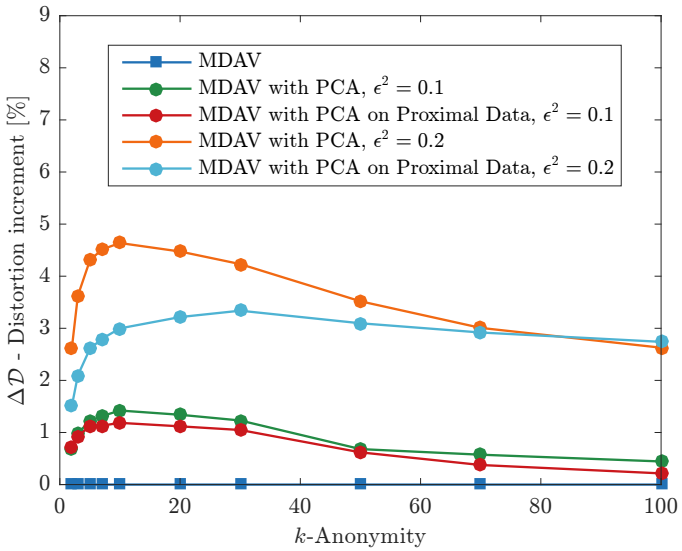


Fig. 16. Distortion increment of the two novel methods MDAV with PCA and MDAV with PCA on proximal data for various values of k-anonymity parameter, with respect to MDAV. Observe that by applying MDAV with PCA on the whole Large Census dataset, with energy loss of 10% and 20%, an average distortion increment of 1% and 4 % approximately has been obtained, respectively. Also, we see that by only applying MDAV with PCA on the proximal dataset with an energy loss of 10% and 20%, an average distortion increment of 1 % and 3% has been obtained, respectively.

The absolute running time of the traditional algorithm and the two novel methods, used in this work will certainly vary depending on both n and k, as well as on the computer and the number of cores employed. However, most of our experiments are in terms of running times relative to the traditional one MDAV. Still, Table I and Table II illustrate the reference times and the normalized distortion values, respectively, for MDAV, MDAV with PCA and MDAV with PCA on proximal data, applied on two standardized datasets, Large Census and Forest. The values of compressed dimensions $\tilde{m}$ and the relative energy loss ϵ^2 employed in the PCA experiments for which the time and distortion are reported in Table I and Table II, are equal to 6 and 0.1, respectively.

Table I
Reference Running Times

Dataset	Samples n	Dimension m	Anonymity k	MDAV	MDAV with PCA	MDAV with PCA on Proximal Data
Large Census	149,642	13	2	16 min 27 sec	10 min 55 sec	5 min 56 sec
			5	6 min 34 sec	4 min 21 sec	2 min 22 sec
			10	3 min 17 sec	2 min 10 sec	1 min 11 sec
			50	40 secs	26 sec	14 sec
			100	20 sec	13 sec	7 sec

Reference running times values for MDAV, MDAV with PCA and MDAV with PCA on proximal data, applied to the standardized dataset, Large Census, with different k-anonymity values. The values of compressed dimensions $\tilde{m}$ and the relative energy loss ϵ^2 employed in the PCA experiments for which the time and distortion are reported, are equal to 6 and 0.1, respectively.

TABLE II
REFERENCE DISTORTIONS

Dataset	Samples n	Dimension m	Anonymity k	MDAV	MDAV with PCA	MDAV with PCA on Proximal Data
Large Census	149,642	13	2	0.0056	0.0126	0.0127
			5	0.0168	0.029	0.028
			10	0.0278	0.042	0.0397
			50	0.0685	0.0754	0.0747
			100	0.0973	0.1018	0.0995

Reference distortion values for MDAV, MDAV with PCA and MDAV with PCA on proximal data, applied to the standardized dataset, Large Census, with different k-anonymity values. The values of compressed dimensions $\tilde{m}$ and the relative energy loss ϵ^2 employed in the PCA experiments for which the time and distortion are reported, are equal to 6 and 0.1, respectively.

B. Forest dataset

Fig. 17 shows a histogram of normalized eigenvalues $(\lambda_i)_{i=1}^m/m$ in order of decreasing dominance, along with the normalized cumulative energy function $\frac{1}{m}\sum_{i=1}^m \lambda_i$, but this time for the Forest dataset. Again, the horizontal axis shows the number $\tilde{m}$ of cumulative principal components, ranging from the use of a single principal component $\tilde{m}=1$ with significant energy loss $\epsilon^2 = 1-\lambda_1$, all the way to $\tilde{m}=m$, for which all energy is preserved and thus $\epsilon^2=0$, but no effective dimensionality reduction occurs. We can clearly observe that 89% of the energy is retained by keeping 6 principal components out of the total of 10 dimensions, 60% of the total amount of dimensions. Remember that in case of the Large Census dataset, we needed only 38% (5 principal components out of 13 dimensions) of the total amount of dimensions to preserve the same amount of information. This is due to the strong linear independence of attributes of Forest dataset.

Fig. 18 and Fig. 19 illustrate the relative time and the distortion increment respectively, both results are per dimension, after applying *MDAV with PCA* on a subset of 75,000 samples randomly selected from the abovementioned dataset, with a k-anonymity parameter considered, $k=10$. As it can be observed, when the projected dataset has 10 dimensions, that is, when the original information stays intact, $\tau = 100\%$, which makes sense as it is exactly the same case of applying only MDAV. Again, as in the Large Census dataset case, as we increase the number $\tilde{m}$ of compressed dimensions of the projected data, running time also increases roughly linearly. However, regarding the extreme case when the dimension of the projected dataset is 1, the relative time is approximately $\tau \approx 39\%$. Nevertheless, Fig. 19 clearly shows that the distortion increment $\Delta\mathcal{D}$ incurred due to the reduction of dimensions into one dimension has increased by more than 69% with respect to the case of using only MDAV, which is an unacceptable increment in high-utility applications, the focus of our work. As we have said before, as our main goal is to preserve the information quality, a tradeoff between relative time τ and distortion increment $\Delta\mathcal{D}$ will be considered, as we did in Large Census dataset case. Again, we can set a tolerance threshold for the distortion increment, for the energy lost in the projection, or for the number of compressed dimensions.

Furthermore, we can clearly see from Fig. 18 and Fig. 19 that by reducing the dimension of the dataset to 6 for example, we obtain a relative time of 77% with an increase of 6.4% in the incurred distortion increment. As there is a strong linear independence of attributes in the Forest dataset, the possibility of obtaining high time gain while maintaining the quality of the released information intact is not possible even though it was possible in the case of Large census dataset which keeps the great part of its relevant information in a few dimensions, contrary to the nature structure of the Forest dataset.

Again and as explained in §IV-B, we proceed to analyze the adequacy of the choice of the fraction ν of distal points and the remaining fraction $1-\nu$ of the records for the proximal part, both relative to n, of the split original Large Census dataset. To such end, and as we did in the Large census dataset case, we must first create the macrocells so we can later compute a suitable number of distal macrocells C_{D} taking into consideration both relative time and distortion increment. Again, the macroaggregation process will be realized with MDAV on the whole Forest dataset with a macrocell size $K = 10,000$, much greater than the anonymity parameter k used later. As the Forest dataset has 581,012 records, 58 macro cells will be created of 10,000 records each, except the last macrocell, which will contain 11012 records.

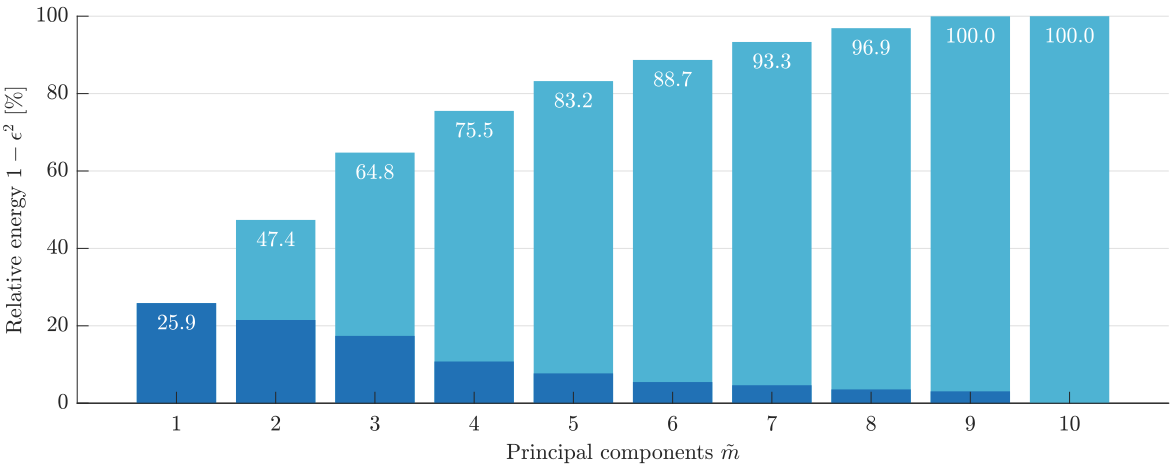


Fig. 17. The graphic shows the normalized energy per dimension and the cumulated energy per dimension for the Forest dataset. The horizontal axis shows the number $\tilde{m}$ of cumulative principal components, ranging from the use of a single principal component $\tilde{m} = 1$ with significant energy loss $\epsilon^2 = 1 - \lambda_1$, all the way to $\tilde{m} = m$, for which all energy is preserved and thus $\epsilon^2 = 0$, but no effective dimensionality reduction occurs. Observe that 93.3% of the energy is retained by keeping only 7 principal components out of the total of 10 dimensions.

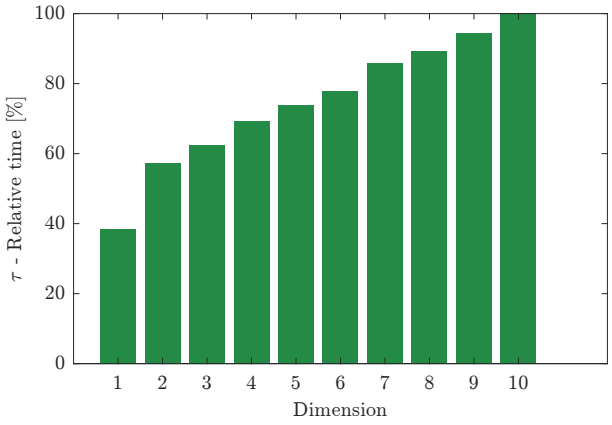


Fig. 18. Relative time after applying *MDAV with PCA* on a subset of 75,000 samples randomly selected from the Forest dataset. Notice that by reducing the dimension of the dataset only from 10 to 6, a gain of 22.2% in time is obtained with respect to the classical MDAV.

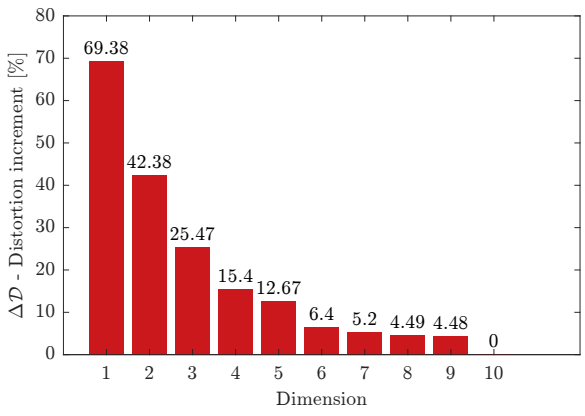


Fig. 19. Distortion increment incurred after applying *MDAV with PCA* on a subset of 75,000 samples randomly selected from the Forest dataset. Obviously, the smaller the dimension of the projected dataset, the higher the distortion increment, since the loss in information is greater.

Fig. 20 shows the distortion value $\mathcal{D}$ for each one of the macrocell created. In general, Fig. 20 illustrates that more or less there is no difference in intracell distortion $\mathcal{D}$ between the created macrocells, which is totally different result from the one obtained when analyzing Large Census dataset, where we detected a huge difference in intracell distortion $\mathcal{D}$ between the created macrocells. This previous result shows that the Forest dataset points are well distributed in $\mathbb{R}^m$ which hinders the process of splitting the Forest dataset into distal dataset and proximal dataset as the $\mathcal{D}$ of the macrocells are quite similar.

As a result, Fig. 21 illustrates relative time τ vs. distortion increment $\Delta\mathcal{D}$, by scanning C_{D} from zero (same case as if we are only applying *MDAV with PCA* to the entire dataset) to the maximum number of possible cells $\lfloor n/K \rfloor = 58$ (same case as if we are only applying *MDAV* to the entire dataset), and in each one of the cases *MDAV with PCA on proximal data* algorithm is applied, with a k-anonymity parameter considered, $k = 100$. From Fig. 21, we have defined an adequate number of distal macrocells $C_{\mathrm{D}} = 39$, that is, $\nu \approx 66\%$. In this specific case, a reduction of time to almost 55% is obtained with a distortion increment $\Delta\mathcal{D}$ of approximately 2%.

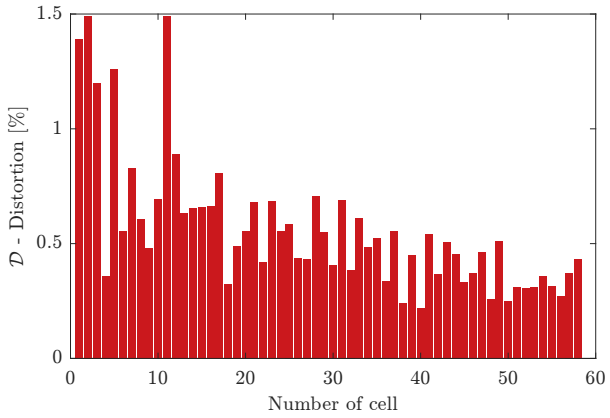


Fig. 20. Distortion $\mathcal{D}$ for each one of the macrocells created by applying MDAV on the whole Forest dataset with a macrocell size $K = 10,000$. Notice that the distortion incurred by macrocells are more or less equal.

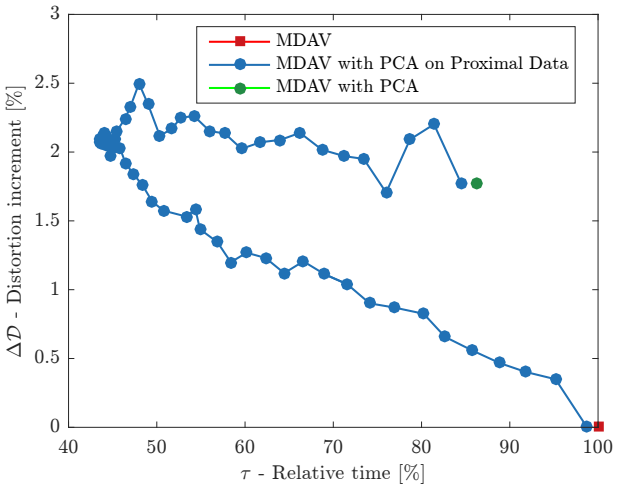


Fig. 21. Relative time vs. distortion increment for each one of the 59 cases (from zero till 58) after applying *MDAV with PCA on proximal data* algorithm. We have defined an adequate number of distal macrocells $C_{\mathrm{D}} = 39$, that is, $\nu \approx 66\%$. With $\nu \approx 66\%$, a reduction of time to almost 55 % is obtained with a distortion increment $\Delta\mathcal{D}$ of approximately 2%. The values of compressed dimensions $\tilde{m}$ and the relative energy loss ϵ^2 employed in the PCA experiments for which the relative time and the distortion increment are reported, are equal to 7 and 0.1, respectively.

As the ν fraction ($\nu \approx 66\%$) has been chosen, the proximal dataset as well as the distal dataset are well created. Fig. 22 shows the relative time for the two novel methods, *MDAV with PCA* and *MDAV with PCA on proximal data*, with respect to MDAV only, for many values of k-anonymity parameter, k. Additionally, two energy loss values ϵ^2 have been used, 0.1 and 0.2, that is, 90% and 80% of the total energy will be preserved, respectively.

Fig. 23 illustrates the distortion increment $\Delta\mathcal{D}$ of the two novel methods *MDAV with PCA* and *MDAV with PCA on proximal data* for various values of k-anonymity parameter, with respect to MDAV. Notice that by applying *MDAV with PCA* on the entire Forest dataset, with energy loss of 10% and 20%, an average time gain of 15% and 25%

approximately has been obtained, respectively, with an average distortion increment of 1.5% and 5.5%. Also, we see that by only applying *MDAV with PCA on the proximal dataset* with an energy loss of 10% and 20%, where the distal dataset is not subject to PCA at all, only MDAV will be applied on it, an average time gain of 48 % and 50% has been obtained, respectively, with an average distortion increment of approximately 1.5% and 2%. Besides, we have also evaluated the case of applying *MDAV without PCA on the proximal dataset* as in case of Large Census dataset, and again effectively, an average time gain of 42% has been obtained, far from 50% obtained by *MDAV with PCA on the proximal dataset*, without any additional cost in distortion.

Additionally, even if most of our experiments are in terms of running times relative to the traditional one MDAV. Still, Table III and Table IV illustrate the reference times and the normalized distortion values, respectively, for MDAV, MDAV with PCA and MDAV with PCA on proximal data, applied to the Forest dataset, with different k-anonymity values. The values of compressed dimensions $\tilde{m}$ and the relative energy loss ϵ^2 employed in the PCA experiments for which the time and distortion are reported in Table III and Table IV, are equal to 7 and 0.1, respectively.

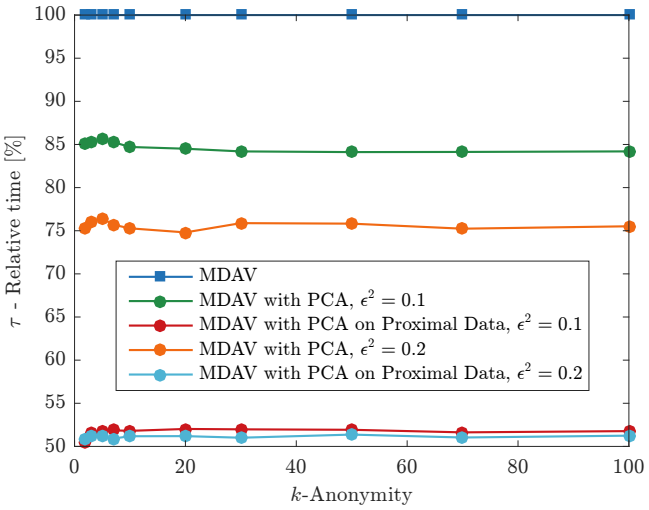


Fig. 22. Relative time of the two novel methods MDAV with PCA and MDAV with PCA on proximal data for various values of k-anonymity parameter, with respect to MDAV. Observe that by applying MDAV with PCA on the entire Forest dataset, with energy loss of 10% and 20%, an average time gain of 15% and 25% approximately has been obtained, respectively. Also, we see that by only applying MDAV with PCA on the proximal dataset, with an energy loss of 10% and 20%, where the distal dataset is not subject to PCA at all, only MDAV will be applied on it, an average time gain of 48% and 50% has been obtained, respectively.

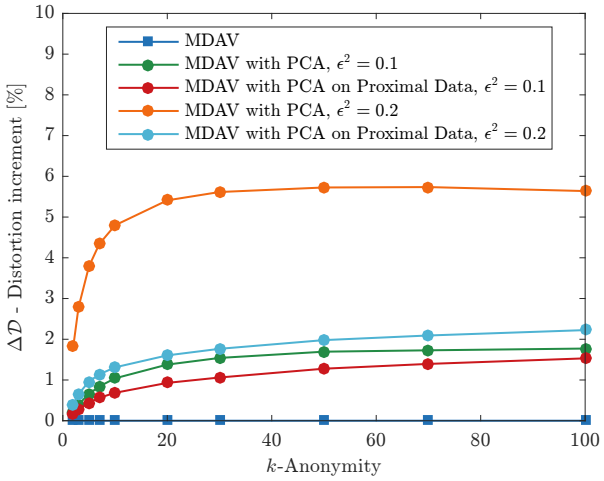


Fig. 23. Distortion increment of the two novel methods MDAV with PCA and MDAV with PCA on proximal data for many values of k-anonymity parameter, with respect to MDAV. Notice that by applying MDAV with PCA on the entire Forest dataset, with energy loss of 10% and 20%, an average distortion increment of 1.5% and 5.5% approximately has been obtained, respectively. Also, we see that by only applying MDAV with PCA on the proximal dataset, with an energy loss of 10% and 20%, where the distal dataset is not subject to PCA at all, only MDAV will be applied on it, an average distortion increment of 1.5% and 2% has been obtained, respectively.

TABLE III
REFERENCE RUNNING TIMES

Dataset	Samples n	Dimension m	Anonymity k	MDAV	MDAV with PCA	MDAV with PCA on Proximal Data
Forest	581,012	10	2	4 hrs	3 hrs 23 min	2 hrs
			5	1 hr 34 min	1 hr 21 min	49 min
			10	47 min 33 sec	40 min 16 sec	24 min 38 sec
			50	9 min 30 sec	8 min	4 min 56 sec
			100	4 min 45 sec	4 min	2 min 28 sec

Reference running times values for MDAV, MDAV with PCA and MDAV with PCA on proximal data, applied to the standardized dataset, Forest, with different k-anonymity values. The values of compressed dimensions $\tilde{m}$ and the relative energy loss ϵ^2 employed in the PCA experiments for which the time and distortion are reported, are equal to 7 and 0.1, respectively.

TABLE IV
REFERENCE DISTORTION

Dataset	Samples n	Dimension m	Anonymity k	MDAV	MDAV with PCA	MDAV with PCA on Proximal Data
Forest	581,012	10	2	0.0025	0.0047	0.004
			5	0.088	0.0154	0.0133
			10	0.0158	0.0262	0.0227
			50	0.0417	0.0586	0.0545
			100	0.058	0.0757	0.0733

Reference distortion values for MDAV, MDAV with PCA and MDAV with PCA on proximal data, applied to the standardized dataset, Forest, with different k-anonymity values. The values of compressed dimensions $\tilde{m}$ and the relative energy loss ϵ^2 employed in the PCA experiments for which the time and distortion are reported, are equal to 7 and 0.1, respectively.

VI. Conclusion and Future Development

This work addresses the problem of computational complexity of k-anonymous microaggregation for large datasets with a substantial amount of numerical quasi-identifiers and records. We proceed by introducing two novel dimensionality-reduction methods in the microaggregation field by means of PCA, named as *MDAV with PCA* and *MDAV with PCA on proximal data*, which greatly reduce the running time of k-anonymous microaggregation of multivariate data, maintaining the quality of the released information.

It is important to remember that even if our work is illustrated with the special case of the widely used algorithm known as MDAV, the two novel methods abovementioned would be easily applicable to other microaggregation techniques. Furthermore, we should stress on that the speed-up achieved through *MDAV with PCA on proximal data* is the resulting synergy of dimensionality reduction and the superadditivity of the running time of microaggregation algorithms with the number of records.

Our experiments indicate that our PCA methods lead to a distortion increment $\Delta\mathcal{D}$ with respect to the total variance of the dataset that does not increase, but rather decreases, with the anonymity factor k, as one can verify, for instance, in Fig. 16. On the other hand, it is clear that the reference distortion $\mathcal{D}_{\text{ref}}$ of conventional MDAV increases with k. These observations imply that the relative distortion increment $\Delta\mathcal{D}/\mathcal{D}_{\text{ref}}$ should decrease with k. This means that in terms of this relative distortion metric, the PCA methods put forth in this work should be particularly efficient for microaggregation with large values of the anonymity parameter k. For small values of k, the admissible energy loss ϵ^2 should be set to a cautiously low value, depending on the relevance of time and distortion in the application at hand.

For any value of k but a low ϵ^2, a dataset with strong linear dependencies can be microaggregated significantly faster in a lower-dimensional subspace, and with a low impact on data utility.

Moreover, experimental results on two standardized datasets, Large Census dataset and Forest dataset, confirm considerably the reduction of k-anonymous microaggregation of multivariate data, maintaining the quality of the released information by means of PCA. Applying *MDAV with PCA* method to both datasets, we achieved significant time gains ($\approx$14–31%) with very little impact on information utility (<2%, with respect to the total variance) with respect to MDAV on the original data. However, when applying *MDAV with PCA on proximal data* method to both datasets again, we obtained even further speed-up ($\approx$48–64%), with mild distortion impact (<3%, with respect to the total variance). As a concluding and essential message of our work, we have shown how useful PCA is in the microaggregation field, which can be used alone as seen in *MDAV with PCA* method, or also it can be combined to other strategies as shown in PCA and MDAV with PCA on proximal data method.

A. Concluding Remarks

As a summary of our work, Table V includes a brief description of the developed algorithms as well as some points that should be taken into account when using them.

TABLE V
MICROAGGREGATION METHODS PROPOSED

Algorithms	Definition	Remarks
MDAV w/ PCA	• Perform PCA on the dataset. • Execute MDAV on the dataset of reduced dimension	• The performance of this algorithm relies on the statistical structure of the information to be treated. Datasets with most of their information conducive to compression into a few dimensions will allow the algorithm to be executed faster.
MDAV w/ PCA on Proximal Data	• Split the original dataset into proximal and distal datasets. • Perform PCA on the proximal dataset only. • Execute MDAV on the dimensionally reduced proximal dataset, and on the original distal dataset, separately.	• The speed-up achieved is the resulting synergy of two effects: first, dimensionality reduction, and secondly, the superadditivity of the running time of microaggregation algorithms in the number of records. • Again, this algorithm relies on the algebraic-statistical redundancy of the information processed, as in MDAV with PCA.

B. Future Directions

Many avenues exist along which we could conduct future research. By means of illustration, we briefly elaborate on our second, more complex method. Recall that said method implemented a prepartition of the dataset into a proximal and distal part, applying PCA only to the former portion in order to keep the distortion overhead in check, be it due to prepartitioning or dimensionality reduction. The procedure in question was advantageous not only because of the reduction of dimensions, but also because of the superadditivity of MDAV. In fact, the time complexity of MDAV is asymptotically quadratic in the number of records, and thus the benefit of the "divide and conquer" effect alone proved significant. Even without PCA, the mere prepartition into a proximal and distal portion to accelerate MDAV with negligible distortion overhead appears to be quite promising. Conceivably, this prepartition method could be combined with any other techniques trading off running time for distortion, aside from PCA. Further, the proximal-distal prepartition itself is not subject to strong restrictions on cell size, and could be carried out with clustering methods alternative to MDAV, for instance with the Lloyd algorithm, also known as the k-means method (where k here represents the number of cells and is unrelated to the notion of anonymity).

Another intriguing direction is the recursive application of the proximal-distal prepartition technique, possibly combined with PCA. We provide an extremely preliminary model to motivate its potential. Suppose that an algorithm required time $t(n) = n^2$ in order to process n records (in units relative to $t(1)$). Rather than executing the algorithm directly on the entire recordset, we might first split it in two parts, one containing a fraction $1-\nu$ of the records, and the other containing the remaining fraction ν, both relative to n. Here, the portion of $1-\nu$ records would play the role of proximal part, less sensitive to distortion changes. Assuming for simplicity in this preliminary analysis that the cost of splitting and recombination were asymptotically negligible, the computation required by this approach would be

$$t'(n) = t((1-\nu)n) + t(\nu n)$$

instead. In our simple, asymptotic argument, we shall disregard cell size restrictions and the fact that they should be integer values instead of idealized real-valued approximations. For any $\nu \in (0,1)$, we conduct a progressively recursive

implementation only on the first (less sensitive) part of the dataset with a fraction $1-\nu$ of the records in $S = \log_{\frac{1}{1-\nu}} n$ additional steps (none for $\nu = 1$). Excluding the residual constant term $t(1) = 1$ at the end of the recursion, asymptotically negligible, the time complexity would be[(b)],

$$\sum_{s=0}^{S-1} t((1-\nu)^s \nu n) < \sum_{s=0}^{\infty} ((1-\nu)^s \nu n)^2 = \frac{(\nu n)^2}{1-(1-\nu)^2} = \frac{\nu}{2-\nu} n^2.$$

This is lower than the original time n^2, in fact vanishing as $\nu \downarrow 0$, and obviously (asymptotically) equal as $\nu \uparrow 1$. For example, a conservative $\nu = 1/3$ would yield an impressive reduction factor $\nu/(2-\nu) = 1/5$. But the distortion overhead of such recursion, potentially prohibitive for aggressively small ν, would demand careful analysis.

Acknowledgment

The authors would like to thank the anonymous reviewers for their thorough suggestions, which greatly helped in improving the readability and contents of this paper. Furthermore,we gratefully thank Dr. Javier Parra-Arnau for his valuable comments on the state of the art of this paper. This manuscript presents some of the results developed through the collaboration of the Universitat Politècnica de Catalunya (UPC) and Scytl Secure Electronic Voting S.A. (Scytl) in the context of the project "Data-Distortion Framework", and in accordance with the guidelines therein. This work is thus partly supported by the Spanish Ministry of Industry, Energy and Tourism (MINETUR) through the "Acción Estratégica Economía y Sociedad Digital (AEESD)" funding plan, through the aforementioned project, "Data-Distortion Framework (DDF)", ref. TSI-100202-2013-23.

Additional funding supporting this work has been granted to UPC by the Spanish Ministry of Economy and Competitiveness (MINECO) through the "Anonymized Demographic Surveys (ADS)" project, ref. TIN2014-58259-JIN, under the funding program "Proyectos de I+D+i para Jóvenes Investigadores", as well as through the projects "MAGOS", ref. TEC2017-84197- C4-3- R and "INRISCO", ref. TEC2014-54335- C4- 1-R. Finally, we acknowledge funding support by the European Commission through the H2020 project "CIPSEC", grant no. 700378.

References

[1] M. Akra and L. Bazzi, "On the solution of linear recurrence equations," *Comput. Optim., Appl.*, vol. 10, no. 2, pp. 195–210, 1998.

[2] J. Byun, Y. Shon, E. Bertino, and N. Li, "Secure anonymization for incremental datasets," in *Proc. VLDB Workshop Secure Data Mgmt. (SDM)*, ser. Lect. Notes Comput. Sci. (LNCS), vol. 4165, Seoul, Korea, Sep. 2006, pp. 48–63.

[3] A. Calviño, "A simple method for limiting disclosure in continuous microdata based on principal component analysis," *J. Official Stat.*, vol. 33, no. 1, pp. 15–41, Feb. 2017.

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

[5] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," in *Proc. IEEE Int. Conf. Data Eng. (ICDE) Workshops*, Brisbane, Australia, Apr. 2013, pp. 88–93.

[6] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, 3rd ed. Cambridge, MA: MIT Press, 2009.

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

[8] 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.

[9] 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, 2002.

[10] 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.

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

[12] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Lang., Program. (ICALP)*, ser. Lect. Notes Comput. Sci. (LNCS), vol. 4052, Venice, Italy, Jul. 2006, pp. 1–12.

[13] A. Hundepool, R. Ramaswamy, P.-P. DeWolf, L. Franconi, R. Brand, and J. Domingo-Ferrer, *μ-ARGUS version 4.1 software and user's manual*, Stat. Neth., Voorburg, Netherlands, 2007. [Online]. Available: http://neon.vb.cbs.nl/casc

[14] H. Jian min, C. Ting ting, and Y. Hui qun, "An improved V-MDAV algorithm for l-diversity," in *Proc. IEEE Int. Symp. Inform. Process. (ISIP)*, Moscow, Russia, May 2008, pp. 733–739.

[15] I. T. Jolliffe, *Principal component analysis*, 2nd ed. New York, NY: Springer-Verlag, 2002.

[16] 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.

[(b)] Although the actual running time will be much lower, the final complexity is still $\Theta(n^2)$, because recursion is applied only to one of the two branches, namely the one corresponding to proximal points. In our simplified analysis without split and recombination cost, suppose further that the distortion overhead were of secondary relevance, so that the recursion could be enforced on the distal part as well. Then, both the master theorem for $t(n) = 2\,t(n/2)$ or the Akra-Bazzi method [1] for the more general setting $t(n) = t((1-\nu)n) + t(\nu n)$ would give linear complexity, $\Theta(n)$.

[17] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Istanbul, Turkey, Apr. 2007, pp. 106–115.

[18] 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.

[19] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 129–137, Mar. 1982.

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

[21] N. Matatov, L. Rokach, and O. Maimon, "Privacy-preserving data mining: A feature set partitioning approach," *Inform. Sci.*, vol. 180, no. 14, pp. 2696–2720, 2010.

[22] J. M. Mateo-Sanz and J. Domingo-Ferrer, "A comparative study of microaggregation methods," *Data Min., Knowl. Discov.*, vol. 11, no. 2, pp. 195–212, 1998.

[23] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inform. Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.

[24] A. Mohamad Mezher, A. García-Álvarez, D. Rebollo-Monedero, and J. Forné, "Computational improvements in parallelized k-anonymous microaggregation of large databases," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS), Workshop Priv., Secur. Big Data (PSBD)*, Atlanta, GA, Jun. 2017, pp. 258–264.

[25] J. J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," in *Proc. Dundee Biennial Conf. Numer. Anal.*, ser. Lect. Notes Math., vol. 630, Dundee, UK, Jun. 1977, pp. 105–116.

[26] J. Nin, J. Herranz, and V. Torra, "On the disclosure risk of multivariate microaggregation," *Data, Knowl. Eng.*, vol. 67, no. 3, pp. 399–412, 2008.

[27] A. Oganian and J. Domingo-Ferrer, "On the complexity of optimal microaggregation for statistical disclosure control," *UNECE Stat. J.*, vol. 18, no. 4, pp. 345–354, Apr. 2001.

[28] J. Panaretos and N. Tzavidis, "Aspects of estimation procedures at Eurostat with some emphasis on over-space harmonisation," in *Proc. Hellenic-Euro. Conf. Comput. Math., Appl. (HErCMA)*, vol. 2, Athens, Greece, Sep. 2001, pp. 853–857.

[29] D. Rebollo-Monedero, J. Forné, E. Pallarès, and J. Parra-Arnau, "A modification of the Lloyd algorithm for k-anonymous quantization," *Inform. Sci.*, vol. 222, pp. 185–202, Feb. 2013. [Online]. Available: http://doi.org/10.1016/j.ins.2012.08.022

[30] D. Rebollo-Monedero, J. Forné, and M. Soriano, "Private location-based information retrieval via k-anonymous clustering," in *Proc. CNIT Int. Workshop Digit. Commun.*, ser. Lect. Notes Comput. Sci. (LNCS), Sardinia, Italy, Sep. 2009, pp. 421–430, invited paper.

[31] ——, "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

[32] D. Rebollo-Monedero, J. Forné, M. Soriano, and J. Puiggalí Allepuz, "k-Anonymous microaggregation with preservation of statistical dependence," *Inform. Sci.*, vol. 342, pp. 1–23, May 2016. [Online]. Available: http://doi.org/10.1016/j.ins.2016.01.012

[33] ——, "p-Probabilistic k-anonymous microaggregation for the anonymization of surveys with uncertain participation," *Inform. Sci.*, vol. 382–383, pp. 388–414, Mar. 2017. [Online]. Available: http://doi.org/10.1016/j.ins.2016.12.002

[34] D. Rebollo-Monedero, C. Hernández-Baigorri, J. Forné, and M. Soriano, "Incremental k-anonymous microaggregation in large-scale electronic surveys with optimized scheduling," *IEEE Access*, vol. 6, no. 1, pp. 60 016–60 044, Dec. 2018.

[35] A. Solanas and A. Martínez-Ballesté, "V-MDAV: Multivariate microaggregation with variable group size," in *Proc. Int. Conf. Comput. Stat. (CompStat)*, Rome, Italy, Aug. 2006, pp. 917–925.

[36] M. Solé, V. Muntés-Mulero, and J. Nin, "Efficient microaggregation techniques for large numerical data volumes," *Int. J. Inform. Secur.*, vol. 11, no. 4, pp. 253–267, Aug. 2012.

[37] J. Soria-Comas, "Improving data utility in differential privacy and k-anonymity," Ph.D. dissertation, Rovira i Virgili Univ. (URV), Apr. 2013.

[38] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, and D. Megías, "Individual differential privacy: A utility-preserving formulation of differential privacy guarantees," *IEEE Trans. Inform. Forensics, Secur.*, vol. 12, no. 6, pp. 1418–1429, Feb. 2017.

[39] X. Sun, H. Wang, J. Li, and T. M. Truta, "Enhanced p-sensitive k-anonymity models for privacy preserving data publishing," *Trans. Data Priv.*, vol. 1, no. 2, pp. 53–66, 2008.

[40] Y. Sun, L. Yin, L. Liu, and S. Xin, "Toward inference attacks for k-anonymity," *Pers., Ubiquit. Comput.*, vol. 18, pp. 1871–1880, Aug. 2014.

[41] L. Sweeney, "Simple demographics often identify people uniquely," Carnegie Mellon Univ., Work. Paper 3, 2000.

[42] ——, "Uniqueness of simple demographics in the U.S. population," Carnegie Mellon Univ., Sch. Comput. Sci., Data Priv. Lab., Pittsburgh, PA, Tech. Rep. LIDAP-WP4, 2000.

[43] M. Templ, "Statistical disclosure control for microdata using the R-package sdcMicro," *Trans. Data Priv.*, vol. 1, no. 2, pp. 67–85, 2008. [Online]. Available: http://cran.r-project.org/web/packages/sdcMicro

[44] T. M. Truta and B. Vinay, "Privacy protection: p-Sensitive k-anonymity property," in *Proc. Int. Workshop Priv. Data Mgmt. (PDM)*, Atlanta, GA, Apr. 2006, p. 94.

David Rebollo-Monedero (david.rebollo@entel.upc.edu) is a senior researcher with the Information Security Group of the Department of Telematic Engineering at the Universitat Politècnica de Catalunya, in Barcelona, Spain, where he investigates the application of information theoretic and operational data compression formalisms to privacy in information systems. He received the M.S. and Ph.D. degrees in electrical engineering from Stanford University, in California, USA, in 2003 and 2007, respectively. Previously, from 1997 to 2000, he was an information technology consultant for PricewaterhouseCoopers in Barcelona. His current research interests encompass data privacy, information theory, data compression, and machine learning.

Ahmad Mohamad Mezher (ahmad.mezher@entel.upc.edu) received the M.S. degree in signals and systems from the Central University of Las Villas, Santa Clara, Cuba, in 2011, and the Ph.D. degree in network engineering from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2016. Currently he holds a Postdoctoral Fellowship position with the Electrical and Computer Engineering department at the University of New Brunswick (UNB). His research interests include smart grid communications, vehicular ad hoc networks, data privacy, and machine learning.

Xavier Casanova Colomé (xavier.casanova.colome@alu-etsetb.upc.edu) received the B.S. and M.S. in telecommunications engineering from the Universitat Politècnica de Catalunya (UPC), in Barcelona, Spain, in 2015 and 2017, respectively. During his bachelor's senior year, he collaborated with the Information Security Group in the development and improvement of methods for anonymous microaggregation, applicable to electronic surveys, among others. He has worked as a developer at Computer Sciences Brand, S.L. and as a presales engineer at Scytl, S.A. His interests include a number of subfields in which engineering and science meet, and which require the use of machine learning techniques.

Jordi Forné (jforne@entel.upc.edu) received the M.S. and the Ph.D. degrees in telecommunications engineering from the Universitat Politècnica de Catalunya (UPC), in Barcelona, Spain, in 1992 and 1997, respectively. He is currently associate professor in the Telecommunications Engineering School of Barcelona at UPC and head of the data privacy team of the Department of Telematic Engineering. From 2007 to 2012, he was coordinator of the Ph.D. program in telematic engineering and director of the master's research program in the same subject. His research interests span a number of subfields within information security and privacy.

Miguel Soriano (soriano@entel.upc.edu) is full professor in the Telecommunications Engineering School of Barcelona, and head of the Information Security Group, both affiliated to the Department of Telematic Engineering at the Universitat Politècnica de Catalunya (UPC), in Barcelona, Spain. Additionally, he works as a researcher at the Centre Tecnològic de Telecomunicacions de Catalunya. He received the M.S. and Ph.D. degrees in telecommunications engineering from UPC, in 1992 and 1996 respectively. His research interests encompass network security, e-voting, and information hiding for copyright protection.