k-Anonymous Microaggregation with Preservation of Statistical Dependence

David Rebollo-Monedero [*,1], Jordi Forné [1], Miguel Soriano [1,2], and Jordi Puiggalí Allepuz [3]

[1] Department of Telematic Engineering
Universitat Politècnica de Catalunya (UPC)
C. Jordi Girona 1–3, E-08034, Barcelona, Spain

[2] Centre Tecnològic de Telecomunicacions de Catalunya (CTTC)
Av. Carl Friedrich Gauss 7, E-08860 Castelldefels, Barcelona, Spain

[3] Scytl Secure Electronic Voting
Pl. Gal·la Placídia 1–3, E-08006 Barcelona, Spain

Abstract—k-Anonymous microaggregation emerges as an essential building block in statistical disclosure control, a field concerning the postprocessing of the demographic portion of surveys containing sensitive information, in order to safeguard the anonymity of the respondents. Traditionally, this form of microaggregation has been formulated to characterize both the privacy attained and the inherent information loss due to the aggregation of quasi-identifiers, which may otherwise be exploited to reidentify the individuals to which a record in a published database refer. Because the ulterior purposes of such databases involves the analysis of the statistical dependence between demographic attributes and sensitive data, we must articulate mechanisms to enable the preservation of the statistical dependence between quasi-identifiers and confidential attributes, beyond the mere degradation of the quasi-identifiers alone.

This work addresses the problem of k-anonymous microaggregation with preservation of statistical dependence in a formal, systematic manner, modeling statistical dependence as predictability of the confidential attributes from the perturbed quasi-identifiers. We proceed by introducing a second mean squared error term in a combined Lagrangian cost that enables us to regulate the trade-off between quasi-identifier distortion and the confidential-attribute predictability. A Lagrangian multiplier enables us to gracefully weigh the importance of each of the two competing objectives.

Keywords—k-Anonymity, microaggregation, statistical disclosure control, statistical dependence, predictability

I. Introduction

According to the International Data Corporation (IDC), a market research, analysis and advisory firm specializing in information technology, the total amount of global data was expected to grow to 2.7 zettabytes during 2012, that is, roughly 270 million times the size of the American Library of Congress print collections. And yet, this is an increase of 48% in the amount of data from 2011. In fact, IBM claims that every day we create 2.5 exabytes of data, and that 90% of the data in the world today has been created in the last two years alone. From an economic perspective, IDC expects the big-data technology and services market to grow at a 26% compound annual growth rate through 2018 to reach $42 billion.

Within this vast universe of digital information, the presence of personal information is undeniable. Examples of information systems or processes in which potentially sensitive data may be linked to specific individuals or companies, directly via identifying attributes or indirectly through demographic attributes, include, among many others, targeted advertising, recommendation systems and collaborative tagging, social networks, e- and m-health, demographic databases of diverse

* Corresponding author. Tel.: +34 93 401 7027, e-Mail: david.rebollo@entel.upc.edu.
Manuscript revised November 11, 2015.

nature, and e-voting. From a conceptual standpoint, we may often construe these systems or processes as examples of electronic surveys, in the sense that they entail the collection, analysis and dissemination of data combining demographic and confidential attributes, with the ulterior purpose of statistical study.

While one cannot object to the appealing potential of computerized data analysis, the inclusion of rich quantities of sensitive data poses privacy risks that cannot simply remain overlooked. Indeed, personal information, explicitly submitted or implicitly inferable from observed behavior, poses evident privacy risks, especially when combined across several information services, and when enriched with metadata indicating size, location, time, frequency, and other contextual information. On the other hand, it is precisely the availability of such sensitive data that enables the intelligent functionality these modern information technologies offer. In all of these technologies, protecting user privacy while maintaining the utility of the data necessarily supplied to possibly untrusted parties emerge as opposed objectives. Concordantly, these developments are giving rise to a growing number of exciting opportunities, leading to a broad array of applications of ever-increasing societal and economic impact; but these opportunities are not without consequential privacy and security risks. Simply put, as technology progresses, any aspect of our lives as individuals and as a society will be more closely reflected in the digital world, with the privacy implications that this entails.

An example of application where anonymity in statistical analysis is of paramount importance is found in the context of electoral processes. In this case, the introduction of electronic voting opens the door to improving the statistical analysis of traditional elections, based on paper ballots, because e-voting can facilitate the segregation of votes before the counting process. Indeed, in traditional elections, the assignment of voters to a specific polling station according to demographic characteristics is done in advance, but merely on the basis of location, that is, according to the district where the voter resides. Any other valuable profile information on the voter, such as age or gender, is unavailable for statistical analysis. By contrast, electronic votes can be segregated after they are cast, but before they are decrypted and counted. Consequently, the correlation between demographics and political preferences can be analyzed without compromising the voter's privacy, for example by assigning encrypted votes of voters of a certain age of gender to the same ballot box. However, an important challenge in the segregation process described consists in preserving the anonymity of the voters once the votes are decrypted. More precisely, the number of votes for a specific demographic profile should be high enough to prevent the disclosure of the political preferences of the respondents within a given profile group, in terms of the estimated probability that any possible candidate has been chosen. This potential privacy risk is increased when several demographic criteria, say both age and gender, are combined in the segregation described, because less votes should fulfill these criteria simultaneously than individually. Clearly, anonymization countermeasures must be taken into consideration in order to adequately implement this segregation of votes, that is, in order to protect the voter's privacy while enabling valuable statistical analyses based on demographics.

In general, the most extensively studied aspects of privacy for any information system deal with unauthorized access to sensitive data, by means of authentication, policies for data-access control and confidentiality, implemented as cryptographic protocols. However, the provision of confidentiality against unintended observers fails to address the practical dilemma when the intended recipient of the information is not fully trusted. Even more so when the database collected is to be made accessible to external parties, or openly published for scientific correlating sensitive information with demographics.

In this regard, it was famously shown in [42] that 87% of the population in the United States may be unequivocally identified solely on the basis of the triple consisting of their date of birth, gender and 5-digit ZIP code, according to 1990 census data. This is in spite of the fact that in that year, the U.S. had a population of over 248 million. This notorious fact illustrates the discriminative potential of the simultaneous combination of a few demographic attributes, which, considered individually, would hardly pose a real anonymity risk. Ultimately, this simple observation means that the mere elimination of identifiers such as first and last name, or social security number (SSN), is

grossly insufficient when it comes to effectively protecting the anonymity of the participants of published statistical studies containing confidential data linked to demographic information.

Statistical disclosure control (SDC) concerns the postprocessing of the demographic portion of the statistical results of surveys containing sensitive personal information, in order to effectively safeguard the anonymity of the participating respondents. In the SDC terminology, a *microdata set* is a database table whose records carry information concerning individual respondents, either people or companies. This database commonly contains a set of attributes that may be classified into identifiers, quasi-identifiers and confidential attributes. Firstly, *identifiers* allow the unequivocal identification of individuals. This is the case of full names, SSNs or medical record numbers, which would be removed before the publication of the microdata set, in order to preserve the anonymity of its respondents. Secondly, *quasi-identifiers*, also called *key attributes*, are those attributes that, in combination, may be linked with external, usually publicly available information to *reidentify* the respondents to whom the records in the microdata set refer. Examples include age, address, gender, job, and physical features such as height and weight. Finally, the dataset contains *confidential attributes* with sensitive information on the respondent, such as salary, political affiliation, religion, and health condition. The classification of attributes as key or confidential may ultimately rely on the specific application and the privacy requirements the microdata set is intended for.

Intuitively, the perturbation of numerical or categorical quasi-identifiers enables us to preserve privacy to a certain extent, at the cost of losing some of the *data utility*, in the sense of accuracy with respect to the unperturbed version. *k-Anonymity* is the requirement that each tuple of key-attribute values be identically shared by at least k records in the dataset. This may be achieved through the *microaggregation* approach illustrated by the simple example depicted in Fig. 1, where gender, age and ZIP code are regarded as quasi-identifiers, and hourly wage and political affiliation as confidential attributes. Rather than making the original table available, we publish a k-anonymous version containing aggregated records, in the sense that all quasi-identifying values within each group are replaced by a common representative tuple. As a result, a record cannot be unambiguously linked to the corresponding record in any external sources assigning identifiers to quasi-identifiers. In principle, this prevents a privacy attacker from ascertaining the identity of an individual for a given record in the microaggregated database, which contains confidential information.

Ideally, microaggregation algorithms strive to introduce the smallest perturbation possible in the quasi-identifiers, in order to preserve the statistical quality of the published data. More technically speaking, these algorithms are designed to find a partition of the sequence of quasi-identifying tuples in k-anonymous cells, while reducing as much as possible the *distortion* incurred when replacing each original tuple by the representative value of the corresponding cell. For numerical key attributes representable as points in the Euclidean space, the *mean-squared error* (MSE) is the

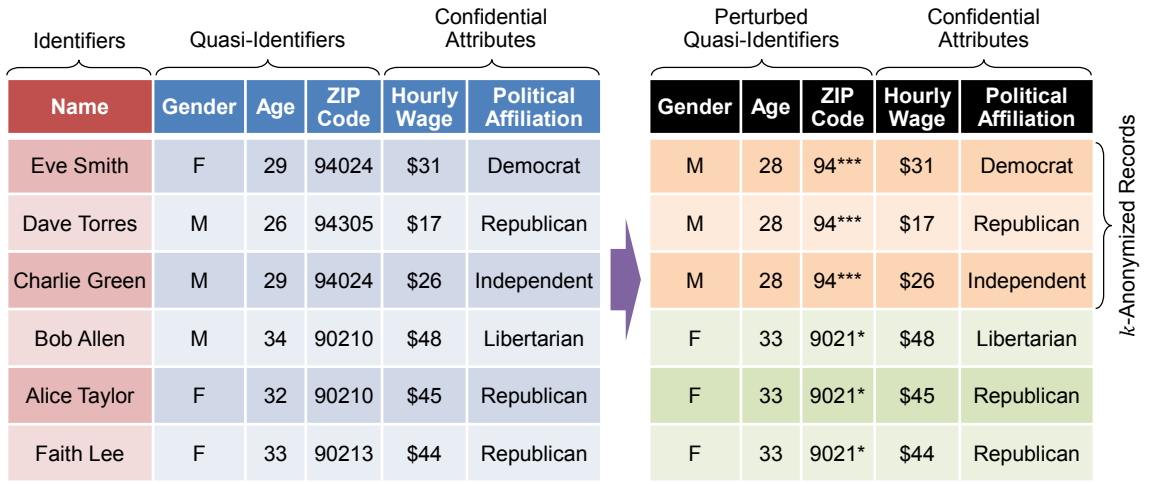


Fig. 1. Example of k-anonymous microaggregation of published data with $k = 3$, showing hourly wages and political affiliation as confidential attributes, in relation to demographic data, gender, age and ZIP code, as quasi-identifiers.

usual criterion to quantify said distortion. Data utility is measured inversely as the distortion resulting from the perturbation of quasi-identifiers.

The state of the art on privacy criteria related to k-anonymity, and on algorithms for k-anonymous microaggregation, is reviewed in §II. Even though, as we shall see, k-anonymity as a quantitative measure of privacy is not without shortcomings, it is still a widely popular criterion in the SDC literature, mainly because of its mathematical and algorithmic tractability.

A. Contribution and Organization

Quantifiable measures of privacy and utility are undoubtedly essential to the assessment, comparison, improvement and optimization of privacy-enhancing mechanisms for information systems in general, and microaggregation algorithms in particular, from both theoretical and numerical perspectives. To a significant extent, controversies in the definition of privacy metrics have eclipsed the intricacies of MSE as a metric for distortion. A particularly important limitation of any measure of loss in data utility consists in neglecting the statistical dependence between quasi-identifiers and confidential attributes. The simplifying assumption that many k-anonymous microaggregation algorithms operate solely on the quasi-identifiers, disregarding the values of confidential attributes, contributes to neglect this dependence further. But we cannot forget that the ulterior purpose of the publication of anonymized data is precisely to enable the study of the statistical relation between demographics and a variety of sensitive data.

In this spirit, the leading object of this paper is to propose a generalization of the widespread measure of loss in data utility, as a mean or sum of squared errors introduced in the perturbation of quasi-identifiers in the k-anonymous microaggregation process.

- More precisely, the proposed measure is a generalization of MSE that extends to the statistical dependence between quasi-identifiers and confidential attributes, and is thus faithful to the intended purposes of database publication for ulterior demographic studies.
- Most conveniently, the new metric remains an MSE for all formal purposes, thereby inheriting the advantages of mathematical and algorithmic tractability offered by the traditional approach.
- Concordantly, our generalized MSE metric is readily applicable to any existing k-anonymous microaggregation algorithms capable of operating with the traditional metric, with no modifications on the algorithm per se. This fact is illustrated experimentally with one of the microaggregation algorithms that constitute the standard the facto in the SDC community.
- Our extensive experimental results on synthetic and standardized datasets carefully relate our generalized utility measure with the traditional understanding of information loss in k-anonymous microaggregation.

From a broader perspective, we must acknowledge that no privacy criterion, including k-anonymity or any of its numerous variants, is the be-all and end-all of database anonymization [32, 2, 14]. Still, thanks to its simplicity, but also in spite of it, k-anonymity is almost universally accepted as a measure of privacy in SDC, at least as a starting point for the design and evaluation of microaggregation algorithms. An entirely analogous argument can be made for MSE as a measure of utility. Far from claiming the absolute exclusivity of MSE to quantify the distortion introduced in the quasi-identifiers in k-anonymous microaggregation, in this work we merely acknowledge the widespread use of MSE for numerical data, and advocate in favor of contemplating the statistical dependence with respect to confidential attributes as well, by generalizing this common metric.

We should hasten to stress that the focus of this work is on k-anonymous microaggregation, in the sense that the risk of disclosure of private information is measured by means of k-anonymity, and the formulation developed in this paper concordantly assumes an underlying k-anonymous microaggregation algorithm. Although the necessarily limited scope of our contribution contemplates only numerical data, MSE and a single yet widely used microaggregation algorithm, many of the ideas put forth here might be further extended to arbitrary metrics for categorical data, and additional microaggregation algorithms. Admittedly, the focus is placed on capturing and illustrating the notion of preserving the statistical dependence between quasi-identifiers and confidential

attributes, in a preliminary albeit systematic, quantifiable manner that offers numerous opportunities for further research.

The remainder of this paper is organized as follows. §II reviews the state of art in anonymity metrics and microaggregation algorithms for SDC. Drawing upon the interplay between statistical dependence and nonlinear predictability, §III formally presents the proposed utility metric, in the context of a quantization model for k-anonymous microaggregation with preservation of the statistical dependence between quasi-identifiers and confidential attributes. We proceed with a theoretical and algorithmic analysis in §IV, while §V numerically illustrates the main results for a variety of synthesized and standardized datasets. Lastly, conclusions are drawn in §VI.

II. Brief Review of the State of the Art on k-Anonymous Microaggregation

Next, we proceed to briefly review the state of the art on microaggregation, with regard to its use and limitations as a measurement of the degree of privacy attained, and the methods and algorithms to construct k-anonymous aggregations with reduced distortion.

A. Widespread Use of k-Anonymity as Privacy Criterion despite its Shortcomings

We mentioned in the introductory section that a specific piece of data on a particular group of respondents is said to satisfy the k-anonymity requirement (for some positive integer k) if the origin of any of its components cannot be ascertained beyond a subgroup of at least k individuals. The original formulation of k-anonymity as a privacy criterion, based on generalization and suppression of key attributes, was modified into the microaggregation-based approach already commented on, in [7, 9, 12, 11]. Both formulations may be regarded as special cases of a more general one utilizing an abstract distortion measure between the unperturbed and the perturbed data, possibly taking on values in rather different alphabets.

We have established that k-anonymity as a privacy criterion ensures that complete reidentification is unfeasible within a group of records sharing the same tuple of perturbed key attribute values. However, if the records in the group also share a common value of a confidential attribute, the association between an individual linkable to the group of perturbed key attributes and the corresponding confidential attribute remains disclosed. More generally, the main issue with k-anonymity as a privacy criterion is its vulnerability against the exploitation of the difference between the prior distribution of confidential data in the entire population, and the posterior conditional distribution of the confidential data within a group given the observed, perturbed key attributes. Simply put, the values of the confidential attributes within a k-anonymous microcell may still be identical, similar, or present a distribution skewed with respect to that of the population, and thus reveal confidential information, in a deterministic or statistical sense. Thus, k-anonymity is inherently vulnerable to several types of well-known attacks, namely the homogeneity or similarity attack, the skewness attack, and the background-knowledge attack [13, 29, 32].

Rapidly since its conception and in spite of the shortcomings already described, this anonymity criterion has gained widespread adoption in the SDC literature [35, 42], mainly because of its mathematical and algorithmic tractability. In fact, in its widespread adoption, the application of the k-anonymity criterion and of the microaggregation methodology goes beyond the publication of databases.

B. Refinements of k-Anonymity, and Alternative Criteria and Metrics

The vulnerabilities of k-anonymity aforementioned motivated the proposal of enhanced privacy criteria, some of which we proceed to sketch briefly, along with modifications in algorithms based on these criteria. A restriction of k-anonymity called *p-sensitive* k-anonymity was presented in [43, 41]. In addition to the k-anonymity requirement, it is required that there be at least p different values for each confidential attribute within the group of records sharing the same tuple of perturbed key attributes. Clearly, large values of p may lead to huge data utility loss. A slight generalization called *l-diversity* [21, 16] was defined with the same purpose of enhancing k-anonymity. The difference with respect to p-sensitivity is that group of records must contain at least l "well-represented"

values for each confidential attribute. Depending on the definition of well-represented, l-diversity can reduce to p-sensitive k-anonymity or be more restrictive. We would like to stress that neither of these enhancements succeeds in completely removing the vulnerability of k-anonymity against skewness attacks. Furthermore, both are still susceptible to similarity attacks, in the sense that while confidential attribute values within a cluster of aggregated records might be p-sensitive or l-diverse, they might also very well be semantically similar for the practical purposes of the attacker.

A privacy criterion aimed at overcoming similarity and skewness attacks is *t-closeness* [19]. An aggregated microdata set satisfies t-closeness if for each group, a predefined measure of discrepancy between the posterior distribution of the confidential attributes within the group, and the prior distribution of the overall population, does not exceed a threshold t. This effectively measures the maximum of the discrepancies for each aggregated group. A particularly useful, information-theoretic metric of discrepancy between probability distributions is the *Kullback-Leibler* (KL) *divergence*, also called *relative entropy* for its relationship with Shannon's entropy. Both Shannon's entropy and KL divergence are also tightly related to the information-theoretic quantity known as *mutual information*, a measure of the uncertainty in one random event unveiled by the outcome of a second, related event [5].

As argued in [13], to the extent to which the within-group distribution of confidential attributes resembles the distribution of those attributes for the entire dataset, skewness attacks will be thwarted. In addition, since the within-group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset. The main limitation of the original t-closeness work [19] is that no general computational procedure to reach t-closeness was specified, with the exceptions of its ready applicability to the Incognito algorithm [18], and the very recent microaggregation procedure proposed in [40].

An information-theoretic privacy criterion, inspired by t-closeness, was proposed in [28, 29]. In the latter work, privacy risk is defined as the conditional KL divergence between the aforementioned posterior and prior distributions, and shown to be equivalent to the mutual information between the confidential attributes and the perturbed quasi-identifiers. This criterion is also tightly related to the concept of *equivocation* introduced by Shannon in 1949 [36], namely the conditional entropy of a private message given an observed cryptogram. A related albeit more conservative criterion, named δ-disclosure, is proposed in [2], and measures the maximum discrepancy between the prior and the posterior distributions.

C. 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 [25], 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 [12, 8], 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 [11], the minimum spanning tree (MST) [17], the variable MDAV (VMDAV) [37] 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 [24].

More recently, an analysis of theoretical optimality in k-anonymous microaggregation [30] extends the necessary (not sufficient) optimality conditions that gave rise to the celebrated Lloyd-Max algorithm [20], 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 con-

ception of the probability-constrained Lloyd (PCL) algorithm [31, 30], which additionally incorporates a variation of the Levenberg-Marquardt algorithm [22], 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.

III. Formal Model Incorporating the Degradation in Statistical Dependence due to Microaggregation

This section formally presents the proposed generalization of MSE as a metric for the loss in data utility due to k-anonymous microaggregation, which takes into account not only the perturbation of quasi-identifiers, but also the degradation of statistical dependence with confidential attributes. This generalization gives rise to the formal statement of the problem of k-anonymous microaggregation with preservation of statistical dependence. Our formal model draws upon a quantization model for general microaggregation, and on the interplay between statistical dependence and nonlinear predictability.

A. Quantization Model for k-Anonymous Microaggregation

The work presented in this paper builds upon our previous formulation of the problem of k-anonymous microaggregation in [30, 31], which formally regards microaggregation as a quantization problem with constraints on the cell probabilities. This subsection adapts the cited model to the problem of microaggregation with preservation of statistical dependence.

Throughout this paper, the measurable space in which a *random variable* (r.v.) takes on values will be called an *alphabet*. We shall follow the convention of using uppercase letters for r.v.'s, and lowercase letters for particular values they take on. *Probability mass functions* (PMFs) are denoted by p and subindexed by the corresponding r.v. In this notation, the probability that a discrete r.v. X takes on the value x is $p_X(x) = \mathrm{P}\{X = x\}$. The *expectation* operator is denoted by E. Expectation can model the special case of averages over a finite set of data points $\{x_1, \dots, x_n\}$, simply by defining an r.v. X uniformly distributed over this set, so that, for instance, $\mathrm{E}\, X = \frac{1}{n}\sum_{j=1}^{n} x_j$.

We shall limit our analysis to the special case of *numerical data*, that is, we shall assume that the quasi-identifiers to be aggregated are represented by n points $x_1, \dots, x_n$ in the Euclidean space $\mathbb{R}^{m_X}$ of dimension m_X, indexed by the corresponding record j, and similarly for the confidential attributes $y_1, \dots, y_n$, in $\mathbb{R}^{m_Y}$. For convenience, we define an r.v. J representing the record index, uniformly distributed on the set of indices $\{1, \dots, n\}$. Note that J may also be regarded as the identity of the respondent. In addition, we introduce an r.v. X representing the quasi-identifiers, whose alphabet consists in the set of m_X-dimensional points $(x_j)_{j=1,\dots,n}$, formally definable as a function $X = x_J$ of J, and an r.v. Y modeling the confidential attributes, taking on values in the set of m_Y-dimensional points $(y_j)_{j=1,\dots,n}$, defined by the indexing function $Y = y_J$. The joint r.v. (X, Y) models the pairs of quasi-identifiers and confidential attributes that form a table $(x_1, y_1), \dots, (x_n, y_n)$ of n records. The notation in terms of r.v.'s will enable us to write averages more compactly as expectations, for instance

$$\frac{1}{n}\sum_{j=1}^{n} f(x_j, y_j) = \mathrm{E}\, f(X, Y),$$

for any function f of the pair of quasi-identifiers and confidential attributes.

The k-anonymous microaggregation algorithm will partition the set of records into microcells of size at least k. In the most traditional form of k-anonymous microaggregation, this partition only takes into account the values of the quasi-identifiers, but in our form of microaggregation with preservation of statistical dependence, the partition design will naturally involve the confidential attributes as well, even if the published table will only effectively perturb the quasi-identifiers and keep the confidential attributes intact.

The resulting microcells will be labeled with a quantization index q. An important subtlety is that the microaggregation process must be formally construed as a quantization function $q(j)$ of the record index j, rather than on the quasi-identifier x, or more generally, on the pairs (x, y) of quasi-identifiers and confidential attributes. The reason is that even though $q(j)$ also induces a partition on the set of quasi-identifiers and confidential attributes, one cannot discard the possibility that some pairs (x, y) might be repeated, and that those repeated values might be assigned to different microcells. Although this could be technically handled with probabilistic microcell assignments, it is simpler and completely general to define a (deterministic) quantization function on the record indices. In our more compact representation with r.v.'s, we define $Q = q(J)$, with finite alphabet $\{1, \dots, |\mathscr{Q}|\}$. The k-anonymity constraint is contemplated by imposing a constraint on cell sizes or, more generally, on the probabilities of the quantization indices $p_Q(q) \geqslant k/n$.

In traditional numerical microaggregation, it is an almost universal convention to measure the distortion introduced in the quasi-identifiers by means of the MSE, and to employ the term distance to refer to its Euclidean definition. Accordingly, unless otherwise stated, the term distance refers to its Euclidean definition. Accordingly, recall that the *centroid* $\hat{x}(q)$ of a subset of n_q points of $x_1, \dots, x_n \in \mathbb{R}^m$ assigned to the q^{th} microcell, is defined as the point that minimizes the MSE with respect to that subset, and that it is, quite simply, the conditional expectation $\mathrm{E}[X|q]$ of X given $Q = q$, which boils down to a vector average, formally,

$$\hat{x}(q) = \arg\min_{\hat{x}} \mathrm{E}[\|X - \hat{x}\|^2 | q] = \mathrm{E}[X|q] = \frac{1}{n_q} \sum_{j \mid q(j)=q} x_j \,.$$

Analogously define the r.v. $\hat{X} = \hat{x}(Q)$, modeling the reconstructed quasi-identifier.

Still in the special case of traditional microaggregation, the entire microaggregation process, which transforms the record index J into the perturbed quasi-identifier $\hat{X}$, can be represented as the composition of two functions, namely the microcell assignment $q(j)$ and the centroid assignment $\hat{x}(q)$, as depicted in Fig. 2. We have mentioned that the problem of microaggregation, may be

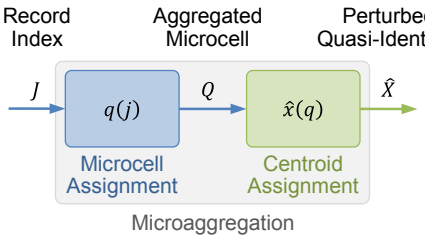


Fig. 2. Traditional microaggregation interpreted as a quantization problem on the record indices j represented by the microcell assignment function $q(j)$, and a centroid assignment function $\hat{x}(q)$ that reconstructs the perturbed version $\hat{x}_j$ of the original quasi-identifier x_j. In the analogous r.v. representation, $Q = q(J)$ and $\hat{X} = \hat{x}(Q)$.

formally understood as a constrained quantization problem, as explained in [30, 31]. This interpretation is particularly intuitive in the special case of traditional microaggregation with numerical quasi-identifiers that do not appear repeatedly. In this special case, the function representing the microcell assignment can be defined directly on the quasi-identifiers, as $q(x)$. The corresponding representation is shown in Fig. 3.

We must recall that it is customary in traditional microaggregation to conduct a *columnwise, unit-variance normalization* of all numerical quasi-identifiers, prior to any manipulation of the data, because it is inherent in the conventional definition of distortion error in SDC. This means that the *total variance* of the data points, that is, the sum of the columnwise variances, will amount to the dimension m of the quasi-identifiers. A zero-mean normalization is also customary, but it bears no theoretical difference in terms of the performance of the microaggregation algorithm, as it merely represents a translation of the data points.

Let us denote the perturbed version of the j^{th} quasi-identifier x_j, that is, its corresponding centroid, by $\hat{x}_j$. The SDC literature conventionally speaks of the *sum of squared errors* (SSE) and the *sum of squares total* (SST). Precisely,

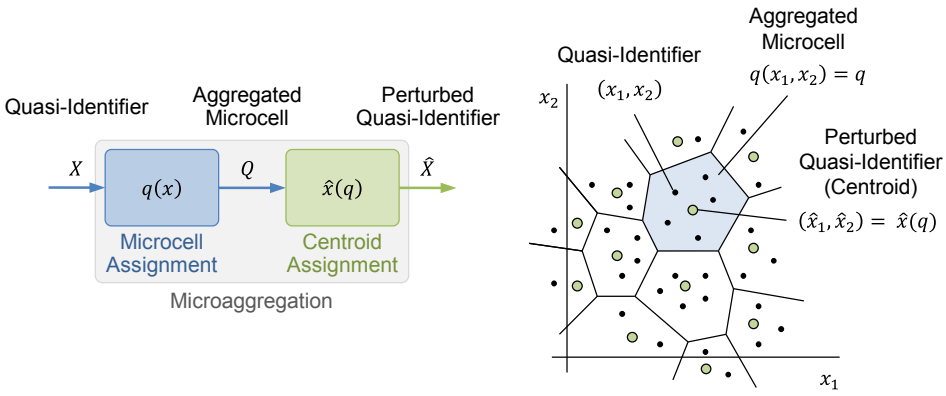


Fig. 3. Traditional microaggregation viewed as a quantization problem on the quasi-identifiers. Although under this interpretation is more intuitive to define the microcell assignment function $q(x)$ directly on the quasi-identifiers x, one cannot discard the possibility that those quasi-identifiers might be repeated across records, and that repeated values might be assigned to different microcells. It is then more rigorous to define a quantization function $q(j)$ on the record index j.

$$\mathrm{SSE}_X = \sum_{j=1}^{n} \|x_j - \hat{x}_j\|^2 ,$$

and since the total variance of the data is the dimension m_X, its unnormalized version becomes $\mathrm{SST}_X = m_X n$. In this work, we formally define the distortion $\mathcal{D}_X$ introduced by the microaggregation algorithm by means of the MSE, implicitly normalized by the number n of samples, and also normalized by the number m of dimensions, that is,

$$\mathcal{D}_X = \frac{1}{m_X} \mathrm{E}\|X - \hat{X}\|^2 = \frac{1}{m_X n} \sum_{j=1}^{n} \|x_j - \hat{x}_j\|^2 = \frac{\mathrm{SSE}_X}{\mathrm{SST}_X} .$$

The performance indicator commonly evaluated in the SDC literature is the quotient between the SSE and the SST, always in the range [0,1], provided that the reconstructions are indeed centroids. This quotient matches our definition of distortion as MSE per dimension.

We now turn back to the more general case of microaggregation with preservation of statistical dependence, the object of this paper. We shall argue later that the suitability of the k-anonymous partition will be assessed in terms of a an additional reconstructed centroid, this time for the confidential attributes, although we should hasten to stress that this additional centroid is merely an intermediate tool, and that the final anonymized table will only contain perturbed quasi-identifiers. With this in mind, analogously to $\hat{X} = \hat{x}(Q)$ for the quasi-identifier r.v. X, let the r.v. $\hat{Y}$ denote the reconstruction of the r.v. Y representing confidential attributes, and let the function $\hat{y}(q)$ model the corresponding centroid assignment. At this point, it suffices to characterize $\hat{y}(q)$ as an arbitrary function, but we shall see that it is in fact a centroid, in the sense defined earlier. A graphical interpretation of the original and the anonymized tables under the probabilistic model proposed is shown in Fig. 4, where we consistently omit $\hat{Y}$.

B. Multiobjective Optimization Criterion for the Loss in Data Utility Incorporating the Degradation in Statistical Dependence

In the following we elaborate on possible optimization criteria corresponding to the objective of reducing the distortion between X and $\hat{X}$, while preserving the statistical dependence originally between X and Y, through $\hat{X}$ and Y. Before proceeding any further with our model, we make a quick digression on the quantification of statistical dependence as a nonlinear-prediction error.

Very early work [39, 38] on the effect of microaggregating quasi-identifiers on their statistical dependence with confidential attributes tentatively resorted to small modifications of traditional algorithms and then to measure the variations incurred on certain correlation coefficients. We would like to acknowledge that our approach is somewhat inspired by the cited work, even though it dramatically departs from any measurement based on correlation coefficients due to the limited connection between correlation and general, nonlinear statistical dependence. Further, our current

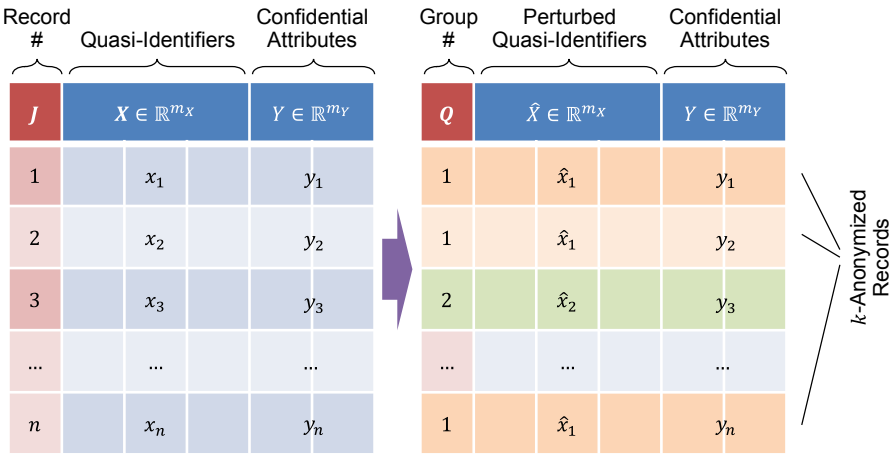


Fig. 4. A graphical interpretation of the original and the k-anonymously microaggregated tables under the formulation proposed.

$$X \sim U\{-2,-1,1,2\} \quad y(x) = \begin{cases} 1, |x| = 1 \\ -1, |x| = 2 \end{cases} \text{ and } \quad Y = y(X)$$

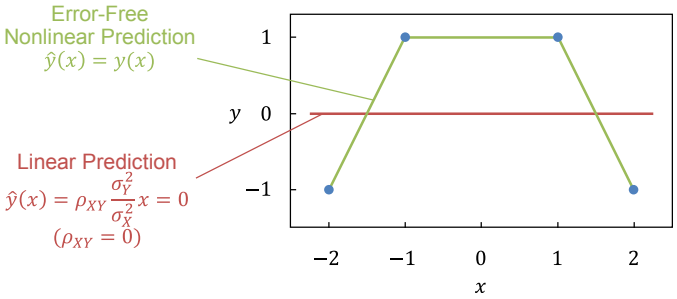


Fig. 5. The two r.v.'s X and Y are uncorrelated yet strongly dependent, as the latter is a function of the former. Consequently, the best nonlinear prediction is a far better estimate than the linear variant.

approach systematically formulates microaggregation functions designed specifically with that dependence in mind, rather than as an afterthought. Finally, we shall argue in favor of quantifying statistical dependence by means of nonlinear predictability.

In order to capture the statistical dependence between the quasi-identifier X and the confidential attribute Y, and that preserved in the published versions, namely the reconstructed quasi-identifier $\hat{X}$ and the original Y, one might consider a variety of crosscorrelation quantities. Recall, however, that two r.v.'s may be uncorrelated yet strongly dependent. For example, let X be uniformly distributed on $\{-2,-1,1,2\}$, and set $Y = 1$ when $|X| = 1$, and $Y = -1$ when $|X| = 2$, as illustrated in Fig. 5. Denote means and covariances with μ and σ^2, appropriately subindexed. It is straightforward to verify that the correlation coefficient ρ_{XY} between such variables is zero. Consequently, the best *linear* prediction of Y from X in the MSE sense is, disappointingly,

$$\hat{Y}_{\text{lin}} = \rho_{XY} \frac{\sigma_Y^2}{\sigma_X^2}(X - \mu_X) + \mu_Y = 0.$$

However, X and Y are an extreme case of strong statistical dependence, as the latter is a function of the former. Hence, the best *nonlinear* prediction in the MSE sense is

$$\hat{Y} = \mathrm{E}[Y|X] = Y,$$

and thus error free.

This simple example suffices to illustrate that a more suitable way of capturing the statistical dependence between the perturbed quasi-identifiers $\hat{X}$ and the confidential attributes Y in the published database is through the *nonlinear* predictability of Y from $\hat{X}$. Because there is a one-to-one correspondence between $\hat{X}$ and the quantization index Q, this is also the predictability of Y

from Q, expressed by the reconstruction function $\hat{y}(q)$. Hence, this predictability may be factored in when measuring the appropriateness of the k-anonymous partition in terms of the degradation in statistical dependence incurred. Of course, the original objective of preserving the values of the quasi-identifiers will have an impact on the choice of the reconstruction $\hat{x}(q)$ of X from Q.

Concordantly, we propose that the quantizer $q(j)$ be designed to minimize the a multiobjective distortion functional, formally a Lagrangian cost incorporating the traditional information loss term,

$$\mathcal{D}_X = \frac{1}{m_X} \mathrm{E}\|X - \hat{X}\|^2 = \frac{\mathrm{SSE}_X}{\mathrm{SST}_X}$$

along with an additional term, analogously defined,

$$\mathcal{D}_Y = \frac{1}{m_Y} \mathrm{E}\|Y - \hat{Y}\|^2 = \frac{\mathrm{SSE}_Y}{\mathrm{SST}_Y}.$$

characterizing the degradation in statistical dependence by means of the error in the predictability of Y from $\hat{X}$, or equivalently as argued, Q. The corresponding Lagrangian distortion is

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

where the Lagrangian multiplier $\lambda \in [0,1]$ enables us to control the trade-off between the two optimization objectives, sweeping the lower convex envelope of possible quantizers in the $\mathcal{D}_X$-$\mathcal{D}_Y$ plane. Recall that the Lagrangian formulation of a multiobjective optimization problem offers a more tractable, unconstrained form of an originally constrained problem, but only constrained solutions that lie on the lower convex envelope can be obtained in this fashion [1]. This potential situation is illustrated in Fig. 6. Still, Lagrangian costs are commonplace in quantizer design in various applications of lossy coding and in the general context of rate-distortion theory [27].

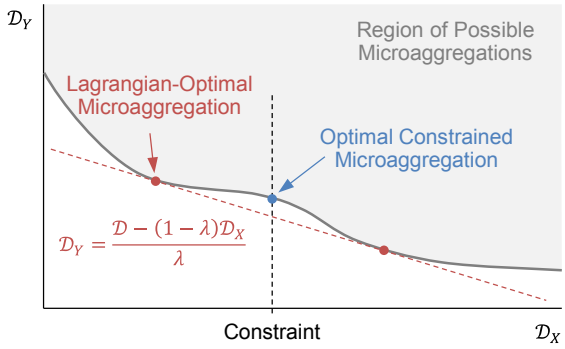


Fig. 6. Lagrangian formulation of the microaggregation problem with the original distortion objective $\mathcal{D}_X$ representing the information loss due to the perturbation of the quasi-identifiers, and the additional distortion objective $\mathcal{D}_Y$ representing the degradation in statistical dependence with the confidential attributes.

In our generalized formulation, traditional k-anonymous microaggregation corresponds to the special case of $\lambda = 0$. In contrast, $\lambda = 1$ corresponds to the case in which we wish to completely favor statistical dependence over quasi-identifier distortion. For any $\lambda > 0$, we must employ the centroid $\hat{Y} = \hat{y}(Q)$ as a prediction of the confidential attribute Y from the microcell label Q, which we introduced earlier in our formulation. Because $\hat{Y}$ must minimize an MSE,

$$\hat{y}(q) = \arg\min_{\hat{y}} \mathrm{E}[\|X - \hat{y}\|^2 | q] = \mathrm{E}[Y|q] = \frac{1}{n_q} \sum_{j \mid q(j)=q} y_j\,,$$

entirely analogously to the centroid $\hat{x}(q)$ for the quasi-identifiers. The only difference is, as we already mentioned, that the reconstruction $\hat{Y}$ is an internal tool for the design of the microaggregation function $q(j)$, but it will not be part of the published database resulting from its application, as we made explicit in Fig. 4.

Our complete formulation of the problem of k-anonymous microaggregation with preservation of statistical dependence is graphically summarized in Fig. 7, in the simpler, more intuitive case

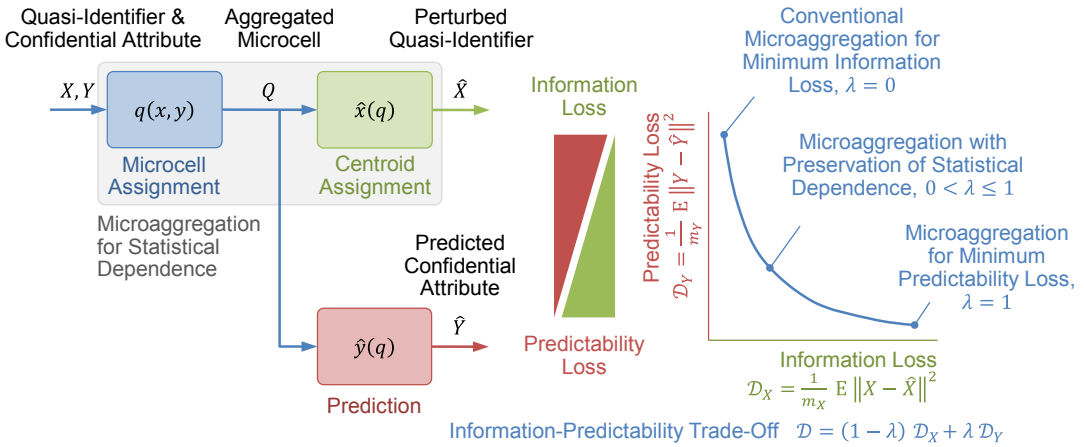


Fig. 7. Microaggregation with preservation of statistical dependence construed as a quantization problem on the pairs (x, y) of quasi-identifiers and confidential attributes. Although it may be more intuitive to define the microcell, assignment function $q(x, y)$ on those data pairs, strictly speaking, such pairs could be repeated across records. Therefore, it is slightly more general to define the quantization function $q(j)$ on the record index j.

when the pairs (x, y) of quasi-identifiers and confidential attributes are not repeated, and the microcell assignment function can thus be defined directly on them. Recall that a more general definition of the quantization function requires that it be defined in terms of the record index j.

Interestingly, minimizing the term $\mathrm{E}\|Y - \hat{Y}\|^2$ turns out to be equivalent to maximizing the (inner) correlation coefficient

$$\frac{\mathrm{E}\left[Y^T \hat{Y}\right]}{\mathrm{E}\|\hat{Y}\|^2 \, \mathrm{E}\|Y\|^2}$$

between the confidential data and its predictions $\hat{Y} = \hat{y}(Q)$ from the microcell label Q. If for any reason we wanted to maximize a cross-correlation between $\hat{X}$ and Y, it must be pointed out that some form of normalization would be needed not to artificially increase it simply by scaling $\hat{X}$. This means that in a scalar version of the problem, a covariance would not be appropriate, but a correlation coefficient might.

Finally, we should hasten to point out that as in many other problems with multiple optimization objectives, said objectives often represent contrasting quantities that pose trade-offs inherent to the problem at hand. In the particular case of k-anonymous microaggregation with preservation of statistical dependence, we wish to maximize privacy, characterized by k, minimize the information loss in the perturbation of the quasi-identifiers, quantified by $\mathcal{D}_X$, and minimize the degradation in statistical dependence with the confidential attributes, measured by $\mathcal{D}_Y$. Naturally, as they constitute contrasting aspects, a compromise must be reached.

Along these lines, an intricate aspect of the work presented here is the choice of criteria, and the fact that improving the predictability of the confidential attributes within the microcells might have a negative impact on the similarity and skewness vulnerabilities already present in traditional microaggregation, as explained in §II.A. This is a possibility that we shall evaluate experimentally later in §V in terms of a certain measure of l-diversity. It is important to realize that any similar criterion for the preservation of statistical dependence will suffer from the same issue. This is nothing but the manifestation of the inherent privacy-utility trade-off in microaggregation, where one strives to minimize and maximize information, commonly, but not necessarily, measured with different criteria, for example worst-case for privacy and average-case for utility.

Although the scope of this manuscript is limited to the case of numerical data, the general case of mixed numerical and categorical data might be addressed by means of distortion measures $d_X(x, \hat{x})$ and $d_Y(y, \hat{y})$ quantifying the level of distortion between original and perturbed or predicted elements, and the generalized Lagrangian distortion

$$\mathcal{D} = (1 - \lambda)\, \mathrm{E}\, d_X\big(X, \hat{X}\big) + \lambda \mathrm{E}\, d_Y\big(Y, \hat{Y}\big).$$

The numerical case tackled here corresponds to the quadratic distortion measures

$$d_X(x,\hat{x}) = \|x-\hat{x}\|^2$$
$$d_Y(y,\hat{y}) = \|y-\hat{y}\|^2.$$

Examples of categorical measures include the Hamming distance, for which the associated distortion amounts to the probability of error, that is,

$$d_Y(y,\hat{y}) = \begin{cases} 0, & y = \hat{y} \\ 1, & y \neq \hat{y} \end{cases},$$
$$\mathcal{D}_Y = \mathrm{E}\, d_Y(Y,\hat{Y}) = \mathrm{P}\{Y \neq \hat{Y}\},$$

or graph distances for ontologies.

IV. Theoretical Analysis of Microaggregation with Preservation of Statistical Dependence

Part of the theoretical elegance and a great deal of the practical applicability of the formulation laid out in the previous section are a consequence of the fact that the general formulation can be reduced back to the special case of traditional microaggregation. Indeed, we claim that the Lagrangian objective with two MSE terms can be expressed as a single MSE term. After showing the formal equivalence between the traditional approach and the one presented in this paper, we shall proceed to reuse theoretical results on the optimality of k-anonymous partitions for the problem at hand.

A. Principle of Formal and Algorithmic Equivalence

For the intermediate case $0 < \lambda < 1$, define

$$\beta = \sqrt{\frac{\lambda}{1-\lambda}\frac{m_X}{m_Y}}.$$

In lieu of the Lagrangian, multiobjective distortion $\mathcal{D}$, we equivalently consider a proportional distortion

$$\frac{m_X}{1-\lambda}\mathcal{D} = \mathcal{D}_X + \beta^2\mathcal{D}_Y = \mathrm{E}\|X-\hat{X}\|^2 + \mathrm{E}\|\beta Y - \beta\hat{Y}\|^2 = \mathrm{E}\left\|\begin{pmatrix} X \\ \beta Y \end{pmatrix} - \begin{pmatrix} \hat{X} \\ \beta\hat{Y} \end{pmatrix}\right\|^2.$$

Observe that the last equality combines quasi-identifiers and confidential attributes into extended points of the form $(X, \beta Y)$ in $\mathbb{R}^{m_X+m_Y}$. Because the centroid corresponding to βY is

$$\widehat{\beta Y} = \mathrm{E}[\beta Y|Q] = \beta\,\mathrm{E}[Y|Q] = \beta\hat{Y},$$

the centroid corresponding to $(X, \beta Y)$ is $\left(\hat{X}, \beta\hat{Y}\right)$. Ergo, the problem of k-anonymous microaggregation with preservation of statistical dependence is equivalent to that of traditional microaggregation, redefined on $m_X + m_Y$-dimensional data points of the form $(X, \beta Y)$, as far as the design of the microaggregation function $q(j)$ is concerned. Clearly, for the practical purposes of database publication, only $\hat{X}$ will be considered, and $\hat{Y}$, merely an intermediate tool, disregarded.

- In other words, for any $0 < \lambda < 1$, any traditional microaggregation algorithm can be enhanced to incorporate the objective related to the preservation of statistical dependence, simply by replacing its input by the extended version $(x_j, \beta y_j)_{j=1,\ldots,n}$.
- Only the extreme cases $\lambda = 0{,}1$ are left. Clearly, for $\lambda = 0$, traditional microaggregation can be carried out as usual.
- The extreme case $\lambda = 1$ in which statistical dependence is completely favored over quasi-identifier accuracy, the problem is formally equivalent to that of traditional microaggregation simply by exchanging the roles of X and Y in the design of the partition function $q(j)$. In this latter case, we would first microaggregate Y traditionally, and then use the resulting groups on X, from which we would finally compute the centroids $\hat{X}$, published in lieu of $\hat{Y}$.

It is important to note that in neither case, that is, for absolutely no value of $\lambda \in [0{,}1]$, does the microaggregation algorithm itself need to be modified at all, only its input. The experimental §V exploits this principle for a well-known microaggregation algorithm applied to a number of datasets.

B. Necessary Optimality Conditions

As in the classical problem of *optimal vector quantization for lossy source coding* [15], the optimal solution is partly characterized in terms of two necessary (but not sufficient) optimality conditions, one for the optimal quantizer given a reconstruction, and another for the optimal reconstruction given a quantizer, both a direct application of *Bayes decision theory*. Still in the classical problem, in practice, these conditions may be applied iteratively as in the *Lloyd algorithm*, or can inspire alternative, partly optimized (but not fully optimal) methods. These fundamental principles were first extended to k-anonymous macro- and microaggregation in [30, 31], further proposing an algorithm, PCL, already introduced in §II.C, capable of outperforming state-of-the-art methods on a wide variety of standardized datasets.

B.1. Centroid Conditions

The principle of equivalence between traditional microaggregation and its variant for preservation of statistical dependence may be exploited to immediately extend the necessary optimality conditions of [30] to the problem formulated in this paper. §III already stated the *centroid conditions*, a simple consequence of the fact that the expectation is the best MSE estimate, which we restate here for convenience:

$$\left.\begin{aligned}\hat{x}^*(q) &= \arg\min_{\hat{x}} \mathrm{E}\,\|X-\hat{x}\|^2 = \mathrm{E}\,[X|q]\\ \hat{y}^*(q) &= \arg\min_{\hat{y}} \mathrm{E}\,\|X-\hat{y}\|^2 = \mathrm{E}\,[Y|q]\end{aligned}\right\}.$$

Unfortunately, this condition does not fully characterize an optimal solution, for two reasons. First, it is a necessary condition, not sufficient. Secondly, the condition is expressed in terms of a given microcell assignment function $q(j)$,which may or may not be optimal.

B.2. Modified Nearest-Neighbor Condition

Similarly, the *modified nearest-neighbor condition* of [30] is also necessary, not sufficient, and it is expressed in terms of the centroids, which may not be optimal. Further, it involves a real-valued Lagrangian regularization function $c(q)$ to ensure that certain constraints on the cell probabilities $p_Q(q)$ are met, including the case of k-anonymous aggregation, for which $p_Q(q) \geqslant k/n$. The principle of equivalence tells us that this condition holds for the problem at hand, so long as the extended input $(X, \beta Y)$ is considered instead. Precisely, under the assumption that a weighting function $c(q)$ exists for the given probability constraints, the optimal microcell assignment must be of the form

$$q^*(j) = \arg\min_q \left\| \begin{pmatrix} x_j \\ \beta y_j \end{pmatrix} - \begin{pmatrix} \hat{x}(q) \\ \beta\hat{y}(q) \end{pmatrix} \right\|^2 + c(q) = \arg\min_q \; \|x_j - \hat{x}(q)\|^2 + \beta^2 \|y_j - \hat{y}(q)\|^2 + c(q)\,.$$

The additive weights $c(q)$ modify the resulting Voronoi partition; larger values shrink cell sizes, while smaller values enlarge it, until the k-anonymity constraints are met. The result on the convex shape of the microcells of [30] (Theor. 3) also holds for the partition on the space of extended points $(x, \beta y)$, as well as for the cross section involving only x.

PCL, the microaggregation algorithm proposed in the cited work [30, 31], resorted to the Levenberg-Marquardt algorithm, a method for solving systems of nonlinear equations, for the adjustment of $c(q)$. The same could be done here. However, this paper emphasizes the objective of preserving statistical dependence over the intricate details of the underlying microaggregation method. This, along with the fastest running time of MDAV, led us to employ this latter algorithm in the experimental analysis of §V.

Incidentally, we would like to offer a simpler proof of the statement regarding the nearest-neighbor condition in [30] (Theor. 4). Consider a regularization function $c(q)$, and a microcell assignment $q^*(j)$ satisfying the nearest-neighbor condition, with cell probabilities $p_Q(q)$. Then, any other assignment $q(j)$ satisfying the same equality constraints must have a larger distortion, be it the distortion corresponding to the traditional case or the general case with statistical dependence.

To see this, denote the distortion of $q^*(j)$ with $\mathcal{D}^*$, and that of $q(j)$ with $\mathcal{D}$. Consider the Lagrangian cost $\mathcal{L} = \mathcal{D} + \mathrm{E}\,c(Q)$ of $q(j)$, and similarly denote $\mathcal{L}^*$for $q^*(j)$. By definition, $q^*(j)$ minimizes the Lagrangian cost, i.e., $\mathcal{L}^* \leqslant \mathcal{L}$. Because both $q^*(j)$ and $q(j)$ have the same value for $\mathrm{E}\,c(Q)$, we can immediately conclude that $\mathcal{D}^* \leqslant \mathcal{D}$, as claimed. In the slightly more general case of

inequality constraints, the solution must still have certain cell probabilities, and replacing the inequality constraints by equality constraints for those probabilities would yield the same result. Consequently, the nearest-neighbor condition is necessary.

As a final remark of theoretical interest, recall that we have defined the microcell assignment function $q(j)$ on the record index j rather than on the data pairs (x, y) of quasi-identifiers and confidential attributes, to accommodate the general case in which those pairs might be repeated, rendering any function of the form $q(x, y)$ ambiguous. Even in the case of numerical data this repetition is possible, especially when the data is a coarse representation with limited precision or within predefined numerical categories. However, in the case when the quasi-identifiers x are not repeated, standard techniques for modified distortion measures in noisy quantization theory [27, 33, 34] may be exploited to reduce the general problem to the traditional variant, with extended points of the form $(x, \beta \mathrm{E}[Y|x])$, yielding a nearest-neighbor condition that can be expressed in terms of a microcell assignment function $q(x)$ on x, precisely,

$$q^*(x) = \arg\min_q \ \|x - \hat{x}(q)\|^2 + \beta^2 \|\mathrm{E}\,[Y|x] - \hat{y}(q)\|^2 + c(q)\,.$$

V. Experimental Results

An essential aspect of our contribution consists in the empirical investigation of the formalisms presented in previous sections. In the necessarily limited extent of this experimental section, we place greater emphasis on the functional verification of the soundness of our approach, over subordinate aspects of algorithmic efficiency. The last subsection preliminarily ventures an experimental assessment of the application of the ideas put forth here, to other microaggregation metrics and mechanisms beyond k-anonymity.

A. Experimental Setup

Precisely, we focus on capturing and illustrating the notion of preserving the statistical dependence between quasi-identifiers and confidential attributes in a systematic, quantifiable manner, for a variety of synthetic and standardized datasets. Our experiments employ one of the microaggregation algorithms that constitute the standard the facto in the SDC community, known as MDAV [12, 8], already introduced in §II.C. The specification of MDAV followed here is that given as Algorithm 5.1 in [12], named "MDAV-generic":

1. Find the centroid of the dataset, find the furthest point P from the centroid, and find the furthest point Q from P.
2. Group the $k-1$ nearest points to P into a group, and then do the same with the $k-1$ nearest points to Q.
3. Repeat steps 1 and 2 on the remaining points until there are less than $2k$ points.
4. If there are k to $2k-1$ points left, form a group with those and finish. Else, if there are 1 to $k-1$ points, adjoin them to the last (hopefully nearest) group.

As for the datasets considered, we first synthesize 1000 samples of 2-dimensional Gaussian data with zero-mean, unit-variance components, and correlation coefficients 0.5 and 0.9 respectively, that is, with covariance matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

and $\rho = 0.5, 0.9$, respectively, signifying different degrees of statistical dependence. The first dimension is taken to represent a scalar quasi-identifier, and the second, a scalar confidential attribute. Additionally, two standardized datasets are considered.

First, the numerical, standardized dataset "Census", was used in the project "computational aspects of statistical confidentiality" (CASC) [3], and has since then been served as a widely spread comparison test in the SDC literature. It contains 1080 records with 13 numerical attributes, from which the first 6 attributes are taken as quasi-identifiers, and the remaining 7 as confidential attributes. In addition to CASC, examples of research studies utilizing this dataset include [9, 17].

The second standardized dataset is "Tarragona", containing 834 records with 13 numerical attributes. Just as before, the first 6 attributes are taken as quasi-identifiers, and the remaining 7 as confidential. Aside from the CASC project, this dataset has been used, for instance, in [10, 6, 44, 12, 17]. An interesting property of this dataset is that it is resistant against microaggregation algorithms that exploit variable-size strategies, natural clusters and heavy skewness of the data, such as μ-Approx [17] or VMDAV [37], making MDAV the best choice.

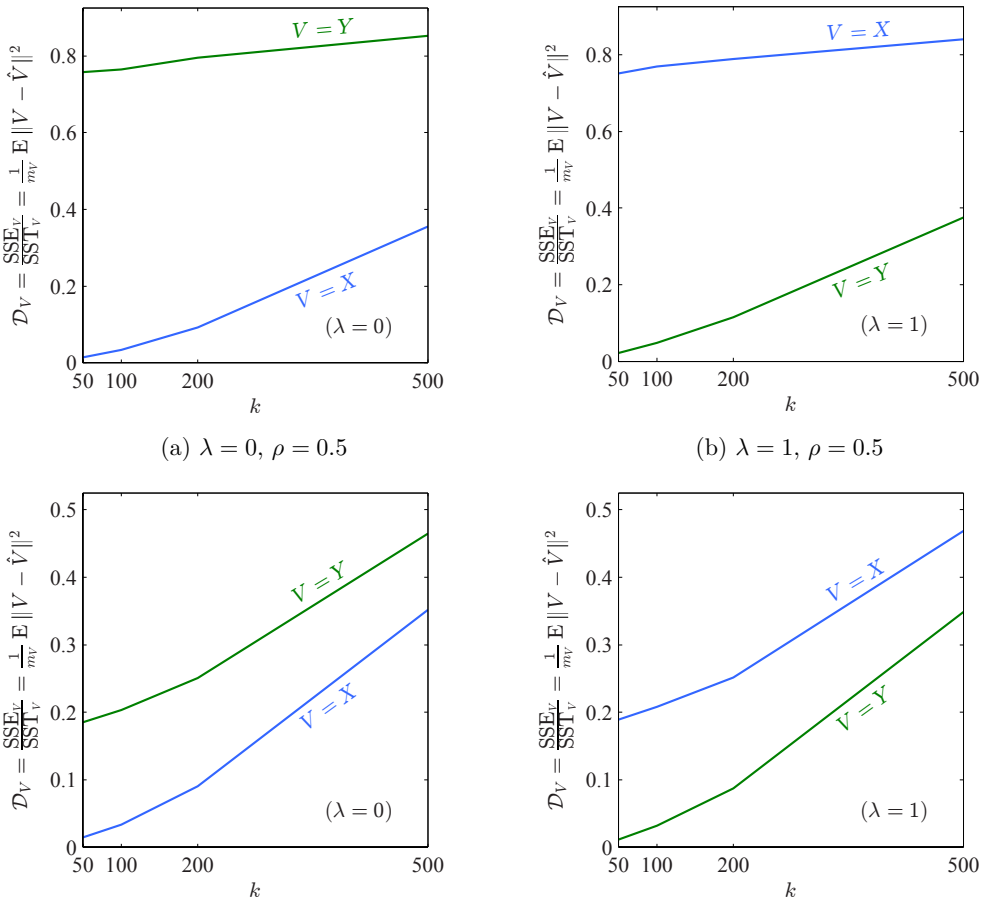
We follow the common practice of normalizing each column of the data set for zero mean and unit variance. We explored a reasonably wide range of target anonymity constraints, $k = 50, 100, 200, 500$ for the synthetic, Gaussian datasets, and $k = 5, 10, 20, 50, 100$ for the standardized datasets, broadly representative of the values in the microaggregation literature. Observe that since every column of the dataset underwent unit-variance normalization, the total variance of the dataset is its number of dimensions. Because our measure of distortion is normalized by the number of dimensions, as we mentioned in §III.A, the distortions $\mathcal{D}_X$ and $\mathcal{D}_Y$ for quasi-identifiers and confidential attributes reported here are equivalent to the popular SDC measure of sum of squared errors (SSE) divided by sum of squared total (SST), always in the range [0,1].

In all cases, the Lagrangian multiplier λ regulating the trade-off between $\mathcal{D}_X$ and $\mathcal{D}_Y$ takes on values from 0 to 1 in steps of 0.1, thus $\lambda = 0, 0.1, 0.2, \ldots, 0.9, 1$. As explained in §IV.A, for any $\lambda \in (0,1)$ MDAV is fed data pairs of the form $(x, \beta y)$, where x represents quasi-identifiers and y confidential attributes, and β is the normalized form of λ defined in that section. The extreme cases $\lambda = 0, 1$ are more simply handled with inputs x and y, respectively, as detailed in the aforementioned section. Bear in mind that because MDAV is a heuristic algorithm, it cannot guarantee that the k-anonymous partition minimize the Lagrangian distortion $\mathcal{D}$. Consequently, in principle, the $\mathcal{D}_X$-$\mathcal{D}_Y$ trade-off curves may be neither convex nor decreasing.

B. Experimental Findings

The first set of plots, in Fig. 8, shows the quasi-identifier distortion $\mathcal{D}_X$ and the confidential-attribute distortion $\mathcal{D}_Y$ as a function of the anonymity parameter k, for the extreme values of the Lagrangian multiplier $\lambda = 0, 1$. Recall that $\lambda = 0$ corresponds to traditional microaggregation, whereas $\lambda = 1$ represents the case in which the preservation of statistical dependence is completely favored over quasi-identifier distortion. The plots confirm the intuition that $\mathcal{D}_X \leqslant \mathcal{D}_Y$ for $\lambda = 0$, and conversely for $\lambda = 1$. Naturally, either distortion increases with k, although not to the same extent. Indeed, the objective distortion, that is, $\mathcal{D}_X$ for $\lambda = 0$ and $\mathcal{D}_Y$ for $\lambda = 1$,drastically improves with lower k. However, the changes in the secondary distortion, that is, $\mathcal{D}_Y$ for $\lambda = 0$ and $\mathcal{D}_X$ for $\lambda = 1$, should be more pronounced and closer in value to the distortion objective in the case of greater statistical dependence, as the plots with synthetic data illustrate.

Figs. 9–12 plot an approximation with MDAV to the trade-off between quasi-identifier distortion $\mathcal{D}_X$.and confidential-attribute distortion $\mathcal{D}_Y$ for various values of the k-anonymity parameter and the Lagrangian multiplier λ. As expected, for a given k, $\mathcal{D}_Y$ as a function of $\mathcal{D}_X$ is roughly decreasing and convex, as one might expect from the considerations made in §III.B, although not exactly so because MDAV is merely a heuristic algorithm which cannot guarantee the optimization of the Lagrangian, combined distortion $\mathcal{D}$. For greater values of the minimum cell size k, both distortions consistently increase. This distortion scaling causes the trade-off to become sharper for lower values of k, and weaker for higher values. Because distortions are normalized in the range [0,1], higher values of the anonymity parameter k will tend to produce distortions closer to the upper bound 1. For extremely large cells, the trade-off would collapse to the point (1,1) on the $\mathcal{D}_X$-$\mathcal{D}_Y$ plane. A sharper trade-off, as that found for low k, means that a significant gain in in $\mathcal{D}_Y$ may be attained at the expense of a small cost in $\mathcal{D}_X$. In all four datasets, the trade-off corresponding to the lowest value of k starts as a steeply descending curve for $\lambda = 0$. This practical implication of this observation is that for microaggregation with low anonymity parameter k may be well worth considering the multiobjective criterion proposed in this work in order to preserve statistical dependence, even if $\lambda \simeq 0$, for a small price in quasi-identifier distortion.



(a) $\lambda = 0$, $\rho = 0.5$ (b) $\lambda = 1$, $\rho = 0.5$

(c) $\lambda = 0$, $\rho = 0.9$ (d) $\lambda = 1$, $\rho = 0.9$

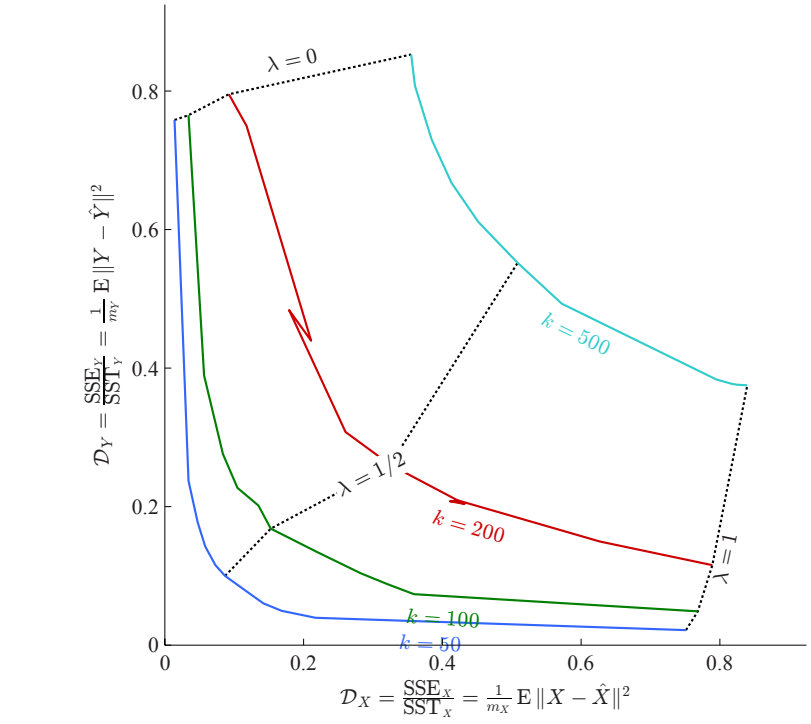
Fig. 8. Quasi-identifier distortion $\mathcal{D}_X$ and confidential-attribute distortion $\mathcal{D}_Y$ versus anonymity k, for extreme values of the Lagrangian multiplier $\lambda = 0, 1$, and for synthetic, Gaussian data, with $\rho = 0.5, 0.9$.

The cost analysis argument just made becomes more apparent in the normalized version of the trade-off plots displayed in Figs. 9–12, in which in lieu of the absolute distortions $\mathcal{D}_X$ and $\mathcal{D}_Y$ we compute the relative increments

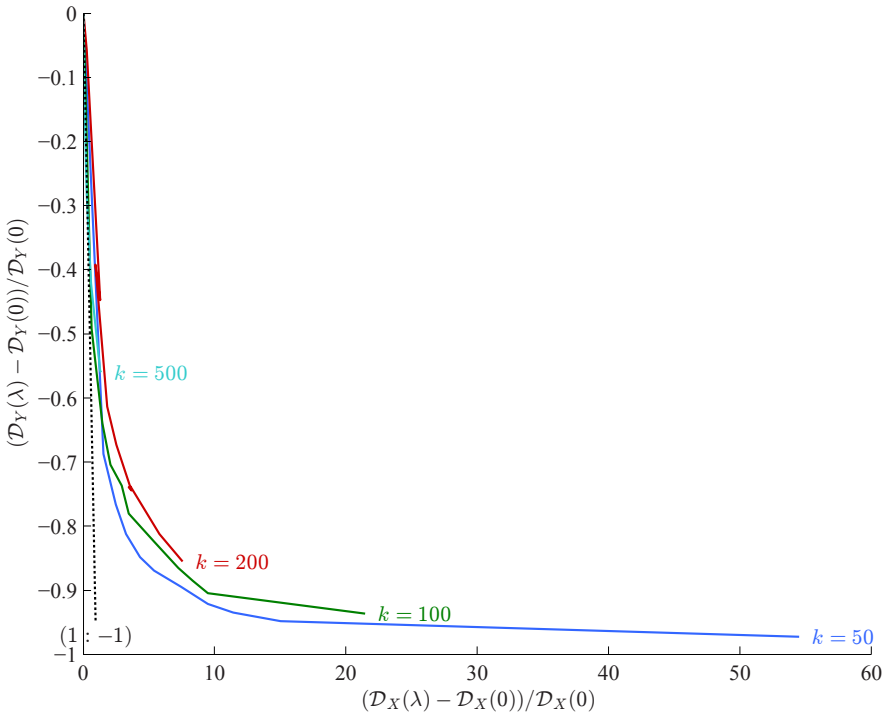
$$\frac{\mathcal{D}_X(\lambda) - \mathcal{D}_X(0)}{\mathcal{D}_X(0)} \quad \text{and} \quad \frac{\mathcal{D}_Y(\lambda) - \mathcal{D}_Y(0)}{\mathcal{D}_Y(0)},$$

with respect to the traditional case $\lambda = 0$, and where distortions are now denoted as functions of the Lagrangian parameter. The dotted line is the line for which the relative improvement (negative increment) in statistical dependence equals the relative degradation (positive increment) in quasi-identifier distortion. As we can see in both of the real-world datasets tested, in terms of the trade-off between quasi-identifier distortion and statistical dependence, the formulation proposed in this work represents quite an advantage with respect to traditional microaggregation, especially for low λ, which would hardly alter the partition obtained, and low anonymity parameter k.

Finally, we remarked in §III.B that preserving statistical dependence might in principle negatively contribute the diversity of confidential attributes within a microcell. That is, improving the predictability of the confidential attributes within the microcells might have a negative impact on the similarity and skewness vulnerabilities already present in traditional microaggregation. In order to investigate this potential effect, we define a measure of *average diversity* ℓ_{avg} as follows. For both "Census" and "Tarragona", our real-world, standardized datasets, after the customary unit-variance normalization and once microaggregation has been carried out, the 7-dimensional confidential attributes are finely quantized with an interval width of 0.02 for each dimension. For "Census", this fine quantization produces 93 distinct categories for the least diverse dimension, and 196 along the

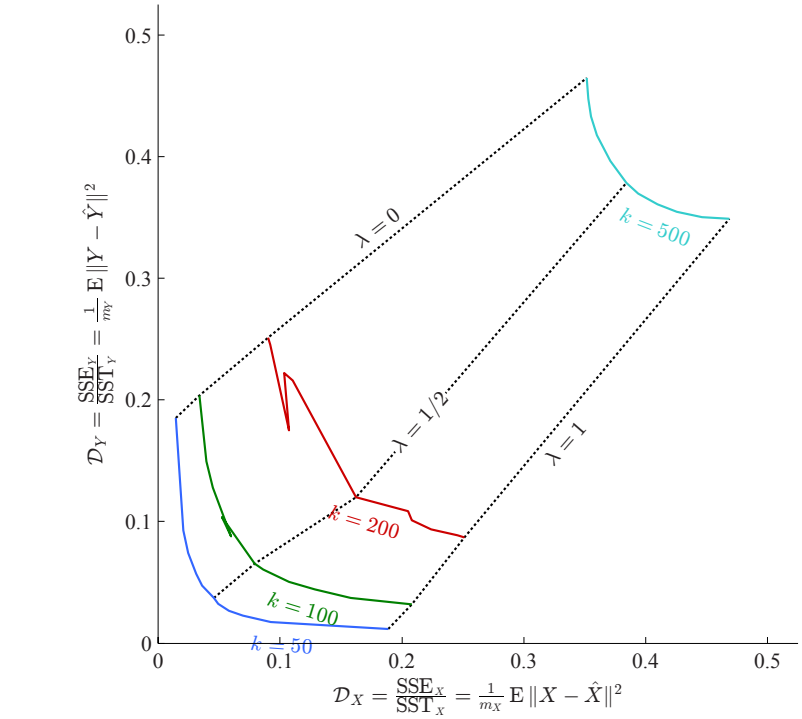


(a)

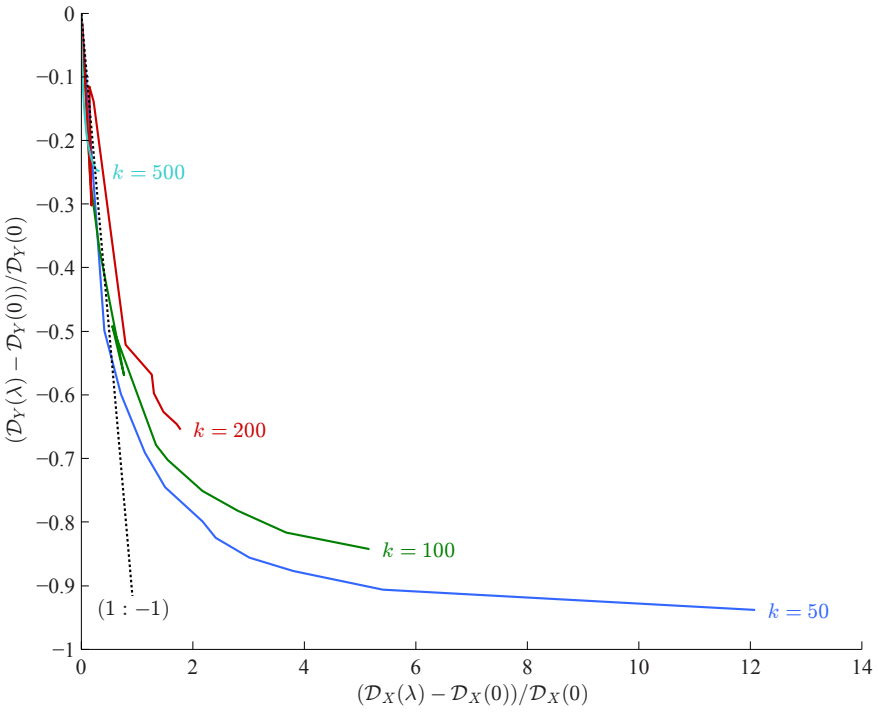


(b)

Fig. 9. Trade-off between quasi-identifier distortion $\mathcal{D}_X$.and confidential-attribute distortion $\mathcal{D}_Y$, for Gaussian data with $\rho = 0.5$, for (a) various values of the k-anonymity parameter and the Lagrangian multiplier λ, and (b) viewed as relative distortion increments with respect to the extreme cases of $\lambda = 0, 1$.

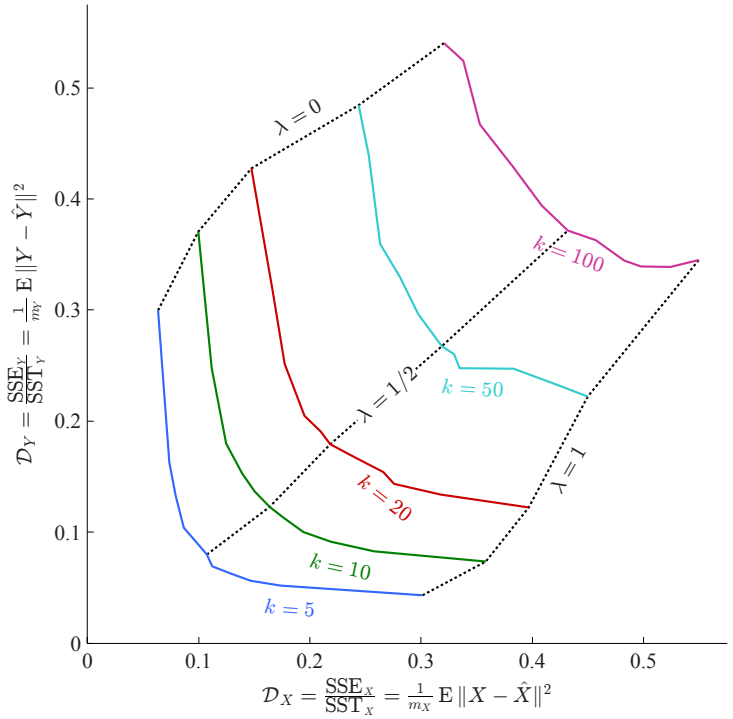


(a)

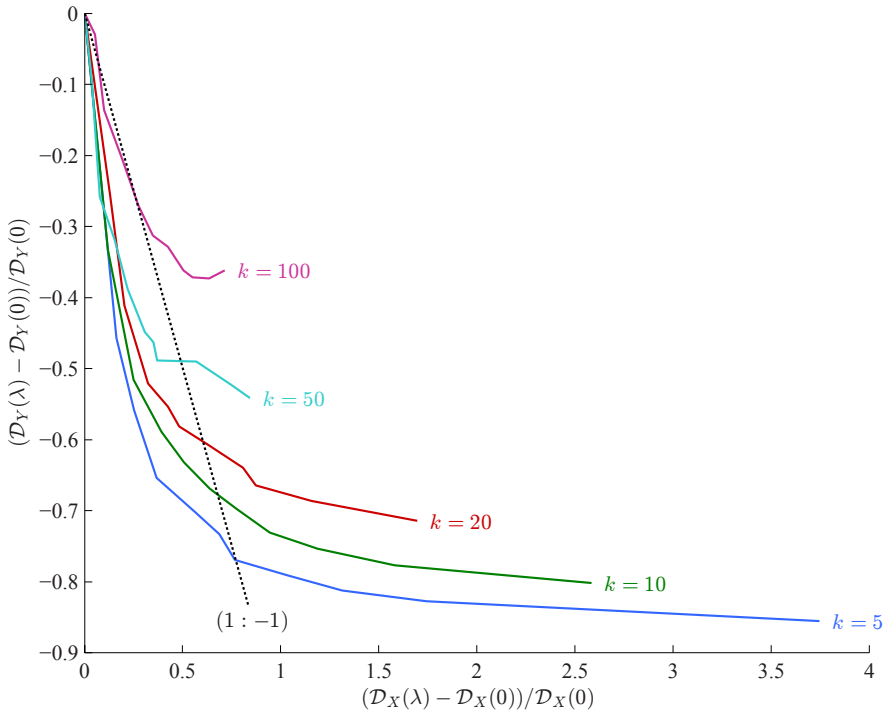


(b)

Fig. 10. Trade-off between quasi-identifier distortion $\mathcal{D}_X$.and confidential-attribute distortion $\mathcal{D}_Y$, for Gaussian data with $\rho = 0.9$, for (a) various values of the k-anonymity parameter and the Lagrangian multiplier λ, and (b) viewed as relative distortion increments with respect to the extreme cases of $\lambda = 0, 1$.

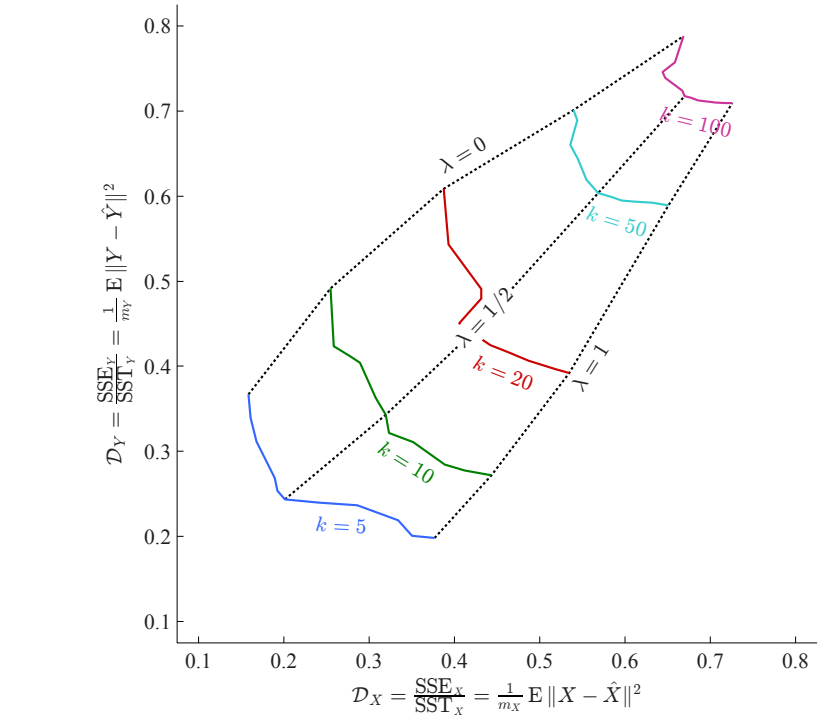


(a)

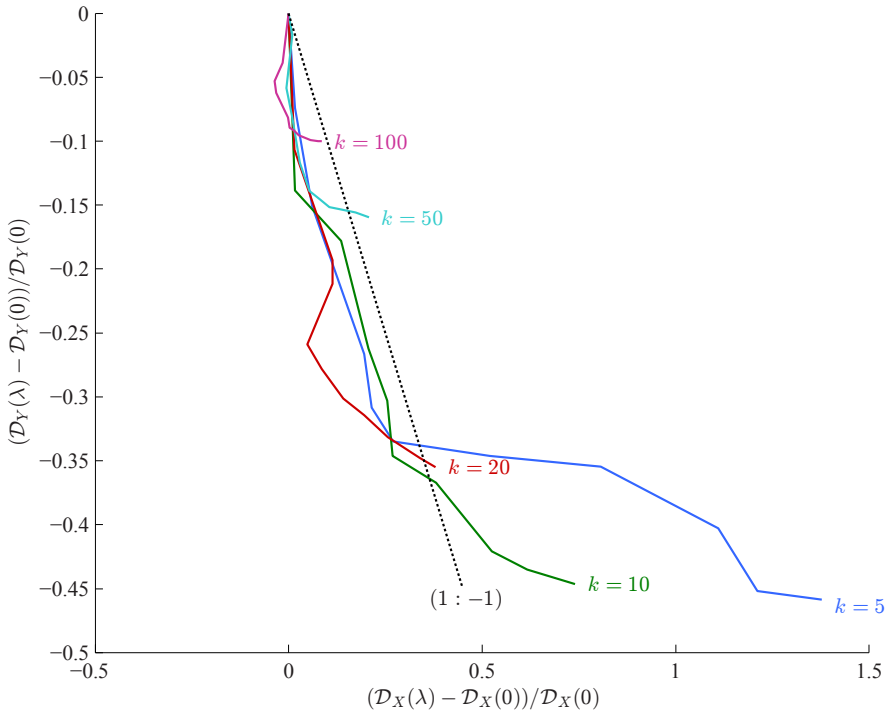


(b)

Fig. 11. Trade-off between quasi-identifier distortion $\mathcal{D}_X$.and confidential-attribute distortion $\mathcal{D}_Y$, for the "Census" dataset, for (a) various values of the k-anonymity parameter and the Lagrangian multiplier λ, and (b) viewed as relative distortion increments with respect to the extreme cases of $\lambda = 0, 1$.

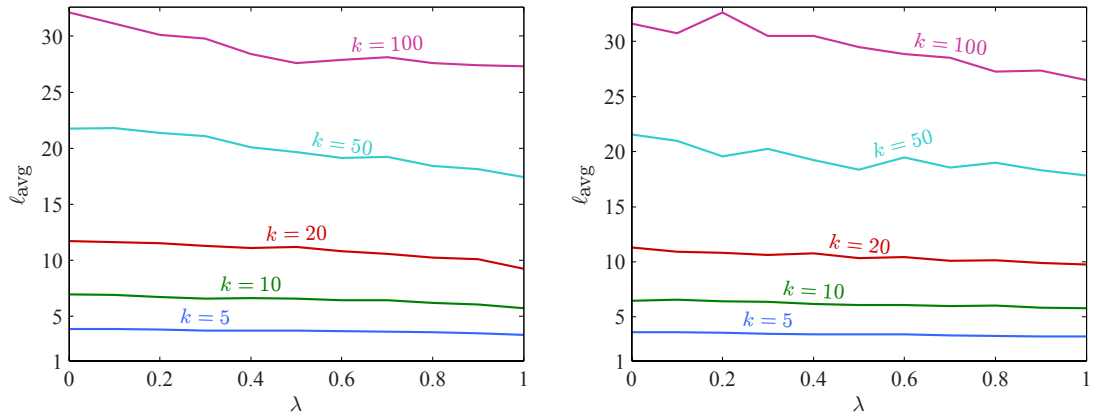


(a)



(b)

Fig. 12. Trade-off between quasi-identifier distortion $\mathcal{D}_X$.and confidential-attribute distortion $\mathcal{D}_Y$, for the "Tarragona" dataset, for (a) various values of the k-anonymity parameter and the Lagrangian multiplier λ, and (b) viewed as relative distortion increments with respect to the extreme cases of $\lambda = 0, 1$.



(a) "Census" (b) "Tarragona"

Fig. 13. Small degradation of the diversity measure ℓ_{avg} as λ increases towards favoring statistical dependence over quasi-identifier distortion, for the standardized datasets "Census" and "Tarragona".

most diverse. For "Tarragona", the number of categories ranges from 106 to 149. Next, for each of the microcells obtained and each of the 7 dimensions, the diversity, that is, the number of different values of the corresponding confidential attribute is computed. The minimum diversity across the dimensions is then found for each cell, and these minima are finally averaged across cells. Thus, ℓ_{avg} is really a between-group average of within-dimension minima; mathematically,

$$\ell_{\text{avg}} = \frac{1}{|\mathscr{Q}|} \sum_{q=1}^{|\mathscr{Q}|} \min_i \left| \{ \lfloor y_{ij}/\Delta \rfloor | q(j) = q \} \right|,$$

with $\Delta = 0.2$. The minimum across dimensions characterizes which confidential attribute is most exposed to a similarity attack. The average across groups measures the typical privacy risk incurred. Clearly, $1 \leqslant \ell_{\text{avg}} \leqslant k_{\max}$, with $k_{\max} \geqslant k$ representing the size of the largest microcell.

As Fig. 13 shows, quite conveniently and somewhat contrary to intuition, there is hardly any negative impact on the diversity measure ℓ_{avg} due to preservation of statistical dependence with $\lambda > 0$, although as one might expect, this impact starts to become marginal towards higher values of the Lagrangian multiplier $\lambda \simeq 1$, for which statistical dependence is favored over quasi-identifier distortion. This behavior is particularly convenient for the case of low k, precisely the case for which the advantage of our formulation over traditional microaggregation is more noticeable. A likely explanation is that while diversity and statistical dependence are conceptually opposing objectives, and any reformulation of the problem of microaggregation should inherit this fundamental challenge, the metrics used for each of these contrasting aspects differ considerably. On the one hand, diversity relates to distinct values, in a more qualitative manner once the quantization interval Δ is fixed, whereas predictability is measured here as a quadratic error, for which the specific quantities involved matter.

C. Quick Glance at Microaggregation beyond k-Anonymity

We have stressed that the focus of this work is on k-anonymous microaggregation, in the sense that the risk of disclosure of private information is measured by means of k-anonymity, and the formulation developed in this paper concordantly assumes an underlying k-anonymous microaggregation algorithm, such as MDAV. The necessarily limited scope of our contribution must leave any exhaustive analysis of other metrics and mechanisms that do not conform with the k-anonymity approach, irrespective of their potential interest, for future investigation. Having stressed that, inspired by [24], we preliminarily venture an experimental assessment of the application of the ideas put forth here, to other microaggregation metrics and mechanisms beyond k-anonymity[(a)]. Said experiments are by no means exhaustive, but merely a quick glance at the question.

[(a)] We gratefully acknowledge the insightful remarks of an anonymous reviewer encouraging us to pursue this avenue.

Specifically, we draw upon the algorithmic variation of MDAV in [24], enforcing the prepartition of quasi-identifiers into contiguous blocks. MDAV is then carried out on each block of quasi-identifiers separately, and the final, protected quasi-identifiers are the direct juxtaposition of those perturbed blocks, preserving the original arrangement. We shall see that prepartitioning along the attribute domain enables us to reduce the distortion introduced in the perturbation of quasi-identifiers. Unfortunately, that gain in utility is not without a considerable impact on anonymity.

We must notice that in the extreme case in which those blocks are narrowed down to scalar attributes, optimal microaggregation techniques may be used, with the caveat that the overall procedure will in general be suboptimal. Because the computational cost of MDAV is affine with the number of attributes aggregated, prepartitioning in the attribute domain, rather than the record domain, offers no better running time.

Most importantly, in microaggregation with prepartition in the attribute domain, k-anonymity is enforced along each block, but not guaranteed on the joint collection of quasi-identifiers. Further, a measure of *disclosure risk* alternative to k-anonymity may be employed, by means of a well-known potential attack against any SDC methods relying on perturbation of the quasi-identifiers.

The attack, called *distance-based record linkage* [26], attempts to reidentify the original quasi-identifier x, for each perturbed quasi-identifier $\hat{x}$, simply by selecting the nearest value in Euclidean distance among all other quasi-identifiers. To better handle shifting and scaling discrepancies, but also to counter the normalization often carried out as part of the microaggregation process, prior to measuring said distances, the attack carries out a zero-mean, columnwise unit-variance normalization of the collections of unperturbed and perturbed values. Mathematically, for any perturbed record $\hat{x}$, the attacker guesses

$$\arg\min_x \left\| \frac{x-\mu}{\sigma} - \frac{\hat{x}-\hat{\mu}}{\hat{\sigma}} \right\|^2$$

as a potential match, where μ and σ denote the mean and standard deviation of the set of original records, and $\hat{\mu}$ and $\hat{\sigma}$ those of the set of perturbed records. Unless the perturbation introduced is significant, the latter normalization is hardly noticeable. The *distance-based linkage disclosure* (DLD) metric is simply the fraction of perturbed records that may be correctly linked in this manner.

For the standardized dataset "Census", introduced in §V.A, once again we use the first $m_X = 6$ attributes as quasi-identifiers, and the remainder $m_Y = 7$ as confidential. We implement a modification of the MDAV algorithm with prepartitioning in the quasi-identifiers in contiguous blocks of sizes $b = 2, 4, 6$, where the case $b = m_X$ formally represents microaggregation without prepartitioning at all, as carried out in §V.B. Each block of quasi-identifiers is individually microaggregated, with a common anonymity parameter $k = 10$, along with the entire set of confidential attributes, with a common value of the Lagrangian multiplier λ weighing the relevance placed on the preservation of statistical dependence. The resulting method is the natural combination of the attribute prepartitioning strategy employed in [24], and the novel formalism introduced in this work.

We measure the quasi-identifier distortion $\mathcal{D}_X$ as a function of the weight λ, for the various block sizes b aforementioned, in Fig. 14. As expected, reducing the block size also reduces the distortion incurred, and as more weight is given to the preservation of statistical dependence, the distortion is increased. The curve for $b = 6$ is perfectly consistent with that for $k = 10$ in Fig. 11, where we observed that a slight increase of λ with respect to traditional microaggregation would hardly make a dent on quasi-identifier distortion, while providing a valuable improvement in preservation of statistical dependence, but extreme values of λ might not be desirable.

For traditional microaggregation without preservation of statistical dependence, that is, $\lambda = 0$, the reduction in distortion achieved by prepartitioning in the attribute domain is significant, as it halves for $b = 3$, and once again for $b = 2$ with respect to $b = 3$. As λ increases, the microcells selected by the modified MDAV algorithm favors the confidential attributes, making this selection identical in the extreme case when $\lambda = 1$.

Unfortunately, the distortion reduction due to prepartitioning, in addition to violating k-anonymity in the strict sense, jointly on all quasi-identifiers, still has a significant price in disclosure

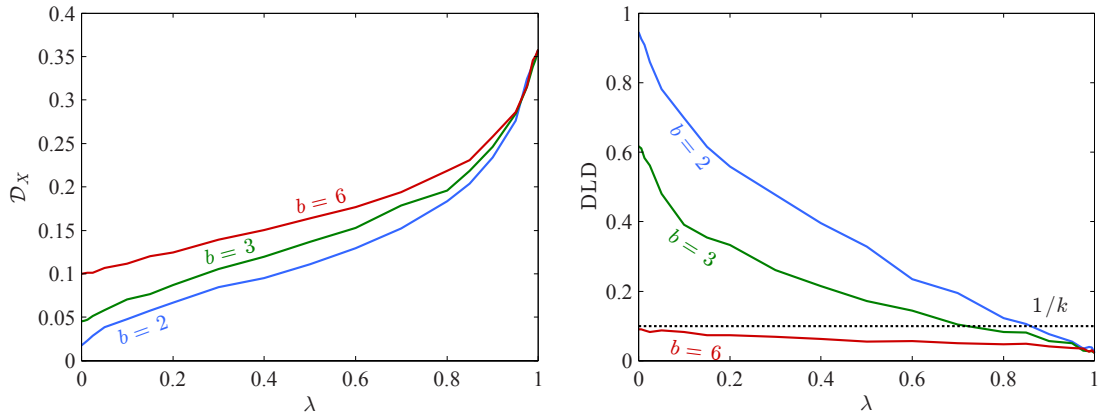


Fig. 14. For the standardized dataset "Census", taking the first $m_X = 6$ attributes as quasi-identifiers and the rest $m_Y = 7$ as confidential, we analyze the effect on the quasi-identifier distortion $\mathcal{D}_X$ of microaggregation with $k = 10$ and prepartitioning in the quasi-identifier domain with various block sizes $b = 2, 4, 6$, as λ increases towards favoring statistical dependence over quasi-identifier distortion. The significant disclosure risk incurred is similarly measured, in terms of DLD.

risk, measured as DLD. As Fig. 14 reports, the ambitious partitioning into blocks of size $b = 2$ leads to a probability of record linkage nearing full disclosure.

The following intuitive argument helps shed some light on the distortion reduction due to attribute prepartitioning, at the expense of disclosure risk. For n records with m_X quasi-identifiers partitioned into blocks of length b, after each block is individually microaggregated with an anonymity parameter k, $\lfloor n/k \rfloor$ cells are obtained for each of the $\lfloor m_X/b \rfloor$ blocks, which partition the joint space of quasi-identifiers into

$$\left\lfloor {}^{n}/_{k} \right\rfloor^{\left\lfloor {}^{m_X}/_{b} \right\rfloor}$$

potential cells, each with its corresponding reconstruction centroid. For the experiment in question, $n = 1080$, $k = 10$, $m_X = 6$, and $b = 2$, so that the number of potential cells is $108^3 = 1\,259\,712$, orders of magnitude larger than the number or records, 1080. Although many of those cells will be empty, the discriminatory potential of the large combination of centroids produced by microaggregation with prepartitioning in the attribute domain suffers from the so-called *curse of dimensionality*, and while reducing distortion, it may also lead to a severe risk of disclosure, with many cells containing very few or even single records. Hence, caution must be exerted when using such algorithms, especially in the traditional case $\lambda = 0$. As the multiplier λ increases and the microaggregation process relies more on the confidential attributes than on the partitioned quasi-identifiers, the end result will resemble that without prepartitioning, that is, that for $b = 6 = m_X$, dampening the utility gain along with the privacy loss, thereby easing the trade-off in between.

As expected for microaggregation without quasi-identifier prepartitioning, that is, $b = 6$, DLD is approximately $1/k$ for $\lambda = 0$, diminishing as λ increases and the quasi-identifier microcells constructed depend less on the quasi-identifiers. If the confidential attributes were statistically independent, which is clearly not the case, DLD would vanish towards $1/n$ and $\mathcal{D}_X$ would grow to 1 as λ increases. Also in k-anonymous microaggregation without prepartitioning and distinct quasi-identifiers, irrespective of λ, because microcells are not Voronoi cells, and the closest point to a centroid may actually be outside the cell, the number of linked records is at most the number $|\mathscr{Q}|$ of (nonempty) cells. Consequently, for any form of unpartitioned microaggregation, that is, $b = m_X$,

$$\mathrm{DLD} \leqslant \frac{|\mathscr{Q}|}{n} \leqslant \frac{1}{k},$$

where $n/|\mathscr{Q}| \geqslant k$ is the average number of points per cell, the average-case analog of k-anonymity, and where the latter inequality would hold with equality if n were a multiple of k. For $\lambda = 0$, one may expect these bounds to constitute a close approximation. For $\lambda = 0$, $m_X = 1$ and n a multiple of k, optimal microaggregation would indeed tighten them.

It was pointed out to us that, as a general principle, the degradation in disclosure risk may be significantly affected by the precise manner in which quasi-identifiers are partitioned, as demonstrated in [23] for traditional microaggregation. In the context of the problem proposed here, where microaggregation further strives to preserve statistical dependence, the study of partitioning strategies acquires special relevance, and certainly constitutes an intriguing avenue for future investigation.

VI. Conclusion

As scientific and technological progress unfolds, our everyday lives are becoming inextricably entangled with the digital world. Modern statistical data analysis technologies with proven success in virtually all fields, including data mining, machine learning and big-data analysis, are becoming gradually ubiquitous in applications as diverse as targeted advertising, recommendation systems and collaborative tagging, social networks, e- and m-health, demographic databases with various scientific purposes, and e-voting. While the technological development is unquestionable, the sustainable availability of any form of potentially sensitive data hinges on the consent of the people involved, and that consent in turn depends on the privacy guarantees offered by such sophisticated information systems.

Traditional k-anonymous microaggregation, arguably an essential building block in the control of the disclosure of statistical databases, has been formulated to characterize both the privacy attained and the inherent information loss incurred due to the necessary perturbation of the quasi-identifiers. Because the ulterior purposes of such databases involves the analysis of the statistical dependence between demographic attributes and sensitive data, we must articulate mechanisms to enable the preservation of the statistical dependence between quasi-identifiers and confidential attributes, beyond the mere degradation of the quasi-identifiers alone.

This work addresses the problem of k-anonymous microaggregation with preservation of statistical dependence in a formal, systematic manner, naturally modeling statistical dependence as predictability of the confidential attributes from the perturbed quasi-identifiers. We proceed by introducing a second MSE term in a combined Lagrangian cost that enables us to regulate the trade-off between quasi-identifier distortion and the confidential-attribute predictability. A Lagrangian multiplier enables us to gracefully control the importance of each of the two competing objectives.

From a broad perspective, we have acknowledged that the formulation of any privacy-utility problem relies on the appropriateness of the criteria optimized. These criteria depend, in turn, on the specific application, on the statistics of the data, on the degree of data utility we are willing to compromise, and on the adversarial model and the mechanisms against privacy contemplated. We have also mentioned that, thanks to its mathematical, algorithmic tractability, k-anonymity is almost universally accepted as a measure of privacy in SDC, at least as a starting point for the design and evaluation of microaggregation algorithms. An entirely analogous argument can be made for our multiobjective criterion for preservation of statistical dependence. Although we cannot claim for it to be appropriate for any and all applications, our criterion inherits the fundamental advantages of the widely accepted MSE for measuring the information loss due to the perturbation of quasi-identifiers.

The formulation of the generalized form of microaggregation developed here enjoys the mathematical, algorithmic appeal of being formally reducible to traditional microaggregation, where the distortion of numerical data is measured by means of a simple MSE term. This permits the ready application of results on necessary optimality of k-anonymous aggregation developed in our own previous work. More pragmatically, it also allows the implementation of this generalized variant without any modification to the underlying microaggregation algorithm, simply by modifying the input vector containing quasi-identifier values to append a weighted version of the confidential attributes.

Experimental results on a number of synthetic and real-world, standardized datasets confirm the practical applicability of our formalism, even if as a candidate mechanism to contemplate for certain types of data. A conservative application of our approach would entail small values of the

Lagrangian multiplier, allowing placing greater emphasis on the objective optimized in traditional microaggregation, while not completely disregarding statistical dependence. Specially for moderately small values of the Lagrangian multiplier and the anonymity parameter k, we attain important gains in the preservation of this dependence, at the expense of a hardly significant degradation of quasi-identifier distortion.

Although the necessarily limited scope of our contribution contemplates only numerical data, MSE and a single yet widely used microaggregation algorithm, MDAV, many of the ideas put forth here might be further extended to additional microaggregation algorithms, such as PCL, and to categorical data, by means of general distortion measures to quantify quasi-identifier distortion and prediction error in confidential attributes. The scope of our work also leaves for future investigation any exhaustive analysis of other metrics and mechanisms that do not conform to the k-anonymity approach. Still, we preliminarily venture a brief experimental assessment of the application of the ideas put forth here to distance-based record linkage and microaggregation with prepartitioning along the attribute domain.

All things considered, we believe that the method proposed here is a formally sound, readily applicable approach for the important yet often neglected problem of preserving statistical dependence in k-anonymous microaggregation. Far from being a closed development, our formalism may very well open a prolific avenue for future investigation, both benefitting from and contributing to, much of the theoretical and algorithmic work on microaggregation carried out to date.

Acknowledgment

We sincerely thank the anonymous reviewers for their valuable, insightful comments. 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", and through the project "INRISCO", ref. TEC2014-54335-C4-1-R, as well as by the Government of Catalonia, under grant 2014 SGR 1504.

References

[1] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, UK: Cambridge Univ. Press, 2004.

[2] J. Brickell and V. Shmatikov, "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," in *Proc. ACM SIGKDD Int. Conf. Knowl. Disc., Data Min. (KDD)*, Las Vegas, NV, Aug. 2008.

[3] "Computational aspects of statistical confidentiality project," European project IST-2000-25069 CASC, 2003. [Online]. Available: http://neon.vb.cbs.nl/casc

[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] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. New York, NY: John Wiley & Sons, 2006.

[6] R. Dandekar, J. Domingo-Ferrer, and F. Sebé, "LHS-based hybrid microdata vs. rank swapping and microaggregation for numeric microdata protection," in *Proc. Infer. Contr. Stat. Databases (ICSD) Conf.*, ser. Lect. Notes Comput. Sci. (LNCS), vol. 2316. Springer, 2002, pp. 153–162.

[7] D. Defays and P. Nanopoulos, "Panels of enterprises and confidentiality: The small aggregates method," in *Proc. Symp. Design, Anal. Longit. Surveys, Stat. Canada*, Ottawa, Canada, 1993, pp. 195–204.

[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, J. M. Mateo-Sanz, and V. Torra, "Comparing SDC methods for microdata on the basis of information loss and disclosure risk," in *Proc. ETK/NTTS Conf.*, vol. 2, Eurostat, Luxemburg, 2001, pp. 807–826.

[11] 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.

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

[13] ——, "A critique of k-anonymity and some of its enhancements," in *Proc. Workshop Priv., Secur., Artif. Intell. (PSAI)*, Barcelona, Spain, 2008, pp. 990–993.

[14] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. ACM Symp. Prin. Database Syst. (PODS)*, San Diego, CA, 2003, pp. 211–222.

[15] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Boston, MA: Kluwer Acad. Publishers, 1992.

[16] 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.

[17] 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.

[18] K. LeFevre, D. J. DeWitt, and R. Ramakrish, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Baltimore, MD, Jun. 2005, pp. 49–60.

[19] 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.

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

[21] 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.

[22] J. J. Moré, "The Levenberg-Marquardt algorithm: Implementation and theory," in *Proc. Dundee Biennial Conf. Numer. Anal.*, ser. Lect. Notes Math., G. A. Watson, Ed., vol. 630. Springer, 1977, pp. 105–116.

[23] J. Nin, J. Herranz, and V. Torra, "How to group attributes in multivariate microaggregation," *Int. J. Uncertain., Fuzz., Knowl.-Based Syst.*, vol. 16, pp. 121–138, 2008.

[24] ——, "On the disclosure risk of multivariate microaggregation," *Data, Knowl. Eng.*, vol. 67, no. 3, pp. 399–412, 2008.

[25] 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.

[26] D. Pagliuca and G. Seri, "Some results of individual ranking method on the system of enterprise accounts annual survey," Esprit SDC Project, Deliverable MI-3/D2, 1999.

[27] D. Rebollo-Monedero, "Quantization and transforms for distributed source coding," Ph.D. dissertation, Stanford Univ., 2007. [Online]. Available: https://sites.google.com/site/davidrebollomonedero/files/-PhDThesis2007-12-13.pdf

[28] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From t-closeness to PRAM and noise addition via information theory," in *Proc. Priv. Stat. Databases (PSD)*, ser. Lect. Notes Comput. Sci. (LNCS). Istambul, Turkey: Springer, Sep. 2008, pp. 100–112.

[29] ——, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, pp. 1623–1636, Nov. 2010. [Online]. Available: http://doi.ieeecomputersociety.org/10.1109/TKDE.2009.190

[30] 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://dx.doi.org/10.1016/j.ins.2012.08.022

[31] D. Rebollo-Monedero, J. Forné, and M. Soriano, "An algorithm for k-anonymous microaggregation and clustering inspired by the design of distortion-optimized quantizers," *Data, Knowl. Eng.*, vol. 70, no. 10, pp. 892–921, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.datak.2011.06.005

[32] D. Rebollo-Monedero, J. Parra-Arnau, C. Diaz, and J. Forné, "On the measurement of privacy as an attacker's estimation error," *Int. J. Inform. Secur.*, vol. 12, no. 2, pp. 129–149, Apr. 2013. [Online]. Available: http://dx.doi.org/10.1007/s10207-012-0182-5

[33] D. Rebollo-Monedero, S. Rane, A. Aaron, and B. Girod, "High-rate quantization and transform coding with side information at the decoder," *EURASIP J. Signal Process., Special Issue Distrib. Source Coding*, vol. 86, no. 11, pp. 3160–3179, Nov. 2006, invited paper.

[34] D. Rebollo-Monedero, S. Rane, and B. Girod, "Wyner-Ziv quantization and transform coding of noisy sources at high rates," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, vol. 2, Pacific Grove, CA, Nov. 2004, pp. 2084–2088.

[35] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, 2001.

[36] C. E. Shannon, "Communication theory of secrecy systems," Bell Syst., Tech. J., 1949.

[37] A. Solanas, A. Martínez-Ballesté, and J. Domingo-Ferrer, "VMDAV: A multivariate microaggregation with variable group size," in *Proc. Comput. Stat. (COMPSTAT)*. Rome, Italy: Springer, 2006.

[38] J. Soria-Comas and J. Domingo-Ferrer, "Probabilistic k-anonymity through microaggregation and data swapping," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Brisbane, Australia, Jun. 2012, pp. 1–8.

[39] J. Soria-Comas, J. Domingo-Ferrer, and D. Rebollo-Monedero, "k-Anonimato probabilístico," in *Proc. Reunión Española sobre Criptología y Seguridad de la Información (RECSI)*, San Sebastián (Donostia), Spain, Sep. 2012, pp. 249–254.

[40] J. Soria-Comas, J. Domingo-Ferrer, D. Sanchez, and S. Martinez, "t-Closeness through microaggregation: Strict privacy with enhanced utility preservation," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3098–3110, May 2015.

[41] 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.

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

[44] W. Yancey, W. Winkler, and R. Creecy, "Disclosure risk assessment in perturbative microdata protection," in *Proc. Infer. Contr. Stat. Databases (ICSD) Conf.*, ser. Lect. Notes Comput. Sci. (LNCS), vol. 2316. Springer, 2002, pp. 135–152.