This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875949, IEEE Access

Date of publication <MMMM d, yyyy>, date of current version October 8, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.<0000000>

Incremental k-Anonymous Microaggregation in Large-Scale Electronic Surveys with Optimized Scheduling

DAVID REBOLLO-MONEDERO[1], CÉSAR HERNÁNDEZ-BAIGORRI[1], JORDI FORNÉ[1], AND MIGUEL SORIANO[1,2]
[1]Department of Telematic Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain
[2]Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), E-08860 Castelldefels, Barcelona, Spain

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

This work is partly supported by the Spanish Ministry of Industry, Energy and Tourism (MINETUR) through the "Acción Estratégica Economía y Sociedad Digital (AEESD)" funding plan, through the project "Data-distortion framework (DDF)", ref. TSI-100202-2013-23. Additional funding supporting this work has been granted to UPC by the Spanish Ministry of Economy and Competitiveness (MINECO) through the project "Anonymized demographic surveys (ADS)", ref. TIN2014-58259-JIN, under the funding program "Proyectos de I+D+i para Jóvenes Investigadores", through the project "INRISCO", ref. TEC2014-54335-C4-1-R, and through the project "Advanced forensic analysis", ref. TEC2015-68734-R. Finally, we gratefully acknowledge funding from the "Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR)" of the Government of Catalonia, under grants SGR 2014-1504 and SGR 2017-782.

ABSTRACT Improvements in technology have led to enormous volumes of detailed personal information made available for any number of statistical studies. This has stimulated the need for anonymization techniques striving to attain a difficult compromise between the usefulness of the data and the protection of our privacy. k-Anonymous microaggregation permits releasing a dataset where each person remains indistinguishable from other $k-1$ individuals, through the aggregation of demographic attributes, otherwise a potential culprit for respondent reidentification. Although privacy guarantees are by no means absolute, the elegant simplicity of the k-anonymity criterion and the excellent preservation of information utility of microaggregation algorithms has turned them into widely popular approaches whenever data utility is critical. Unfortunately, high-utility algorithms on large datasets inherently require extensive computation. This work addresses the need of running k-anonymous microaggregation efficiently with mild distortion loss, exploiting the fact that the data may arrive over an extended period of time. Specifically, we propose to split the original dataset into two portions that will be processed subsequently, allowing the first process to start before the entire dataset is received, while leveraging the superlinearity of the microaggregation algorithms involved. A detailed mathematical formulation enables us to calculate the optimal time for the fastest anonymization, as well as for minimum distortion under a given deadline. Two incremental microaggregation algorithms are devised, for which extensive experimentation is reported. The theoretical methodology presented should prove invaluable in numerous data-collection applications, including large-scale electronic surveys in which computation is possible as the data comes in.

INDEX TERMS data privacy, statistical disclosure control, k-anonymity, microaggregation, electronic surveys, large-scale datasets

I. INTRODUCTION

DATA ANALYSIS continues to acquire prominence in the scientific and technological advances of recent years. The vast amount of digital information stored on a daily basis in conjunction with major breakthroughs in networking, data storage, and processing capabilities enable the possibility of retrieving information from numerous systems, ever more swiftly and in greater abundance than before. In the white paper "The Zettabyte Era" [10], Cisco claimed that global IP traffic was expected to surpass the zettabyte by the end of 2016 and predicted that it reach two zettabytes in 2019. The opportunity of transmitting and analyzing such astounding quantities of information has led to the use of new ways to advertise, communicate, create, and consume digital content. In the same vein, it has been reported by the market research firm International Data Corp (IDC) that the amount of data in the world is presumed to grow at a rate of 40% each year, going from 4.4 zettabytes in 2013 to ten times this amount by 2020.

Clearly, this fast-paced technological scenario makes it exceedingly difficult for privacy rights and regulatory laws to stay up to speed, despite increasingly stricter legislation such as the recent General Data Protection Regulation (GDPR) [20] of the European Union. Naïve anonymization protocols merely consisting in removing identifiers in demographic surveys and databases are notoriously insufficient. An experiment conducted in the 1990s by L. Sweeney [45] famously demonstrated that 87% of the population in the U.S. could be unequivocally identified using only three parameters: date of birth, gender and ZIP code. Despite the legal boundaries, under certain circumstances, being identified in a medical or political study could lead to the loss of a job offer and many other forms of discrimination.

A. BRIEF PRIMER ON k-ANONYMITY AND k-ANONYMOUS MICROAGGREGATION

The field of statistical disclosure control (SDC) emerged to address this type of privacy problems in the publication of data for statistical analysis. SDC strives to reduce the risk of confidential information being disclosed while maintaining the usefulness of the data, permitting the release of an effectively anonymized dataset, invaluable for any number of demographic studies.

In this field, sensitive attributes in a dataset are commonly classified as identifiers, quasi-identifiers, and confidential attributes depending on the level of information contained.

- Identifiers can unequivocally lead to the recognition of an individual, including attributes such as the name or the social security number (SSN). These are customarily removed in anonymized data.
- On the other hand, quasi-identifiers do not suffice to identify an individual when considered individually, but in combination with other quasi-identifiers, and in the context of publicly available information, effectively narrow the identity of the respondents to whom the records in the dataset refer. Examples of quasi-identifiers include most demographic attributes such as address, gender, age, birthdate, job type, height, and weight.
- Finally, confidential attributes contain sensitive information on the respondent.

There exist several well-known methods resorting to the modification of quasi-identifiers in an attempt to gain effective anonymity. They do so at the expense of a loss in data utility, and with significant computational cost. This work is directed towards the release of the results of electronic surveys with particular emphasis on preserving the utility of the data contained. As we shall argue later on, we accordingly choose k-anonymous microaggregation as the method employed, over alternatives offering stronger privacy guarantees albeit at significant cost in data utility.

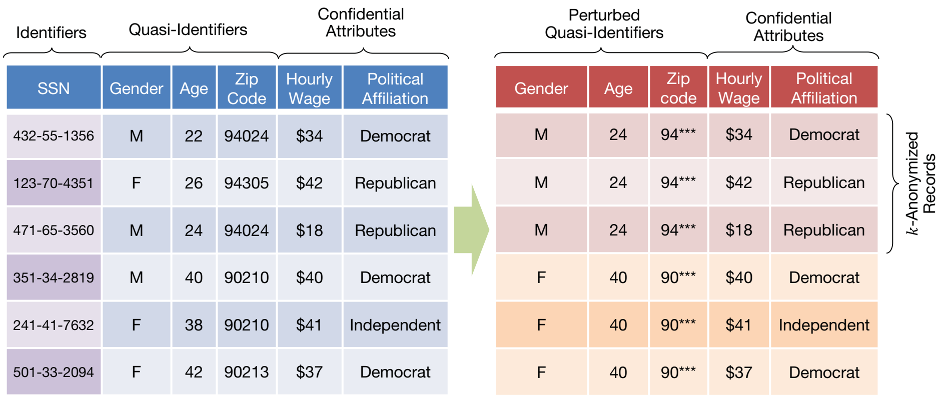
Recall that the k-anonymity criterion ensures that each person cannot be identified due to the existence of, at least, $k-1$ identical tuples of demographic attributes within the processed dataset. More precisely, this technique consists in clustering k or more records according to the values of the quasi-identifiers, which are then replaced with a common reconstruction tuple, usually the arithmetic mean when numerical data is involved. To minimize the distortion loss, these clusters are assembled by similarity of the quasi-identifiers. A simple example of k-anonymous microaggregation is shown in Fig. 1, representing an electronic survey containing confidential attributes such as hourly wages and political preferences. Once a dataset is aggregated in this manner, a specific respondent cannot be reidentified on the basis of their quasi-identifiers. Safeguarding the identity of the individuals participating in the released dataset or electronic survey hinders a privacy attacker in their effort to gain access to confidential attributes. Evidently, a larger k-anonymity parameter reduces the probability of reidentification, but it does so at the cost of further distorting the data.

B. CONCEPTUAL PROBLEM STATEMENT

Even though the basic privacy guarantees of traditional k-anonymous microaggregation are conducive to efficient anonymization with low distortion in the data released, the computation requirements on large-scale datasets are by no means negligible. The leading objective of this contribution is to offer a fast k-anonymization strategy to address the computational requirements in the special case of electronic surveys in which the respondents participate over an extended period of time.

We shall reasonably assume in the sequel that the data from our electronic survey is collected through a significant span of time, in relation to the duration of the anonymization process. Traditional k-anonymous microaggregation would start once all the data has been collected. The incremental method for k-anonymous microaggregation introduced in this work proposes microaggregating in two algorithmic steps instead, operating on two portions of the data. The first algorithm, which we call base algorithm, would start before all the data is available, say one hour before finishing the data collection process, and operate on the incomplete portion of data available at that point. Subsequently, the second microaggregation process, called here incremental algorithm, would start once all the data has been collected, and operate on the portion of the data yet unprocessed. In principle, as we shall see, the second process could also exploit the results computed by the first to its advantage.

Naturally, owing to the fact that the anonymization process starts earlier, the two-step approach should finish faster. Far less obvious is the fact that this approach also



Identifiers	Quasi-Identifiers			Confidential Attributes	
SSN	Gender	Age	Zip Code	Hourly Wage	Political Affiliation
432-55-1356	M	22	94024	$34	Democrat
123-70-4351	F	26	94305	$42	Republican
471-65-3560	M	24	94024	$18	Republican
351-34-2819	M	40	90210	$40	Democrat
241-41-7632	F	38	90210	$41	Independent
501-33-2094	F	42	90213	$37	Democrat

Perturbed Quasi-Identifiers			Confidential Attributes	
Gender	Age	Zip code	Hourly Wage	Political Affiliation
M	24	94***	$34	Democrat
M	24	94***	$42	Republican
M	24	94***	$18	Republican
F	40	90***	$40	Democrat
F	40	90***	$41	Independent
F	40	90***	$37	Democrat

FIGURE 1. Synthetic example of k-anonymous microaggregation in published data with $k = 3$, relating various demographic attributes acting as quasi-identifiers, namely genc

benefits from a remarkable property of most low-distortion microaggregation algorithms. It turns out that the running time of such algorithms is not linear in the number of records, but superadditive. This means that the running time $t(n+m)$ on a dataset of $n+m$ records is in general greater than or equal to the sum of running times corresponding to the individual operation of the parts, that is,

$$t(n+m) \geqslant t(n) + t(m).$$

That is the case of the maximum distance to average vector (MDAV) algorithm [14, 18], for example, which indeed runs in (asymptotically) quadratic time. The underlying reason is that forming k-anonymous clusters with similar quasi-identifying values to minimize distortion requires comparisons between records in one way or another. Superlinearity is to be expected from any algorithm striving for high data utility.

As a consequence, the two-step process split into base and incremental microaggregation will have a shorter total running time than the traditional one, in addition to the time saved from starting earlier. This becomes of capital importance for the anonymization of large-scale electronic surveys, which are severely affected by these typically quadratic running times. We shall see that the work presented here exploits both advantages in a synergic manner, from a mathematically optimized perspective.

C. EXAMPLE OF INCREMENTAL MICROAGGREGATION

We illustrate our proposal with an example of electronic survey represented in Fig. 2. The figure shows that the base algorithm starts when most of the data is available. Next, the incremental algorithm starts when the base one finishes and all the data is available. Our method, consisting in running the traditional algorithm in two portions, allows anonymization to be completed in 39 minutes after the completion of the electronic survey, instead of two hours. The incremental algorithm is typically ran on a much smaller dataset, therefore, even if the same microaggregation algorithm were used, it would still finish earlier than the traditional one. This permits the two-step process to considerably outperform the traditional method in time efficiency. The time saved is possible not only thanks to the head start, but also to the aforementioned property of superadditivity of the running time of most high-utility microaggregation algorithms.

The two-step approach proposed may be perfectly suitable in most forms of electronic surveys, but we must caution that it may not be applicable to processes such as elections in which ballot boxes must remain closed until the end.

D. SPECIFIC CONTRIBUTIONS

Specifically, the leading object of this paper is a novel, optimized methodology to employ k-anonymous microaggregation in a convenient, incremental fashion. We investigate how to implement incremental microaggregation efficiently, taking also into consideration the superlinearity of the microaggregation algorithms involved, through a number of mathematical models. In addition, we thoroughly analyze the impact on data distortion from an empirical perspective.

- Different from microaggregation streaming, our incremental approach proposes two distinct stages, optimized according to the availability of data over time and the protection algorithms employed, ultimately striving to reduce the overall computation time in application scenarios that demand it, at the expense of an acceptable loss in distortion.
- Precisely, the initial motivation of this work is to take advantage of the time gained, head start, when processing the initial data partition before receiving all data. This strategy by itself can result in a large amount of time saved. Additionally, the prevailing methods to perform k-anonymous microaggregation run in superlinear time, typically quadratic. Whenever the running

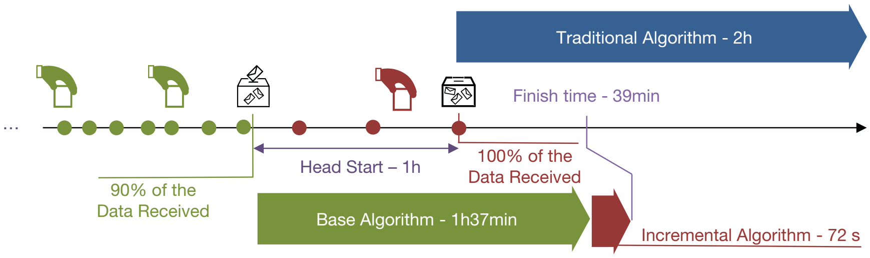
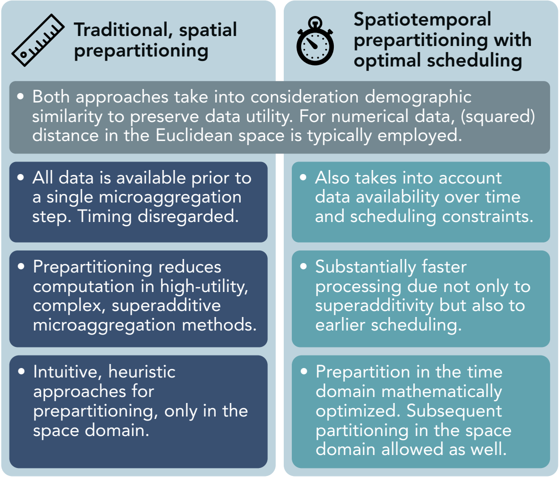


FIGURE 2. Example of incremental approach versus traditional one in a 10-hour electronic survey. Note that the traditional approach would not be usable if the result were requ

time is superadditive, any sort of partition will yield a faster running time.

- In practice, these methods are directed towards time-critical applications, where the time needed to publish the data is more relevant than a limited degradation in distortion. We shall demonstrate that publishing an anonymized version of the results of a survey involving a significant number of participants can be done in minutes instead of hours, at the expense of quadratic distortion increments well under 10%.
- Accordingly, we use two microaggregation algorithms based on the aforementioned MDAV algorithm: a two-step MDAV application, and a nearest neighbor strategy, after an initial application of MDAV. It shall become apparent that the proposed approach can actually be adapted to other microaggregation algorithms and privacy criteria.
- However, our extensive experimental results show that this method provides better running times while not introducing high rates of distortion loss, even without a head start, when used on datasets containing a high number of attributes and records. As we focus on large-scale data, a synthetic dataset and an extended version of 'Census', popular in the SDC literature, have been used.

From a conceptual perspective, prepartitioning is a well-established strategy to reduce computation in high-utility, superadditive algorithms. Traditionally, prepartitioning takes into consideration demographic similarity to preserve data utility. For numerical data, (square) distance in the Euclidian space is typically employed. In this work, we take a first step towards extending this spatial strategy along the time domain. We shall demonstrate, both theoretically and empirically, that privacy-preserving algorithms highly benefit from our mathematically optimized scheduling framework. The relative value of our proposal with respect to traditional prepartitioning is summarized in Fig. 3.


Traditional, spatial prepartitioning

Spatiotemporal prepartitioning with optimal scheduling

- Both approaches take into consideration demographic similarity to preserve data utility. For numerical data, (squared) distance in the Euclidean space is typically employed.

- All data is available prior to a single microaggregation step. Timing disregarded.
- Also takes into account data availability over time and scheduling constraints.
- Prepartitioning reduces computation in high-utility, complex, superadditive microaggregation methods.
- Substantially faster processing due not only to superadditivity but also to earlier scheduling.
- Intuitive, heuristic approaches for prepartitioning, only in the space domain.
- Prepartition in the time domain mathematically optimized. Subsequent partitioning in the space domain allowed as well.

FIGURE 3. Our work proposes to consider data availability over time in addition to demographic similarity for substantially faster microaggregation, due not only to superadditivity, but also to mathematically optimized scheduling.

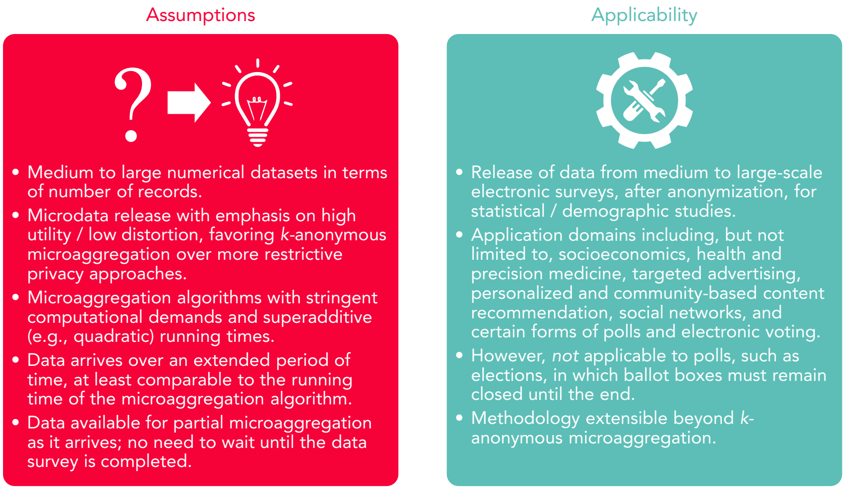
The formulation of the problem as a way to exploit prepartitioning in the time domain rather than in the space domain, both incremental microaggregation algorithms proposed and their variations, all of the mathematics in our theoretical analysis, including those in the main text as well as in the appendix, and the experimental results, are entirely novel work and thus our own contribution.

E. ASSUMPTIONS AND APPLICABILITY

The potential applicability of this work encompasses information systems designed for the collection, analysis, or dissemination of large amounts of anonymized data over extended periods of time in relation to the running time of a one-step microaggregation method. The ulterior purpose is permitting the swift release of data for statistical study, in contexts including, but not limited to, socioeconomics, healthcare, targeted advertising, personalized content recommendation, social networks, and certain forms of electronic voting.

We must caution that the underlying application must make part of the data available for anonymization before all of it is gathered. This may be possible in most forms of electronic surveys, but may not be applicable to certain electoral processes where ballot boxes must remain closed until the end of the election. A conceptually summarized list of assumptions and applicability of this work is provided in Fig. 4.

It is important to stress that the use of k-anonymous microaggregation is not essential to our methodology, which



Assumptions

- Medium to large numerical datasets in terms of number of records.
- Microdata release with emphasis on high utility / low distortion, favoring *k*-anonymous microaggregation over more restrictive privacy approaches.
- Microaggregation algorithms with stringent computational demands and superadditive (e.g., quadratic) running times.
- Data arrives over an extended period of time, at least comparable to the running time of the microaggregation algorithm.
- Data available for partial microaggregation as it arrives; no need to wait until the data survey is completed.

Applicability

- Release of data from medium to large-scale electronic surveys, after anonymization, for statistical / demographic studies.
- Application domains including, but not limited to, socioeconomics, health and precision medicine, targeted advertising, personalized and community-based content recommendation, social networks, and certain forms of polls and electronic voting.
- However, *not* applicable to polls, such as elections, in which ballot boxes must remain closed until the end.
- Methodology extensible beyond *k*-anonymous microaggregation.

FIGURE 4. Conceptually summarized list of assumptions and wide applicability of this work.

would be readily applicable to a number of privacy enhancements (at the cost of utility) for offline data release, such as l-diversity or t-closeness, approaches discussed in §II. The key assumptions in this work are quite general: the availability of data over time, and the nonlinearity of the anonymization process; hence its wide applicability.

F. CONTENTS AND ORGANIZATION

The rest of this paper is organized as follows. We briefly review in §II the current state of the art in k-anonymous microaggregation metrics and algorithms. §III formalizes the problem investigated, and introduces the metrics that characterize the effectiveness and viability of the methods proposed later. Specific algorithms are detailed and theoretically analyzed in §IV, whilst §V presents the experimental analysis and outcomes of these heuristic methods. Finally, conclusions are drawn in §VI.

II. BRIEF REVIEW OF THE STATE OF THE ART ON k-ANONYMOUS MICROAGGREGATION

This section provides an overview of the state of the art on those aspects of the field of statistical disclosure control more pertinent to this work. We focus on the methods and algorithms used to perform k-anonymous microaggregation while mitigating data utility loss. Additionally, we revise several attacks against anonymous using microaggregation, from an engineering perspective, arguing in favor of k-anonymity when data utility is critical, over alternatives offering stronger privacy guarantees at a higher cost in terms of information loss. An extensive survey of legal, socioeconomic aspects on the field of privacy can be found in [12].

A. k-ANONYMOUS MICROAGGREGATION IN THE CONTEXT OF STATISTICAL DISCLOSURE CONTROL

The field of SDC has been briefly presented and motivated in the introductory section. One of the methods initially proposed in the field of SDC was additive noise, studied in [4], consisting in adding random noise to the original dataset. Aside from the strong impact on distortion loss being, the noise introduced could be statistically dependent on the confidential attributes, posing a high disclosure risk. Since then, methods have been proposed from a more systematic perspective in order to anonymize datasets under certain constraints, data utility, privacy risk, and time efficiency.

A few years later, the k-anonymity model was presented in [39, 45]. We have explained that this privacy criterion ensures that reidentification based on quasi-identifiers is unfeasible, as there are, at least, k records sharing the same tuple of quasi-identifiers. Although there exist newly proposed methods aiming to enhance this privacy criterion [23], k-anonymity can still be viewed as a baseline criterion with high data utility. Original computational methods to achieve k-anonymity were based on the generalization and suppression of the quasi-identifiers. This was later modified into the k-anonymization criterion based on a microaggregation approach [1, 13, 15, 16, 18], used in the early nineties by the Eurostat agency. This showed that a similar outcome could be achieved without introducing or suppressing data. This model offers algorithmic and mathematical tractability without incurring in unmanageable computational complexity or prohibitive loss in data utility. Recent work [38] shows that k-anonymous microaggregation offers excellent data utility in machine learning applications. Specifically, it preserves the statistical dependence between demographics and confidential attributes, as macrotrends inferable by machine learning algorithms. Widely recognized in the SDC literature, the concept of k-anonymity through microaggregation is employed in other fields, beyond the publication of anonymized databases, such as artificial intelligence [17].

Different measures for data utility loss for numerical data can be found in the SDC literature, mainly relying on some form of distance, such as Euclidean distance or Minkowski distance. The nearly universally accepted measure is the sum of squared errors (SSE) [15, 22, 25, 29], defined as the sum of the squared distances from a common reconstructing value, ideally the mean of the points in the microaggregation cell.

B. A CRITICAL VIEW OF k-ANONYMITY AND OF ITS VARIANTS

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

In some cases, confidential attributes may be repeated or too similar. Revisiting the example presented in Fig. 1, an attacker who may know the age, gender, and ZIP code of one of the two females belonging to the second cluster knows that their hourly wage is in the range of \$37 and \$41. This inference is known as homogeneity or similarity attack. The attack is often formulated in qualitative terms as a privacy deficiency of k-anonymity. Observe however that in practice, the severity of the attack depends on the prevalence of the sensitive values of the confidential attributes, and the microcell size k. For example, the prevalence of type-2 diabetes in the general population in the U.S. is close to 9%, but for senior citizens 65 years and older that figure may rise to more than 25%. For $k = 10$, the risk of homogeneity attack in a microcell corresponding to aged individuals, with $p = {}^1\!/_4$, can be coarsely estimated as $p^k = 1/1,048,576$, less than once in a million. For microcells representative of the average population, with p=9%, the risk is even lower. And even for high prevalence nearing $p = {}^1\!/_2$ in symmetric studies, $p^k = 1/1024$.

Certain countermeasures, such as p-sensitive k-anonymity [43, 46], have been proposed to mitigate this kind of attacks. This stronger requirement advocates for at least p different values for each confidential attribute within each microcell. Although privacy is improved, it comes at the price of data utility. A slight generalization of this concept was introduced in [24, 28], and termed l-diversity. It requires at least l "well-represented" confidential attributes. Depending on the definition of well represented, l-diversity can be reduced to p-sensitive or be more restrictive, again at the expense of higher information loss.

Attacks against k-anonymity of a more probabilistic nature, known as skewness attacks, exploit the discrepancy between the distribution of confidential attributes of the entire table, or the population, and the distribution within a given k-anonymous cell. In the hypothetical example in Fig. 1, suppose that it is widely known in the country of reference that 33% of the entire population usually votes for the democratic party. A privacy attacker looks for a female individual aged 39 and resident in the area with ZIP code 9213. The attacker notes that there is a 66% probability that this individual is favorable to the democrat party, which is well above the population's average. In order to address this risk, the t-closeness criterion [27] requires that the distribution of a confidential attribute in a concrete cluster be similar to the distribution of the overall dataset. Even though differential privacy [11, 19] was conceived for online querying, and this work deals exclusively with microdata release for offline use, the criterion may be implemented as a form of t-closeness, as described in [42].

Strongly restrictive privacy criteria such as t-closeness, or differential privacy under the representation in [42], require that the within-cell probability be similar to that of the table or the general population. However, unveiling the absence or low prevalence of a sensitive condition below the population's average may pose no privacy risk. In the above diabetes example, a cell comprising only healthy individuals may be acceptable from a privacy perspective.

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

Finally, the attacker can gain insight if he is equipped with certain side information. In the synthetic example of Fig. 1, imagine that the attacker knows that the individual is an African-American male aged 22 who lives in the area with ZIP code 94024. Suppose that external demographic studies pointed out that African-Americans of this age were unlikely to support the republican party. The attacker could discard 2 out of 3 records and guess the individual's hourly wage. This form of statistical inference is known as background knowledge attack. These kind of attacks are studied in [44], where the authors propose strategies based on graph theory and inference paths.

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

C. ON THE COMPUTATIONAL COMPLEXITY OF k-ANONYMITY

A k-partition is said to be optimal when the SSE is minimum, but as shown in [33], attaining optimality is NP-hard. This fundamental result was later refined in [2], showing that the problem remains NP-hard even in the substantially simplified case of a ternary alphabet, $k = 3$, and length of the rows unbounded. In order to provide a lower bound, [3] showed that the problem is APX-hard when using a binary alphabet and $k = 3$. Unsurprisingly, current methods for microaggregation are heuristic, and strive to form partitions of k to $2k-1$ records per cluster [15], minimizing the SSE. The problem of constructing k-anonymous clusters with low SSE has been widely studied in the k-anonymity literature. Because they involve record comparison in one way or another, many heuristics for low-distortion k-anonymous microaggregation are quadratic, which makes dealing with large datasets computationally challenging.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875949, IEEE Access

Several effects affecting the complexity of k-anonymous microaggregation have been studied. In order to characterize the effect of input and output homogeneity, [6] introduces two new parametrizations: t_{in} defines the number of different input rows and t_{out} the number of different output rows. It is shown that k-anonymity is fixed-parameter tractable for t_{in} and that the problem becomes solvable in polynomial time when there is only one output row t_{out} for full homogeneity. However, the problem is still NP-hard and not fixed-parameter tractable for more than one output row.

Pattern-guided k-anonymity [5] aims to reduce the running time by letting the users express the differing importance of attributes. The complexity of the algorithm is relaxed by suppressing combinations of attributes that contain less information. In a demographical study of vote intention that collects data about the house income and the marital status of the respondents, the income input could be related to the marital status since the majority of participants surpassing a certain amount of earnings are married, defined as a pattern vector. Therefore, the marital status attribute would be suppressed and associated to the house income. This approach achieves a faster running time by reducing the data to process at the cost of an information loss, for instance, the information on an individual who is single and earns more than the threshold amount could be lost.

From a more practical perspective, recent efforts to tackle the stringent computational requirements of k-anonymous microaggregation resort to parallelization [31].

D. ALGORITHMS FOR k-ANONYMOUS MICROAGGREGATION

Different forms of microaggregation have been proposed in the SDC field, mainly driven by the similarity criterion used for clustering the records. These can be categorized into fixed-size and variable-size methods. The maximum distance (MD) algorithm [15], and its variation, the aforementioned MDAV algorithm [14, 18], more efficient in terms of computational complexity, are fixed-size algorithms, which implies that all groups but one, usually the last, contain k records. These algorithms are particularly efficient in terms of data utility for many data distributions, while requiring a relatively simple implementation.

On the other hand, variable-size methods try to exploit the possibility that different cell sizes lead to lower distortion, as long as all groups contain at least k records. Some popular implementations of variable-size algorithms are the μ-Approx [16], the minimum spanning tree (MST) [26], the variable MDAV (VMDAV) [40], and the two fixed reference points (TFRP) [9] algorithms. Attempts to bypass the complexity of multivariate microaggregation focus on projections onto one dimension, but are reported to yield a much higher disclosure risk [32].

Last but not least, we would like to cite two recent examples of k-anonymous microaggregation algorithms as an illustration of current research avenues in the field, beyond the functionality and applicability of the more traditional approaches aforementioned. One of them [36] explores an elegant extension of the usual SSE metric that contemplates not only the distortion of the quasi-identifiers due to aggregation, but also the valuable statistical dependence between quasi-identifiers and confidential attributes, in order to improve the statistical reliability of demographic studies. The second extension [37] explores in great detail a probabilistic variant of the k-anonymity criterion to address surveys with uncertain respondent participation.

E. RELATION WITH OTHER INCREMENTAL APPROACHES

In closer relation with the incremental method proposed in this paper, unsurprisingly, the basic notion of processing data as it arrives is by no means new in the field of SDC. A representative example is the brief yet insightful study on protection of dynamic databases presented in [7]. The authors address the challenge of microaggregating a continuously growing database, in order to release a series of frequently updated versions. Whilst the specific solution proposed deals exclusively with l-diversity, it can be immediately modified to incorporate k-anonymity, just as our own method can be readily extended to various flavors of anonymity criteria. There exist, however, great differences with respect to our approach, reaching even the most fundamental level.

- The incremental microaggregation process in [7] operates in a continuous stream of data samples, queuing, at most, l different values for the confidential attributes, and publishing marginally larger, protected versions of a dynamic database. Our proposal splits the data coarsely into two batches, and a single resulting database is published. The streaming approach addresses the functional requirement of frequent publication updates, whereas our proposal to split the data in batches obeys to a computational demand.
- Because [7] publishes several updates, the privacy criterion employed, l-diversity, must be satisfied not only for each individual database, but for the entire historic sequence of databases accessible by a privacy attacker with memory, that is, capable of cross-referencing records among updates. This is not a concern in our study.
- Yet another difference, this time of less significance, is the measure of distortion employed. The microaggregation algorithm in [7] constructs hyperrectangular microcells, and accordingly the authors advocate for a measure of information loss based on the sum of interval lengths for each dimension, relative to the dynamic range of each attribute. This measure cannot distinguish, for example, between data samples uniformly distributed along a microinterval, and identical samples with a single outlier. Our measure of distortion is effectively a form of statistical variance, characterizing the dispersion of a distribution on any microcell shape.

In general, we should stress that despite the rather lax use of the term incremental in the literature, the specific focus of this paper is radically different from any existing methods. We shall see that our proposal resorts to microaggregating data in two batches, in a manner carefully timed to optimally advance the end result, as well as to reduce the degradation in terms of data utility. Further, our method is a mathematically founded running-time strategy that releases the data only once, so that no additional privacy risks arise from cross-referencing several instances of published tables. Taking into consideration those fundamental differences, to the best of our knowledge, the specific approach of spatiotemporal prepartitioning with optimal scheduling put forth in this work is entirely novel.

III. FORMAL STATEMENT OF THE PROBLEM OF INCREMENTAL MICROAGGREGATION

This section formally presents the proposed model for k-anonymous incremental microaggregation, which aims to reduce the amount of time needed to anonymize a dataset, having as a reference the traditional alternative for microaggregation once all the data is available, following any of the approaches mentioned in §II.

A. FORMULATION OF MULTIVARIATE MICROAGGREGATION AS VECTOR QUANTIZATION

We briefly review the formulation of multivariate microaggregation as a vector quantization problem previously described in [34, 35]. The traditional k-anonymous microaggregation algorithm partitions a set of quasi-identifiers into cells of at least k samples. The scope of our analysis is limited to numerical data, meaning that, we assume that the quasi-identifiers aggregated are represented by n points $X = x_1, \ldots,\ x_n$ placed in the Euclidean space $\mathbb{R}^m$ of dimension m. These points are grouped into cells indexed by $q = 1, \ldots,\ Q \geqslant n/k$. Let x_j define the j^{th} record and $\hat{x}_q$ the mean value or centroid of the samples aggregated in the cell where x_j is assigned. The construction of cells, gathering at least k nearby samples, is represented by the cell-assignment or quantization function $q(j)$. Before releasing the dataset, each quasi-identifier tuple x_j is replaced by its perturbed version $\hat{x}_j$, the mean of the corresponding cell $q = q(j)$. This determines a centroid-assignment or reconstruction function $\hat{x}(q)$. This process is conceptually depicted in Fig. 5.

B. FORMULATION FOR INCREMENTAL k-ANONYMOUS MICROAGGREGATION

Consider a fixed number of records n, fixed dimension m, and a fixed k-anonymity parameter k. Let t represent the time required to microaggregate the entirety of the data using a traditional algorithm as MDAV.

Suppose now that we proceed to microaggregate the data in two portions. A large portion of the data, consisting in n_0 records, is microaggregated first, with a conventional microaggregation algorithm, which we shall call base algorithm. Next, a smaller portion of n_+ records is processed with an incremental algorithm that aggregates the new data to the old one, in such a manner that the overall result is k-anonymous.

Let $\nu \in [0, 1)$ denote the fraction of the data processed incrementally, so that $\nu = n_+/n$. In practice, ν may be small not to degrade the overall distortion with respect to the traditional one-step approach, formally representable with $\nu = 0$. For convenience, the complement $1 - p$ of any expression p is occasionally denoted by $\bar{p}$. In this notation, $\bar{\nu} = n_0/n$ is the fraction of base data. In our analysis, n_0 and n_+ are bijectively expressed as n and ν. This notation is represented in Fig. 6.

1) Running Times

For consistency, all times τ are relative to the running time t of the traditional method. For instance, if t_0 represents the running time of the baseline algorithm, the relative time is $\tau_0 = t_0/t$. Following this notation, let τ_+ represent the relative running time of the incremental algorithm on the incremental portion ν of the data. The head start of the base algorithm is defined as τ_-, also taking the duration t of the traditional method as time unit.

Assume for simplicity that at least toward the end of the electronic survey, data arrivals are approximately uniformly paced. We introduce a head start coefficient ς representing the amount of time, always in our normalized units, required for a fraction ν of samples to arrive. Thus, $\tau_- = \varsigma\nu$. Although such uniformity is really only required by our analysis for a subrange of ν of interest, extrapolating it to all the data would give $\tau_- = \varsigma$ for $\nu = 1$. In this case, the head start coefficient ς would represent the time required for the entire survey, in relation to the time required by the traditional microaggregation algorithm. For instance, in an eight-hour electronic survey followed by a traditional algorithm that needs two hours to run, this coefficient would be $\varsigma = 4$. Concordantly, $\varsigma \ll 1$ indicates very fast data availability, or slow traditional microaggregation. The opposite case, $\varsigma \gg 1$, for which the data arrives very slowly in terms of the time required for microaggregation, should allow the greatest gain when resorting to incremental microaggregation instead. We shall demonstrate that both cases will in fact yield a significant time gain, and that indeed, the case of slow arrivals, will be more convenient.

Although for mathematical tractability this work is limited to uniform arrivals, more broadly, we could encounter the type of arrivals profiled in Fig. 7, in which ς would play the role of derivative of τ_- at $\nu = 0$, offering a linear approximation to its real trend.

Finally, $\Delta\tau$ characterizes the relative time gain due to incremental microaggregation with respect to traditional microaggregation on the entire data. If the base algorithm were to finish before the available head start period had run out, that is, if $\tau_0 \leqslant \tau_-$, then $\Delta\tau = 1 - \tau_+$. Otherwise, if the head start is insufficiently generous, that is, $\tau_0 \geqslant \tau_-$, then,

$$\Delta\tau = \tau_- + 1 - \tau_0 - \tau_+.$$

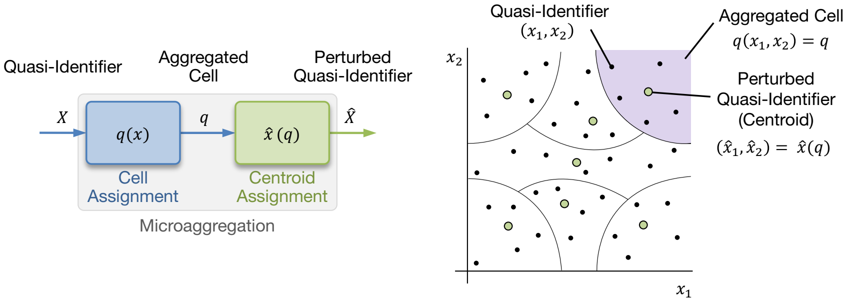

Quasi-Identifier
Aggregated Cell
Perturbed Quasi-Identifier
X
$q(x)$
q
$\hat{x}(q)$
$\hat{X}$
Cell Assignment
Centroid Assignment
Microaggregation
Quasi-Identifier (x_1, x_2)
Aggregated Cell $q(x_1, x_2) = q$
Perturbed Quasi-Identifier (Centroid)
$(\hat{x}_1, \hat{x}_2) = \hat{x}(q)$
x_2
x_1

FIGURE 5. Microaggregation interpreted as a quantization problem on the quasi-identifiers. The shape of the cells attempts to illustrate how a real microaggregation algorithm

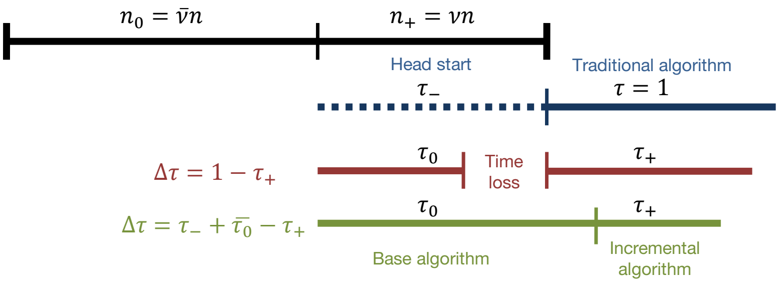


FIGURE 6. Notation in the formulation of the incremental microaggregation problem. τ defines the relative time of each computational process in comparison with the time nee

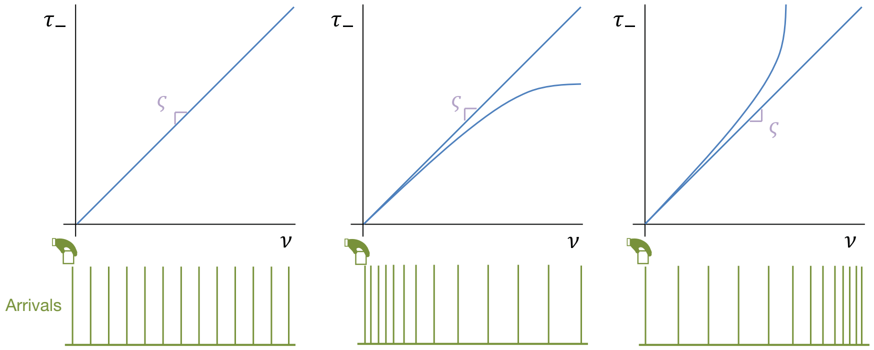


FIGURE 7. The relative head start time is directly related to the pace with which data arrives and the incremental data ratio ν. To characterize this dependence in the simplest r

Consequently,

$$\Delta\tau = \min\{\tau_- + \bar{\tau}_0, 1\} - \tau_+.$$

We wish to achieve a large relative time gain $\Delta\tau$, which will depend on the design of the incremental algorithm, the portion ν of incremental data chosen, and the head start τ_-, which in turn depends on the arrivals coefficient ς.

2) Critical and Optimal Data Ratio

In the conceivable event that $\tau_0 \leqslant \tau_-$, it is intuitively clear that the time from the end of the base algorithm until the start of the incremental algorithm is effectively wasted, with no computation being carried out. In this work, ν_- denotes the critical data ratio for which $\tau_- = \tau_0$, that is, when the running time of the base algorithm matches the head start time. .

Let $\nu^\star$ denote the optimal data ratio for incremental microaggregation, defined as the incremental data ratio maximizing the time gain:

$$\nu^\star = \arg\max_{\nu} \Delta\tau.$$

As expected, $\nu^\star \leqslant \nu_-$, and consistently, $\tau_0 \geqslant \tau_-$. Otherwise, part of the head start would be time wasted without computation. This can also be seen by arguing that $\Delta\tau$ is decreasing for $\nu \geqslant \nu_-$.

3) Data Ratio for a Deadline

Aiming to maximize the time gain may not be necessary if the requirements imposed by the application at hand are simply expressed by means of a deadline. We shall see in §V-B3, that larger data increments ν will generally have a distortion impact. Suppose that we wish to choose the smallest ν possible in order to satisfy the given deadline.

Let $\tau_{\text{end}} = t_{\text{available}}/t_{\text{traditional}}$ define the amount of relative time available and $\tau^\star < \tau_{\text{end}}$ define the relative time required using the optimal data ratio $\nu^\star$, the earliest our method can finish. Let ν_{end} define the data ratio that has to be selected to fill the span of time available and reduce the negative effect on data utility. As we argued before, not to waste head start time without computation, $\nu_{\text{end}} \leqslant \nu_-$, or equivalently, $\tau_0 \geqslant \tau_-$. To meet the deadline exactly, for the smallest possible ν with the smallest distortion impact, we wish to satisfy

$$\tau_0 + \tau_+ = \tau_{\text{end}} + \tau_- .$$

Since $\nu_{\text{end}} \leqslant \nu_-$, we may reuse the previous definition of the relative time gain before the saturation region,

$$\Delta\tau = \bar{\tau}_0 + \tau_- - \tau_+ ,$$

with $\tau_{\text{end}} = 1 - \Delta\tau$.

4) Data Utility Loss

Splitting the data for microaggregation may have an impact on the distortion incurred, as the two-step process cannot group samples belonging to different time segments, even if the quasi-identifiers are spatially close in Euclidean distance. This motivates the distortion metrics introduced next, and the empirical analysis in §V-B3.

Recall that is usual in traditional microaggregation to conduct a attribute-wise, unit-variance normalization prior to any manipulation of the data, inherent to the conventional definition of distortion error in SDC.

In the SDC literature distortion is conventionally evaluated as the quotient between the sum of squared errors (SSE) and the sum of squares total (SST). The SSE is defined as the sum of squared differences between each sample and its cell mean (centroid). Similarly, SST is defined as the sum of squared differences between each sample and the mean of all samples. Owing to the attribute-wise unit-variance normalization on the data, the total variance of the data is the number of dimensions m, and SST $= mn$. Therefore, the customary distortion criterion SSE/SST in the SDC literature is identical to the total variance per sample widely employed in the field of vector quantization:

$$\mathcal{D} \overset{\text{def}}{=} \frac{\text{SSE}}{\text{SST}} = \frac{1}{mn} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2 .$$

When performing microaggregation as the two-step process described, we should carefully average the distortion of each step to compare fairly with the distortion $\mathcal{D}$ on all data using the traditional algorithm. Let $\mathcal{D}_0$ and $\mathcal{D}_+$ be the distortions of the base algorithm and the incremental algorithm, respectively. We define the total distortion $\mathcal{D}_T$, when computing the k-anonymous microaggregation incrementally, as the weighted sum of both base and incremental distortion, precisely, $\mathcal{D}_T \overset{\text{def}}{=} \bar{\nu}\mathcal{D}_0 + \nu\mathcal{D}_+$, which gives us exactly the average distortion per sample faithful to the original definition.

To facilitate comparisons, the measure considered will be relative to the traditional algorithm distortion. That is, define $\delta_T \overset{\text{def}}{=} \mathcal{D}_T/\mathcal{D}$, and finally, let δ represent the relative distortion loss measured as

$$\delta \overset{\text{def}}{=} \frac{\mathcal{D}_T - \mathcal{D}}{\mathcal{D}}, \quad \text{or equivalently,} \quad \delta = \delta_{\text{T}} - 1.$$

It is natural to expect that the time partition effectively carried out by our proposal should have an impact on the overall distortion, as data points spatially close may end up split apart in the two time segments. Indeed, this will be the case, and we shall expect that, in general the relative overhead δ in quadratic distortion be positive, although hopefully, only by a small percentage. This extra distortion will be the price to pay for the considerable time gains obtained with the optimal and deadline-based strategies previously outlined.

5) Effect of the Number of Dimensions on the Overall Distortion

This section merely offers an extremely informal argument, by no means a rigorous proof, which anticipates a somewhat counterintuitive yet extremely convenient experimental finding regarding the distortion overhead δ just defined. The argument will also help us understand the real impact of our proposal and interpret the experimental results from a better-informed perspective. We have argued that δ should typically increase with ν, and constitute the price to pay for achieving substantial time gains. We shall however find quite fortunately that for datasets with a large number m of nonredundant quasi-identifiers, the distortion penalty will be unexpectedly small.

We assume that the reader has certain conceptual familiarity with the fundamental principles of vector quantization [21]. Informally, suppose first the simple case of one-dimensional microaggregation, that is, $m = 1$. For samples approximately uniformly distributed, traditional microaggregation will produce a certain distortion $\mathcal{D}$. Suppose further that the set of samples were split at random into two, corresponding to an extreme value of $\nu = \frac{1}{2}$, with the proviso that the arrival times were independent from the demography of the respondent. Then, on each partition, samples would be on average twice more distant from each other, the average quadratic distortion for each part should be roughly four times higher, and the same would apply to the overall distortion $\mathcal{D}_T$ averaging the two parts. In conclusion,

$$\mathcal{D}_T/\mathcal{D} \approx 4 \quad \text{for} \quad m = 1.$$

This suggests that for very low-dimensional microaggregation we should avoid at all costs a large value of ν, or avoid this approach altogether if time is not of the essence. What about $m \to \infty$? We claim that in that case, the distortion overhead would be negligible.

We provide here a justification in terms of m-dimensional hyperballs roughly representing microcells in the space of m-dimensional data records. Recall that spherical balls minimize the moment of inertia (normalized squared distortion) of convex polytopes approximately tessellating a high-dimensional Euclidean space. For this reason, cells in high-dimensional vector quantization should approach this ideally shape [21].

Accordingly suppose that k-anonymous microcells containing m-dimensional quasi-identifiers resulting from a microaggregation algorithm with excellent performance in terms of low distortion could be suitably approximated by m-balls, with k points roughly uniformly distributed inside them. The volume of a ball in m dimensions is

$$V_m(R) = \frac{\pi^{\frac{m}{2}}}{\Gamma(\frac{m}{2}+1)} R^m,$$

where R is the radius of the ball and Γ denotes the Euler gamma function. In many dimensions, most of the volume of an m-dimensional ball is concentrated on its crust, because the volume increases with the m^{th} power of the radius. To see this, imagine two concentric 10-dimensional balls with radius $R = 0.9$ and $R' = 1$, respectively, and let $V_{\text{crust}} = V' - V$ denote the volume of the crust. Then,

$$V/V' = (R/R')^{10} = 0.9^{10} \approx 0.349,$$

which means that approximately 65% of the volume is concentrated in the ball's crust. If the number of dimensions were increased further, the amount of volume concentrated in the balls crust would eventually approach 100%.

But if most of the volume is concentrated in the ball's crust, most of the data samples should be close to the surface, and the difference between each sample and its centroid would by roughly the radius of the m-dimensional ball:

$$\mathcal{D} = \frac{1}{mn} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2 \approx \frac{1}{m} R^2.$$

Additionally, if we randomly reduce the points in the dataset from n to $n/2$ with $\nu = \frac{1}{2}$, maintaining a uniform distribution and keeping k points per cell, the volume of the m-dimensional ball will double. Strikingly, the effect of doubling the volume will leave the radius almost unaffected, since for $V' = 2V$ the radius is $R' = \sqrt[m]{2}R$, but $\sqrt[m]{2} \xrightarrow[m\to\infty]{} 1$. In the previous example of two ten-dimensional balls, the effect of doubling the volume increases the radius by a meager $\sqrt[10]{2} \approx 1.07$. And we have seen that the distortion is a quadratic function of the radius, regardless of the dimension m. Therefore, if the radius grows only slight, so will the distortion. Precisely,

$$\mathcal{D}_T/\mathcal{D} \approx 2^{\frac{2}{m}} \xrightarrow[m\to\infty]{} 1,$$

as claimed.

6) Inertial Coefficient

A goal of our project consists in designing algorithms to process large datasets in two steps enabling the possibility of starting the process before receiving the whole dataset. Consider the following incremental microaggregation strategy, which reuses the cells and centroids constructed by the base algorithm. For each incremental data point, we simply select a previously formed nearby cell to which the point could be adjoined. The simplest approach to adjoin the new point would consist in choosing the nearest centroid. However, adding a new sample will alter the cell's centroid, affecting the distortion of rest of samples in cell. A slightly better strategy follows, which contemplates updating the centroid to minimize the within-group distortion, and the consequent effect on the other points originally assigned to the corresponding cell.

Consider a set of points $x_1, \ldots, x_n \in \mathbb{R}^d$, representing in this subsection the points of a cell rather than the entire dataset, despite the reuse of n, and define the random variable (r.v.) X, uniformly distributed over this set. Their centroid is the mean $\mu = \operatorname{E} X = \frac{1}{n}\sum_j x_j$, and denote the variance by σ^2. In this notation, the SSE of the set is $n\sigma^2$. Let $y \in \mathbb{R}^d$ be the new point to be added to the set, resulting in a modified centroid μ', and a modified SSE$'$. Without the centroid update, the resulting error would in general be greater. By iterated expectation, the new centroid is

$$\mu' = \frac{n}{n+1}\mu + \frac{1}{n+1}y.$$

Directly from its definition, the resulting SSE is

$$\text{SSE}' = n \operatorname{E} \|X - \mu'\|^2 + \|y - \mu'\|^2,$$

where

$$\operatorname{E} \|X - \mu'\|^2 = \sigma^2 + \|\mu - \mu'\|^2.$$

But $y - \mu' = \frac{n}{n+1}(y - \mu)$ and $\mu - \mu' = \frac{1}{n+1}(\mu - y)$. Consequently,

$$\begin{aligned}\text{SSE}' &= n\sigma^2 + \frac{n}{(n+1)^2}\|\mu - y\|^2 + \left(\frac{n}{n+1}\right)^2 \|y - \mu\|^2 \\ &= \text{SSE} + \frac{n}{n+1}\|y - \mu\|^2,\end{aligned}$$

giving the SSE increment

$$\Delta\text{SSE} = \frac{n}{n+1}\|y - \mu\|^2,$$

lesser than the increment $\|y - \mu\|^2$ incurred if the centroid had not been updated, particularly for small cell sizes n.

The above reasoning leads to the following strategy to adjoin a new point y to an anonymized microdata set, through selection of the optimal (quantization) cell $q^\star$ among all possible cells, indexed by q, of size n_q, with centroid $\hat{x}_q$:

$$q^\star = \arg\min_q \frac{n_q}{n_q + 1}\|y - \hat{x}_q\|^2.$$

Note that the squared distance to the nearest centroid is weighted by a coefficient that increases with the cell size, to account for the effect of the centroid update on other points in the cell. The centroid update will produce a final SSE smaller than if the centroid were left unchanged. Evidently, for this latter, suboptimal strategy, the best choice would have been the naïve one, namely, to select the closest centroid without regard for the size of the corresponding cell, that is, $\arg\min_q \|y - \hat{x}_q\|^2$. The naïve strategy would also be approximately optimal in the case of macroaggregation, where $n_q \geqslant k \gg 1$ and $\frac{n_q}{n_q+1} \approx 1$.

In either case, the computation required is negligible with respect to the alternative of microaggregating the $n+1$ points from scratch, at least for static algorithms such as MDAV. Finally, because we have assumed that the centroids are stored for this type of procedures, we point out that computing and storing the updated centroid $\hat{x}'_q$ is a trivial task. It suffices to reuse the iterated-expectation equation in the above reasoning,

$$\hat{x}'_q = \frac{n}{n+1}\hat{x}_q + \frac{1}{n+1}y,$$

a mere convex combination of the old centroid with the new point.

The presumably small amount of points that can be adjoined in this fashion, before a fresh microaggregation would significantly improve the overall distortion, remains to be assessed experimentally. Evidently, cells of sizes $2k$ or larger created by adjoining points could be split to lower the distortion.

IV. THEORETICAL ELEMENTS AND ALGORITHMIC PROPOSALS

While the essential aspect of our contribution is the reduction of computational cost, a high distortion loss must also be avoided. We shall seek the best compromise between time efficiency and minimum statistical change of the data when anonymizing. Bearing this in mind, two incremental microaggregation algorithms are devised here and evaluated experimentally later on.

Our analysis and comparisons will be based on the use of MDAV [14, 18], as both traditional and base algorithm, one of the most widely accepted microaggregation algorithms in the SDC literature. The specification of MDAV used in this paper is the one provided as Algorithm 5.1 in [18], named "MDAV-generic":

1. Find the centroid C of the n records, find the furthest point P from the centroid C, and find the furthest point Q from P.
2. Group the $k-1$ nearest points to P into a group and the 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.

It is routine to show that the running time of MDAV in terms of the anonymity parameter k and the number of records n is approximately n^2/k, for $n \gg k$.

In the following, we study a series of incremental methods and the tools employed in their conception. We preface our presentation with Table 1, summarizing the incremental algorithms and variations proposed in this theoretical section. In short, the first approach, termed 2MDAV, resorts to using MDAV on each of the two partitions of the data. The second approach employs MDAV once on the base partition, and considers variations of the strategy for adjoining new points to previously formed cells outlined in §III-B6. Cells resulting in $2k$ records or more can always be split for a distortion reduction.

A. MDAV AS INCREMENTAL ALGORITHM

The simplest strategy, which will serve as reference, is simply two perform the traditional algorithm on each individual time partition. Here, we run MDAV in two consecutives steps, directly on the two portions of the data, and concatenate them for publication:

1. Wait until a portion $\bar{\nu}$ of the data is available, that is, wait for the n_0 base records.
2. Run MDAV on the n_0 base records.
3. Once MDAV has finished and when the remaining portion ν of the data is available, proceed to run MDAV on the new n_+ incremental records.
4. Join the two resulting partitions directly.

According to the formulation introduced in §III, the approximate running time n^2/k of MDAV enables us to immediately determine the base running time τ_0 in terms of the incremental data ratio ν. Recall that $\bar{\nu} \overset{\text{def}}{=} 1-\nu$, and that $\nu \in [0,1)$. Precisely,

$$\tau_0 = \frac{t_0}{t} = \frac{(\bar{\nu}n)^2/k}{n^2/k} = \bar{\nu}^2.$$

Even though, in general, running times depend on n, ν and k, the advantage of the normalization in traduced in §III is that leads to simpler expressions that do not explicitly depend on k or n, or the specific machine the algorithms are executed on. If MDAV is also used as an incremental algorithm, then a similar argument shows that $\tau_+ = \nu^2$. Therefore, the running time of microaggregating data in two portions using MDAV is $\frac{1}{2} \leqslant \bar{\nu}^2 + \nu^2 \leqslant 1$, where the lower bound expresses the full power of subadditivity for $\nu = \frac{1}{2}$, but disregards the head start effect, and might have too strong an impact on distortion. Due to subadditivity alone, running MDAV in two steps will always be more efficient than a single run on the whole dataset.

B. OPTIMAL DATA RATIO

Recall that we defined the critical data ratio ν_- as that satisfying $\tau_0 = \tau_-$, that is, the data ratio for which the running time of the base algorithm is equal to the available head start. Recall also that we aim to achieve $\tau_- \geqslant \tau_0$ so there is no time loss when collecting data. Under the uniformity assumption made in §III-B2, expressed by means of $\tau_- = \varsigma\nu$, where ς is the time for all the data to become available, relative to the running time of the traditional microaggregation algorithm. Using MDAV as a base algorithm, or any other quadratic method, the relative computational cost of the base algorithm is $\tau_0 = \bar{\nu}^2$. It is routine to verify that the value ν_- for which $\tau_0 = \tau_-$ is

$$\nu_- = \frac{2+\varsigma-\sqrt{\varsigma(4+\varsigma)}}{2},$$

TABLE 1. Incremental Algorithms

Algorithm	Variants	Description
2MDAV	-	Run two consecutive MDAV and aggregate the results.
Find Nearest Neighbor	**No Splitting**	Assign incremental records to its nearest centroid of those found by the base algorithm.
	Splitting in the Middle	Split cells bigger than $2k-1$ using MDAV while assigning incremental records to its nearest centroid of these found by the base algorithm.
	Splitting at the End	Assign incremental records to its nearest centroid of those found by the base algorithm. Once all incremental points have been assigned, split cells bigger than $2k-1$ using MDAV.

whenever $\tau_0 = \bar{\nu}^2$ and $\tau_- = \varsigma\nu$, that is, whenever the base algorithm is quadratic and the arrivals uniform. The negative square root is consistent with the constraint $\nu_- \in (0, 1]$. Note also that viewing the square root as a geometric mean, it must be greater than the minimum but lesser than the arithmetic mean, in keeping with the range of the solution. We proceed to explore approximations to the critical data ratio ν_- for increasing values of the arrival time coefficient ς, that is, for slow arrivals.

To that end, we make a brief digression and introduce the concept of strong approximation, which will enable us to rewrite some of the results in this work in a more intuitive form, yet remarkably accurate. These approximations will exhibit trends in the particularly interesting case of slow arrivals, for which the entirety of the data is available over a period of time far greater than the running time of the traditional algorithm, case characterized by $\varsigma \to \infty$. For readability, the basic concepts regarding this strong approximation are introduced below, but the derivation of results is relegated to the Appendix.

Let f and g be real-valued functions of a common real-valued argument, and consider their limit as the argument approaches a given value, or infinity. We write $\lim_{x\to x_0} f(x) = l$ for some x_0, possibly infinity, more compactly as $f \to l$. We seek an approximation stronger than the approximation in absolute error

$$f \simeq g \overset{\text{def}}{\Leftrightarrow} f - g \to 0,$$

and the approximation in relative error

$$f \sim g \overset{\text{def}}{\Leftrightarrow} f/g \to 1 \Leftrightarrow \frac{f-g}{g} \to 0.$$

An adequate definition for our purposes follows. We shall say that f is a strong approximation for g, or vice versa, when the following two conditions are satisfied:

$$f \doteq g \overset{\text{def}}{\Leftrightarrow} \begin{cases} f - g \to 0 \\ 1/f - 1/g \to 0 \end{cases}.$$

We show in the Appendix that the approximation employed in this work is stronger than both the absolute and relative approximations holding simultaneously. In the event that $f \to a$ for some constant $a \neq 0, \infty$, then clearly $f \doteq a$.

Back to our analysis on incremental microaggregation, we apply the results described in the Appendix to the previous expression for ν_-, obtaining an accurate approximation for slow arrival times in terms of a more intuitive expression:

$$\nu_- = \frac{2 + \varsigma - \sqrt{\varsigma(4+\varsigma)}}{2} \doteq \frac{1}{2+\varsigma} \quad \text{as} \quad \varsigma \to \infty.$$

Both the exact values and the approximation are shown in Fig. 8.

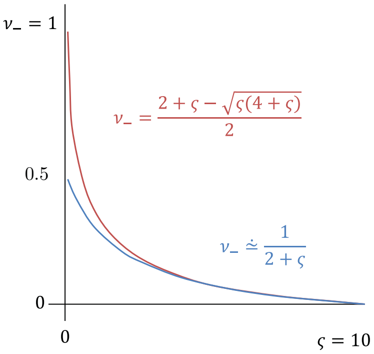


FIGURE 8. Comparison of the critical data ratio ν_- with its strong approximation, suitable for large values of ς.

As previously defined, $\Delta\tau$ characterizes the relative amount of time gained using the incremental method. In order to select the optimal data ratio, we aim to maximize

$$\Delta\tau = \min\{\bar{\tau}_0 + \tau_-, 1\} - \nu^2,$$

where a quadratic base algorithm gives

$$\bar{\tau}_0 = 1 - (1-\nu)^2 = \nu(2-\nu),$$

and uniform arrivals imply

$$\bar{\tau}_0 + \tau_- = \nu(2 + \varsigma - \nu).$$

As $\Delta\tau$ is non increasing past the critical data ratio, without loss of generality the maximization is restricted to $\nu \leqslant \nu_-$, where $\Delta\tau = (2 + \varsigma - 2\nu)\nu$. From this, we can deduce that the value of the optimal data ratio is $\nu^\star = (2+\varsigma)/4$ for small values of ς. In order to completely determine $\nu^\star$ we shall verify the constraint $\nu^\star \leqslant \nu_-$, and find when ν_- matches this preliminary maximum, that is,

$$\frac{2 + \varsigma - \sqrt{\varsigma(4+\varsigma)}}{2} = \frac{2+\varsigma}{4},$$

which gives the solution

$$\varsigma_- = 2(2/\sqrt{3} - 1) \approx 0.309,$$

and which we shall refer to as critical head start coefficient.

We may finally completely specify the optimal data ratio under the above assumptions, namely quadratic base algorithm, quadratic incremental algorithm, and uniform arrivals, as

$$\nu^\star = \begin{cases} \frac{2+\varsigma}{4} \leqslant \nu_-, & \varsigma \leqslant \varsigma_- = 2\left(\frac{2}{\sqrt{3}} - 1\right) \\ \frac{2+\varsigma-\sqrt{\varsigma(4+\varsigma)}}{2} = \nu_- \simeq \frac{1}{2+\varsigma}, & \varsigma \geqslant \varsigma_- = 2\left(\frac{2}{\sqrt{3}} - 1\right) \end{cases},$$

where the approximation is for slow arrivals $\varsigma \to \infty$. At the threshold $\varsigma_- = 2(2/\sqrt{3} - 1)$, $\nu^\star = 1/\sqrt{3}$. The solution is plotted in Fig. 9.

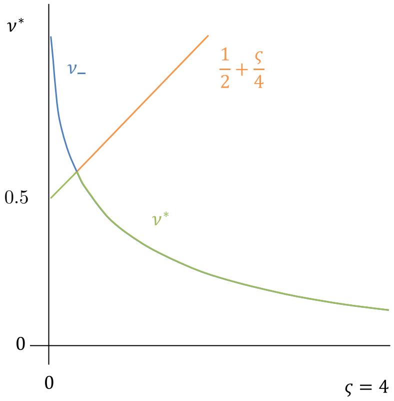


FIGURE 9. Representation of the optimal data ratio $\nu^\star$ under the assumption of uniform arrivals and using MDAV as both base and incremental algorithm. Observe that once we surpass the saturation point, $\nu^\star = \nu_-$.

We already established that, for efficiency, the head start should be entirely consumed in the computation of the microaggregation of the base data, that is, $\tau_0 \geqslant \tau_-$, and we argued that consistently, $\nu^\star \leqslant \nu_-$. For the special case of 2MDAV and uniform arrivals, our conclusions can be made more precise. The above analysis shows that the critical data ratio ν_- becomes the optimal data ratio $\nu^\star$ for sufficiently slow arrivals timed according with $\varsigma \geqslant \varsigma_-$, so that $\tau_0 = \tau_-$. Fig. 10 illustrates this coincidence. On the other hand, fast arrivals determined by $\varsigma < \varsigma_-$ require $\nu^\star < \nu_-$, that is, $\Delta\tau$ attains its maximum before the saturation point ν_-, and accordingly, the running time of the base algorithm τ_0 will exceed the head start τ_-.

The corresponding optimal relative time gain, still under the assumptions of uniform arrivals and quadratic microaggregation in both incremental stages,

$$\Delta\tau^\star = (2 + \varsigma - 2\nu^*)\nu^\star =$$

$$= \begin{cases} \frac{(2+\varsigma)^2}{8}, & \varsigma \leqslant \varsigma_- = 2\left(\frac{2}{\sqrt{3}} - 1\right) \\ \frac{\sqrt{\varsigma(4+\varsigma)}\,\left(2+\varsigma-\sqrt{\varsigma(4+\varsigma)}\,\right)}{2} \simeq 1, & \varsigma \geqslant \varsigma_- = 2\left(\frac{2}{\sqrt{3}} - 1\right) \end{cases}.$$

It is routine to check that at the threshold $\varsigma = \varsigma_-$, we have $\Delta\tau^\star = 2/3$.

For extremely fast arrivals, in the limit of the head start coefficient $\varsigma = 0$, $\nu^\star = 1/2 = \Delta\tau^\star$. Each of the two steps anonymizes half of the data and thus runs in $1/4$ of the time corresponding to the traditional approach. The overall time gain comes only from the superadditivity of the quadratic algorithm, as the head start $\tau_- = \varsigma\nu^\star \to 0$ becomes irrelevant. This is the best we could do with two-stage spatial prepartitioning when all data is available at the same time instant. On the other hand, in the limit of $\varsigma \to \infty$ for extremely slow arrivals, we have $\nu^\star \to 0$, but interestingly $\Delta\tau^\star \to 1$ (thus $\Delta\tau^\star \simeq 1$), which is not immediately obvious. Our approach clearly excels in this case. In general, the relative gain $\Delta\tau^\star$ monotonically increases with the relative arrival coefficient ς, in other words, slower arrivals yield better time gains, owing not only to superadditivity but also to head starts, as one would expect. In all cases, $\Delta\tau^\star \geqslant 1/2 > 0$, which indicates that this formalism is applicable to a wide range of arrival profiles. The above optimality analysis is summarized in Fig. 11.

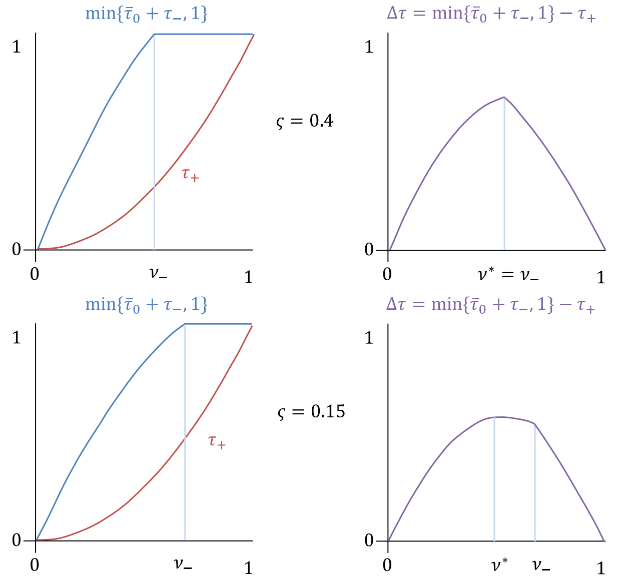


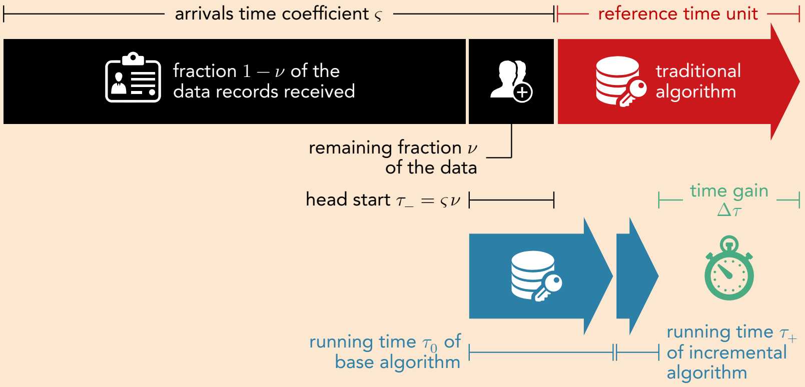
FIGURE 10. Effect of ς on the critical data ratio $\nu^\star$ and the relative time gain $\Delta\tau$. The relative time gain reaches its maximum where $\tau_0 + \tau_- = 1$, saturation region, for values bigger than $\varsigma = \varsigma_-$. However, when using values of ς smaller than the threshold the maximum, and optimal data ratio, is reached before the saturation point.

C. DATA RATIO FOR A DEADLINE

The previous analysis of the optimal data ratio aims to maximize the time gained by selecting the appropriate amount of incremental data. However, as already outlined in §III-B3, certain applications may have a more lenient time constraint. We have mentioned that using this time to enlarge the span of time taken to run the incremental approach has a positive effect on the data utility loss. In that regard, we further study the relative time measure proposed in §III-B3 for the particular case of MDAV as an incremental algorithm and under the assumption of uniformly distributed arrivals.

Recall that we wish to finish the overall anonymization process right at the deadline $\tau_{\text{end}} = 1 - \Delta\tau$. To avoid wasting time for computation, we first consider the maximum relative time gain before the saturation region, which is the earliest our method can finish, precisely $\Delta\tau_{\max} = \nu^\star(2+\varsigma-2\nu^*)$. We reasonably assume the upper bound $\tau_{\text{end}} < 1$; otherwise, the traditional approach would be perfectly suitable. Also

Brief Recapitulation of the Optimality Analysis of Incremental Microaggregation


arrivals time coefficient ς
reference time unit
fraction $1-\nu$ of the data records received
traditional algorithm
remaining fraction ν of the data
head start $\tau_- = \varsigma\nu$
time gain $\Delta\tau$
running time τ_0 of base algorithm
running time τ_+ of incremental algorithm

- Under the two-step microaggregation approach proposed in this work, the incremental data ratio ν represents the fraction of the data records to be microaggregated incrementally. The base algorithm operates on the initial portion $1-\nu$ of the data.
- All times τ are relative to the running time of the traditional algorithm, carried out in a single step once all data is available. For the proposed two-step alternative, τ_0 represents the relative running time of the base algorithm, and τ_+ the relative running time of the incremental step.
- The arrivals coefficient ς relates the head start time τ_- with the fraction ν of the data arrived over that period. Under the assumption of uniform arrivals during the head start toward the end of the survey, $\tau_- = \varsigma\nu$. The coefficient may be interpreted as the time for the entire data to arrive, extrapolating the pace at the end. As usual, this time is relative to the duration of the traditional microaggregation algorithm.
- The critical data ratio ν_- corresponds to the data ratio for which the running time τ_0 of the base algorithm matches the head start τ_-.
- The optimal data ratio ν^* maximizes the relative time gain $\Delta\tau$ of the two-step approach with respect to the traditional one, and it will depend on the arrivals coefficient ς. We show that $\nu^* \leqslant \nu_-$. For convenience, denote the transition arrivals coefficient by $\varsigma_- = 2\left(\frac{2}{\sqrt{3}}-1\right) \approx 0.309$. All approximations are for slow arrivals, that is, $\varsigma \to \infty$. Assuming uniform arrivals during the head start, and quadratic running times for both the base and the incremental algorithm,

$$\nu^* = \begin{cases} \dfrac{2+\varsigma}{4} \leqslant \nu_-, & \varsigma \leqslant \varsigma_- \\ \dfrac{2+\varsigma-\sqrt{\varsigma(4+\varsigma)}}{2} = \nu_- \doteq \dfrac{1}{2+\varsigma}, & \varsigma \geqslant \varsigma_- \end{cases}.$$

- Under the same assumptions, the optimal relative time gain is

$$\Delta\tau^* = (2+\varsigma-2\nu^*)\nu^* = \begin{cases} \dfrac{(2+\varsigma)^2}{8}, & \varsigma \leqslant \varsigma_- \\ \dfrac{\sqrt{\varsigma(4+\varsigma)}\,\left(2+\varsigma-\sqrt{\varsigma(4+\varsigma)}\right)}{2} \doteq 1, & \varsigma \geqslant \varsigma_- \end{cases}.$$

- For extremely fast arrivals, in the limit of the head start coefficient $\varsigma = 0$, $\nu^* = {}^1\!/_2 = \Delta\tau^*$. On the other hand, in the limit of $\varsigma \to \infty$ for extremely slow arrivals, we have $\nu^* \to 0$, but quite conveniently, $\Delta\tau^* \to 1$. Our two-step method excels in applications where $\varsigma \gg 1$.
- Slower arrivals yield better time gains, owing not only to a more generous head start, but also to the superadditivity of the running times of the algorithms involved. For any ς, the time gain $\Delta\tau^* \geqslant {}^1\!/_2 > 0$, showing that our approach is helpful in a wide range of arrival profiles.

FIGURE 11. Quick summary of the optimality part of our theoretical analysis on incremental microaggregation for electronic surveys, formulated in §III-B2 and presented in §IV

quite naturally, the deadline cannot be more restrictive than the optimal, which gives the lower bound

$$1 > \tau_{\text{end}} \geqslant \tau_{\text{end min}} = \begin{cases} 1-\frac{1}{8}(2+\varsigma)^2, & \varsigma \leqslant \varsigma_- \\ \frac{(2+\varsigma-\sqrt{\varsigma(4+\varsigma)})^2}{4}, & \varsigma > \varsigma_- \end{cases},$$

represented in Fig. 12.

We need to find the appropriate data ratio ν_{end} to meet exactly the deadline τ_{end}. Not to waste valuable time for computation, ν_{end} must satisfy $\tau_0 + \tau_+ = \tau_{\text{end}} + \tau_-$. Under the constraints of uniform arrivals and quadratic base and incremental algorithm, we shall solve

$$(1-\nu_{\text{end}})^2 + \nu_{\text{end}}^2 = \tau_{\text{end}} + \varsigma\nu_{\text{end}}.$$

Routine manipulation leads to the solution

$$\nu_{\text{end}} = \frac{2+\varsigma-\sqrt{(2+\varsigma)^2-8(1-\tau_{\text{end}})}}{4} \doteq \frac{1-\tau_{\text{end}}}{2+\varsigma},$$

where the approximation for slow arrivals $\varsigma \to \infty$ follows from the direct application of the results in the Appendix.

Note that the discriminant $(2+\varsigma)^2 - 8(1-\tau_{\text{end}})$ is nonnegative, thus the solution is well defined under the aforementioned range for τ_{end}. Since τ_{end} changes with the transition point ς_-, so does ν_{end}. As for the selection of the positive or negative options for the solution to the quadratic equation, observe that the positive solution increases while the negative one decreases when ς grows. Since ν_{end} is by definition smaller than one and we aim to minimize the data ratio processed incrementally to reduce data utility loss, we may safely select the negative solution as the valid one. As confirmed in Fig. 13, the discriminant is zero for values smaller than the threshold ς_- and grows past this point.

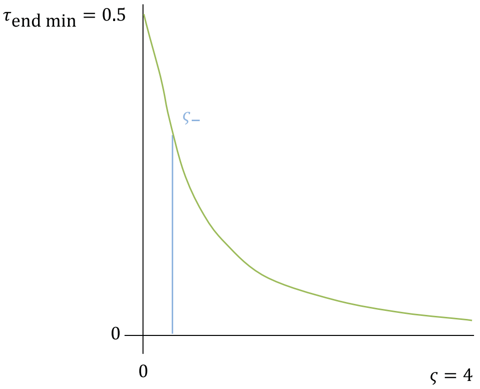


FIGURE 12. In order to establish a lower bound for the relative deadline τ_{end} we define it assuming the maximum relative time gain, concretely the relative time gain obtained using the optimal data ratio $\nu^\star$.

Accordingly, we may bound the incremental data ratio as

$$1 > \nu_{\text{end}} \geqslant \nu_{\text{end min}} = $$
$$= \begin{cases} \nu^* = \frac{2+\varsigma}{4}, & \varsigma \leqslant \varsigma_- \\ \frac{2+\varsigma-\sqrt{(2+\varsigma)^2-8(1-\tau_{\text{end min}})}}{4}, & \varsigma > \varsigma_- \end{cases}.$$

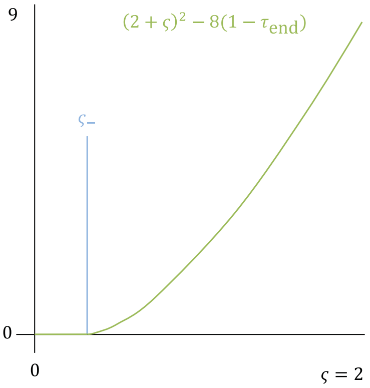


FIGURE 13. The discriminant $(2+\varsigma)^2 - 8(1-\tau_{\text{end min}})$ of the data ratio needed to end at the deadline is zero before reaching saturation point $\varsigma = \varsigma_-$ and grows past it.

D. NEAREST-NEIGHBOR METHOD

Two implementations of this algorithm are proposed. The first approach consists in simply assigning the incremental data to the centroids found by the base algorithm, in our case MDAV, precisely:

1. Wait until a portion $\bar{\nu}$ of the data is available and then proceed to run MDAV on the base data.
2. Once MDAV has finished and the remaining portion ν of the data is available, assign the incremental records to the nearest centroids of those found running MDAV.
3. Update the centroids of the enlarged cells.

In terms of time efficiency this algorithm should be fast. Its running time would be proportional to the product of the number of centroids $\bar{\nu}n/k$, and the number of incremental points νn, so that the relative time complexity would be $\tau_+ = \alpha\bar{\nu}\nu$ for some constant α depending on the implementation.

However, this algorithm is not meant to be used with large ν, as it would reduce the number of centroids found by the base algorithm creating bigger cells. This would lead to an increase of the SSE, described in §III-B4, resulting in unmanageable distortion. Expecting an improvement in the data utility of the anonymized release, the algorithm is modified introducing the concept of cell splitting. As commented on the lines above, bigger cells lead to worse performance in terms of distortion, which points to the creation of new cells when adding points as the key to reduce the distortion loss. In this contribution, MDAV is used to introduce new centroids ran on cells having more than $2k - 1$ samples. Two approaches are proposed. First, split cells while adding points:

1. Wait until a portion $\bar{\nu}$ of the data is available and then proceed to run MDAV on it.
2. Once MDAV has finished and the remaining portion of the data ν is available, assign the incremental records to the nearest centroids of those found running MDAV. While there still are incremental points to be assigned,
3. assign points to the nearest centroid of those found running MDAV, and
4. if the cell where the point was assigned reaches $2k$ records, then
5. split the cell into smaller cells with at least k records by running MDAV inside this cell.
6. Update the centroids of the modified cells.

And secondly, split cells at the end:

1. Wait until a portion $\bar{\nu}$ of the data is available and then proceed to run MDAV on it.
2. Once MDAV has finished and the remaining portion of the data ν is available, assign incremental records to nearest centroids of those found running MDAV.
3. Once all incremental records have been assigned, un MDAV locally on each cell that contains $2k$ samples or more, and aggregate the results.
4. Update the centroids of the modified cells.

Running MDAV on $2k$ points while adding them, segregates the cell into two of size k which will result in more available centroids when adding the remaining incremental points, however there might be points that have not been assigned to the nearest cell at the end since the closest centroid had not been created when the point was added. On the other hand, splitting cells bigger than $2k-1$ at the end avoids cell overlapping but increases the MDAV running time when splitting since its computational complexity depends on the number of samples that are microaggregated. Note that this last variation will behave as two-step MDAV when the base data is smaller than $3k - 1$ but bigger than $2k - 1$ leading to similar outcomes for high values of ν.

E. ON ASYMPTOTIC RUNNING TIMES

For the main algorithm for incremental microaggregation studied here, 2MDAV, which employs MDAV for both the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875949, IEEE Access

base records and the incremental portion, the theoretical analysis presented in this section already provides an accurate characterization of the relative time gain $\Delta\tau$ attained with respect to conventional microaggregation in a single MDAV run. We derive here an immediate consequence on asymptotic equivalence in the limit of the number n of records.

Denote the reference running time of traditional microaggregation as t_{ref}, that is, the time required for the traditional microaggregation algorithm to run on all data, starting as soon as all records are made available, in any suitable absolute time unit. Denote by t_{inc} the finishing time of the incremental procedure in two stages, in the same time units, also measured from the instant all data is available, thereby subtracting the head start. We have stressed that both subadditivity and the head start contribute to the reduction of t_{inc}. We may immediately express the absolute finishing time t_{inc} of the incremental strategy in terms of the relative time gain $\Delta\tau$, as

$$t_{\text{inc}} = (1-\Delta\tau)t_{\text{ref}} \quad \text{(equivalently, } \Delta\tau = (t_{\text{ref}} - t_{\text{inc}})\,/t_{\text{ref}}).$$

Under the assumption of uniform arrivals, recall that the relative head start $\tau_- = \varsigma\nu$ is proportional to the incremental data ratio ν, that is, the fraction of records processed in the second, incremental stage. In §III-B1, we interpreted the arrivals coefficient ς as the head start τ_- in the extreme case when $\nu = 1$, that is, ς represents the relative duration of the entire process of data arrivals or survey, always with respect to t_{ref}. For the optimal data ratio $\nu^\star$ obtained in §IV-B we obtain the corresponding optimal relative time gain $\Delta\tau^\star$, completely characterized by the arrivals coefficient ς. Hence, the optimal finishing time of the incremental algorithm is simply $t^\star_{\text{inc}} = (1-\Delta\tau^\star)t_{\text{ref}}$.

Suppose further that the microaggregation algorithm employed in the conventional, single-stage process, and in each of the two stages of the incremental approach is the same, and that its running time is quadratic in the number of records, or more specifically, proportional to n^2/k, where n is the number of records and k the anonymity parameter or cluster size. For notational compactness, we may do away with the proportionality constant simply by choosing an appropriate time unit, and simply write $t_{\text{ref}} = n^2/k$. Recall that this is an excellent characterization of the running time of MDAV, at least for $n \gg k$.

As customary, the statement $f(n) \sim g(n)$ denotes asymptotic equality between two positive sequences $f(n)$ and $g(n)$, often read as "$f(n)$ is of the order of $g(n)$", and defined by the condition

$$\lim_{n\to\infty} f(n)/g(n) = 1.$$

Recall that asymptotic equality implies asymptotic bounding from above and below, in the usual big theta notation $f(n) = \Theta(g(n))$, which is in turn more informative than asymptotic bounding merely from above, written in big-O notation as $f(n) = O(g(n))$.

Finally, in order to retrieve the asymptotic behavior of the absolute running time $t^\star_{\text{inc}}$ of the incremental algorithm with optimized scheduling, we consider an arbitrarily large number n of records, but a fixed absolute time t_{dat} allotted for the survey, during which all data records taken into consideration must arrive. Therefore,

$$t_{\text{ref}} \xrightarrow[n\to\infty]{} \infty \quad \text{implies} \quad \varsigma = \frac{t_{\text{dat}}}{t_{\text{ref}}} = \frac{k t_{\text{dat}}}{n^2} \xrightarrow[n\to\infty]{} 0,$$

which is precisely the case of extremely fast arrivals analyzed at the end of §IV-B (where $\varsigma \leqslant \varsigma_-$). Consequently,

$$\Delta\tau^\star = (2+\varsigma)^2/8 = \left(2 + \frac{k t_{\text{dat}}}{n^2}\right)^2/8 \xrightarrow[n\to\infty]{} 1/2,$$

leading to the asymptotic equivalence

$$t^\star_{\text{inc}} = \left(1 - \left(2 + \frac{k t_{\text{dat}}}{n^2}\right)^2/8\right) t_{\text{ref}} \sim \frac{t_{\text{ref}}}{2} = \frac{n^2}{2k} = \Theta(n^2).$$

This means that the complexities of the incremental and the traditional approach are both quadratic; the difference lies in the coefficient $t^\star_{\text{inc}}/t_{\text{ref}} = 1 - \Delta\tau^\star = 1/2$.

We must hasten to remind the reader that extremely fast arrivals, represented by the limiting case $\varsigma \to 0$, yield the worst optimal time gain performance $\Delta\tau^\star \to 1/2$. The overall time gain comes from the superadditivity of the quadratic algorithm with optimal data ratio $\nu^\star = 1/2$ and vanishing head start $\tau_- = \varsigma\nu^\star \to 0$. However, as argued also in §IV-B, extremely slow arrivals, in the limit of $\varsigma \to \infty$, yield arbitrarily large time gains, mathematically corresponding to $\Delta\tau^\star \to 1$. In other words, even in the worst case, 2MDAV will half the running time of the conventional approach, but we should expect far better performance in practical cases with slower arrivals. In the quantitative remarks of §VI-B, we show a numerical example where traditional microaggregation requires 2 hours, but our incremental strategy finishes roughly 2 minutes and 33 seconds after the survey is closed and all data is available. In that case, data arrives over a period of 10 hours, and thus, $\varsigma = 5$.

This addresses the incremental algorithm 2MDAV, for the special case of optimal data ratio for the fastest incremental running time. The case of a deadline, analyzed in §IV-C, has by definition a limited running time. As our contribution is necessarily limited, the case of the nearest-neighbor algorithm NN is left for future investigation. We already mentioned that $\tau_+ = \alpha\bar{\nu}\nu$ (quadratic), but without further inspection, we may merely claim that $t_{\text{inc}} = O(n^2)$, on account of the practical constraint $t_{\text{inc}} \leqslant t_{\text{ref}}$.

V. EXPERIMENTAL RESULTS

The essential aspect of our contribution relies in the empirical investigation of the formalism presented in the previous sections. In experimental section, we aim to confirm the algorithmic efficiency of our approach in terms of time gain, and the consequent price in distortion, in comparison with traditional methods. Bear in mind that the leading object of this work is to propose a general method rather than a specific algorithm, and the use of MDAV is merely illustrative in §IV.

A. EXPERIMENTAL SETUP

The algorithms for traditional, base, and incremental microaggregation were detailed in §IV. The values of the k-anonymity parameter considered, $k = 10$ and 100, are roughly representative of the lower and upper end of the values typically selected in the literature. As for the datasets considered, since this contribution aims to reduce the computational complexity of microaggregating large volumes of data, many small datasets commonly used in SDC literature, merely in the thousands of records or even under a thousand, were immediately discarded. Our experiments were targeted toward 50 000 records, a value deemed high enough to adequately illustrate the time gain that our incremental approach offers, and to retrieve reliable measurements of time and distortion. Recall from §IV that the relative time efficiency $\Delta\tau$ of the incremental algorithms tested does not depend explicitly on the number n of samples processed or on the k-anonymity parameter, although the absolute running times most certainly do.

We first synthesize 50 000 of 15-dimensional Gaussian data with independent, zero-mean and unit-variance components, taken as quasi-identifiers. Gaussian datasets typically represent a challenge for microaggregation algorithms, due to the profile of dispersion of the data. In order to assess the effect of the dimension m, three versions of the synthetized dataset are used, one containing only the first five dimensions, considered as quasi-identifiers, another one containing the first ten dimensions, and the last one containing all fifteen dimensions, also taken as quasi-identifiers.

Secondly, we incorporate a standardized dataset, known as "(Very) Large Census", previously documented and used in [14, 41][(a)]. This dataset was chosen to adhere to the de facto convention in the area, for fairer comparison and reference to previous work on microaggregation. It contains 149 642 records with 13 numerical attributes, regarded here as quasi-identifiers. In our experiments, we subsample 50 000 records of this dataset at random, consistently with the synthetic dataset. The corresponding three views with varying dimensionality contain the first three dimensions, the first six dimensions, and all thirteen dimensions. We adhere to the common practice of normalizing each column of the dataset for unit variance.

All of the experiments described here were implemented in Matlab, explicitly disabling any form of multithreading or parallelization for appropriate reporting of absolute and relative running times.

With the objective of making the graphs more readable, and also as a reminder, we offer in Table 2 a summary of all the variables used, most of them presented in §III-B and Fig. 6, in order to characterize data amounts, running times, and loss in data utility.

B. EXPERIMENTAL FINDINGS

As stated above, the number n of records and the k-anonymity parameter do not explicitly affect the relative time measurements τ, however we use datasets large enough to be able to provide accurate time measurements and two values of k different enough to demonstrate this effect. The absolute running time of the traditional algorithm used in this work will certainly vary depending on both n and k, as well as on the computer employed.

Not to delve into second-order implementation aspects, most of our experiments are in terms of running times relative to the traditional, single-stage use of MDAV. Absolute times can be easily retrieved from the normalizing values in Table 3.

1) MDAV as an Incremental Algorithm

The first set of plots, in Fig. 14, show the relative running time of the algorithm proposed in §IV-A, which consists in running MDAV in two consecutive steps. We confirmed that k affects both the traditional and the base algorithm in the same way since the normalized times τ do not depend explicitly on it. We measure $\tau_0 + \tau_+$ in order to verify the superadditivity property and the theoretical characterization of running times under the quadratic approximation for MDAV. These first plots do not yet incorporate the head start τ_-, which would further increase the relative time gain. Due to superadditivity alone, the incremental algorithm outperforms, as expected, traditional MDAV, confirming the bounds $\frac{1}{2} \leqslant \tau_0{+}\tau_+ \leqslant 1$. It can also be readily verify that the highest time reduction, considering only the superadditivity effect, is obtained, as predicted, for $\nu = 1/2$.

To further study the expected behavior of τ_+ in §IV-A, two more plots are presented for both the synthetic and the standardized datasets, using $k = 10$ and 100. It can be seen in Fig. 15 that the relative time gain does not depend on k and that the assumption of $\tau_+ = \nu^2$ is satisfied. Since we require $1 + \tau_- > \tau_0 + \tau_+$ to actually have a gain in time efficiency, we establish the boundary $\bar{\tau}_0 + \tau_- > \tau_+$ as a reference to illustrate the advantage of the incremental algorithm. According to the definition provided in §III-B2 we assume a head start coefficient $\varsigma = 1$, which results in $\tau_- = \nu$.

Next, we study the effects on distortion introduced by incremental microaggregation Fig. 16. Surprisingly, we observe a relative distortion reduction markedly decreasing with the number m of dimensions. A plausible explanation for this striking, extremely advantageous finding was given in §III-B5. Observe that the effect of augmenting the number of dimensions has less influence when processing "Large Census"; this owes to the fact that the quasi-identifiers in the database are statistically dependent and have a minor effective dimension that that of the synthetic data.

[(a)] The "(Very) Large Census" dataset is, strictly speaking, a synthetic dataset, generated with the procedure described in [30], but maintaining the same covariance matrix of the smaller dataset "Census", used in CASC [8], a well-known project in the SDC arena, within the FP5 European program. The latter dataset, however, contains real census data, obtained in 2000 via the Data Extraction System of the U.S. Census Bureau. We are most thankful to Jordi Nin, one of the authors of [41], for kindly providing us with the data file, back in 2015.

TABLE 2. Summary of Symbols

Symbol	Description (See also Fig. 6)
ν	Incremental data ratio, that is, the relative number of records processed incrementally.
n	Total number of records within the dataset.
n_0	Number of records processed by the base algorithm. Thus, $n_0 = \bar{\nu}n$ (where $\bar{\nu} = 1 - \nu$).
n_+	Number of records processed by the incremental algorithm, with $n_+ = \nu n$, $n_0 + n_+ = n$.
τ	Normalized running time of the traditional algorithm, taking the running time t of the traditional algorithm itself as reference, so that, trivially, $\tau = t/t = 1$.
τ_0	Running time t_0 of the base algorithm, normalized according to the running time t of the traditional algorithm, i.e., $\tau_0 = t_0/t$.
τ_+	Running time t_+ of the incremental algorithm, relative to the running time t of the traditional algorithm, i.e., $\tau_+ = t_+/t$.
τ_-	Relative head start normalized according to the running time t of the traditional algorithm. More precisely, $\tau_- = t_-/t$, where t_- defines the (absolute) head start of the base algorithm on the first portion of the data, before all the data is available.
$\Delta\tau$	Relative time gain of incremental microaggregation with respect to traditional microaggregation on the entire data. If the base algorithm were to finish before the available head start period has run out, that is, if $\tau_0 \leqslant \tau_-$, then $\Delta\tau = 1 - \tau_+$. Otherwise, if the head start is insufficiently generous, that is, $\tau_0 \geqslant \tau_-$, then, $\Delta\tau = \tau_- + 1 - \tau_0 - \tau_+$.
ς	Arrivals coefficient relating the head start time τ_- with the fraction ν of the data arrived over that period. Under the assumption of uniform arrivals, at least during the head start toward the end of the survey, $\tau_- = \varsigma\nu$. May be interpreted as the time for the entire data to arrive, extrapolating the pace at the end. As usual, this time is relative to the duration of the traditional microaggregation algorithm.
ν_-	Critical data ratio, for which $\tau_- = \tau_0$, that is, when the running time of the base algorithm matches the head start time.
$\nu^\star$	Optimal data ratio maximizing the time saving with respect to the traditional algorithm. Optimality implies that $\nu^\star \leqslant \nu_-$.
$\mathcal{D}$	Distortion loss introduced when running the traditional microaggregation algorithm on the entire dataset.
$\mathcal{D}_0$	Distortion loss introduced by the base algorithm on the first portion of the data.
$\mathcal{D}_+$	Distortion loss introduced by the incremental algorithm on the remainder of the data.
$\mathcal{D}_\text{T}$	Overall distortion loss in our incremental approach, satisfying $\mathcal{D}_\text{T} = \bar{\nu}\mathcal{D}_0 + \nu\mathcal{D}_+$.
δ	Relative distortion increment incurred by our incremental approach, with respect to the traditional microaggregation method. Specifically, $\delta_T = \mathcal{D}_\text{T}/\mathcal{D}$ denotes the distortion loss introduced in the two-step process, with relative increment $\delta = \delta_T - 1$.

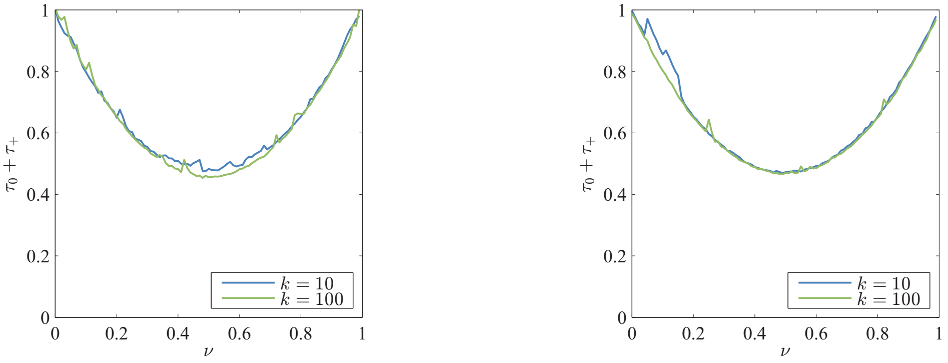
TABLE 3. Reference Times Required by the Traditional MDAV algorithm, without Applying our Incremental Approach

Dataset	Samples n	Dimension m	Anonymity k	Original Time $t_\text{traditional}$
"Large Census" (random subsampling)	50×10^3	3	10	16.82 sec
		6		24.22 sec
		13		40.53 sec
		3	100	1.71 sec
		6		2.49 sec
		13		4.09 sec
Gaussian (i.i.d.)	50×10^3	5	10	22.41 sec
		10		33.05 sec
		15		46.12 sec
		5	100	2.23 sec
		10		3.38 sec
		15		3.32 sec
	500×10^3	15	10	47 min 30 sec
	10^6	15	10	3 hrs 7 min

2) Nearest-Neighbor Algorithm

We initially implemented this algorithm simply adjoining each incremental point to its nearest centroid, of those found when running the base algorithm. Despite the good performance of this naïve approach in terms of computational cost, it is shown in Fig. 17 that it introduces a significant relative distortion in comparison with the traditional MDAV algorithm. As the incremental data ratio ν tends to 1, the number of centroids found, when running the base algorithm MDAV, decreases up to the point where maximum distortion loss is introduced. Precisely, when all the incremental records n_+ are assigned to the same centroid when $\lceil n_0^2/k \rceil = 1$ for certain values of k and n_0.

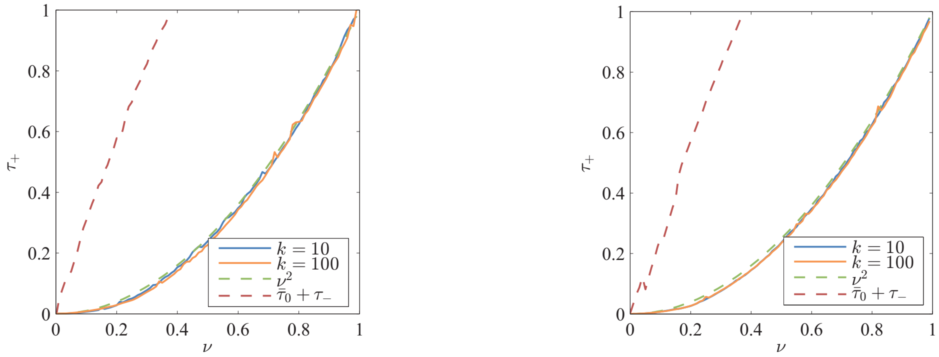
In order to address the unmanageable behavior in terms of relative distortion loss, we introduce the concept of cell splitting, as detailed in §IV-D. As previously shown, two variants have been proposed for cell splitting, split cells while adding incremental points, Split mid in the figures, and split cells once all the records have been added to the nearest centroid, Split end in the figures. Obviously, the



(a) Gaussian data with 50 000 samples and 15 dimensions.

(b) 50 000 samples randomly subsampled from “Large Census”.

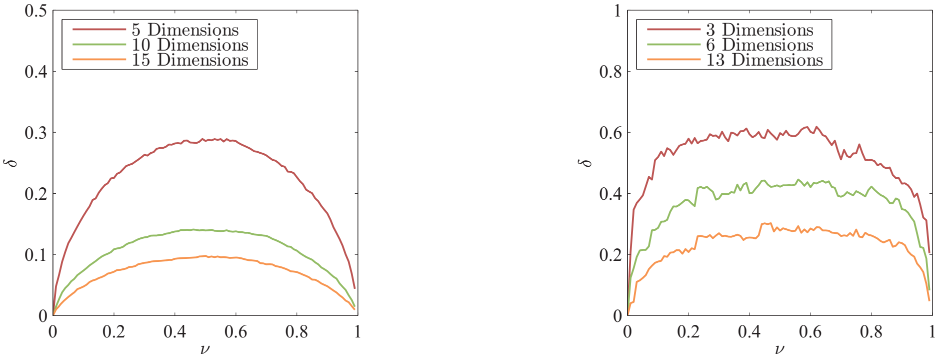
FIGURE 14. Relative time gain of MDAV as an incremental algorithm versus traditional MDAV due to superadditivity alone, without a head start.



(a) Gaussian data with 50 000 samples and 15 dimensions.

(b) 50 000 samples randomly subsampled from “Large Census”.

FIGURE 15. Evaluation of τ_+ as a function of ν using MDAV as an incremental algorithm in comparison with the expected outcome ν^2 and the boundary $\bar{\tau}_0 + \tau_- > \tau_+$.

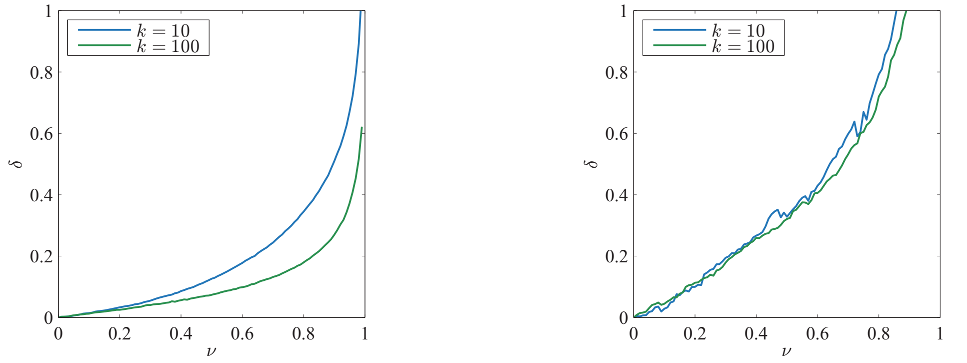


(a) Gaussian data with 50 000 samples and 15 dimensions.

(b) 50 000 samples randomly subsampled from “Large Census”.

FIGURE 16. Relative distortion loss of MDAV as an incremental algorithm versus traditional MDAV using $k = 10$. It can be seen that the relative distortion loss decreases as the number of dimensions increases.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2875949, IEEE Access



(a) Gaussian data with 50 000 samples and 15 dimensions.

(b) 50 000 samples randomly subsampled from “Large Census”.

FIGURE 17. Relative distortion loss of Find Nearest-Neighbors as an incremental algorithm versus traditional MDAV using $k = 10$ and 100. The performance of this approach in distortion is merely displayed as a comparison with the improved version, which uses cell splitting, but it is not a practical method.

huge improvement in terms of data utility shown in Fig. 18 introduces a slightly higher computational cost which is displayed in Fig. 20. Note that splitting cells at the end has a slightly better performance in terms of distortion for high values of $\nu \in [0, 1)$.

Another approach proposed to reduce distortion loss is the use of the inertial coefficient in §III-B6 when adding incremental points, but is shown in Fig. 19 that due to the high amount of records, the relative distortion loss does not effectively improve, and in some cases worsens, in comparison with only cell splitting. Therefore, the algorithm finally employed to measure time and distortion will be split cells without the inertial coefficient.

For the method previously analyzed, employing MDAV as incremental algorithm, the k-anonymity parameter does not affect the relative time gain since it affects both the traditional and the incremental approaches in the same way. In this implementation, increasing k reduces the running time of the traditional and base MDAV algorithms but also reduces the amount of distances that need to be computed by the incremental algorithm. Indeed, there will be less centroids supplied by the base algorithm, and less time required to perform cell splitting, which is computed using MDAV. This effect, despite the $1/k$ factor in the absolute running time of MDAV, makes the sum of relative time gains $\tau_0 + \tau_+$ slightly worse for bigger k-anonymity parameters, as shown in Fig. 20. Once more, thus far, these experiments merely assess the superadditive effect in isolation, without exploiting the head start.

Next, as previously detailed in §V-B1, Fig. 21 plots τ_+ in relation to the boundary $\bar{\tau}_0 + \tau_- > \tau_+$ and the theoretical $\tau_+ = \nu^2$ of MDAV as an incremental algorithm. As expected, due to MDAV having a quadratic running time, Find Nearest-Neighbors outperforms MDAV for high values of ν but has worse performance for small incremental data ratios. We can also note that splitting cells at the end presents better behavior for high values of ν. This, considering also that presents better performance in data utility, makes splitting cells at the end the better choice.

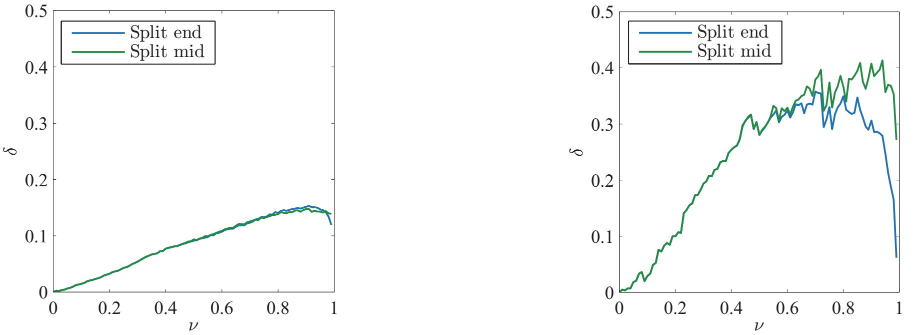
3) Incremental Algorithms Comparison

Because Find Nearest-Neighbors with cell splitting at the end (NN-SE) presents the best performance in running time and distortion among both variants proposed, it will be the candidate to compete against the two-step MDAV (2MDAV) approach. Since the base algorithm in both approaches is MDAV, τ_0 is identical. Hence, we only need to compare τ_+, the relative running time of the incremental algorithm, to measure computational time efficiency, and the relative quadratic distortion overhead δ to measure performance in terms of data utility. As shown in Fig. 22 and Fig. 23, and as could be expected from the fact that the running time of MDAV is quadratic, 2MDAV is faster than NN-SE for low values of ν. However, as ν increases, so does the running time of 2MDAV, but as ν^2, while NN-SE does not surpass the 10% in terms of relative running time τ_+. On the flip side, NN-SE introduces lower relative distortion loss for low values of ν.

In the first plot, Fig. 22, using 50 000 records of 15 dimensional Gaussian data and $k = 10$, from approximately $\nu \approx 30\%$ to $\nu \approx 50\%$ there is a range where NN-SE outperforms 2MDAV both in time and distortion performance, however this is a consistent behavior for different values of k and both datasets, obviously with changing ν ranges. Finally, for high values of ν NN-SE presents better relative time gain than 2MDAV at the cost of slightly higher distortion loss. Therefore, NN-SE would be the better choice for values of ν where NN-SE outperforms 2MDAV in time, in this case $\nu \approx 30\%$ and upwards. For values of ν lower below this point, we would select 2MDAV when pressed for running time, or NN-SE if distortion constitutes a higher concern.

VI. CONCLUSION

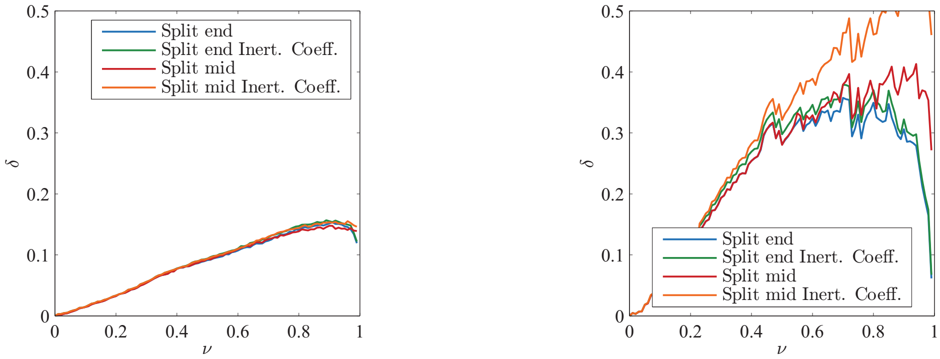
New capabilities in the areas of computation and storage enable the possibility of amassing vast quantities of potentially sensitive information. Many of today’s modern information systems incorporate processes that may be construed as electronic surveys, in the sense that they entail



(a) Gaussian data with 50 000 samples and 15 dimensions.

(b) 50 000 samples randomly subsampled from "Large Census".

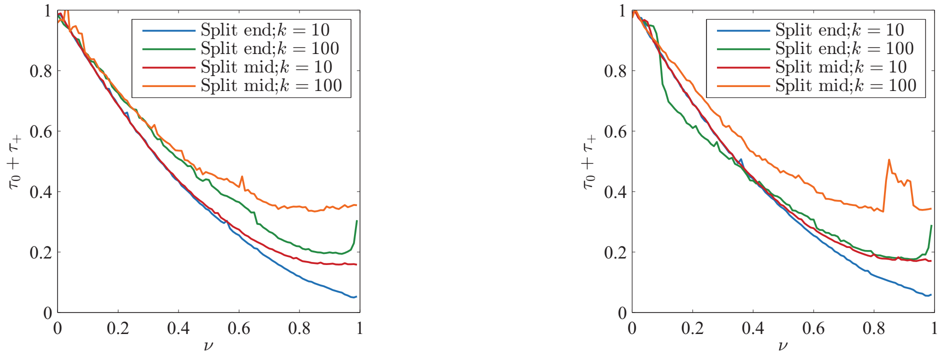
FIGURE 18. Relative distortion loss of Find Nearest-Neighbors, with cell splitting using MDAV, as an incremental algorithm versus traditional MDAV using $k = 10$.



(a) Gaussian data with 50 000 samples and 15 dimensions.

(b) 50 000 samples randomly subsampled from "Large Census".

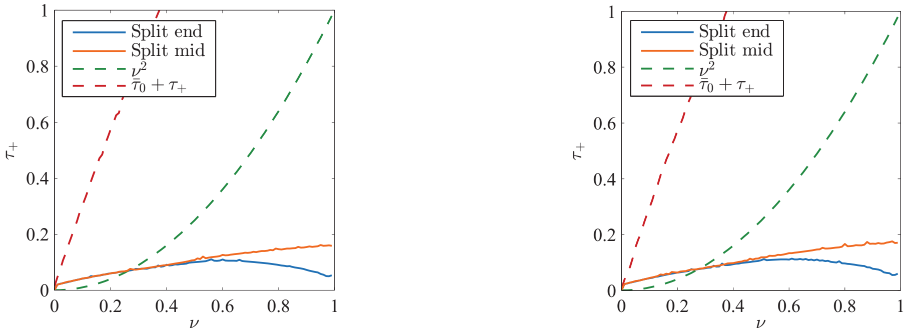
FIGURE 19. Relative distortion loss of Find Nearest-Neighbors, with cell splitting using MDAV and with and without the inertial coefficient being considered, as an incremental algorithm versus traditional MDAV using $k = 10$.



(a) Gaussian data with 50 000 samples and 15 dimensions.

(b) 50 000 samples randomly subsampled from "Large Census".

FIGURE 20. Relative time gain of Find Nearest-Neighbors, with cell splitting using MDAV, as an incremental algorithm versus traditional MDAV without a head start. This plot compares the time efficiency of the two approaches proposed, split cells while adding records and split cells at the end.



(a) Gaussian data with 50 000 samples and 15 dimensions.

(b) 50 000 samples randomly subsampled from “Large Census”.

FIGURE 21. Evaluation of τ_+ as a function of ν using Find Nearest-Neighbors, with cell splitting using MDAV, as an incremental algorithm in comparison with the expected MDAV outcome ν^2 and the boundary $\bar{\tau}_0 + \tau_- > \tau_+$.

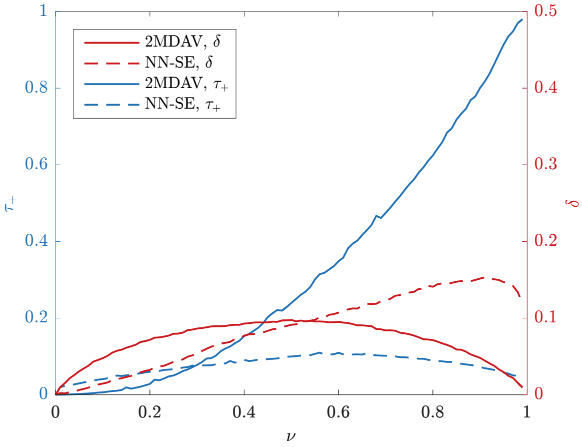


FIGURE 22. Comparison of our two incremental microaggregation algorithms for the Gaussian dataset, in terms of running time and distortion, in a chart with double Y axis. Specifically, we compare MDAV as an incremental algorithm (2MDAV) with Find Nearest-Neighbor with cell splitting at the end (NN-SE), on a dataset consisting of 50 000 records of 15-dimensional Gaussian data, and for an anonymity parameter $k = 10$. The horizontal axis shows the incremental data ratio ν, that is, the number of records processed incrementally in a second step, as a fraction of the total number of records to be anonymized. The double vertical axis measures the relative running times τ_+ of the incremental algorithms on the left, in blue, and the relative distortion loss δ introduced on the right, in red, always with respect to the traditional MDAV algorithm running in a single pass on all data.

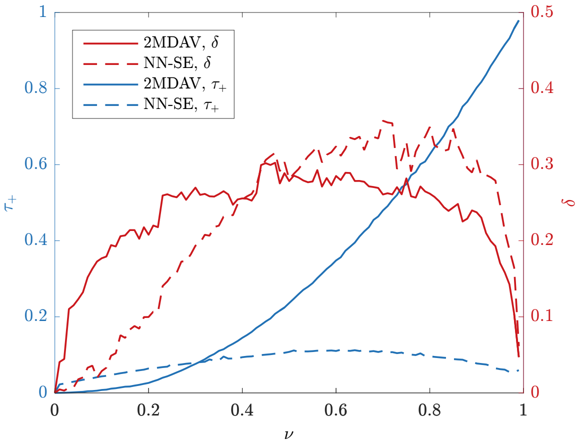


FIGURE 23. Comparison of our two incremental microaggregation algorithms for the “Large Census” dataset, in terms of running time and distortion, in a chart with double Y axis. Specifically, we compare MDAV as an incremental algorithm (2MDAV) with Find Nearest-Neighbor with cell splitting at the end (NN-SE), on a dataset consisting of 50 000 randomly subsampled records of “Large Census”, and for an anonymity parameter $k = 10$. The horizontal axis shows the incremental data ratio ν, that is, the number of records processed incrementally in a second step, as a fraction of the total number of records to be anonymized. The double vertical axis measures the relative running times τ_+ of the incremental algorithms on the left, in blue, and the relative distortion loss δ introduced on the right, in red, always with respect to the traditional MDAV algorithm running in a single pass on all data.

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 such technologies, the inclusion of rich quantities of sensitive data poses privacy risks that cannot simply remain overlooked. This raises the need for statistical disclosure control, specifically through the anonymization of datasets, in a manner protective of the data utility contained in the original copy, in order to allow demographic studies of diverse nature.

SDC through k-anonymous microaggregation continues to be a baseline procedure in confidential attribute release with high-utility preservation. This is applicable to systems where data accuracy is critical, or where utility may be sacrificed for stricter privacy criteria, but said criteria are but built upon the underlying substratum of k-anonymity, such as l-diversity, t-closeness (or differential privacy for microdata release implemented through t-closeness).

A. QUALITATIVE REMARKS

Prepartitioning is a well-established strategy to reduce computation in high-utility, superadditive algorithms. Traditionally, prepartitioning takes into consideration demographic similarity to preserve data utility. For numerical data, (square) distance in the Euclidian space is typically employed, whilst ontological distances may be preferred for categorical variables.

In this work, we take a first step towards extending this spatial strategy along the time domain. Regardless of the specific anonymization algorithm, inessential to our methodology, this work is concerned with sophisticated mechanisms striving to attain the optimal privacy-utility trade-off, which generally require superlinear computational complexity in the number of records of potentially large datasets. We demonstrate, both theoretically and empirically, that complex, high-utility privacy-preserving algorithms would highly benefit from our mathematically optimized scheduling framework.

The method for database anonymization introduced in this work exploits the fact that in electronic surveys, the data is often available over a substantial period of time. Our work considers breaking down the anonymization process in two algorithmic steps on two portions of the data. The first algorithm, called base algorithm, would start before all the data is available, say one hour before finishing the data collection process. Subsequently, the second anonymization process, called incremental algorithm, would start once all the data has been collected. On the one hand, certainly starting earlier helps. On the other, the superadditivity of the anonymization algorithm benefits from the proposed partition.

This double-edged strategy of "divide and conquer" is analyzed in thorough mathematical detail. In particular, we characterize the relationship between the data partition point and the overall time gain due to the head start and the superadditivity of the algorithms involved. We find the mathematically optimal partition point for the fastest possible anonymization. Finally yet importantly, we analyze the problem of finishing before a given deadline.

The direct applicability of this work was summarized in Fig. 4. But the real applicability of the ideas proposed here reach beyond the specific algorithms analyzed, which should be seen as merely illustrative of a general methodology that could be employed on a variety of superlinear computationally demanding processes.

The synergic advantage provided by the head start and the superadditivity of the algorithms analyzed translates into remarkable time gains, allowing in some cases the anonymization of large-scale datasets in minutes instead of hours. Nevertheless, this improved performance in running time also comes at a price in data utility, just as stricter privacy would. The compromise between computation and distortion is also the object of this work, this time from an empirical perspective.

An additional factor playing to our advantage stems from the distortion phenomenon in high-dimensional microaggregation informally described in §III-B5, and later rigorously verified in our experiments. As the number m of statistically independent quasi-identifiers grows, the distortion overhead δ tends to diminish. This means that for large databases, not only in terms of the number of records, but also in terms of the number of quasi-identifiers, our proposal is particularly beneficial.

B. QUANTITATIVE REMARKS

Having summarized the main conclusions extracted from our investigation in a conceptual manner, we proceed to remark in greater detail on a few numerical results on the example that served as motivation in the introductory section, illustrated in Fig. 2. These results are presented in this section as additional conclusions of a quantitative nature, in terms of time gain and distortion, rather than as a discussion subsection. We believe these numbers validate and speak highly of the ideas introduced in this paper.

Recall that the introductory example shown in Fig. 2 employed an incremental data ratio of approximately $\nu \approx 10\%$, merely for motivation purposes. Particularizing for a 15-dimensional Gaussian dataset with 50 000 records, the relative quadratic distortion increase using MDAV as an incremental algorithm is found to be $\delta \approx 5\%$. On the other hand, Find Nearest-Neighbors presents a lower increase in relative distortion $\delta \approx 2\%$ at the cost of a slightly higher running time.

We revisit the example of Fig. 2, assuming a 10-hour electronic survey with 2MDAV versus traditional MDAV, for a head start of 1 hour. A slow, approximately uniform rate of data arrivals over 10 hours, relative to a duration of 2 hours for the traditional algorithm, corresponds to an arrivals coefficient $\varsigma = 5$. Rather than employing the arbitrary data ratio $\nu \approx 10\%$ originally chosen, we consider the optimal data ratio $\nu^\star \approx 14.6\%$ introduced in §III-B2 and derived in §IV-B. The result, nothing short of impressive, is reported in Fig. 24. The base algorithm, with a head start of 1 hour and 27 minutes, finishes exactly at the end of the head start, when the deadline for the electronic survey expires. The incremental algorithm, operating now on only a fraction $\nu^\star$ of the data, ends in just 153 seconds once all the data has been received, well before the 2 hours required by the traditional method. Evidently, due to the nonlinearity of running times, the time required is smaller than the fraction of data processed, advantage that our mathematical optimization certainly exploits. Furthermore, this enormous time gain, from hours to minutes, comes at the expense of a relative quadratic distortion increment of only $\delta \approx 6\%$ for the dataset aforementioned.

The idea behind all this is somewhat counterintuitive because running times are quadratic, not additive. With $\nu^\star \approx 14.6\%$, the majority of the data, precisely $1-\nu^\star \approx 85.4\%$, is microaggregated first during the head start of 1 hour and 27 minutes, while the rest of the data comes in. And the small incremental portion $\nu^\star$ requires a relative time $\nu^{\star 2} \approx 2.1\%$ quadratically small to run in about 2 and a half minutes instead of 2 hours. Because the majority of the data was microaggregated traditionally, the impact in distortion is of only $\delta \approx 6\%$, more than reasonable for such impressive time gain. The exact $\nu^\star$ is chosen via mathematical optimization, from the data arrival coefficient ς, and the running time for microaggregation of the base data may or may not exceed the head start.

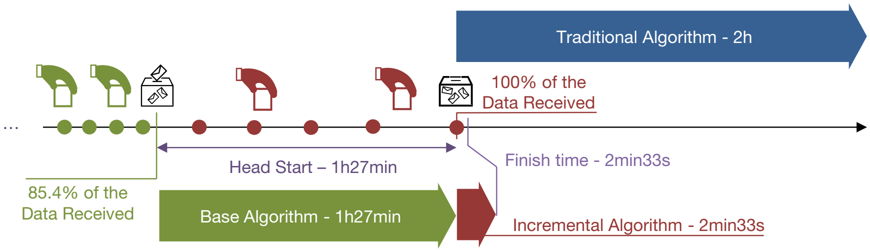

Traditional Algorithm - 2h
100% of the Data Received
Head Start – 1h27min
Finish time - 2min33s
85.4% of the Data Received
Base Algorithm - 1h27min
Incremental Algorithm - 2min33s

FIGURE 24. Example of incremental microaggregation using 2MDAV where the optimal data ratio $\nu^* \approx 14.6\%$ is used. This allows the entire process to finish in just 2 min 33

Regarding the price in distortion of time-based partitioning of the data, we verified in §V-B1 the very intuitive fact that lower incremental data ratios yield lower distortion. We consider introducing a time constraint of one hour on the running example presented of Fig. 2. In that regard, we use the deadline data ratio ν_{end}, formulated in §III-B3 and derived in §IV-C, to adjust it to the time constraint. Striving only to comply with the deadline but not for optimal time gain, data utility should be better preserved. Indeed, while the optimal approach yielded a relative increase in quadratic distortion of $\delta \approx 6\%$, the deadline method presents a lower negative impact, of $\delta \approx 3\%$, while halving the waiting time required by traditional microaggregation. The running example using the deadline approach is illustrated in Fig. 25.

Both incremental algorithms are suitable alternatives to traditional microaggregation. On the one hand, MDAV as an incremental algorithm presents better relative running time for low values of ν and less degradation in relative distortion loss for high values of the incremental data ratio. On the other hand, there is a certain middle range of ν values where Find Nearest-Neighbors as an incremental algorithm consistently provides better performance in both relative running time and relative data utility loss. We illustrate this comparison in Fig. 26.

C. FINAL CONSIDERATIONS

We may thus conclude that the incremental method proposed in this work offers a most valuable alternative to traditional microaggregation when dealing with large-scale electronic surveys where data is available over an extended period of time, but anonymization must be conducted under reasonable time constraints. To better preserve utility, low values of the data ratio ν should be employed, whenever possible. The choice of an incremental algorithmic procedure, be it 2MDAV or Find Nearest Neighbor, depends on the constraints on running time and data utility.

We view the main contribution of this work as the general methodology of optimized incremental anonymization, above the proposal and evaluation of specific algorithms. Striving for applicability, our design criteria encompass both maximum time gain and meeting a given deadline, but resort to mathematical formalisms for optimal performance, of which Fig. 11 constitutes a brief, partial recapitulation. Despite the significant length and detail of this document, we must acknowledge that it merely constitutes a preliminary analysis of a particularly complex problem with a number of intricate aspects requiring further study. Most assuredly, additional experimentation with real large-scale survey data should prove extremely useful, particularly with regard to data arrivals and distortion impact, in future work refining our research results or oriented toward development. Our method is certainly open to future investigation along a number of research avenues, including but not limited to additional partition steps, alternative incremental microaggregation algorithms, non-quadratic asymptotic time complexities, a theoretical rather than experimental analysis of the distortion incurred, and its application to related problems of time-consuming data processing.

APPENDIX. BRIEF NOTE ON APPROXIMATIONS BEYOND THE ACCURACY OF VANISHING ABSOLUTE AND RELATIVE ERRORS

Let f and g be real-valued functions of a common real-valued argument x. For certain ratios to be well defined, we shall require that $f(x),\ g(x) \neq 0$ for all x. We consider the limiting trends of these functions as the argument x approaches a given value, or infinity. As mentioned earlier, we write $\lim_{x\to x_0} f(x) = l$ for some x_0, possibly infinity, more compactly as $f \to l$. We denote the approximation in absolute error by

$$f \simeq g \overset{\text{def}}{\Leftrightarrow} f - g \to 0 \;\Leftrightarrow\; g \simeq f \Leftrightarrow \\ \Leftrightarrow\; g - f \to 0 \;\Leftrightarrow\; f = g + o(1),$$

and the approximation in relative error by

$$f \sim g \overset{\text{def}}{\Leftrightarrow} f/g \to 1 \;\Leftrightarrow\; g \sim f \;\Leftrightarrow\; g/f \to 1 \;\Leftrightarrow \\ \Leftrightarrow\; \frac{f-g}{g} \to 0 \;\Leftrightarrow\; f = g + o(g).$$

Clearly, either type of approximation induces an equivalence relation, satisfying reflexivity, symmetry, and transitivity.

For instance, as $x \to \infty$, certainly $x + 1 \sim x$ holds, but it is not true that $x + 1 \simeq x$. The function x does succeed in approximating $x + 1/x$ both absolutely and relatively. If $f, g \to \infty$, then $f \simeq g$ would entail $f \sim g$. But when $f, g \to 0$, the statement $f \simeq g$ holds trivially, and only $f \sim g$ remains informative. For example, $2/x \simeq 1/x$, but the relative approximation fails to hold. Consider now

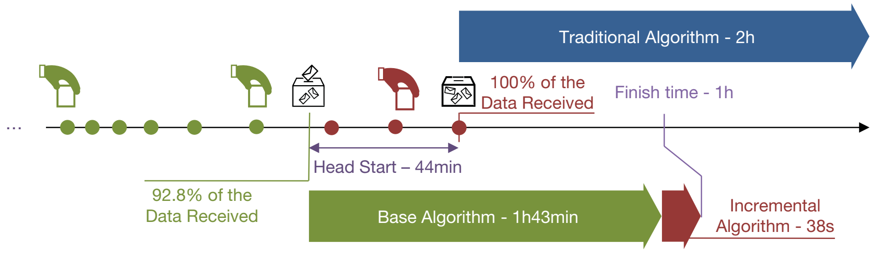
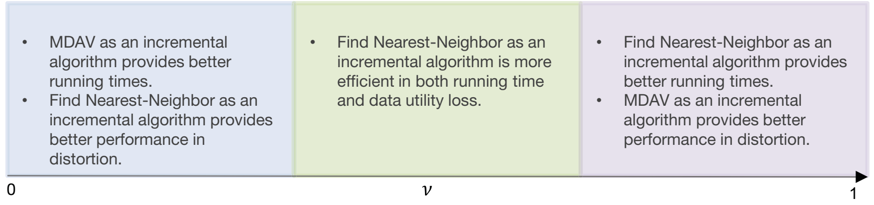

Traditional Algorithm - 2h
100% of the Data Received
Finish time - 1h
...
Head Start – 44min
92.8% of the Data Received
Base Algorithm - 1h43min
Incremental Algorithm - 38s

FIGURE 25. Example of incremental microaggregation using 2MDAV, where the data ratio is adjusted to finish right at the one-hour deadline. In this case, $\nu_{\text{end}} = 7.2\%$, below


- MDAV as an incremental algorithm provides better running times.
- Find Nearest-Neighbor as an incremental algorithm provides better performance in distortion.

- Find Nearest-Neighbor as an incremental algorithm is more efficient in both running time and data utility loss.

- Find Nearest-Neighbor as an incremental algorithm provides better running times.
- MDAV as an incremental algorithm provides better performance in distortion.

FIGURE 26. Comparison of the performance in terms of running time and distortion loss introduced by the proposed incremental algorithms, as a function of the values of ν.

the inverses of the first example, $f(x) = 1/(x+1)$ and $g(x) = 1/x$. Both the absolute and the relative approximation criteria are satisfied. However, it is not true that $1/f \simeq 1/g$. When $f, g \to 0$, the criterion $1/f - 1/g \to 0$ is stronger than $f \sim g$, the latter equivalent to $1/f \sim 1/g$. In the following, we propose a criterion stronger than the simultaneous occurrence of both the absolute and relative approximations. Further, when $f, g \to 0$, the new criterion will offer better accuracy than the relative approximation.

This approximation criterion has been specifically devised for this work, but should prove useful in other arenas. We shall say that f is a strong approximation for g, or vice versa, when

$$f \mathrel{\dot{\simeq}} g \overset{\text{def}}{\Leftrightarrow} \begin{cases} f - g \to 0 \\ 1/f - 1/g \to 0 \end{cases} \overset{\text{def}}{\Leftrightarrow} \begin{cases} f \simeq g \\ 1/f \simeq 1/g \end{cases}.$$

The first defining component makes this approximation trivially stronger than the absolute version. The second condition, $1/f - 1/g \to 0$, is equivalent to $f = g + o(fg)$. In the special case when $f \geqslant g$, the conditions on the inverses would be implied by $f = g + o(g^2)$. The second statement in the next proposition reaches beyond our observation that when $f, g \to 0$, the condition $1/f \simeq 1/g$ is stronger than $f \sim g$, as the example with $1/(x+1)$ and $1/x$ illustrated.

Proposition 1. *For any f, g such that $f(x), g(x) \neq 0$ for all x, (i) if $\liminf |f| > 0$ or $\liminf |g| > 0$, then $f \simeq g$ implies $f \sim g$, (ii) if $\limsup |f| < \infty$ or $\limsup |g| < \infty$, then $1/f \simeq 1/g$ implies $f \sim g$.*

Proof: (i) By symmetry, it suffices to assume either condition on the limit inferior. Suppose for instance that $\liminf |g| = \mathrm{l} > 0$. The statement follows directly from the epsilon characterization of the limit inferior ($|g|$ is eventually above $l/2$), and the limit implicit in the absolute approximation ($|f - g|$ is eventually below $\epsilon l/2$). Bound $|f - g|/|g|$ (eventually below ϵ). (ii) is simply an application of (i) to the inverse functions. ■

The next proposition holds yet more generally; unlike the previous one, no assumptions are made on the inferior or superior limits. However, both approximations defining the strong criterion are employed simultaneously.

Proposition 2. *For any f, g such that $f(x), g(x) \neq 0$ for all x, the strong approximation $f \mathrel{\dot{\simeq}} g$ implies both the absolute approximation $f \simeq g$ and the relative counterpart $f \sim g$.*

Proof: The proof employs the epsilon characterization of the two limits in the definition of the strong approximation, writing $1/f - 1/g = \frac{g-f}{fg}$ ($|f - g|$ is eventually below ϵ, and $\frac{|f-g|}{|f||g|}$ is eventually below $\epsilon/(1+\epsilon)$). Consider the two cases $|g(x)| \geqslant 1$ and $|g(x)| \leqslant 1$ (for which $|f(x)| < 1 + \epsilon$) separately, bounding $|f - g|/|g|$ (eventually below ϵ). ■

The following statements will prove particularly useful to the work developed here, which assumes quadratic running times for the microaggregation algorithms. The difference between an arithmetic mean and a geometric mean arises in the solution

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

of the quadratic equation $ax^2 + bx + c = 0$, since

$$-b = \tfrac{1}{2}((-b + 2\sqrt{ac}) + (-b - 2\sqrt{ac})),$$
$$\sqrt{b^2 - 4ac} = \sqrt{(-b + 2\sqrt{ac})(-b - 2\sqrt{ac})}.$$

More specifically, our next proposition analyzes the difference

$$\phi \overset{\text{def}}{=} \frac{f+g}{2} - \sqrt{fg},$$

as the functions involved increase, and the quality of the approximation

$$\psi \overset{\text{def}}{=} \frac{\frac{1}{8}(f-g)^2}{\frac{f+g}{2}}.$$

Notice that $\phi \geqslant 0$ by the arithmetic-mean geometric-mean inequality on f and g, by Young's inequality on $\sqrt{f}$ and $\sqrt{g}$, or simply from the fact[b] that $\phi = \frac{1}{2}(\sqrt{f} - \sqrt{g})^2$. Finally, we should point out that the very last statement on $o(\psi^2)$ is stronger than the limit result on the difference of the inverses, equivalent to $\phi = \psi + o(\phi\psi)$, since $\psi \leqslant \phi$.

Proposition 3. *For any $f, g > 0$, suppose that $f, g \to \infty$ and*

$$\frac{(f-g)^2}{f+g} \to 0.$$

We may then conclude that the nonnegative difference ϕ between the arithmetic and the geometric mean vanishes in the limit, that is,

$$\phi \overset{\text{def}}{=} \frac{f+g}{2} - \sqrt{fg} \to 0,$$

and that this difference can be strongly approximated according to

$$\phi \overset{\text{def}}{=} \frac{f+g}{2} - \sqrt{fg} \;\doteq\; \psi \overset{\text{def}}{=} \frac{\frac{1}{8}(f-g)^2}{\frac{f+g}{2}},$$

for which the right-hand side constitutes a lower bound, i.e., $\phi \geqslant \psi$. Further,

$$\phi = \psi + o(\psi^2).$$

Proof: The proof is mainly routine algebraic manipulation, once we realize that

$$\frac{f+g}{2} - \sqrt{fg} = \frac{\left(\frac{f-g}{2}\right)^2}{\frac{f+g}{2} + \sqrt{fg}}.$$

The bounds follow from the arithmetic mean-geometric mean inequality, but also from the explicit computation of the differences involved. ■

The following corollary readily applies the previous proposition to $f(x) = x+a$ and $g(x) = x+b$, the functional forms encountered in the main text.

Corollary 4. *For any $a, b \in \mathbb{R}$, as $x \to \infty$,*

$$x + \tfrac{a+b}{2} - \sqrt{(x+a)(x+b)} \to 0,$$

$$x + \tfrac{a+b}{2} - \sqrt{(x+a)(x+b)} \doteq \frac{\frac{(a-b)^2}{8}}{x + \frac{a+b}{2}},$$

$$x + \tfrac{a+b}{2} - \sqrt{(x+a)(x+b)} = \frac{\frac{(a-b)^2}{8}}{x + \frac{a+b}{2}} + o(1/x^2).$$

The last formula asserts that the error in the approximation to the arithmetic mean-geometric mean difference in terms of the simpler rational fraction vanishes faster than x^{-2} as $x \to \infty$, that is, faster than the asymptotic square of the functions approximated. Taking $b = -a$ gives our last immediate corollary.

Corollary 5. *For any $a \in \mathbb{R}$, as $x \to \infty$,*

$$x - \sqrt{x^2 - a^2} \to 0,$$

$$x - \sqrt{x^2 - a^2} \doteq \frac{a^2}{2x},$$

$$x - \sqrt{x^2 - a^2} = \frac{a^2}{2x} + o(1/x^2).$$

ACKNOWLEDGMENT

The authors gratefully acknowledge the patient, thorough suggestions of the anonymous reviewers, which greatly assisted in improving the readability and contents of this document. 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.

REFERENCES

[1] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, "Achieving anonymity via clustering," ACM Trans. Alg., vol. 6, Jun. 2010.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Anonymizing tables," in Proc. Int. Conf. Database Theory (ICDT), vol. 3363, Edinburgh, UK, Jan. 2005, pp. 246–258.

[3] P. Bonizzoni, G. della Vedova, and R. Dondi, "Anonymizing binary and small tables is hard to approximate," J. Comb. Optim., vol. 22, no. 1, pp. 97–119, 2009.

[4] R. Brand, "Microdata protection through noise addition," in Inference control in statistical databases: From theory to practice, ser. Lect. Notes Comput. Sci. (LNCS), J. Domingo-Ferrer, Ed. Berlin/Heidelberg, Germany: Springer-Verlag, 2002, vol. 2316, pp. 97–116.

[5] R. Bredereck, A. Nichterlein, and R. Niedermeier, "Pattern-guided k-anonymity," Algorithms, vol. 6, pp. 678–701, Jun. 2013.

[6] R. Bredereck, A. Nichterlein, R. Niedermeier, and G. Philip, "The effect of homogeneity on the computational complexity of combinatorial data anonymization," Data Min., Knowl. Discov., vol. 28, pp. 65–91, Oct. 2012.

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

[8] "Project CASC: Computational aspects of statistical confidentiality," EU, Framework Program 5, ref. IST-2000-25069, 2003. [Online]. Available: http://neon.vb.cbs.nl/casc

[9] 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.

[10] Cisco VNI Group, "The zettabyte era: Trends and analysis," Cisco Syst., White Paper, May 2015. [Online]. Available: www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

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

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

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

[14] 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.

[15] 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.

[16] 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.

[b] Incidentally, the form $\phi = \frac{1}{2}(\sqrt{f} - \sqrt{g})^2$ is the Bregman divergence associated with Fenchel's inequality, of which Young's is a special case. Precisely, it is the divergence induced by the self-conjugate function $t \mapsto \frac{1}{2}t^2$, between the arguments $\sqrt{f}$ and $\sqrt{g}$.

[17] J. Domingo-Ferrer and V. Torra, “On the connections between statistical disclosure control for microdata and some artificial intelligence tools,” Inform. Sci., vol. 151, pp. 153–170, May 2003.

[18] ——, “Ordinal, continuous and heterogeneous k-anonymity through microaggregation,” Data Min., Knowl. Discov., vol. 11, no. 2, pp. 195–212, 2005.

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

[20] “General Data Protection Regulation (GDPR),” Regul. (EU) 2016/679, Eur. Parliam., Apr. 2016. [Online]. Available: http://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04

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

[22] P. Hansen, B. Jaumard, and N. Mladenovic, “Minimum sum of squares clustering in a low dimensional space,” J. Classif., vol. 15, pp. 37–55, 1998.

[23] D. Ikarashi, R. Kikuchi, K. Chida, and K. Takahashi, “k-Anonymous microdata release via post randomisation method,” in Proc. Int. Workshop Secur. (IWSEC), vol. 9241, Nara, Japan, Aug. 2015, pp. 225–241.

[24] 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.

[25] Y. Jung, H. Park, D.-Z. Du, and B. L. Drake, “A decision criterion for the optimal number of clusters in hierarchical clustering,” J. Global Optim., vol. 25, pp. 91–111, 2003.

[26] 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.

[27] 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.

[28] 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.

[29] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. Berkeley Symp. Math. Stat., Prob., vol. I (Stat.), Berkeley, CA, Jun. 1965, pp. 281–297, 1965–1966 Symp., 1967 Proc.

[30] J. M. Mateo-Sanz, A. Martínez-Ballesté, and J. Domingo-Ferrer, “Fast generation of accurate synthetic microdata,” in Proc. Int. Conf. Priv. Stat. Databases (PSD), Barcelona, Spain, Jun. 2004, pp. 298–306.

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

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

[33] 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.

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

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

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

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

[38] A. Rodríguez-Hoyos, J. Estrada-Jiménez, D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, “Does k-anonymous microaggregation affect machine-learned macrotrends?” IEEE Access, vol. 6, pp. 28 258–28 277, May 2018. [Online]. Available: http://doi.org/10.1109/ACCESS.2018.2834858

[39] P. Samarati, “Protecting respondents’ identities in microdata release,” IEEE Trans. Knowl., Data Eng., vol. 13, no. 6, pp. 1010–1027, 2001.

[40] A. Solanas and A. Martínez-Ballesté, “V-MDAV: Multivariate microaggregation with variable group size,” in Proc. Int. Conf. Comput. Stat. (CompStat), Rome, Italy, 2006.

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

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

[43] 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.

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

[45] L. Sweeney, “k-Anonymity: A model for protecting privacy,” Int. J. Uncertain., Fuzz., Knowl.-Based Syst., vol. 10, no. 5, pp. 557–570, 2002.

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

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

CÉSAR HERNÁNDEZ-BAIGORRI received the B.S. degree in telecommunications engineering from the Universitat Politècnica de Catalunya (UPC), in Barcelona, Spain, in 2016. During his senior year, he collaborated with the Information Security Group in the development of novel optimization methods for anonymous microaggregation applicable to electronic surveys. Currently, he is IT network manager in Technology 2 Client, a private IT company based in Barcelona. Although the main scope of his current work consists in optimizing and maintaining infrastructure, he is transitioning towards the proposal and deployment of innovative solutions to the challenges of today’s businesses.

JORDI FORNÉ received the M.S. and the Ph.D. degrees in telecommunications engineering from the Universitat Politècnica de Catalunya (UPC), in Barcelona, Spain, in 1992 and 1997, respectively. He is currently associate professor in the Telecommunications Engineering School of Barcelona at UPC. From 2007 to 2012, he was coordinator of the Ph.D. program in telematic engineering and director of the master’s research program in the same subject. His research interests span a number of subfields within information security and privacy.

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