p-Probabilistic k-Anonymous Microaggregation for the Anonymization of Surveys with Uncertain...

Article · December 2016

DOI: 10.1016/j.ins.2016.12.002

CITATIONS
0

READS
9

4 authors, including:

Jordi Forné
Universitat Politècnica de Catalunya
129 PUBLICATIONS **718** CITATIONS
SEE PROFILE

Miguel Soriano
Universitat Politècnica de Catalunya
131 PUBLICATIONS **693** CITATIONS
SEE PROFILE

Jordi Puiggalí
Scytl Secure Electronic Voting SA
16 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Preprint Submitted to

Elsevier Journal

2016-10

p-Probabilistic k-Anonymous Microaggregation for the Anonymization of Surveys with Uncertain Participation

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

[1]Department of Telematics Engineering, Universitat Politècnica de Catalunya (UPC), E-08034 Barcelona, Spain
[2]Centre Tecnològic de Telecomunicacions de Catalunya (CTTC), E-08860 Castelldefels, Barcelona, Spain
[3]Scytl Secure Electronic Voting, E-08006 Barcelona, Spain

Article Info

Article History:
Submitted Feb. 18, 2016
Revised Oct. 26, 2016

Keywords:
k-Anonymity
Microaggregation
Probabilistic anonymity
Surveys

Abstract

We develop a probabilistic variant of k-anonymous microaggregation, which we term p-probabilistic, resorting to a statistical model of respondent participation in order to aggregate quasi-identifiers in such a manner that k-anonymity is concordantly enforced with a parametric probabilistic guarantee. Succinctly, owing the possibility that some respondents may not finally participate, sufficiently larger cells are created striving to satisfy k-anonymity with probability at least p. The microaggregation function is designed before the respondents submit their confidential data. More precisely, a specification of the function is sent to them, which they may verify and apply to their quasi-identifying demographic variables, prior to submitting the microaggregated data along with the confidential attributes to an authorized repository.

We propose a number of metrics to assess the performance of our probabilistic approach in terms of anonymity and distortion, which we proceed to investigate theoretically in depth, and empirically with synthetic and standardized data. We stress that in addition to constituting a functional extension of traditional microaggregation, thereby broadening its applicability to the anonymization of statistical databases in a wide variety of contexts, the relaxation of trust assumptions is arguably expected to have a considerable impact on user acceptance, and ultimately, on data utility, through mere availability.

I. Introduction

Big data is better data. Abundant quantities of precise, frequent information may provide qualitatively superior insight into challenges and opportunities that may otherwise remain unseen amidst the intricacies of any sufficiently complex system. Peter Norvig, director of research at Google, in his brilliant paper "The unreasonable effectiveness of data" [15], insightfully acknowledges that a massive wealth of data may radically improve the effectiveness of a machine-learning algorithm, to the point of turning a hopeless computer model into an expert system exceeding human performance. Certainly, the availability of large quantities of data to modern information systems of ever-increasing power and sophistication may offer extraordinary potential.

But for every auspicious opportunity that modern information technologies offer, a daunting challenge to protect the privacy of their users arises, as the inclusion of rich quantities of sensitive data poses privacy risks that cannot simply remain overlooked. Personal information, explicitly submitted or implicitly inferable from observed behavior, poses evident privacy risks, especially when combined across several information services, and when enriched with metadata indicating size, location, time, frequency, and other contextual information.

Fortunately, privacy-enhancing technologies have today reached far beyond the more traditional approaches of encryption and granular access-control policies envisioned in days of yore. When the intended recipients of sensitive information are themselves untrusted, modern privacy mechanisms may resort to advanced data-perturbative strategies

* Corresponding author.
sites.google.com/site/davidrebollomonedero, david.rebollo@entel.upc.edu.

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

in order to attain a desired measure of privacy, at the expense of an acceptable loss in data utility. Some of these leading-edge privacy technologies draw upon mathematical formalisms originally intended for information-theoretic and operational data compression, as well as convex optimization, in an attempt to systematically measure and attain the optimal trade-off between a potentially sizeable gain in privacy and a calculated loss in data utility [29, 28, 37].

While researchers strive to reduce the information loss of our privacy-enhancing algorithms by introducing complex mathematics and ingenious heuristics —and those improvements are certainly most welcome— we often neglect the crucial aspect of data availability. By relaxing the trust assumptions imposed on the users of a system, the consequent increase in users willing to provide additional data may very well represent a far greater gain in utility than that from algorithmic improvements alone. Not entirely unlike the way massive amounts of data may far outweigh the efforts to programmatically improve a machine-learning algorithm, pointed out earlier.

A. Contribution and Organization

The technical contents of this paper fall within the area of statistical disclosure control and k-anonymous microaggregation, the fundamentals of which are succinctly reviewed in §II. Although we opted to explain the object of our work without further ado, readers less familiar with the subject, or more immediately interested in the illustrative medical example provided there, may prefer to skip over to the referred section.

Bearing in mind the profound impact of the trust model on the utility of privacy-enhancing mechanisms, in this work, we formulate a functional extension of the method of k-anonymous microaggregation for anonymization of datasets containing confidential data linked to quasi-identifying demographic variables that may be exploited to reidentify the individuals involved. Traditionally, users are required to provide their information in full to a party responsible for such microaggregation, thereby fully entrusting it with the data-anonymization process. In our functional extension, we conceive a variant of k-anonymous microaggregation that enables users to anonymize their own quasi-identifiers, thus considerably relaxing the trust assumptions demanded from the participating respondents.

Concisely, we develop a probabilistic variant, which we term p-probabilistic k-anonymous microaggregation, resorting to a statistical model of respondent participation in order to aggregate quasi-identifiers in such a manner that k-anonymity is concordantly enforced with a parametric probabilistic guarantee. The microaggregation function is designed before the respondents submit their confidential data. More precisely, a specification of the function is sent to them, which they may verify and apply to their quasi-identifying demographic variables, prior to submitting the microaggregated data along with the confidential attributes to an authorized repository. In addition to constituting a functional extension of traditional microaggregation, thereby broadening its applicability to the anonymization of statistical databases in a wide variety of contexts, the relaxation of trust assumptions is arguably expected to have a considerable impact on user acceptance, and ultimately, on data utility, through mere availability.

The remainder of this paper is organized as follows. A succinct introduction to k-anonymous microaggregation is offered in §II. Next, §III reviews the state of art in anonymity metrics and microaggregation algorithms for statistical disclosure control. We present a conceptual formulation of p-probabilistic k-anonymous microaggregation in §IV, formalized in §V. We proceed with a theoretical analysis in §VI, while §VII numerically illustrates the main results.

II. Background on k-Anonymous Microaggregation with an Example of Application to the Medical Sciences

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

It was famously shown in [45] that 87% of the population in the United States might be unequivocally identified solely on the basis of the triple consisting of their date of birth, gender and 5-digit ZIP code, according to 1990 census data. This is in spite of the fact that in that year, the U.S. had a population of over 248 million. This notorious fact illustrates the discriminative potential of the simultaneous combination of a few demographic attributes, which, considered individually, would hardly pose a real anonymity risk. Ultimately, this simple observation means that the mere elimination of identifiers such as first and last name, or social security number (SSN), is grossly insufficient when it comes to effectively protecting the anonymity of the participants of published statistical studies containing confidential data linked to demographic information.

Statistical disclosure control (SDC) concerns the postprocessing of the demographic portion of the statistical results of surveys containing sensitive personal information, in order to effectively safeguard the anonymity of the participating respondents. In the SDC terminology, a *microdata set* is a database table whose records carry information concerning individual respondents, either people or companies. This database commonly contains a set of attributes that may be classified into identifiers, quasi-identifiers and confidential attributes. Firstly, *identifiers* allow the unequivocal identification of individuals. This is the case of full names, SSNs or medical record numbers, which would be removed before the publication of the microdata set, in order to preserve the anonymity of its respondents. Secondly, *quasi-identifiers*, also called *key attributes*, are those attributes that, in combination, may be linked with external, usually publicly available information to *reidentify* the respondents to whom the records in the microdata set refer. Examples include

age, address, gender, job, and physical features such as height and weight. Finally, the dataset contains *confidential attributes* with sensitive information on the respondent, such as salary, political affiliation, religion, and health condition. The classification of attributes as key or confidential may ultimately rely on the specific application and the privacy requirements the microdata set is intended for.

A primary field of application, in which confidential information linkable to demographic variables may provide enormous data utility and at the same time require special privacy measures, is that of medical sciences. We illustrate some fundamental aspects of SDC with an example on Hashimoto's thyroiditis, one of the first diseases to be recognized as an autoimmune disorder, in which the immune system causes antibodies to mistakenly attack normal tissue.

Hashimoto's thyroiditis, or chronic lymphocytic thyroiditis, is an autoimmune disease in which the thyroid gland is attacked by a variety of cell- and antibody-mediated immune processes, causing primary hypothyroidism. Thyroperoxidase (TPO) is an enzyme involved in thyroid hormone synthesis. The determination of anti-TPO antibodies levels in blood is the most sensitive test for detecting autoimmune thyroid disease, and detectable concentrations of anti-TPO antibodies are observed in roughly 90% of patients with Hashimoto's. Another common indicator of Hashimoto's is an increased level of thyroid-stimulating hormone (TSH), as the pituitary tries to compensate for decreased thyroxine. Because of the association between hypothyroidism and weight gain, another variable that might alert of the presence of the disease is the *body-mass index* (BMI). The BMI serves as a rough estimate of the amount of body fat in an individual, a metric routinely employed by the WHO as the standard means for estimating adiposity and recording obesity statistics since the early 1980s, now commonplace to categorize a person as underweight, normal, or overweight.

Owing to the relevance of anti-TPO and TSH levels in the diagnosis of thyroid disorders, and the potential impact on the patient's health and general wellbeing, they undoubtedly constitute prime examples of confidential attributes. Because BMI is immediately derived from readily observable quantities, it may be rightfully construed as a quasi-identifier, just as weight or height would. Gender and age also play an important role in the diagnosis of Hashimoto's thyroiditis, being far more common in women, with onset typically at 30 to 50 years of age.

Intuitively, the perturbation of numerical or categorical quasi-identifiers enables us to preserve privacy to a certain extent, at the cost of losing some of the *data utility*, in the sense of accuracy with respect to the unperturbed version. *k-Anonymity* is the requirement that each tuple of key-attribute values be identically shared by at least k records in the dataset. This may be achieved through the *microaggregation* approach illustrated by the synthetic example depicted in Fig. 1. Rather than making the original table available, we publish a k-anonymous version containing aggregated

Identifiers	Quasi-Identifiers			Confidential Attributes	
Name	Sex	Age	BMI	Anti-TPO	TSH
Alice Adams	♀	32	29.3	++	8.01
Bob Brown	♂	34	26.9	−	2.56
Chloe Carter	♀	33	32.1	+++	14.41
Dave Diaz	♂	43	25.7	++	11.32
Eve Ellis	♀	47	21.4	+	0.94
Frank Fisher	♂	45	22.0	−	3.29

Microaggregated Quasi-Identifiers			Confidential Attributes		
Sex	Age	BMI	Anti-TPO	TSH	
♀	33	29.4	++	8.01	k-Anonymized Records
♀	33	29.4	−	2.56	
♀	33	29.4	+++	14.41	
♂	45	23.0	++	11.32	
♂	45	23.0	+	0.94	
♂	45	23.0	−	3.29	

Fig. 1. Example of k-anonymous microaggregation of published data with $k = 3$, showing indicators of Hashimoto's thyroiditis (anti-TPO antibodies and TSH levels) as confidential attributes, in relation to demographic variables (gender, age and body-mass index) as quasi-identifiers.

records, in the sense that all quasi-identifying values within each group are replaced by a common representative tuple. As a result, a record cannot be unambiguously linked to the corresponding record in any external sources assigning identifiers to quasi-identifiers. In principle, this prevents a privacy attacker from ascertaining the identity of an individual for a given record in the microaggregated database, which contains confidential information.

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

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

Our study of the probabilistic generalization of k-anonymous microaggregation for large-scale demographic surveys, in which respondent participation is uncertain, is, to the best of our knowledge, entirely novel. Rather unsurprisingly, the term "probabilistic k-anonymity" has been used in the literature [42] although with a different meaning from that

intended here. Precisely, the cited work slightly relaxes the definition of the term k-anonymity to require only that the probability of reidentification be $1/k$. In any event, we deem it relevant to briefly review the state of the art on traditional, deterministic microaggregation, with regard to its use and limitations as a measurement of the degree of privacy attained, and the methods and algorithms to construct k-anonymous aggregations with reduced distortion.

A. Shortcomings of k-Anonymity as Privacy Criterion

We mentioned in the introductory section that a specific piece of data on a particular group of respondents is said to satisfy the k-anonymity requirement if the origin of any of its components cannot be ascertained beyond a subgroup of at least k individuals. We also said that the concept of k-anonymity, originally proposed by the SDC community [38, 45], is a widely popular privacy criterion, partly due to its mathematical and algorithmic tractability.

Despite the popularity of k-anonymity as a measure of privacy, it is not without shortcomings [13, 31, 36]. Indeed, while this criterion prevents identity disclosure, it may fail against the full disclosure of the confidential attribute. Strictly speaking, the criterion is defined exclusively on the basis of the quasi-identifiers, completely disregarding the specific values that the confidential attributes take on, consequently neglecting the possibility of groups where the confidential values are accidentally repeated. Even if such repetition could be discarded, several deficiencies remain.

More specifically, the *homogeneity* or *similarity attack* exploits the possibility that values of a confidential attribute within a group may turn out to be quantitatively or qualitatively similar. From a more general, probabilistic perspective, the *skewness attack* exploits the difference between the prior distribution of confidential data in the entire population, and the posterior conditional distribution of the confidential data within a group given the observed, perturbed key attributes in the table. Further, the *background-knowledge attack* resorts to any deterministic or statistical background knowledge or side information available to the attacker, in addition to the published records.

B. Use of k-Anonymity in Database Publication and Other Domains

The original formulation of k-anonymity as a privacy criterion, based on generalization and suppression of key attributes, was modified into the microaggregation-based approach already commented on, in [6, 9, 12, 10]. Both formulations may be regarded as special cases of one utilizing an abstract distortion measure between the unperturbed and the perturbed data, possibly taking on values in rather different alphabets. Rapidly since its conception, this anonymity criterion has gained widespread adoption in the SDC literature, in spite of the shortcomings already described.

Indeed, the application of the k-anonymity criterion and of the microaggregation methodology goes beyond the publication of databases. Further recent investigation has also been conducted on the scenario of online data collection, where a data miner queries a set of users, each of whom responds with a piece of data. For example, [47] proposes a cryptographic method that allows users to submit their data anonymously. Namely, the authors present a protocol that eliminates the restriction of using unidentified communication channels and allows users to include identifying information in their responses. Still in this context, [11] presents a set of protocols and methods aimed to protect the privacy of users that query Web search engines. Lastly, in the scenario of streaming data, [3] proposes a cluster-based approach that k-anonymizes data streams and, in addition, guarantees the freshness of the anonymized data by imposing a restriction on the delay.

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

The vulnerabilities of k-anonymity already explained motivated the proposal of enhanced privacy criteria, some of which we proceed to sketch briefly, along with modifications in algorithms based on these criteria. A restriction of k-anonymity called *p-sensitive k-anonymity* was presented in [46, 44]. In addition to the k-anonymity requirement, it is required that there be at least p different values for each confidential attribute within the group of records sharing the same tuple of perturbed key attributes. Clearly, large values of p may lead to huge data utility loss. A slight generalization called *l-diversity* [22, 16] was defined with the same purpose of enhancing k-anonymity. The difference with respect to p-sensitivity is that group of records must contain at least l "well-represented" values for each confidential attribute. Depending on the definition of well-represented, l-diversity can reduce to p-sensitive k-anonymity or be more restrictive. We would like to stress that neither of these enhancements succeeds in completely removing the vulnerability of k-anonymity against skewness attacks. Furthermore, both are still susceptible to similarity attacks, in the sense that while confidential attribute values within a cluster of aggregated records might be p-sensitive or l-diverse, they might also very well be semantically similar for the practical purposes of the attacker. Consider for example confidential attributes indicating similar salaries, political affiliations or diseases.

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

As argued in [13], to the extent to which the within-group distribution of confidential attributes resembles the distribution of those attributes for the entire dataset, skewness attacks will be thwarted. In addition, since the within-

group distribution of confidential attributes mimics the distribution of those attributes over the entire dataset, no semantic similarity can occur within a group that does not occur in the entire dataset. The main limitation of the original t-closeness work [19] is that no general computational procedure was specified, with the exceptions of its ready applicability to the Incognito algorithm [18], and the very recent microaggregation procedure proposed in [43].

An information-theoretic privacy criterion, inspired by t-closeness, was proposed in [30, 31]. In the latter work, privacy risk is defined as the conditional KL divergence between the aforementioned posterior and prior distributions, and shown to be equivalent to the mutual information between the confidential attributes and the perturbed quasi-identifiers. This criterion is also tightly related to the concept of *equivocation* introduced by Shannon in 1949 [39], namely the conditional entropy of a private message given an observed cryptogram. Conceptually, the privacy risk thus defined may be regarded as an averaged version of the t-closeness requirement, over all aggregated groups. It is important to notice as well that this *average privacy risk*, in spite of its convenient mathematical tractability, as any criterion based on averages, may not be adequate in all applications [14]. A related albeit more conservative criterion, named δ-disclosure, is proposed in [2], and measures the maximum discrepancy between the prior and the posterior distributions. These three criteria, namely average privacy risk, t-closeness, and δ-disclosure, may all be formally defined in terms of averages and maxima of KL divergences, as explained in [36].

D. Algorithms for k-Anonymous Microaggregation

A number of algorithms for microaggregation have been developed, with the goal of minimizing the perturbation of the key attributes with accordance to a variety of distortion measures, while meeting a given k-anonymity constraint. As multivariate microaggregation is known to be NP-hard [27], several heuristic methods have been proposed, which can be categorized into fixed-size and variable-size methods, according to whether all aggregated groups but one have exactly k elements. The maximum distance (MD) algorithm [9] and its less computationally demanding variation, the maximum distance to average vector (MDAV) algorithm [12, 8], are fixed-size algorithms that perform particularly well in terms of the distortion they introduce, for many data distributions. Popular variable-size algorithms include the μ-Approx [10], the minimum spanning tree (MST) [17], the variable MDAV (VMDAV) [40] and the two fixed reference points (TFRP) [4] algorithms. Efforts to circumvent the complexity of multivariate microaggregation exploit projections onto one dimension, but are reported to yield a much higher disclosure risk [26].

Research on microaggregation algorithms has continued recently. In particular, an approach recommends creating clusters of k records according to their densities [20]. Still in the case of perturbative algorithms, [23] contemplates the partition of the original dataset into several projections such that each projection satisfies the k-anonymity requirement, with the help of genetic algorithms. A well-known alternative to perturbative algorithms is the generation of synthetic data that preserves some pre-established statistics of the original dataset. A combination of perturbed and synthetic data is followed by [7].

More recently, an analysis of theoretical optimality in k-anonymous microaggregation [32] extends the necessary (not sufficient) optimality conditions that gave rise to the celebrated Lloyd-Max algorithm [21, 24], a celebrated quantization method for lossy data compression, also known as the k-means method in the areas of statistics and computer science. The properties of theoretical optimality and the excellent behavior of the Lloyd-Max algorithm in practice motivated the conception of the probability-constrained Lloyd (PCL) algorithm [34, 35, 32], which additionally incorporates a variation of the Levenberg-Marquardt algorithm [25], in order to adjust cell sizes. PCL is capable of outperforming even the popular MDAV in terms of distortion, typically by a reduction in MSE of roughly 10–30% [35]. Unfortunately, the distortion improvement offered by PCL comes at the expense of increased mathematical sophistication, which translates into a significantly costlier implementation and a substantially longer running time.

IV. Conceptual Formulation of p-Probabilistic k-Anonymous Microaggregation

Next, we address the generalization of the notion of k-anonymous microaggregation to account for statistical models of respondent participation, in which anonymity is concordantly enforced with a probabilistic guarantee.

A. The Notion of p-Probabilistic k-Anonymity

The fundamental notion is that along with the quasi-identifiers, we are in possession of a statistical model of respondent participation, and are given a desired probability with which k-anonymity is to be enforced. We must stress that in this type of probabilistic microaggregation, the partition of the quasi-identifiers into microcells must be carried out before knowing which respondents will in fact participate in the corresponding survey, on the basis of a statistical participation model. Traditional, deterministic microaggregation would correspond to the special case in which probabilities of participation are all equal to 1, or from a conceptually equivalent perspective, microaggregation is carried out after knowing which respondents are finally included in the published dataset.

In this type of microaggregation, k-anonymity might conceivably be violated for some cells depending on the actual respondent participation and the accuracy of the underlying statistical model, but the micropartition is carefully designed to make such event highly unlikely. However, the larger size of the microcells required in this approach will undoubtedly translate into higher distortion. The notion of p-probabilistic k-anonymous microaggregation is conceptually represented in Fig. 2.

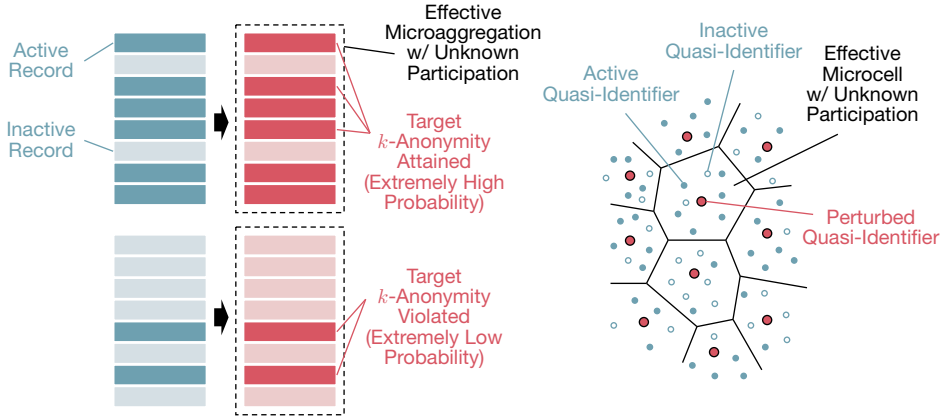


Fig. 2. Fundamental notion of p-probabilistic k-anonymous microaggregation. Based on a statistical model of respondent participation, microaggregation is carried out with larger microcells, and k-anonymity is concordantly enforced with a probabilistic guarantee.

B. Modes of Operation of Probabilistic Microdata Anonymization, in Practice

In the context of real applicability, we would like to propose several trust models, with varying modes of operation in the anonymization of microdata sets. Their adequacy will of course depend on the specific application at hand, but all three should represent valid alternatives, and present exciting business opportunities.

- Remotely trusted anonymization. In this case, all respondents trust a common party responsible for collecting and microaggregating the data set. Raw data is simply sent over and traditional k-anonymous microaggregation remains a perfectly suitable approach. Among all the approaches discussed here, this imposes the most stringent constraints on trust, but offers the lowest distortion and therefore the highest data utility.
- *Locally trusted* anonymization. On the basis of potential quasi-identifier values and a statistical model of respondent participation, a common party computes a microaggregation that can only guarantee k-anonymity with a desired target probability. The partition function, that is, the assignment of possible quasi-identifiers to cells or centroids is made available to all respondents. The respondent may verify that the microaggregation is sound, makes the actual assignment, and merely sends back the resulting cell index or centroid, in lieu of his original, unperturbed quasi-identifier, accompanied with the confidential data. The response is transmitted through a confidential, anonymized channel [37, 33]. It is assumed that the assignment specified by the microaggregation function is carried out by open-source software running locally. Hence, a reasonable degree of trust of the respondent on his local computer and on the network is required. To accommodate the improbability of k-anonymity violations, cells are effectively larger than those in traditional microaggregation, with a consequent impact on distortion.
- *Untrusted* anonymization. The operation is similar to that of locally trusted anonymization, with the caveat that the respondent does not wish to place his trust on the client software or hardware, and is only willing to implement the assignment specified by the microaggregation function manually. This means that the microaggregation function must be extremely simple. Concretely, we limit it to the specification of ranges and categories for each individual quasi-identifier. The respondent is asked to select the matching ranges and categories for the demographic data of interest, along with the confidential information required by the application at hand. The extreme simplicity in the specification of the microaggregation function would allow its inclusion in surveys in physical paper. In the case of numerical data in the d-dimensional Euclidean space, this translates into the specification of d-dimensional orthotopes (hyperrectangles), significantly constraining the type of microaggregation allowed. It is reasonable to expect that the lack of trust requirements of this mode of operation come at a considerable cost in terms of distortion.

Fig. 3 represents these three modes of operation.

In this context, the object of this work is the modification of existing microaggregation algorithms, primarily MDAV, to enforce p-probabilistic k-anonymity in the special case of locally trusting respondents. This case, defined above, does not impose any constraints on the shapes of the cells constructed, merely on their size. The design of microaggregation algorithms to enforce p-probabilistic k-anonymity in the special case of untrusted respondents is left for future investigation. As explained above, the severe constraints imposed on the shapes of the microcells, and not just their sizes, preclude the strategy of directly adapting an existing microaggregation algorithm.

We must stress that the gradual relaxation of trust requirements in these approaches will naturally come at an increasingly high price in distortion. We already commented in the introductory section the nearly ubiquitous existence of an inherent trade-off between privacy and utility in modern information systems. It is important to consider that under some circumstances, even with privacy-enhancing technologies designed for the most conservative degree of data sensitivity, and with severe impact on data utility, the absence of such technology or its replacement by more lenient ones may constitute a far worse alternative. The reason is that such lack of adequate protection would discourage the use of the underlying information system by those rightfully wary of their privacy. In other words, respondents that do

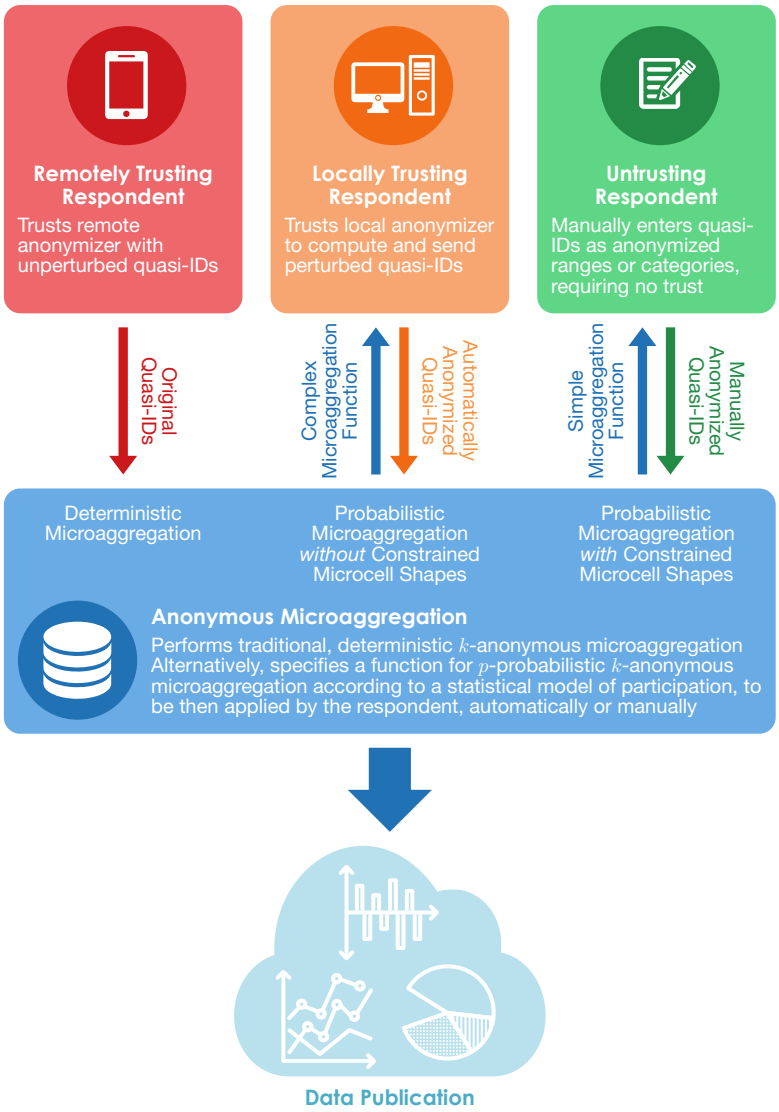


Fig. 3. Modes of operation for microdata anonymization, in practice, with varying degrees of trust. Remotely trusting respondents merely require traditional, deterministic microaggregation. Locally trusted respondents require probabilistic microaggregation, but without any constraints on the cell shape, which may be effectively reduced to traditional methods. Untrusting respondents demand probabilistic microaggregation with constraints on the cell shapes, and therefore attain the strongest level of privacy, but at the expense of the highest loss in data utility.

not accept the trust model assumed or imposed may choose not share their data, and in that case, the data utility would be, by any reasonable measure, zero.

V. Formal Problem Statement

This section formally presents the proposed p-probabilistic extension of k-anonymous microaggregation conceptually introduced earlier, for the mode of operation involving locally trusting respondents.

A. Mathematical Preliminaries and Quantization Model for Traditional k-Anonymous Microaggregation

The work presented here builds upon a formulation of the problem of k-anonymous microaggregation in [32, 35], which formally regards microaggregation as a quantization problem with constraints on the cell probabilities. Throughout this paper, the measurable space in which a *random variable* (r.v.) takes on values will be called an *alphabet*. We shall follow the convention of using uppercase letters for r.v.'s, lowercase letters for particular values they take on, and script letters for sets of such values. *Probability mass functions* (PMFs) are denoted by lowercase p, and *cumulative mass functions* (CMFs), by a capital P. An upright P denotes a general probability measure. For example, in this notation, the probability that a discrete r.v. X takes on the value x in an alphabet $\mathscr{X}$ is $p_X(x) = \mathrm{P}\{X = x\}$. Similarly, $P_X(x) = \mathrm{P}\{X \leqslant x\}$. The *expectation* operator is denoted by E. Expectation can model the special case of averages over a finite set of data points $\{x_1, \ldots, x_n\}$, simply by defining an r.v. X uniformly distributed over this set, so that, for instance, $\mathrm{E}X = \frac{1}{n}\sum_{j=1}^{n} x_j$.

We shall limit our analysis to the special case of *numerical data*, that is, we shall assume that the quasi-identifiers to be aggregated are represented by n points $x_1, \ldots, x_n$ in the Euclidean space $\mathbb{R}^m$ of dimension m, indexed by the corresponding record j. For convenience, we define an r.v. J representing the record index, uniformly distributed on

the set of indices $\{1,\dots,n\}$. Note that J may also be regarded as the identity of the respondent. In addition, we introduce an r.v. X representing the quasi-identifiers, whose alphabet consists in the set of m-dimensional points $(x_j)_{j=1,\dots,n}$, formally definable as a function $X = x_J$ of J. The r.v. X models tuples of quasi-identifiers of a table $x_1,\dots,x_n$ of n records. The notation in terms of r.v.'s will enable us to write averages more compactly as expectations, for instance $\frac{1}{n}\sum_{j=1}^{n}\varphi(x_j) = \mathrm{E}\,\varphi(X)$, for any functional $\varphi\colon \mathbb{R}^m \to \mathbb{R}$ of the tuple of quasi-identifiers.

The k-anonymous microaggregation algorithm will partition the set of records into microcells of size at least k. In the most traditional form of k-anonymous microaggregation, this partition only takes into account the values of the quasi-identifiers, as the published table will only effectively perturb the quasi-identifiers and keep the confidential attributes intact. The resulting microcells will be labeled with a quantization or microcell index c. An important subtlety is that the microaggregation process must be formally construed as a quantization function $c(j)$ of the record index j, rather than on the quasi-identifier x. The reason is that even though $c(j)$ also induces a partition on the set of quasi-identifiers and confidential attributes, one cannot discard the possibility that some tuples x might be repeated, and that those repeated values might be assigned to different microcells. Although this could be technically handled with probabilistic microcell assignments, it is simpler and completely general to define a (deterministic) quantization function on the record indices. In our more compact representation with r.v.'s, we define $C = c(J)$, with finite alphabet $\{1,\dots,|\mathscr{C}|\}$. The k-anonymity constraint is contemplated by imposing a constraint on cell sizes or, more generally, on the probabilities of the quantization indices $p_C(c) \geqslant k/n$.

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

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

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

The entire microaggregation process, which transforms the record index J into the perturbed quasi-identifier $\hat{X}$, can be represented as the composition of two functions, namely the microcell assignment $c(j)$ and the centroid assignment $\hat{x}(q)$, as depicted in Fig. 4. We have mentioned that the problem of microaggregation, may be formally understood as

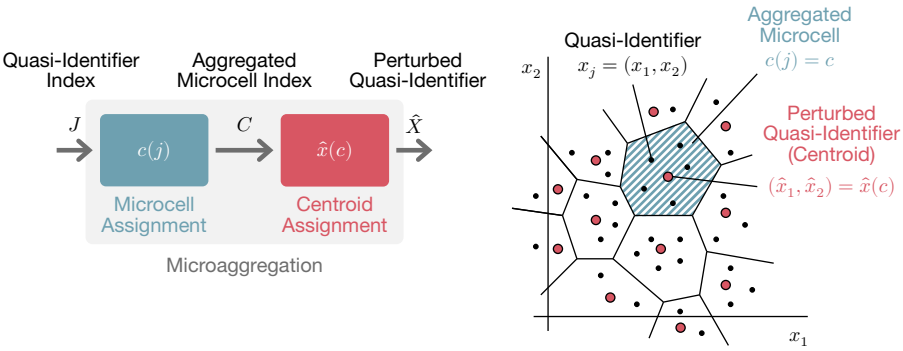


Fig. 4. Traditional microaggregation interpreted as a quantization problem on the record indices j represented by the microcell assignment function $c(j)$, and a centroid assignment function $\hat{x}(c)$ that reconstructs the perturbed version $\hat{x}_j$ of the original quasi-identifier x_j. In the analogous r.v. representation, $C = c(J)$ and $\hat{X} = \hat{x}(C)$.

a constrained quantization problem, as explained in [32, 35]. This interpretation is particularly intuitive in the special case of traditional microaggregation with numerical quasi-identifiers that do not appear repeatedly. In this special case, the function representing the microcell assignment can be defined directly on the quasi-identifiers, as $c(x)$.

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

Let us denote the perturbed version of the j^{th} quasi-identifier x_j, that is, its corresponding centroid, by $\hat{x}_j$. The SDC literature conventionally speaks of the *sum of squared errors* (SSE) and the *sum of squares total* (SST). Precisely, $\mathrm{SSE} = \sum_{j=1}^{n}\|x_j - \hat{x}_j\|^2$, and since the total variance of the data is the dimension m, its unnormalized version becomes $\mathrm{SST} = mn$. In this work, we formally define the distortion $\mathcal{D}$ introduced by the microaggregation algorithm by means of the MSE, implicitly normalized by the number n of samples, and also normalized by the number m of dimensions:

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

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

B. Mathematical Formulation of the Problem of p-Probabilistic k-Anonymous Microaggregation

We are now ready to proceed with the more general case of k-anonymous microaggregation with a probabilistic guarantee for locally trusting respondents, the object of this paper. Suppose that along with a collection of quasi-identifiers $x_1,\ldots,x_N$ in a predefined order, we are given probabilities $\pi_1,\ldots,\pi_N$ of respondent participation, and a desired probability $p\in(0,1)$ with which k-anonymity is to be enforced. Precisely, we are to microaggregate the data, merely knowing that a record $j=1,\ldots,N$ will be active, and count towards the k-anonymity requirement, with probability π_j, independently of the participation of others. Intuitively, we shall create cells of size n larger than the minimum k, in order to provide the guarantee, with at least probability p, that any given cell will be in fact k-anonymous. We shall refer to this type of k-anonymous microaggregation as p-probabilistic.

Precisely, our participation model consists of a collection of Bernoulli (binary) r.v.'s I_j with parameter π_j representing the probability of participation of an individual, that is, $I_j\sim\mathrm{Ber}(\pi_j)$, assumed for simplicity to be statistically independent, for $j=1,\ldots,N$, from which we may select n candidate respondents up to a maximum of N available. We also define $K_n=\sum_{j=1}^{n}I_j$ as the sum of the participation indicators of a cell containing $j=1,\ldots,n\leqslant N$, for which the selection order in general matters. In the special case of identical participation $\pi_j=\pi$, the collection of indicators I_j is *independent and identically distributed* (i.i.d.), and K_n becomes a binomial r.v. with parameters n and π, which we may denote by $K_n\sim\mathrm{Bin}(n,\pi)$. When the participation parameters are not identical, the associated distribution is often called Poisson binomial. Formally, a cell with probabilistically active records will be considered k-anonymous if k records or more are active, but also if no records at all are active. The latter condition owes to the fact that if no records are active, no anonymity leak is possible. Conversely, k-anonymity will be considered violated if, and only if, the *number of active records* K_n is between 1 and $k-1$.

For any probability expression a, we often write $1-a$ more conveniently as $\bar{a}$, a notation reminiscent of set complementation; indeed, for any event A, $\mathrm{P}(\bar{A})=\overline{\mathrm{P}(A)}$. Concordantly, the probabilistic parameter p may be conversely regarded as the *acceptable (cell) failure* probability $\bar{p}=1-p$, and the probability that k-anonymity is violated, as the *attained (cell) failure* probability $\bar{q}=1-q$, formally defined as $\bar{q}\stackrel{\text{def}}{=}\mathrm{P}\{0<K_n<k\}$. For a specific participation model represented by $(\pi_j)_j$, we define a given cell to be *p-probabilistically k-anonymous* when the probability q that k-anonymity is satisfied is at least p, that is, if, and only if, $\bar{q}\leqslant\bar{p}$.

For a predefined order of the respondents, we may define the *effective anonymity* $n_{\min}$ as the smallest integer, greater than or equal to k, for which a cell of at least that size will be p-probabilistically k-anonymous, that is,

$$n_{\min}\stackrel{\text{def}}{=}\min\{n\geqslant k\,|\,\bar{q}\leqslant\bar{p}\}\,,$$

where we may additionally impose $n\leqslant N$, when convenient, understanding that the additional constraint may cause $\bar{q}>\bar{p}$. Under this model, traditional microaggregation becomes the special case $\pi_j=\pi=1$, for which $n_{\min}=k$.

The constraint $n_{\min}\geqslant k$ in this definition is a technicality to circumvent the pathological case in which impractical values of π_j and p would make the event in which no records are active in the cell sufficiently likely. Assume for example that $\pi_j=\pi$ and that $\bar{\pi}=1-\pi>p$, and consider a cell with just a single record. The probability that such cell is empty, and therefore formally anonymous, would exceed p. Practical values of π and p should never be a concern, particularly because π will rarely approach 0, and p will typically be nearly 1. The main concepts just defined are summarized in Table I.

In the case of locally trusting recipients defined in §IV.B, no additional constraints to the problem exist, and cell shapes are arbitrary. Assuming further identical participation $\pi_j=\pi$, once the effective anonymity $n_{\min}$ is determined, the p-probabilistic problem is reduced to the traditional, deterministic version, with $n_{\min}$ in lieu of k. As represented in Fig. 5, in essence, our work identifies the possibility of procedurally reducing p-probabilistic k-anonymous microaggregation to traditional microaggregation, for certain practical purposes, and sets its main theoretical focus on the relationship between the target anonymity k and its effective counterpart $n_{\min}$. The theoretical analysis of such relationship is carried out for a given participation π and acceptable cell failure $\bar{p}$, in terms of additional failure rates presented later in §VI. How the effective anonymity $n_{\min}$ will translate into an increased distortion $\mathcal{D}$ will ultimately depend on the dataset and on the choice of microaggregation algorithm, which will create cells of size $n_{\min}\geqslant k$. In order to demonstrate the practical applicability of our work, in the referred section, we propose a natural modification of the traditional MSE-based distortion metric, suitable for p-probabilistic k-anonymous microaggregation, and in the experimental section §VII, we evaluate the entire dependence chain, encompassing k, $n_{\min}$, and $\mathcal{D}$.

In the general case of (possibly) uneven participation, informally denoted as $\pi_j\neq\pi$, the order in which we aggregate the available records will in general be a relevant issue. The simplest approach may consist in sorting records in ascending distance from a given reference. In any event, in the determination of the effective anonymity parameter for uneven participation, the remainder of this work assumes that a predefined order exists, and leaves for further investigation the possibility of optimizing the order for minimum overall distortion subject to the probabilistic anonymity constraint.

Table I. Summary of Main Definitions

Concepts	Symbols	Definitions
Participation Model	$(\pi_j)_{j=1}^N$	• Individual $j = 1, \ldots, N$ participates in the survey with probability π_j. • The unindexed symbol π is used for the simpler case of constant participation probability.
Number of Active Records	K_n	• For a given cell containing a subset of $n \leqslant N$ records, a total of K_n are active, representing the number of individuals that actually participate in the survey. • k-Anonymity is violated whenever $0 < K_n < k$ (but not when $K_n = 0$).
Cell-Failure Probabilities	$\bar{p}$, $\bar{q}$	• For a given cell containing n records, the attained (cell) failure probability is $1 - q \stackrel{\text{def}}{=} \bar{q} \stackrel{\text{def}}{=} \mathrm{P}\{0 < K_n < k\}$. • The acceptable (cell) failure probability is $1 - p \stackrel{\text{def}}{=} \bar{p}$. We wish that $\bar{q} \leqslant \bar{p}$.
p-Probabilistic k-Anonymity	p, k	• For a specific participation model represented by $(\pi_j)_j$, we define a given cell to be p-probabilistically k-anonymous when the probability that k-anonymity is satisfied is $q \geqslant p$, that is, if, and only if, $\bar{q} \leqslant \bar{p}$.
Effective Anonymity	$n_{\min}$	• The effective anonymity $n_{\min}$ is defined as the smallest integer (greater than or equal to k) for which a cell of at least that size will be p-probabilistically k-anonymous, that is, $n_{\min} \stackrel{\text{def}}{=} \min\{n \geqslant k \,|\, \bar{q} \leqslant \bar{p}\}$.

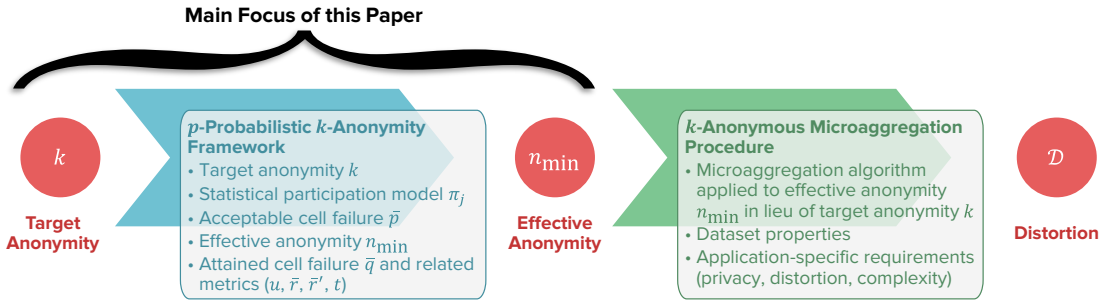


Fig. 5. In essence, our work identifies the possibility of procedurally reducing p-probabilistic k-anonymous microaggregation to traditional microaggregation, for certain practical purposes, and sets its necessarily limited theoretical focus on the relationship between the target anonymity k and its effective counterpart $n_{\min}$. However, in order to illustrate the practical applicability of our work, a suitable measure of distortion is proposed and the entire dependence chain is evaluated experimentally, in the following sections.

C. Reference Microaggregation Algorithm and Specification of the Microaggregation Function

Although the formulation in this work is readily extensible to any k-anonymous microaggregation algorithm, we take as reference, and make extensive use of, one of the algorithms that constitute the standard the facto in numerical microaggregation, known as MDAV [12, 8]. In any event, in the mode of operation assumed in this work, locally trusting respondents are assumed to be supplied with a specification of a potentially complex microaggregation function, mapping quasi-identifiers to microcells. Such specification will necessarily depend on the underlying k-anonymous microaggregation algorithm.

- For microaggregation algorithms without a simple clustering geometry, we may still define a table of assignments of quasi-identifiers in the statistical model, indexed by a record number j, to a microcell, indexed by c, thereby giving an *extensional specification* of $c\colon j \mapsto c(j)$. If the actual records are not exactly those in the statistical model, the local aggregation function may simply chose the nearest neighbor. In fact, it may be possible to reduce this exhaustive specification by exploiting the geometry of the induced Voronoi partition.
- An *intensional specification*, that is, a definition based on the geometrical properties of the microcell clusters, may be possible for certain algorithms. For the MDAV algorithm, one may give a list of the reference points to which other $k - 1$ points are adjoined, in the same order they were selected, along with the distance to the last point added. For the PCL algorithm, cited in §III.D, a list of centroids $\hat{x}(c)$ and their corresponding weights could be employed. In either example, particularly for categorical data, it might be necessary to introduce an ordering rule for samples at the same exact distance from a given reference.

VI. Theoretical Analysis

We analyze the problem of p-probabilistic k-anonymous microaggregation from the perspective of the anonymity attained, and the distortion introduced. Specifically, for anonymity, we analyze the effective anonymity constraint and the required cell size, along with several measures of failure related to the acceptable and attained rates $\bar{p}$ and $\bar{q}$.

A. Anonymity Analysis

Let $p_n(k) \stackrel{\text{def}}{=} \mathrm{P}\{K_n = k\}$ and $P_n(k) \stackrel{\text{def}}{=} \mathrm{P}\{K_n \leqslant k\}$ denote the PMF and the CMF of the number of active records K_n of a cell with potential participants $j = 1, \ldots, n$ in a predefined order (out of a maximum of N). For convenience, we define $K_0 = 0$ with probability 1. In the general case of uneven participation,

$$p_n(k) = \sum_{\mathscr{S}_n^k} \prod_{j \notin \mathscr{S}_n^k} \bar{\pi}_j \prod_{j \in \mathscr{S}_n^k} \pi_j \,,$$

where $\mathscr{S}_n^k$ denotes a subset of $\{1, \ldots, n\}$ of size k, and $p_n(0) = P_n(0) = \prod_{j=1}^n \bar{\pi}_j$. In the special case of identical participation $\pi_j = \pi$,

$$p_n(k) = \binom{n}{k} \bar{\pi}^{n-k} \pi^k, \quad P_n(k) = \sum_{j=0}^{k} \binom{n}{j} \bar{\pi}^{n-j} \pi^j \,, \quad p_n(0) = P_n(0) = \bar{\pi}^n.$$

Again in the general case, with our convention for $n = 0$, $P_0(k-1) = 1$. For any other $n \geqslant 1$,

$$\bar{P}_n(k-1) = \mathrm{P}\{K_n \geqslant k\} = \mathrm{P}\{K_{n-1} \geqslant k\} + \mathrm{P}\{K_{n-1} = k-1\}\pi_n = \bar{P}_{n-1}(k-1) + \pi_n p_{n-1}(k-1),$$
$$P_n(k-1) = P_{n-1}(k-1) - \pi_n p_{n-1}(k-1).$$

Similarly, $p_0(0) = 1$, $p_0(k) = 0$ for $k > 0$, and for $n \geqslant 1$,

$$p_n(k) = \mathrm{P}\{K_n = k\} = \mathrm{P}\{K_{n-1} = k\}\bar{\pi}_n + \mathrm{P}\{K_{n-1} = k-1\}\pi_n = \bar{\pi}_n p_{n-1}(k) + \pi_n p_{n-1}(k-1).$$

We may rewrite the latter recursion more compactly with the help of a few definitions. Namely,

$$p_n \stackrel{\text{def}}{=} \begin{pmatrix} p_n(0) \\ \vdots \\ p_n(k-1) \end{pmatrix}, \quad p_0 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^k,$$

and a ShiftDown function implementing the linear operator in $\mathbb{R}^k$ consisting in shifting down the entries of a vector, filling out the top with a zero, that is,

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix} \mapsto \begin{pmatrix} 0 \\ x_1 \\ \vdots \\ x_{k-1} \end{pmatrix}, \quad \text{so that} \quad p_n = \bar{\pi}_n p_{n-1} + \pi_n \textsc{ShiftDown}(p_{n-1}).$$

Note also that $P_n(k-1) = 1$ and $p_n(k-1) = 0$ for $n < k$, and

$$\bar{q} = \mathrm{P}\{0 < K_n < k\} = P_n(k-1) - p_n(0).$$

The two recursions introduced thus far will enable us to build an algorithm for the computation of the effective anonymity $n_{\min}$. But first, we would like to establish a few additional metrics of anonymity failure.

The *expected number of unprotected records* when k-anonymity is violated is

$$u \stackrel{\text{def}}{=} \mathrm{E}[K_n \,|\, 0 < K_n < k] \leqslant k - 1,$$

and may be computed from p_n and $\bar{q}$, simply by observing that

$$u\bar{q} = \mathrm{E}[K_n \,|\, 0 < K_n < k]\mathrm{P}\{0 < K_n < k\} = \sum_{j=0}^{k-1} \mathrm{P}\{K_n = j\} j = \sum_{j=0}^{k-1} p_n(j) j \,, \quad \text{or more compactly,} \quad u\bar{q} = p_n^T \begin{pmatrix} 0 \\ 1 \\ \vdots \\ k-1 \end{pmatrix}.$$

The case of $\pi_j = \pi$ and $k = 2$ is particularly simple, since

$$\{0 < K_n < k\} = \{K_n = 1\}, \quad \bar{q} = p_n(1) = n\bar{\pi}^{n-1}\pi, \quad u = \mathrm{E}[K_n \,|\, K_n = 1] = 1 = k - 1.$$

Yet another interesting failure metric is the probability that a record is unprotected in a cell of size n, which is the probability that the record is active, and that the cell is unprotected, precisely,

$$\bar{r}_j \stackrel{\text{def}}{=} \mathrm{P}\big\{I_j = 1, 0 < K_n < k\big\} = \mathrm{P}\big\{I_j = 1 \big| 0 < K_n < k\big\}\bar{q}.$$

Since $K_n = \sum_{j=1}^n I_j$, the expectation operator is linear, and the expectation of a binary indicator is the probability of the event indicated,

$$\sum_{j=1}^{n} \mathrm{P}\big\{I_j = 1 \big| 0 < K_n < k\big\} = \mathrm{E}[K_n \,|\, 0 < K_n < k] = u.$$

Therefore, the probability that a record, uniformly chosen from a cell of size n, will turn out to be (active and) unprotected is,

$$\bar{r} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} \bar{r}_j = \frac{u}{n}\bar{q} = \frac{1}{n} p_n^T \begin{pmatrix} 0 \\ 1 \\ \vdots \\ k-1 \end{pmatrix} \leqslant \frac{k-1}{n}\bar{q},$$

which we may consistently call *(attained) record-failure* probability. In the special case of $\pi_j = \pi$, by symmetry, $r_j = r$. Clearly, the expected number of unprotected records in a table with a total of N (potentially active) records is Nr.

A twist in this record failure metric, considering the point of view of the users, is the probability that a participating respondent will not be successfully protected, that is, the probability that a record is unprotected given that it is active,

Algorithm A. Effective k-anonymity, failure indicators

function EffectiveAnonymity

input k, $(\pi_j)_{j=1}^N$, $\bar{p}$ ▷Anonymity parameter k, participation π_j in predefined order $j=1,\dots,N$ (or constant π), acceptable cell failure $\bar{p}$

output $n_{\min}$, $\bar{q}$, u, $\bar{r}$ ▷Effective k-anonymity $n_{\min} \stackrel{\text{def}}{=} \min\{n \geqslant k \,|\, \bar{q} \leqslant \bar{p}\}$ (N if insufficient), attained cell failure $\bar{q} \stackrel{\text{def}}{=} \mathrm{P}\{0 < K_n < k\}$, average # of unprotected records $u \stackrel{\text{def}}{=} \mathrm{E}[K_n \,|\, 0 < K_n < k]$, attained record failure $\bar{r} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n \mathrm{P}\{I_j = 1, 0 < K_n < k\}$

1. $n \leftarrow 0$, $p_0 \leftarrow \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^k$, $P_0(k-1) \leftarrow 1$ ▷Initialize recursion
2. **repeat** ▷Assume at least $N = 1$ participants available
3. $n \leftarrow n+1$ ▷Increase cell size
   ▷Recursion on PMF $p_n(k) \stackrel{\text{def}}{=} \mathrm{P}\{K_n = k\}$ and CMF $P_n(k) \stackrel{\text{def}}{=} \mathrm{P}\{K_n \leqslant k\}$, the former represented by $p_n \stackrel{\text{def}}{=} \begin{pmatrix} p_n(0) \\ \vdots \\ p_n(k-1) \end{pmatrix}$
4. $P_n(k-1) \leftarrow P_{n-1}(k-1) - \pi_n p_{n-1}(k-1)$
5. $p_n \leftarrow \bar{\pi}_n p_{n-1} + \pi_n \textsc{ShiftDown}(p_{n-1})$ ▷Shift p_{n-1} down, set top entry to 0
6. $\bar{q} \leftarrow P_n(k-1) - p_n(0)$
7. **until** $n \geqslant k$ **and** $\bar{q} \leqslant \bar{p}$ **or** $n = N$ ▷Ensure $n \geqslant k$ and $\bar{q} \leqslant \bar{p}$ if N large
8. $u \leftarrow p_n^T \begin{pmatrix} 0 \\ 1 \\ \vdots \\ k-1 \end{pmatrix} \Big/ \bar{q}$, $\bar{r} \leftarrow p_n^T \begin{pmatrix} 0 \\ 1 \\ \vdots \\ k-1 \end{pmatrix} /n$
9. **return** n, $\bar{q}$, u, $\bar{r}$

Recursive procedure on the PMF and the CMF of the r.v. K_n devised to accurately compute the effective k-anonymity parameter $n_{\min}$, the attained cell-failure probability $\bar{q}$, the average number u of unprotected records, and the attained record-failure probability $\bar{r}$, from the level of k-anonymity required, the participation model $(\pi_j)_{j=1}^N$ of up to N available records in a predefined order, and the acceptable cell-failure probability $\bar{p}$. The same exact procedure is to be employed even in the simpler case of constant participation probability $\pi_j = \pi$. The ShiftDown function implements the linear operator $(x_1,\dots,x_k)^T \mapsto (0, x_1, x_2, \dots, x_{k-1})^T$.

$$\bar{r}'_j \stackrel{\text{def}}{=} \mathrm{P}\{0 < K_n < k \,\big|\, I_j = 1\} = \frac{\bar{r}_j}{\pi_j} \geqslant \bar{r}_j,$$

on average, $\bar{r}' \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n \bar{r}'_j$. Under constant participation $\pi_j = \pi$, by symmetry $\bar{r}'_j = \bar{r}'$, and

$$\bar{r}' = \frac{\bar{r}}{\pi} = \frac{u}{n\pi}\bar{q} \leqslant \frac{k-1}{n\pi}\bar{q}.$$

The recursions obtained earlier permit devising an algorithm for the computation of all of the metrics established, along with the effective anonymity $n_{\min}$. We present the algorithm in question as Algorithm A. It may be worth noting that $P_n(k-1) = 1$, $p_n(k-1) = 0$ for $n < k$, which suggests an alternative implementation, slightly faster, with two phases $n = 1,\dots,k-1$, and $n \geqslant k$. In the former phase, only p_n would need to be recursively computed. The latter phase would be identical to the more compact pseudocode preferred here.

Finally, for a table of $c = 1,\dots,|\mathscr{C}|$ cells, and attained success $\bar{q}_c \leqslant \bar{p}$, the probability that a table contains no failing cells is $\prod_c q_c$. Accordingly, the probability of table success is $t \stackrel{\text{def}}{=} \prod_{c=1}^{|\mathscr{C}|} q_c \geqslant p^{|\mathscr{C}|}$, and we consistently refer to $\bar{t}$ as the *(attained) table-failure* probability. When $\pi_j = \pi$, $|\mathscr{C}| = \lfloor \frac{N}{n} \rfloor$ cells in total, $\lfloor \frac{N}{n} \rfloor - 1$ of size n with attained cell success q and one last cell of size

$$n' = n + N \bmod n = N - (|\mathscr{C}| - 1)n \geqslant n,$$

with attained cell success $q' \geqslant q \geqslant p$, and table success probability $t = q^{\lfloor \frac{N}{n} \rfloor - 1} q' \geqslant q^{\lfloor \frac{N}{n} \rfloor} \geqslant p^{\lfloor \frac{N}{n} \rfloor}$, where q' can be computed with our recursive algorithm, Algorithm A, simply by setting $N = n'$ and $\bar{p} = 0$.

We conclude with a couple of quick considerations to bound $n_{\min}$. From an average-case perspective, the level of anonymity attained may be measured by the expected cell size $\mathrm{E}K_n = \sum_{j=1}^n \pi_j$. For constant participation $\pi_j = \pi$, $\mathrm{E}K_n = n\pi$. The application of Markov's inequality enables us to conclude that

$$\mathrm{P}\{K_n \geqslant k\} \leqslant \frac{n\pi}{k}, \quad \bar{p} \geqslant \bar{q} = \mathrm{P}\{K_n < k\} - \mathrm{P}\{K_n = 0\} \geqslant 1 - \frac{n}{k}\pi - \bar{\pi}^n, \quad p \leqslant q \leqslant \frac{n}{k}\pi + \bar{\pi}^n.$$

Since $n \geqslant k$,

$$p \leqslant \frac{n}{k}\pi + \bar{\pi}^k, \quad \frac{n}{k} \geqslant \frac{p - \bar{\pi}^k}{\pi},$$

which gives a simple lower bound on the effective anonymity $n_{\min}$. On the other hand, Chebyshev's inequality implies

$$\mathrm{P}\{K_n \geqslant k\} \leqslant \mathrm{P}\{|K_n - n\pi| \geqslant k - n\pi\} \leqslant \frac{n\bar{\pi}\pi}{(k-n\pi)^2},$$

a quadratic inequality that can be similarly used to obtain un upper bound on n.

B. Distortion Analysis

In this subsection, n now denotes all available records, not just the cell size, and $c = 1, \ldots, |\mathscr{C}|$ indexes the cells in a published table. Define $n_c \stackrel{\text{def}}{=} |\{j \,|\, c(j) = c\}|$, so that $n = \sum_{c=1}^{|\mathscr{C}|} n_c$. We should hasten to observe that in p-probabilistic k-anonymous microaggregation, the SSE within a microcell

$$\mathrm{SSE}_c = \sum_{j \,|\, c(j)=c} I_j \|x_j - \hat{x}_c\|^2$$

is in fact an r.v., with expectation

$$\mathrm{E}[\mathrm{SSE}_c] = \sum_{j \,|\, c(j)=c} \pi_j \|x_j - \hat{x}_c\|^2 .$$

Naturally, we may select a centroid minimizing the expected SSE, that is,

$$\hat{x}_c = \arg\min_{\hat{x}} \mathrm{E}[\mathrm{SSE}_c] = \frac{1}{\sum_{j \,|\, c(j)=c} \pi_j} \sum_{j \,|\, c(j)=c} \pi_j x_j .$$

For constant participation $\pi_j = \pi$, $\hat{x}_c = \frac{1}{n_c} \sum_{j \,|\, c(j)=c} x_j$, as with deterministic participation.

Alternatively, one could define centroid conditioned on the only information known to the probabilistic microaggregation method at the time of publication, namely the actual number of participants in the cell, which would induce a probability distribution on the participants related to π_j. With constant participation, by symmetry, the end result is unchanged. In general,

$$\mathrm{P}\{I_j = 1 \,|\, K_{n_c} = k\} = \frac{\mathrm{P}\{K_{n_c} = k \,|\, I_j = 1\}\pi_j}{\mathrm{P}\{K_{n_c} = k\}},$$

where by independence, $K_{n_c} \,|\, I_j = 1$ may be regarded as a smaller case of K_{n_c} for $k-1$ active records.

In traditional, deterministic microaggregation, the distortion is normalized not only by the number n of records, but also by the dimension m, that is,

$$\mathcal{D} = \frac{1}{mn} \sum_{c=1}^{|\mathscr{C}|} \sum_{j \,|\, c(j)=c} \|x_j - \hat{x}_c\|^2 .$$

With probabilistic participation, the analogous measure of distortion,

$$\frac{1}{m \sum_{j=1}^n I_j} \sum_{c=1}^{|\mathscr{C}|} \sum_{j \,|\, c(j)=c} I_j \|x_j - \hat{x}_c\|^2 ,$$

is a random quantity that cannot be computed unless the actual participation is known.

To tackle the issue, we propose a metric based on the expectation of the SSE and the normalization factor, reminiscent of the law of large numbers. Precisely, we replace amount of active records $\sum_j I_j$ with its expectation

$$\mathrm{E}\left[\sum_{j=1}^n I_j\right] = \sum_{j=1}^n \pi_j ,$$

and replace the SSE also with its expected value

$$\mathrm{E}\left[\sum_{c=1}^{|\mathscr{C}|} \sum_{j \,|\, c(j)=c} I_j \|x_j - \hat{x}_c\|^2\right] = \sum_{c=1}^{|\mathscr{C}|} \sum_{j \,|\, c(j)=c} \pi_j \|x_j - \hat{x}_c\|^2 .$$

Define the normalized participation $\pi'_j = \pi_j / \sum_{j=1}^n \pi_j$, a PMF proportional to π_j. The distortion metric proposed is then

$$\mathcal{D} = \frac{1}{m \sum_{j=1}^n \pi_j} \sum_{c=1}^{|\mathscr{C}|} \sum_{j \,|\, c(j)=c} \pi_j \|x_j - \hat{x}_c\|^2 = \frac{1}{m} \sum_{c=1}^{|\mathscr{C}|} \sum_{j \,|\, c(j)=c} \pi'_j \|x_j - \hat{x}_c\|^2 = \mathrm{E}\|X - \hat{X}\|^2,$$

with X an r.v. on the alphabet of quasi-identifiers $\mathscr{X} = (x_j)_{j=1}^n$, with probability distribution $(\pi'_j)_{j=1}^n$.

In the special case of constant participation $\pi_j = \pi$, the normalized participation $\pi'_j = \frac{1}{n}$ is uniform and

$$\mathcal{D} = \frac{1}{mn} \sum_{c=1}^{|\mathscr{C}|} \sum_{j \,|\, c(j)=c} \|x_j - \hat{x}_c\|^2 = \mathrm{E}\|X - \hat{X}\|^2,$$

exactly as in the case of deterministic participation, with X uniformly distributed on $\mathscr{X} = (x_j)_{j=1}^n$, which also measures expected distortion normalized per sample (and dimension), even if with probabilistic participation the number of active records is smaller.

C. Approximations for Large Effective Anonymity

In practice, we shall typically require an extremely low failure rate $\bar{p}$, or we shall be interested in the asymptotic characterization of the problem for a large anonymity parameter k. Either situation translates into a large effective anonymity $n_{\min}$, which suggests bringing our attention to an approximation analysis under this working hypothesis. For simplicity, we restrict our analysis to the case of identical participation $\pi_j = \pi$, with a fixed, common participation probability π. Our approximations are based on the connection of the problem of p-probabilistic k-anonymity with the method of types and large deviation theory. We first require a few mathematical preliminaries on approximations.

C.1. Mathematical Preliminaries on Relative, Absolute and Exponential Approximations

A minor subtlety is that we use three different symbols denote three slightly different types of approximations. Throughout the paper, in the context of real values x and y, we reserve the usual notation $x \approx y$ to indicate that the relative error $(x-y)/y$ is small within a given precision. In §VI.C.4, in the context of mathematical expressions, we shall employ this symbol for informal, heuristic approximations.

Throughout §VI.C, and in the context of the limits of the functions f and g, the notation $f \sim g$ indicates a *relative approximation*, that is, that the relative error $(f-g)/g \to 0$, or equivalently, $f/g \to 1$, usually read as "f is of the *order* of g". On the other hand, we enforce the notation $f \simeq g$ to indicate an *absolute approximation*, that is, that the absolute error $f - g \to 0$.

Provided that $f, g \to 0$, $f \simeq g$ follows trivially, being $f \sim g$ the only informative option. However, when $f, g \to \infty$, $f \simeq g$ entails $f \sim g$. Trivially, for any fixed value α, $f \to \alpha$ may be equivalently written as $f \sim \alpha$ or $f \simeq \alpha$. Note that either type of approximation induces an equivalence relation (satisfying reflexivity, symmetry and transitivity). Recall that the Landau little-oh notation $f = \mathrm{o}(g)$ indicates that $f/g \to 0$. Clearly, for any f and g,

$$f = g + \mathrm{o}(g) \quad \text{if, and only if,} \quad f \sim g, \quad f = g + \mathrm{o}(1) \quad \text{if, and only if,} \quad f \simeq g,$$

the direct implication of which may be expressed in a slightly more compact fashion as

$$g + \mathrm{o}(g) \sim g, \quad g + \mathrm{o}(1) \simeq g, \quad \text{for any } g.$$

Linearity properties, namely additivity and homogeneity, hold with mild restrictions. Precisely, for any function h, the assumption $f \sim g$ implies $f + h \sim g + h$ as long as the limit of h/g exists and is different from -1, possibly infinite, discarding the counterexample $f = x + 2$, $g = x + 1$, $h = -x$ for $x \to \infty$. Provided that $f \simeq g$, suppose that in addition the limit of h exists and is not $\pm\infty$. Then, $fh \simeq gh$.

In information theory, it is often convenient to characterize the exponential trend of a sequence. Recall [5] (§3, §12) that two sequences a_n and b_n are said to be *equal to the first order in the exponent* whenever $\frac{1}{n}\log\frac{a_n}{b_n} \to 0$, as $n \to \infty$, which in the notation we have introduced may be alternatively expressed as $\frac{1}{n}\log a_n \simeq \frac{1}{n}\log b_n$ or as $\sqrt[n]{a_n} \sim \sqrt[n]{b_n}$. Suppose that $f \sim g$, both positive, and that g has a limit different from 1, possibly infinite. Then, $\log f \sim \log g$. The additional requirement accounts for cases such as $f = e^{2x}$ and $g = e^x$ in the limit of $x \to 0$. The converse is not true, that is, $f \sim g$ does not imply $e^f \sim e^g$, as readily shown by $f = x + 1$ and $g = x$ as $x \to \infty$. On the other hand, for positive functions, $f \sim g$ holds if, and only if, $\log f \simeq \log g$.

C.2. Connection with the Method of Types and Large Deviation Theory

We explore connection between p-probabilistic k-anonymity and the *method of types*, a powerful technique in *large deviation theory* lying at the heart of the intersection between information theory and statistics. Some of the results enumerated here are a review of [5] (§12), but many others were specifically derived for this work. Let p_k^n denote the probability of k successes for a binomial r.v. K_n with n trials and probability of success π, and *type* or empirical parameter $t = \frac{k}{n}$. In other words, the type t is a relative representation of the empirical number K_n of successful outcomes, in contrast with the theoretical success probability π.

Under these definitions, $p_k^n = \mathrm{P}\{K_n = k\}$. Further,

$$p_k^n = \binom{n}{k}\bar{\pi}^{n-k}\pi^k = \binom{n}{k}2^{-n\mathrm{H}(t\|\pi)}, \quad \text{where} \quad \mathrm{H}(t\|\pi) = -\bar{t}\log\bar{\pi} - t\log\pi = \mathrm{H}(t) + \mathrm{D}(t\|\pi) \geqslant \mathrm{H}(t)$$

denotes the *cross-entropy* of the binary distribution with probability of success t, with respect to that with π, and where $\mathrm{D}(t\|\pi)$ denotes the KL divergence between the same quantities. Concordantly define $T_n \stackrel{\text{def}}{=} \frac{1}{n}K_n$ and $p_T(t) \stackrel{\text{def}}{=} \mathrm{P}\{T_n = t\}$.

Unless otherwise stated, limits are in terms of $n \to \infty$, but k remains fixed (instead of fixing t), and accordingly, $t \to 0$. The limit in n is consistent with the assumption of vanishing acceptable, and therefore attained cell-failure rates, $\bar{q} \leqslant \bar{p} \to 0$, with fixed anonymity parameter k. Occasionally, it will be useful to think of the dual case of $n \to \infty$ with fixed t, where the limit is now caused by $n \geqslant k \to \infty$ instead, and $\bar{q}$ remains fixed, but sufficiently small.

A fundamental result of the method of types [5] (Theor. 12.1.4, employing a tighter version of the bound with $|\mathscr{X}| - 1$ in lieu of $|\mathscr{X}|$) asserts that

$$\tfrac{1}{n+1}2^{-n\mathrm{D}(t\|\pi)} \leqslant p_T(t) \leqslant 2^{-n\mathrm{D}(t\|\pi)}, \quad \text{or equivalently,} \quad 0 \leqslant -\tfrac{1}{n}\log p_k^n - \mathrm{D}(t\|\pi) \leqslant \tfrac{\log(n+1)}{n},$$

which, in our notation, implies $-\frac{1}{n}\log p_k^n \simeq \mathrm{D}(t\|\pi)$ as $n \to \infty$. For fixed k, we have $t \to 0$ and

$$-\frac{1}{n}\log p_k^n \to \mathrm{D}(0\|\pi) = -\log\bar{\pi}\,.$$

Note however that approximating $\mathrm{D}(t\|\pi)$ by its value at $t = 0$ may be fairly inaccurate in practice, because $\frac{\partial \mathrm{D}}{\partial t}\big|_{t=0} = -\infty$, and n may not be large enough when compared to k.

We show that $\mathrm{P}\{K_n = k\}$ dominates the event $\mathrm{P}\{K_n \leqslant k\}$, when k remains fixed. Since

$$\frac{p_{k-1}^n}{p_k^n} = \frac{k}{n-k+1}\frac{\bar{\pi}}{\pi} = \frac{t}{\bar{t}+\frac{1}{n}}\frac{\bar{\pi}}{\pi} < \frac{t/\bar{t}}{\pi/\bar{\pi}},$$

$p_{k-1}^n = \mathrm{o}(p_k^n)$ as $n \to \infty$, for fixed k, and consequently,

$$\mathrm{P}\{K_n \leqslant k\} = \sum_{j=0}^{k} p_j^n \sim p_k^n = \mathrm{P}\{K_n = k\}.$$

In light of our definition of k-anonymity violation, the cumulative event could have excluded the case $K_n = 0$, but the asymptotic dominance of $K_n = k$ makes this distinction irrelevant. Also, to be precise, application of these preliminaries requires $k-1$ in lieu of k.

At this point, it is relatively straightforward to show that the monotonicity of the probability terms under the assumption that $t \leqslant \pi$ in the cumulative probability permits extending the previous bounds to

$$-\tfrac{\log(k+1)}{n} \leqslant -\tfrac{1}{n}\log\sum_{j=0}^{k} p_j^n - \mathrm{D}(t\|\pi) \leqslant \tfrac{\log(n+1)}{n},$$

which, in our notation, implies

$$-\tfrac{1}{n}\log\sum_{j=0}^{k} p_j^n \simeq \mathrm{D}(t\|\pi) \simeq -\tfrac{1}{n}\log p_k^n\,,$$

and for fixed k, $t \to 0$ gives $-\frac{1}{n}\log\sum_{j=0}^{k} p_j^n \to -\log\bar{\pi}$, as before. In terms of p-probabilistic k-anonymity, more specifically in terms of the attained cell-failure rate $\bar{q}$ and the participation probability π, this means that $-\frac{1}{n}\log\bar{q} \to -\log\bar{\pi}$.

Consider, on the other hand, a marginal increase in the number of trials, corresponding to the potential discrepancy between accepted and attained cell-failure rates. We have

$$\frac{p_k^{n+1}}{p_k^n} = \frac{n+1}{n+1-k}\,\bar{\pi} = \frac{1+\frac{1}{n}}{\bar{t}+\frac{1}{n}}\,\bar{\pi},$$

which leads us to observe that

$$\frac{p_k^{n+1}}{p_k^n} \to \bar{\pi} \quad \text{and} \quad \frac{\sum_{j=0}^{k} p_j^{n+1}}{\sum_{j=0}^{k} p_j^n} \to \bar{\pi}$$

as $n \to \infty$, for fixed k, which in turn implies that $\bar{q}/\bar{p}$ is asymptotically bounded from below by $\bar{\pi}$, precisely,

$$\bar{\pi} = \liminf\frac{\bar{q}}{\bar{p}} \leqslant \max\frac{\bar{q}}{\bar{p}} = 1, \quad \text{or} \quad 0 = \min\frac{\bar{p}-\bar{q}}{\bar{p}} \leqslant \limsup\frac{\bar{p}-\bar{q}}{\bar{p}} = \pi,$$

in terms of the relative difference.

However, we proved that in general, regardless of whether k or t is fixed,

$$-\tfrac{1}{n}\log\sum_{j=0}^{k} p_j^n \simeq \mathrm{D}(t\|\pi),$$

so that the largest probability dominates the cumulative event at least in this logarithmic sense. This enables us to relate the asymptotic behavior of the acceptable and attained cell-failure probabilities, $\bar{p}$ and $\bar{q}$ respectively, at least in their logarithm. Precisely, for either $t \to 0$ or $t \neq \pi$, observing that $\frac{k}{n+1} = \frac{nt}{n+1} \to t = \frac{k}{n}$, we may conclude that

$$\log\sum_{j=0}^{k} p_j^{n+1} \sim \log\sum_{j=0}^{k} p_j^n\,, \quad \log\bar{p} \sim \log\bar{q}\,, \quad -\tfrac{1}{n}\log\bar{p} \simeq -\tfrac{1}{n}\log\bar{q} \simeq \mathrm{D}(t\|\pi).$$

C.3. Case when $k = 2$

We apply the considerations of the previous subsection, §VI.C.2, to the problem of p-probabilistic microaggregation with identical participation $\pi_j = \pi$, and small acceptable failure rate $\bar{p} \to 0$, which requires large effective anonymity $n_{\min} \to \infty$. We tackle for now the extreme case of the smallest possible anonymity parameter, $k = 2$. In this case,

$$n_{\min} = \{n \geqslant 2 \,|\, n\bar{\pi}^{n-1}\pi \leqslant \bar{p}\}, \quad n_{\min}\bar{\pi}^{n_{\min}} = \frac{\bar{\pi}}{\pi}\bar{q} \leqslant \frac{\bar{\pi}}{\pi}\bar{p}.$$

At this point, we make a quick digression to investigate the transcendental equation $xb^x = a$ for positive real numbers a and b, with $b \in (0,1)$, and $0 < a \leqslant -\frac{1}{e\ln b}$, which implies that at least a solution exists in the real variable x, and in most cases, two. In fact, the solution for x in the above equation satisfying $x \geqslant -\frac{1}{\ln b}$ may be expressed in terms of the lower branch W_{-1} of the *Lambert W function* or *product logarithm*, defined as the solution to $W_{-1}(t)e^{W_{-1}(t)} = t$ restricted to $W_{-1}(t) \leqslant -1$. Precisely, $x = W_{-1}(a\ln b)/\ln b$. In general, no closed-form solution exists in terms of more conventional functions, but we may readily obtain an approximate solution for $a \ll 1$, or equivalently, $x \gg 1$.

In the logarithmic form of the equation at hand, $\ln x + x\ln b = \ln a$, the approximation $x\ln b \gg \ln x$ yields the simple approximate solution

$$x \sim \tilde{x}_0 \stackrel{\text{def}}{=} \frac{\ln a}{\ln b} = \log_b a\,,$$

with $b \in (0,1)$ a suitable logarithmic base. To see that $x \sim \tilde{x}_0$ formally, recall from §VI.C.1 that $f + \mathrm{o}(f) \sim f$. Direct application of this observation to the logarithmic form of the equation divided by $\ln b$,

$$x + \frac{\ln x}{\ln b} = \frac{\ln a}{\ln b} = \tilde{x}_0,$$

proves the claim. This form of the equation also demonstrates that the absolute version of the approximation does not hold, since the absolute error $x - \tilde{x}_0 = -\frac{\ln x}{\ln b}$ diverges.

For our practical purposes, the simple approximation $\tilde{x}_0$ is insufficiently accurate, as it will require an extremely low value for a. However, we may linearize $\ln x$ in the equation around x_0, which gives the more accurate solution

$$x \sim \tilde{x}_1 \stackrel{\text{def}}{=} \left(1 - \frac{\ln \frac{\ln a}{\ln b}}{1 + \ln a}\right) \frac{\ln a}{\ln b} = \left(1 - \frac{\ln \log_b a}{1 + \ln a}\right) \log_b a\,.$$

A swift application of L'Hôpital's rule confirms that $(\ln \frac{\ln a}{\ln b})/(1 + \ln a) \underset{a \downarrow 0}{\longrightarrow} 0$, which means that $\tilde{x}_1 \sim \tilde{x}_0 \sim x$ as $a \downarrow 0$, as expected. Incidentally,

$$\tilde{x}_1 \sim (1 - \log_a \log_b a) \log_b a\,.$$

The same method, applied now to the quadratic Taylor expansion of $\ln x$ around x_0 yields

$$x \sim \tilde{x}_2 \stackrel{\text{def}}{=} \left(2 + \ln a + \sqrt{(1 + \ln a)^2 + 2 \ln \frac{\ln a}{\ln b}}\right) \frac{\ln a}{\ln b}$$

instead (for $b > 1$ and $a \gg 1$ the approximation remains valid with a negative square root). It is routine to check that

$$1 + \ln a + \sqrt{(1 + \ln a)^2 + 2 \ln \frac{\ln a}{\ln b}} = \frac{-2 \ln \frac{\ln a}{\ln b}}{1 + \ln a - \sqrt{(1 + \ln a)^2 + 2 \ln \frac{\ln a}{\ln b}}} \underset{a \downarrow 0}{\longrightarrow} 0,$$

thereby concluding that $\tilde{x}_2 \sim \tilde{x}_0$ as $a \downarrow 0$. Back to the problem of p-probabilistic 2-anonymous microaggregation, in the preliminary results of §VI.C.2, we showed that $\log \bar{p} \sim \log \bar{q}$. We approximate the solution to the equation $n \bar{\pi}^n = \frac{\bar{\pi}}{\pi} \bar{q}$ in n observing that

$$n \sim \frac{\log(\frac{\bar{\pi}}{\pi} \bar{q})}{\log \bar{\pi}}, \quad \text{which gives} \quad n_{\min} \sim \tilde{n}_0 \stackrel{\text{def}}{=} \frac{\log(\frac{\bar{\pi}}{\pi} \bar{p})}{\log \bar{\pi}} = \log_{\bar{\pi}} \left(\frac{\bar{\pi}}{\pi} \bar{p}\right) = 1 + \log_{\bar{\pi}} \frac{\bar{p}}{\pi} = \frac{-\log \bar{p} + \log \frac{\pi}{\bar{\pi}}}{-\log \bar{\pi}},$$

where $\log \frac{\pi}{\bar{\pi}}$ may be readily identified as the log odds of π. The above analysis offers the more accurate solutions

$$n_{\min} \sim \tilde{n}_1 \stackrel{\text{def}}{=} \left(1 - \frac{\ln \frac{\ln(\frac{\bar{\pi}}{\pi} \bar{p})}{\ln \bar{\pi}}}{1 + \ln(\frac{\bar{\pi}}{\pi} \bar{p})}\right) \frac{\ln(\frac{\bar{\pi}}{\pi} \bar{p})}{\ln \bar{\pi}}, \quad n_{\min} \sim \tilde{n}_2 \stackrel{\text{def}}{=} \left(2 + \ln \left(\frac{\bar{\pi}}{\pi} \bar{p}\right) + \sqrt{\left(1 + \ln \left(\frac{\bar{\pi}}{\pi} \bar{p}\right)\right)^2 + 2 \ln \frac{\ln(\frac{\bar{\pi}}{\pi} \bar{p})}{\ln \bar{\pi}}}\right) \frac{\ln(\frac{\bar{\pi}}{\pi} \bar{p})}{\ln \bar{\pi}},$$

with $\tilde{n}_2 \sim \tilde{n}_1 \sim \tilde{n}_0 \sim n_{\min}$ as $\bar{p} \downarrow 0$.

C.4. Cases of Arbitrary and Large k

Maintaining the assumption of small cell-failure rate $\bar{p}$, we do away with the restriction $k = 2$, and now proceed to explore the general case of arbitrary anonymity parameter k. We contemplate as well the case of large k. Either small $\bar{p}$ or large k translates into a cell size n and an effective anonymity $n_{\min}$ also large. In short, the analysis in this subsection assumes $\bar{p} \ll 1$ or $k \gg 1$, $n \geqslant n_{\min} \gg 1$, and $\pi_j = \pi$ fixed.

Define $\kappa \stackrel{\text{def}}{=} (k-1)/n_{\min}$. As $n_{\min} \to \infty$, $\bar{\pi}^{n_{\min}} \downarrow 0$, and the bound in §VI.A based on Markov's inequality gives

$$1 \simeq p \leqslant q \leqslant \frac{n_{\min}}{k} \pi + \bar{\pi}^{n_{\min}} \simeq \frac{n_{\min}}{k} \pi,$$

meaning that $n_{\min}/(k-1) > n_{\min}/k$ should typically be greater than $\frac{1}{\pi}$, and accordingly we should expect $\kappa \leqslant \pi$ for sufficiently small $\bar{p}$ and sufficiently large $n_{\min}$.

We established in §VI.C.2 that for fixed k, as $n \to \infty$, $-\frac{1}{n} \log \bar{q} \to -\log \bar{\pi}$, warning of the practical inaccuracy of the approximation due to the infinite partial derivative of the KL divergence with respect to its first argument at zero. It was also shown that for either fixed k or fixed ratio $\frac{k}{n_{\min}}$, $\log \bar{p} \sim \log \bar{q}$, which finally implies that

$$n_{\min} \sim \frac{\log \bar{q}}{\log \bar{\pi}} \sim \frac{\log \bar{p}}{\log \bar{\pi}},$$

which in spite of its formal correctness, may be of questionable accuracy for practical values of n.

Fortunately, we may resort to the more sophisticated form of the arguments made in §VI.C.2, drawing upon the method of types of information theory, to more adequately approximate $n_{\min}$. Indeed, we showed at the end of the referred subsection that, as $n \to \infty$, and consistently replacing k with $k-1$ and t with κ,

$$-\tfrac{1}{n_{\min}} \log \bar{p} \simeq -\tfrac{1}{n_{\min}} \log \bar{q} \simeq \mathrm{D}(\kappa \| \pi).$$

In order to produce an approximation to the effective anonymity $n_{\min}$ of practical value, we shall immediately become less formal and propose the heuristic quantity $\tilde{n}_1$ as the solution to the equation associated with the approximation, that is,

$$\tilde{n}_1 \stackrel{\text{def}}{:} \; -\tfrac{1}{\tilde{n}_1}\log\bar{p} = \mathrm{D}\left(\tfrac{k-1}{\tilde{n}_1}\middle\|\,\pi\right).$$

Intuition suggests that small $\bar{p}$ and large $n_{\min}$, leading to small KL divergence, should make κ close to π. Inspired by this intuition, a further twist in this proposal, which will also prove adequate in our experiments, resorts to the second-order Taylor approximation to the KL divergence viewed as a function of κ around π, that is,

$$\mathrm{D}(\kappa\|\pi) = \frac{1}{2}\frac{(\kappa-\pi)^2}{\bar{\pi}\pi} + O(|\kappa-\pi|^3).$$

We simply replace $\mathrm{D}(\kappa\|\pi)$ by $\frac{1}{2}\frac{(\kappa-\pi)^2}{\bar{\pi}\pi}$ in the previous approximation to $n_{\min}$, which immediately suggests

$$\tilde{n}_1' \stackrel{\text{def}}{:} \; -\frac{1}{\tilde{n}_1'}\ln\bar{p} = \frac{1}{2}\frac{\left(\frac{k-1}{\tilde{n}_1'}-\pi\right)^2}{\bar{\pi}\pi},$$

and finally leads to

$$\tilde{n}_1' = \frac{k-1-\bar{\pi}\ln\bar{p}+\sqrt{-\bar{\pi}\ln\bar{p}\,(2(k-1)-\bar{\pi}\ln\bar{p})}}{\pi}.$$

Interestingly, we may rewrite this last expression in terms of the functionals A and G implementing the arithmetic and geometric averages

$$\begin{matrix}\mathrm{A}\colon \binom{x_1}{x_2} \mapsto \dfrac{x_1+x_2}{2},\\ \mathrm{G}\colon \binom{x_1}{x_2} \mapsto \sqrt{x_1x_2}\end{matrix} \quad \text{as} \quad \tilde{n}_1' = \frac{1}{\pi}(\mathrm{A}+\mathrm{G})\left(\binom{2(k-1)}{0} - \bar{\pi}\ln\bar{p}\right).$$

On account of the fact that

$$\sqrt{x+a}-\sqrt{x} = \frac{a}{\sqrt{x+a}+\sqrt{x}} \underset{x\to\infty}{\longrightarrow} 0$$

this last approximation admits a slightly simpler form as a quadratic polynomial of $\sqrt{k-1}$, precisely,

$$\tilde{n}_1'' \stackrel{\text{def}}{=} \frac{1}{\pi}\big(k-1+\sqrt{-2\bar{\pi}\ln\bar{p}}\,\sqrt{k-1}-\bar{\pi}\ln\bar{p}\big).$$

We should hasten to point out that similar approximations may be obtained using the Gaussian approximation given by the de Moivre-Laplace theorem, or using Hoeffding's inequality, among other well-known approximations and bounds in the context of the binomial distribution. Yet another simple quadratic approximation to $\mathrm{D}(\kappa\|\pi)$ determined by the constraints $\mathrm{D}(0\|\pi) = -\log\bar{\pi}$ and $\mathrm{D}(\pi\|\pi) = 0$ is

$$\mathrm{D}(\kappa\|\pi) \approx -\log\bar{\pi}\,\left(1-\frac{\kappa}{\pi}\right)^2,$$

where we recall that the symbol $\approx$ denotes an informal approximation. This suggests our final heuristic approximation to the effective anonymity,

$$\tilde{n}_2 \stackrel{\text{def}}{=} \frac{2(k-1)\log\bar{\pi}+\pi\log\bar{p}-\sqrt{\pi\log\bar{p}\,(\pi\log\bar{p}\;+4(k-1)\log\bar{\pi})}}{2\pi\log\bar{\pi}}.$$

In addition to the relationship between probabilities and divergences, we explored in §VI.C.2 the dominance of the PMF $\mathrm{P}\{K_n=k\}$ over the CMF $\mathrm{P}\{K_n\leqslant k\}$, in the limit of increasing n, for fixed k. In the problem analyzed in this paper, conveniently replacing k by $k-1$, this means that for any cell size $n\to\infty$, and in particular for $n_{\min}\to\infty$,

$$\bar{q} \stackrel{\text{def}}{=} \mathrm{P}\{0<K_n<k\} \sim \mathrm{P}\{K_n=k-1\}, \quad \text{which implies} \quad u \stackrel{\text{def}}{=} \mathrm{E}[K_n\,|\,0<K_n<k] \simeq k-1,$$

and, in turn,

$$\bar{r} = \frac{u}{n_{\min}}\bar{q} \sim \frac{k-1}{n_{\min}}\bar{q} = \kappa\bar{q} \leqslant \pi\bar{q} \leqslant \pi\bar{p}, \quad \bar{r}' = \frac{\bar{r}}{\pi} = \frac{u}{n_{\min}\pi}\bar{q} \sim \frac{\kappa}{\pi}\bar{q} \leqslant \bar{q} \leqslant \bar{p}.$$

As far as the attained table failure, reasonably, $\bar{p}\ll 0$, and consistently, $\bar{q}\ll 0$. For large datasets with $\frac{N}{n_{\min}}\gg 1$, we may informally approximate

$$\bar{t} \approx 1-(1-\bar{q})^{\left\lfloor\frac{N}{n_{\min}}\right\rfloor} \approx 1-e^{-\frac{N}{n_{\min}}\bar{q}} \leqslant 1-e^{-\frac{N}{n_{\min}}\bar{p}}, \quad \text{where} \quad \frac{N}{n_{\min}}\bar{q} \sim \frac{N}{k-1}\bar{r} \leqslant \frac{N}{k-1}\pi\bar{p}.$$

Provided that, in addition, $\frac{N}{n_{\min}}\bar{p}\approx 0$, for example for extremely low $\bar{p}$, which will increase $n_{\min}$, $\bar{t}\approx\frac{N}{n_{\min}}\bar{q}$.

VII. Numerical and Experimental Analysis

We numerically verify the most important results stated theoretically in §VI, and present experimental results for p-probabilistic k-anonymous microaggregation of synthetic and standardized data. The numerical analysis in this section

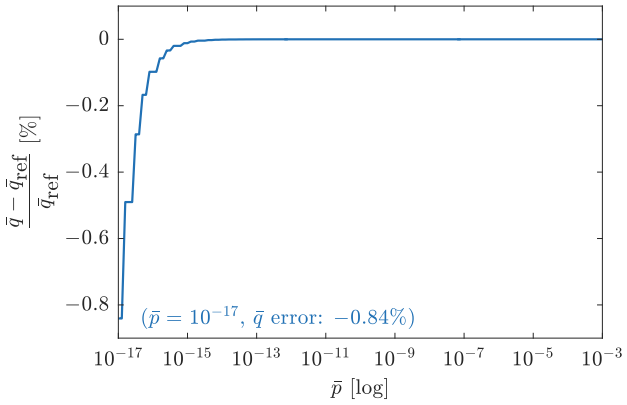


Fig. 6. Numerical stability of Algorithm 1 in the computation of $\bar{q}$ with respect to Matlab's function `binocdf`, taken as the reference value $\bar{q}_{\text{ref}}$, for $k=20$ and $\pi=0.5$. The recursion in the algorithm is remarkably accurate all the way down to $\bar{p}=10^{-17}$, with a maximum relative error of −0.84%.

TABLE II. SIMPLE VERIFICATION OF FAILURE METRICS VIA SIMULATION

Variable	Algorithm	Simulation	Relative Error [%]
$\bar{q}$	0.09671	0.09606	−0.67
u	17.85	17.83	−0.13
$\bar{r}$	0.03597	0.03605	0.23

Verification via Monte Carlo simulation of the failure parameters computed with Algorithm 1. Precisely, we verify the attained cell failure $\bar{q}$, the average number u of unprotected records under failure, and the attained record failure $\bar{r}$, for an anonymity parameter of $k=20$, an acceptable cell failure $\bar{p}=0.1$, a participation $\pi=0.5$, an effective anonymity $n_{\min}=48$, and $|\mathscr{C}|=10^5$ simulated cells.

was carried out in its entirety with Matlab (R2015b). The traditional microaggregation algorithm on which these experiments are based is MDAV, introduced in §III.D, in accordance with the formulation in §V.C.

A. Numerical Stability and Functional Verification of our Recursive Algorithm

Our first verification is the numerical stability of the recursion in Algorithm A, by recomputing the attained cell-failure probability $\bar{q}$ with the complex implementation of the binomial cumulative distribution function `binocdf` provided with Matlab, based on [1]. Specifically, $n_{\min}$ and $\bar{q}$ were computed with our algorithm, whereas the reference value $\bar{q}_{\text{ref}}$ was computed as

$$\bar{q}_{\text{ref}} = \texttt{binocdf}(k-1, n_{\min}, \pi) - \bar{\pi}^{n_{\min}},$$

for $k=20$ and $\pi=0.5$. Fig. 6 indicates that the recursion in our algorithm is remarkably accurate all the way down to $\bar{p}=10^{-17}$, with a maximum relative error $(\bar{q}-\bar{q}_{\text{ref}})/\bar{q}_{\text{ref}} \approx -0.84\%$.

Additionally, we carried out a simple Monte Carlo simulation to verify most of the values computed with our algorithm. We set $k=20$, $\bar{p}=0.1$, $\pi=0.5$, which gives $n_{\min}=48$, and synthesized $|\mathscr{C}|=10^5$ microcells with those parameters, measuring participation. The high value of the acceptable cell failure rate facilitated the simulation, whose mere purpose is a functional verification of Algorithm A. Table II presents the results and the relative error of the simulation with respect to the values computed with the recursive algorithm.

B. Numerical Examples of Effective Anonymity and Failure Metrics

In order to gain some quantitative understanding of the problem investigated, we employ our recursive algorithm to compute the effective anonymity $n_{\min}$, the attained cell-failure probability $\bar{q}$, the average number u of unprotected records in case of failure, and the record-failure metrics $\bar{r}$ and $\bar{r}'$, for various synthetic combinations of anonymity k, participation π, and acceptable cell-failure probability $\bar{p}$. The results are presented in Table III.

Additionally, for a fixed participation of $\pi=0.75$ and various table sizes N, we report the corresponding table-failure rate $\bar{t}$ in Table IV. Recall that the table-failure metric entails a rather steep requirement, as absolutely none of the active records can be unprotected, in any of the $|\mathscr{C}|=\left\lfloor \frac{N}{n_{\min}} \right\rfloor$ cells created, and might be considered as a rather pessimistic, worst-case metric. The unconditional record-failure metric $\bar{r}$ and the conditional version $\bar{r}'$ given participation, far more lenient metrics, are bounded by $\bar{p}$, as argued in §VI.C.3. Recall that the average number of unprotected records is $N\bar{r}$.

C. The Price of High-Quality Design

Undoubtedly, the most pressing question is whether demanding failure rates come at the expense of large microcells, which may then introduce a significant price in distortion. We give a partial answer at this point, in terms of cell size, and explore the distortion impact shortly. In Fig. 7, we plot the effective anonymity $n_{\min}$ for acceptable cell-failure

Table III. Effective Anonymity $n_{\min}$ and Failure Metrics $\bar{q}$, u, $\bar{r}$, $\bar{r}'$

k	π	$\bar{p}$	$n_{\min}$	$\bar{q}$	u	$\bar{r}$	$\bar{r}'$
10	0.75	10^{-4}	25	$4.31 \cdot 10^{-5}$	8.80	$1.52 \cdot 10^{-5}$	$2.02 \cdot 10^{-5}$
		10^{-5}	27	$6.05 \cdot 10^{-6}$	8.82	$1.98 \cdot 10^{-6}$	$2.64 \cdot 10^{-6}$
		10^{-6}	29	$7.95 \cdot 10^{-7}$	8.84	$2.42 \cdot 10^{-7}$	$3.23 \cdot 10^{-7}$
	0.5	10^{-4}	43	$8.51 \cdot 10^{-5}$	8.69	$1.72 \cdot 10^{-5}$	$3.44 \cdot 10^{-5}$
		10^{-5}	48	$7.61 \cdot 10^{-6}$	8.73	$1.38 \cdot 10^{-6}$	$2.77 \cdot 10^{-6}$
		10^{-6}	53	$6.1 \cdot 10^{-7}$	8.77	$1.01 \cdot 10^{-7}$	$2.02 \cdot 10^{-7}$
50	0.75	10^{-4}	88	$6.2 \cdot 10^{-5}$	48.4	$3.41 \cdot 10^{-5}$	$4.54 \cdot 10^{-5}$
		10^{-5}	91	$9.82 \cdot 10^{-6}$	48.4	$5.22 \cdot 10^{-6}$	$6.97 \cdot 10^{-6}$
		10^{-6}	95	$7.14 \cdot 10^{-7}$	48.5	$3.64 \cdot 10^{-7}$	$4.86 \cdot 10^{-7}$
	0.5	10^{-4}	144	$7.86 \cdot 10^{-5}$	48.1	$2.62 \cdot 10^{-5}$	$5.25 \cdot 10^{-5}$
		10^{-5}	151	$9.64 \cdot 10^{-6}$	48.2	$3.08 \cdot 10^{-6}$	$6.15 \cdot 10^{-6}$
		10^{-6}	159	$7.35 \cdot 10^{-7}$	48.3	$2.23 \cdot 10^{-7}$	$4.46 \cdot 10^{-7}$

Computation of the effective anonymity $n_{\min}$, the attained cell-failure probability $\bar{q}$, the average number u of unprotected records in case of failure, and the record-failure metrics $\bar{r}$ and $\bar{r}'$, with Algorithm 1, for various synthetic combinations of anonymity k, participation π, and acceptable cell-failure probability $\bar{p}$.

Table IV. Record and Table Failure Metrics $\bar{r}$, $\bar{r}'$, $\bar{t}$

k	N	$\bar{p}$	$\bar{r}$	$\bar{r}'$	$\bar{t}$
10	10^4	10^{-4}	$1.52 \cdot 10^{-5}$	$2.02 \cdot 10^{-5}$	0.0171
		10^{-5}	$1.98 \cdot 10^{-6}$	$2.64 \cdot 10^{-6}$	0.00223
		10^{-6}	$2.42 \cdot 10^{-7}$	$3.23 \cdot 10^{-7}$	0.000273
	10^5	10^{-4}	$1.52 \cdot 10^{-5}$	$2.02 \cdot 10^{-5}$	0.158
		10^{-5}	$1.98 \cdot 10^{-6}$	$2.64 \cdot 10^{-6}$	0.0221
		10^{-6}	$2.42 \cdot 10^{-7}$	$3.23 \cdot 10^{-7}$	0.00274
	10^6	10^{-4}	$1.52 \cdot 10^{-5}$	$2.02 \cdot 10^{-5}$	0.822
		10^{-5}	$1.98 \cdot 10^{-6}$	$2.64 \cdot 10^{-6}$	0.201
		10^{-6}	$2.42 \cdot 10^{-7}$	$3.23 \cdot 10^{-7}$	0.027
50	10^4	10^{-4}	$3.41 \cdot 10^{-5}$	$4.54 \cdot 10^{-5}$	0.00692
		10^{-5}	$5.22 \cdot 10^{-6}$	$6.97 \cdot 10^{-6}$	0.00106
		10^{-6}	$3.64 \cdot 10^{-7}$	$4.86 \cdot 10^{-7}$	$7.42 \cdot 10^{-5}$
	10^5	10^{-4}	$3.41 \cdot 10^{-5}$	$4.54 \cdot 10^{-5}$	0.0679
		10^{-5}	$5.22 \cdot 10^{-6}$	$6.97 \cdot 10^{-6}$	0.0107
		10^{-6}	$3.64 \cdot 10^{-7}$	$4.86 \cdot 10^{-7}$	0.00075
	10^6	10^{-4}	$3.41 \cdot 10^{-5}$	$4.54 \cdot 10^{-5}$	0.505
		10^{-5}	$5.22 \cdot 10^{-6}$	$6.97 \cdot 10^{-6}$	0.102
		10^{-6}	$3.64 \cdot 10^{-7}$	$4.86 \cdot 10^{-7}$	0.00748

For a fixed participation of $\pi = 0.75$, and various table sizes N, we compute the corresponding table-failure rate $\bar{t}$, and compare it to the record-failure rates $\bar{r}$ and $\bar{r}'$.

rates $\bar{p}$ logarithmically ranging from 1, for which trivially $n_{\min} = k$, down to 10^{-8}, with fixed participation $\pi = 0.5$, and target anonymity $k = 50$.

As a reference, consider that a confidence interval of 5 standard deviations around the mean contains 99.9999427% of the probability mass of a normal distribution, leaving a chance of 1 in 1.74 million of being on either of the two tails outside the interval, or a chance of 1 in 3.49 million of being on a given tail. The latter probability, roughly $2.87 \cdot 10^{-7}$, is the p-value standardly used in physics to accept the discovery of a particle, such as the Higgs boson. The "six sigma" high-quality manufacturing standard —somewhat of a misnomer, as it really corresponds to 4.5 standard deviations— tolerates a percentage of defects of 0.00034% or $3.40 \cdot 10^{-6}$. A striking observation is that quite conveniently, extremely demanding values of $\bar{p}$ did not impose prohibitive values of $n_{\min}$. The plot suggests that $n_{\min}$ rapidly grows from k as $\bar{p}$ departs from the trivial acceptance of a 100% of failures, and then continues to grow very slowly, even over a logarithmically scaled $\bar{p}$.

D. Verification of the Approximations for Small $\bar{p}$ and Large $n_{\min}$

Next, we verified the main approximations of §VI.C numerically. Our first graph, in Fig. 8, verifies the approximation to the effective anonymity $n_{\min}$ based on the method of types and the KL divergence between $\kappa = \frac{k-1}{n_{\min}}$ and the probability of participation π. On the one hand, we computed $n_{\min}$ with Algorithm A, and then numerically solved the equation $-\frac{1}{\tilde{n}} \ln \bar{p} = \mathrm{D}(\frac{k-1}{\tilde{n}} \| \pi)$ in the approximation variable $\tilde{n}$ with Brent's method (implemented in Matlab with `fzero`), for $\pi = 0.5$ and $\bar{p} = 10^{-6}$. The validity of this approximation, along with the quadratic, explicit forms derived from it, is tested in Fig. 9, in the form of a relative error of each approximation with respect to the reference computed with

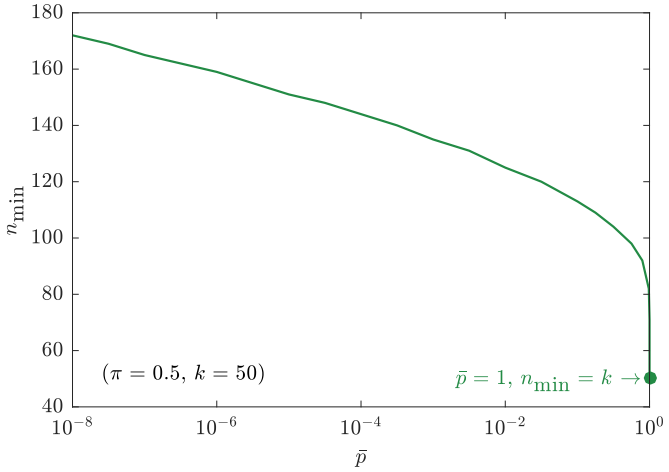


Fig. 7. Effective anonymity $n_{\min}$ versus acceptable cell-failure probability $\bar{p}$. Quite strikingly, extremely demanding (low) values of $\bar{p}$, even by engineering standards of the highest quality such as those used in "six sigma" design ($3.40 \cdot 10^{-6}$), or in hypothesis testing in particle physics ($2.87 \cdot 10^{-7}$), do not impose prohibitive (high) values of $n_{\min}$.

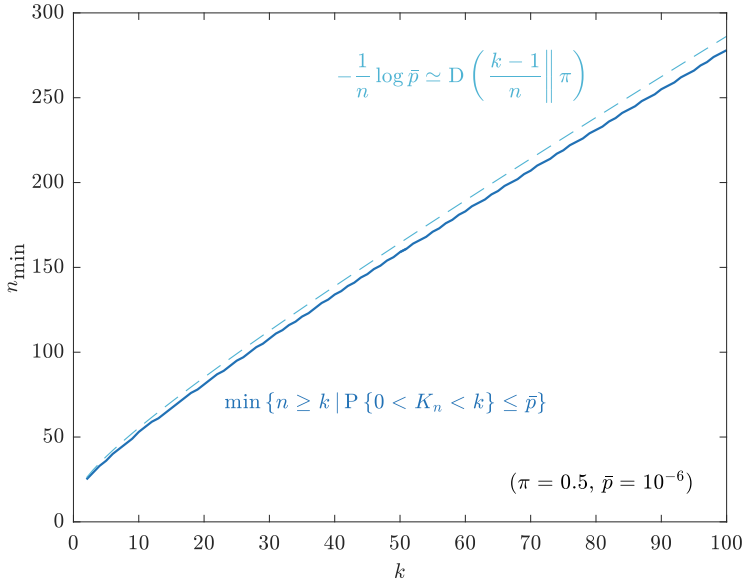


Fig. 8. Numerical verification of the approximation to the effective anonymity $n_{\min}$ in terms of the target anonymity k, for a participation probability $\pi = 0.5$, and an acceptable cell failure $\bar{p} = 10^{-6}$.

our recursive algorithm. Interestingly, the approximation based on the theoretical limit $\mathrm{D}(\frac{k-1}{n}\|\pi) \rightarrow -\log\bar{\pi}$ required such small values of $\bar{p}$ that it proved of little use for any practical purpose. We opted for the divergence form of the approximation to $-\frac{1}{n}\ln\bar{p}$, which exhibits the dependence with k, not negligible when compared to actual values of $n_{\min}$.

The logarithmic approximations for $k = 2$ are verified in Fig. 10. According to these results, the ceiling of the logarithmic approximation based on a first-order Taylor expansion is an upper bound, whereas the one based on the second-order expansion is optimistic but more often accurate.

E. Microaggregation of Synthetic and Standardized Datasets

Last but not least, we microaggregated two datasets with the MDAV algorithm, setting the minimum microcell size to the effective anonymity parameter $n_{\min}$ in lieu of the target anonymity k. Each dataset contains a total of $N = 10^4$ records. The first dataset consists of zero-mean, unit-variance, independent Gaussian vectors of dimension $m = 10$. The second is a uniformly random subsampling without replacement of an extension of the standardized dataset "Census", called "Large Census", employed in [41], originally containing 149 642 records with $m = 13$ numerical attributes, taken here as quasi-identifiers. The subsampled standardized dataset, which we call "Census 10k", was generated once and reused in all experiments.

We assumed an identical participation probability of π, but considered the traditional, deterministic case of $\pi = 1$ as well, for the purposes of comparison. The experiments, reported in Fig. 11 and Fig. 12, respectively, were designed to investigate the behavior of the normalized distortion metric $\mathcal{D}$ for p-probabilistic k-anonymous microaggregation defined in §VI.B, which for identical participation amounts to the distortion of traditional microaggregation with $n_{\min}$ in lieu of k, as a function of the target anonymity k, the participation probability π, and the acceptable cell-failure probability $\bar{p}$.

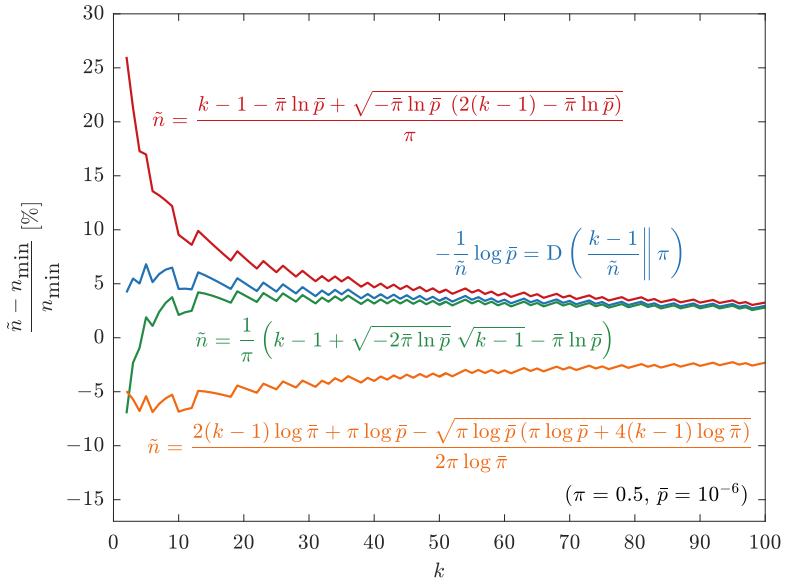


Fig. 9. Relative error in the approximations, based on the KL divergence and the quadratic variants, to the effective anonymity $n_{\min}$ as a function of the target anonymity k, for a participation $\pi = 0.5$ and an acceptable cell-failure $\bar{p} = 10^{-6}$. The jagged shape of the lines owes to the fact that the reference $n_{\min}$ is a positive integer, whereas the approximations $\tilde{n}$ are double-precision numbers.

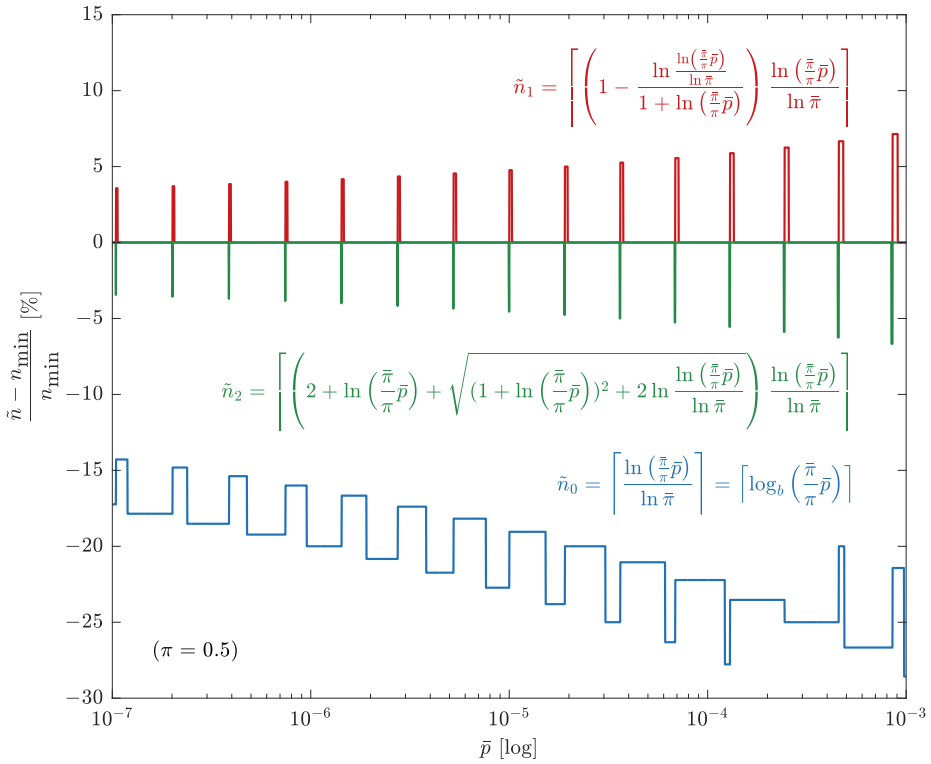


Fig. 10. Relative error in the logarithmic approximations to $n_{\min}$ as a function of $\bar{p}$ ranging from 10^{-7} to 10^{-3}, with exponent steps of 0.001 to highlight the pulsating shape of the error. In this case, we use the ceiling function to convert the double-precision approximation $\tilde{n}$ into an integer.

For both datasets, the introduction of uncertain participation substantially increased the resulting distortion $\mathcal{D}$, as one might have anticipated, gradually for high participation π at first, as we depart from certainty, but more sharply later for very low participation and high uncertainty. Far more surprising is the lack of sensitivity of $\mathcal{D}$ with respect to the cell-failure rate $\bar{p}$. Because in p-probabilistic k-anonymous microaggregation $n_{\min}$ effectively plays the role of k, the slow growth of $\mathcal{D}$ with k for deterministic microaggregation manifests as a slow growth of $\mathcal{D}$ with $n_{\min}$. Additionally, as we saw earlier, $n_{\min}$ also increases slowly with $\bar{p}$. Putting these observations together explains the extremely convenient, slow increase of distortion $\mathcal{D}$ as we impose extremely low cell-failure rates $\bar{p}$, even on a logarithmic scale. Of course, the practical significance of this observation is that we may enjoy an extremely robust probabilistic design with a manageable price in distortion.

A final set of experiments are aimed to emphasize the difference between traditional k-anonymous microaggregation and our proposal, p-probabilistic k-anonymous microaggregation. Firstly, in Fig. 13(a) we compute the distortion

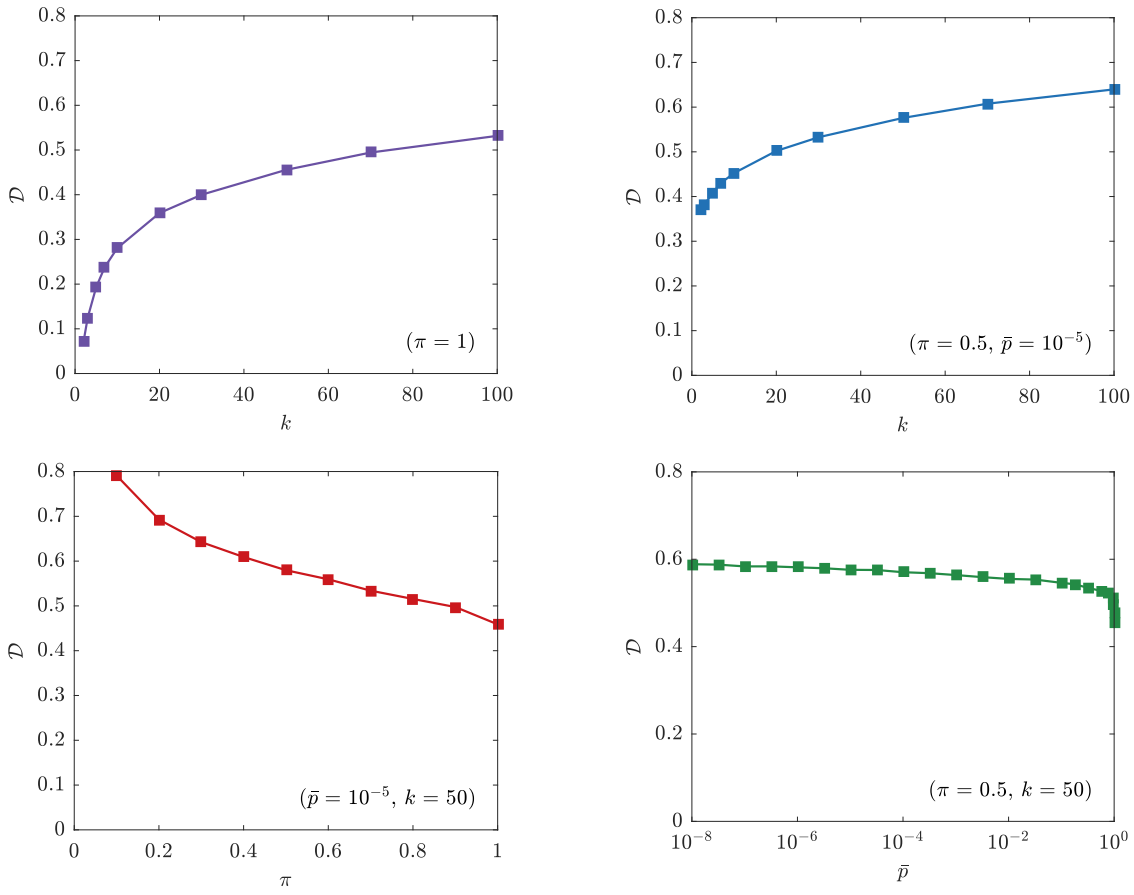


Fig. 11. Gaussian dataset, zero-mean, unit-variance, independent samples, $N = 10^4$ records, $m = 10$ quasi-identifiers.

$\mathcal{D}(\pi = 0.75)$ for a probability of participation $\pi = 0.75$ and an acceptable failure rate $\bar{p} = 10^{-5}$, relative to the distortion in the traditional case, $\mathcal{D}(\pi = 1)$, and plot the quotient as a function of the target anonymity parameter k. We employ the Gaussian dataset, with zero-mean, unit-variance, independent samples, arranged as $N = 10^4$ records, and $m = 2, 5, 10$ quasi-identifiers. The most glaring effect is the price in distortion that must be paid to ensure k-anonymity, particularly for lower values of k. Interestingly, as either k or the dimension m increases, the relative increment in distortion is reduced.

Secondly, also as a though experiment to emphasize the necessity of p-probabilistic k-anonymous microaggregation over its traditional counterpart, suppose we employ traditional microaggregation to construct cells of size $n = k$, completely disregarding that some of the records could be inactive and k-anonymity thus violated. Specifically, for $n = k$, the attained rate of cell failure would be $\bar{q} = 1 - \mathrm{P}\{K_k = 0 \text{ or } K_k = k\} = 1 - (\bar{\pi}^k + \pi^k)$, plotted in Fig. 13(b). By symmetry, the worst failure rate corresponds to $\pi = {}^1\!/_2$, namely $\bar{q} = 1 - 2^{1-k}$. Even for the smallest possible k, any practical requirement of the form $\bar{q} \leqslant \bar{p} \ll 1$ would translate into a participation rate $\pi \approx 1$ (or, theoretically, $\pi \approx 0$).

VIII. Conclusion

We contend that by relaxing the trust assumptions imposed on the users of a system, the consequent increase in users willing to provide additional data may very well represent a far greater gain in utility than that from algorithmic improvements alone. In this spirit, this paper develops a probabilistic variant of k-anonymous microaggregation, which we term p-probabilistic, resorting to a statistical model of respondent participation in order to aggregate quasi-identifiers in such a manner that k-anonymity is concordantly enforced with a probabilistic guarantee. The users themselves may perturb their own quasi-identifiers according to a verifiable, predetermined microaggregation function, without imposing the requirement that they completely trust an external data anonymizer. This may vastly broaden the potential range of applications of k-anonymous microaggregation in particular, and statistical disclosure control in general.

In our theoretical analysis, we formally view p-probabilistic k-anonymity as microaggregation with an effective anonymity parameter $n_{\min} \geqslant k$ leading to a probability $\bar{p}$ of k-anonymity violation at the microcell level. For a statistical model of participation given by a series of participation probabilities π_n, we present a recursive algorithm for the efficient calculation of the effective anonymity $n_{\min}$, with great numerical precision. We also develop a number of metrics associated with the probabilistic violation of k-anonymity, including the expected number u of unprotected records, the attained record-failure rate $\bar{r}$, the related probability $\bar{r}'$ that a participating respondent will not be successfully protected, and the failure rate $\bar{t}$ at the table level, establishing a number of bounds and approximations on all of them.

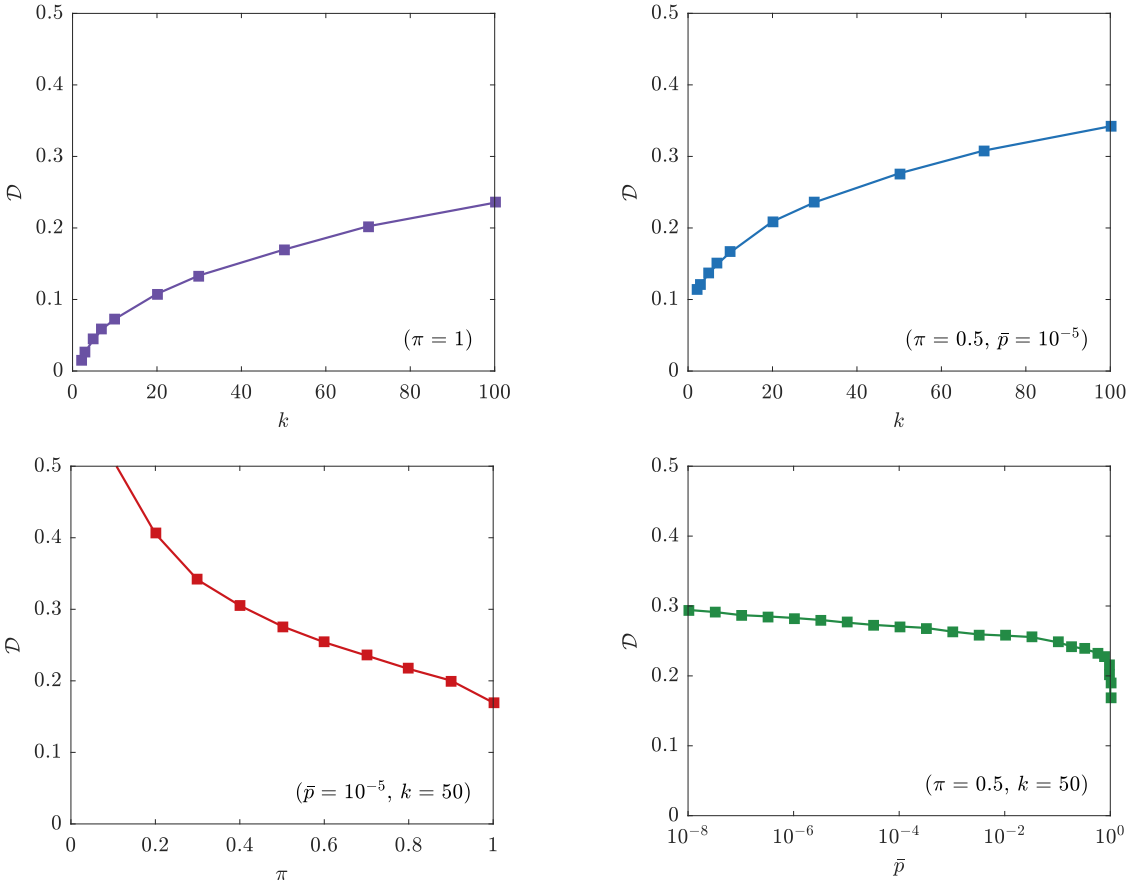
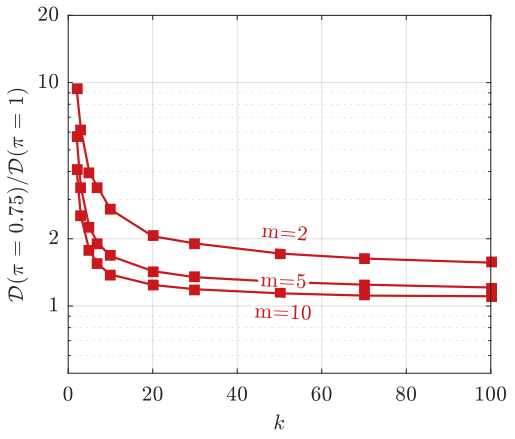
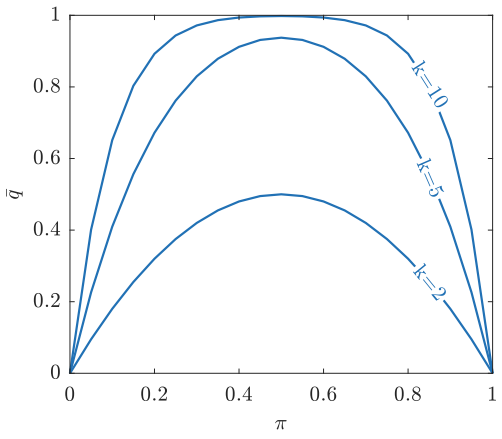


Fig. 12. "Census 10k" dataset, $N = 10^4$ records, $m = 13$ quasi-identifiers.



(a) Distortion $\mathcal{D}(\pi = 0.75)$ for a probability of participation $\pi = 0.75$ and failure rate $\bar{p} = 10^{-5}$, relative to the traditional distortion $\mathcal{D}(\pi = 1)$, both computed for the Gaussian dataset, with i.i.d. samples, $N = 10^4$ records, and $m = 2, 5, 10$ quasi-identifiers.



(b) We employ traditional microaggregation to construct cells of size $n = k$, disregarding the fact that some records may not be active and thus k-anonymity violated, resulting in a cell-failure rate $\bar{q}$ prohibitively large.

Fig. 13. Additional experiments to emphasize the difference between p-probabilistic k-anonymous microaggregation and its traditional counterpart.

Some of our approximate characterizations draw upon the method of types, a powerful technique in large deviation theory lying at the heart of the intersection between information theory and statistics, which enables to provide substantive insight into the anonymity properties of our microaggregation model. In addition, we propose a specific distortion metric, naturally derived from SSE for traditional k-anonymous microaggregation.

Extensive numerical analysis and experimentation confirms and illustrates the theoretical analysis, further revealing that the impact in distortion due to the cautious provision of cells larger than those required by traditional microaggregation is quite manageable. The reason is the combination of two dampening effects. First, the slow increase in effective anonymity $n_{\min}$ with even extremely low cell-failure rates $\bar{p}$, and secondly, the slow growth of distortion $\mathcal{D}$ with $n_{\min}$. This offers a convenient answer to the fundamental question of whether the gain in utility due to the availability of

additional data —in turn due to a relaxation of the trust model— might be diminished by the inherently larger cell sizes of our probabilistic microaggregation model.

We must recognize that the necessarily limited scope of this work is but a humble portion of the theoretical and practical extent to which our proposal of p-probabilistic k-anonymous microaggregation may reach. The compelling theoretical results developed and the promising experimental outcomes observed provide ample encouragement to continue building upon the preliminary work carried out here. Concrete directions primarily include the study of the untrusted mode of operation, in which cells must be defined as the Cartesian product of intervals or simple collections of quasi-identifier values, the specification of complex microaggregation functions in the locally trusting mode of operation analyzed here, and more sophisticated statistical models of participation.

Acknowledgment

The authors gratefully acknowledge the invaluable assistance of Irene Carrión Barberà, M.D., in the preparation of the example on Hashimoto's thyroiditis illustrating the background section. We would also like to thank the anonymous reviewers for their immensely helpful suggestions to improve the readability and contents of this paper.

This manuscript presents some of the results developed through the collaboration of the Technical University of Catalonia (UPC) and Scytl Secure Electronic Voting (Scytl) in the context of the project "Data-Distortion Framework (DDF)", and in accordance with the guidelines therein. This work is thus partly supported by the Spanish Ministry of Industry, Energy and Tourism (MINETUR) through the "Acción Estratégica Economía y Sociedad Digital (AEESD)" funding plan, through the aforementioned project, ref. TSI-100202-2013-23.

Additional funding supporting this work has been granted to UPC by the Spanish Ministry of Economy and Competitiveness (MINECO) through the "Anonymized Demographic Surveys (ADS)" project, ref. TIN2014-58259-JIN, under the funding program "Proyectos de I+D+i para Jóvenes Investigadores", and through the project "INRISCO", ref. TEC2014-54335-C4-1-R, as well as by the Government of Catalonia, under grant 2014 SGR 1504.

References

[1] M. Abramowitz and I. A. Stegun, Eds., *Handbook of mathematical functions: With formulas, graphs, and mathematical tables*. New York, NY: Dover Publ., 1965.

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

[3] J. Cao, B. Carminati, E. Ferrari, and K. Tan, "CASTLE: Continuously anonymizing data streams," *IEEE Trans. Depend., Secure Comput.*, vol. 99, 2009.

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

[5] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY: John Wiley & Sons, 1991.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[22] 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.

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

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

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

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

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

[28] J. Parra-Arnau, D. Rebollo-Monedero, and J. Forné, "Measuring the privacy of user profiles in personalized information systems," *Future Gen. Comput. Syst.*, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.01.001

[29] D. Rebollo-Monedero and J. Forné, "Optimal query forgery for private information retrieval," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4631–4642, 2010. [Online]. Available: http://dx.doi.org/10.1109/TIT.2010.2054471

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

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

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

[33] D. Rebollo-Monedero, J. Forné, E. Pallarès, J. Parra-Arnau, C. Tripp, L. Urquiza, and M. Aguilar, "On collaborative anonymous communications in lossy networks," *Secur., Commun. Netw., Special Issue Secur. Completely Interconn. World*, 2013. [Online]. Available: http://dx.doi.org/10.1002/sec.793

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

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

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

[37] D. Rebollo-Monedero, J. Parra-Arnau, and J. Forné, "Optimizing the design parameters of threshold pool mixes for anonymity and delay," *Comput. Netw.*, vol. 67, pp. 180–200, Jul. 2014. [Online]. Available: http://dx.doi.org/10.1016/-j.comnet.2014.04.007

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

[39] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, no. 4, pp. 656–715, Oct. 1949.

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

[41] M. Solé, V. Muntés-Mulero, and J. Nin, "Efficient microaggregation techniques for large numerical data volumes data volumes," *Int. J. Inform. Secur.*, vol. 11, pp. 253–267, Aug. 2012.

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

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

[44] 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.

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

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

[47] S. Zhong, Z. Yang, and T. Chen, "*k*-Anonymous data collection," *Inform. Sci.*, vol. 179, no. 172, pp. 2948–2963, Aug. 2009.