WCCI 2012 IEEE World Congress on Computational Intelligence
June, 10-15, 2012 - Brisbane, Australia FUZZ IEEE

Probabilistic k-Anonymity through Microaggregation and Data Swapping

Jordi Soria-Comas and Josep Domingo-Ferrer
UNESCO Chair in Data Privacy
Dept. of Computer Engineering and Maths
Universitat Rovira i Virgili
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
E-mail {jordi.soria,josep.domingo}@urv.cat

***Abstract*—k-Anonymity is a privacy property used to limit the risk of re-identification in a microdata set. A data set satisfying k-anonymity consists of groups of k records which are indistinguishable as far as their quasi-identifier attributes are concerned. Hence, the probability of re-identifying a record within a group is $1/k$. We introduce the probabilistic k-anonymity property, which relaxes the indistinguishability requirement of k-anonymity and only requires that the probability of re-identification be the same as in k-anonymity. Two computational heuristics to achieve probabilistic k-anonymity based on data swapping are proposed: MDAV microaggregation on the quasi-identifiers plus swapping, and individual ranking microaggregation on individual confidential attributes plus swapping. We report experimental results, where we compare the utility of original, k-anonymous and probabilistically k-anonymous data.**

***Index Terms*—Computational intelligence; anonymization; clustering; microaggregation; swapping; statistical disclosure control; k-anonymity; differential privacy.**

I. Introduction

A microdata file is composed of records that contain information specific to individuals (who may be citizens, companies, etc.) in the data set. These records contain, for each specific individual, the values corresponding to a list of attributes. Microdata files are the result of data collection processes carried out by national statistical offices, healthcare systems, electronic commerce, etc. They are a valuable resource for analysts and researchers, but also a threat to the individuals' privacy. Direct publication of microdata files results in an unacceptable privacy breach for the individuals therein contained. Therefore, before being released, microdata files must undergo a process of anonymization that dissociates the identity of the individuals from specific records.

The two main aspects that any anonymization method must address are disclosure risk and information loss. There is an extensive literature about methods used to provide anonymity for microdata releases. Some good surveys on microdata anonymization are [1], [2], [3].

Among the several approaches to disclosure risk limitation for microdata files, we focus on k-anonymity. This is not really an anonymization method, but a privacy property that the published data set must satisfy. If k-anonymity is judged to be a sufficient guarantee for the privacy of the individuals in the data set, then the focus goes to the selection of a method that produces a data set satifying k-anonymity with minimal information loss.

In a k-anonymous microdata set, for each combination of values of the quasi-identifier attributes present in the data set, there must be at least k records sharing that combination. In other words, a record must be indistinguishable within a set of k records as far as their quasi-identifier attributes are concerned. To fulfill this requirement, the data granularity of the quasi-identifiers is reduced, usually by generalization, suppression or micro-aggregation. The strict indistinguishability requirement of k-anonymity may lead to a substantial amount of information loss, especially if there is a large number of quasi-identifiers [4].

Our goal is to achieve the same level of disclosure risk limitation that k-anonymity provides, while improving the data quality of the released data set. Our proposal is based on a relaxation of the indistinguishability requirement of k-anonymity. Instead of requiring records to be indistinguishable within sets of k records in terms of the quasi-identifiers, we focus on the probability of re-identification. By requiring this probability to be at most $1/k$, we achieve the same level of protection against re-identification provided by k-anonymity, but the range of applicable methods is wider and hence the information loss can be reduced.

A. Contribution and plan of this paper

We introduce in this paper the concept of probabilistic k-anonymity, which like k-anonymity yields a re-identification probability at most $1/k$ but with much better data quality preservation. This is especially relevant when dealing with a data set that contains many quasi-identifier attributes.

Section II introduces some background concepts that are required for later sections. Section III presents probabilistic k-anonymity. Section IV describes a computational procedure based on microaggregation and swapping to achieve probabilistic k-anonymity. Experimental results comparing the data quality loss caused by standard k-anonymity and probabilistic k-anonymity are reported in Section V. Conclusions and future research are summarized in Section VI.

II. Background

A microdata set can be modeled as a table where each row refers to a different individual and each column contains information regarding one of the attributes collected. We use the notation $T(A_1, \ldots, A_n)$ to denote a microdata set with information about attributes $A_1, \ldots, A_n$.

The attributes in a microdata set are usually classified in the following non-exclusive categories according to the sensitiveness of the information they convey and the risk of record re-identification they imply:

- *Identifiers*. An attribute is an identifier if it provides unambiguous re-identification of the individual to which the record refers. Some examples of identifier attributes are the social security number, the passport number, etc. If a record contains an identifier, any sensitive information contained in other attributes may immediately be linked to a specific individual. To avoid direct re-identification of an individual, identifier attributes are usually removed or encrypted. We assume in the rest of this paper that the microdata set $T(A_1, \ldots, A_n)$ does not contain any identifier attribute.
- *Quasi-identifiers*. Unlike an identifier, a quasi-identifier attribute alone does not lead to record re-identification. However, in combination with other quasi-identifier attributes, it may allow unambiguous re-identification of some individuals. For example, [5] shows that 87% of the population in the U.S. can be unambiguously identified by combining a 5-digit ZIP code, birth date and sex. Removing quasi-identifier attributes, as proposed for the identifiers, is not possible, because quasi-identifiers are required to perform any useful analysis on the data. Moreover, any attribute is potentially a quasi-identifier, depending on the external information available to the intruder; hence, to make sure all quasi-identifiers have been removed, one should remove all attributes (!).
- *Confidential attributes*. Confidential attributes hold sensitive information on the individuals that took part in the data collection process (*e.g.* salary, health condition, sex orientation, etc.). The primary goal of microdata protection techniques is to prevent intruders from learning confidential information about a specific individual. This goal involves not only preventing the intruder from determining the exact value a confidential attribute takes for some individual, but preventing inferences on the value of that attribute (like bounding it).
- *Non-confidential attributes*. Non-confidential attributes are those that do not belong to any of the previous categories. As they do not contain sensitive information about individuals and cannot be used for record re-identification, they do not affect our discussion on disclosure limitation for microdata sets. We assume for the rest of the paper that none of the attributes in $T(A_1, \ldots, A_n)$ belongs to this category.

When publishing a microdata file, the data collector must guarantee that no sensitive information about specific individuals is disclosed. To do so, the data collector does not publish the original microdata set $T(A_1, \ldots, A_n)$, but a modified version $T'(A_1, \ldots, A_n)$ where the quasi-identifiers and/or the confidential attributes have been masked. Disclosure can be classified in two categories [6]:

- *Identity disclosure*. The intruder is able to determine the true identity of the individual to which a record in the microdata file corresponds; the intruder can subsequently associate to this individual the values of the confidential attributes for that record.
- *Attribute disclosure*. Even if identity disclosure does not happen, it may be possible for an intruder to infer some information for a specific individual based on the published microdata set. For example, imagine that the salary is one of the confidential attributes and the job is a quasi-identifier attribute; if an intruder is interested in a specific individual whose job he knows to be "accountant" and there are several accountants in the data set (including the target individual), the intruder will be unable to re-identify the individual's record based only on her job, but he will be able to lower-bound and upper-bound the individual's salary (which lies between the minimum and the maximum salary of accountants in the data set).

An intruder who wants to re-identify an individual usually exploits some external information to perform a record linkage attack. A record linkage attack tries to link the records in an external non-anonymous data set back to the records in the published data set, thereby associating an identity to them. Assume that the intruder knows that some individual is in the published microdata set, and also knows a set of quasi-identifier attributes regarding this individual. To perform the record linkage attack, the intruder tries to match the quasi-identifier attributes he knows to some record in the published data set. If the intruder performs the right linkage to the published data set, the attack succeeds, and the intruder learns the confidential attribute values associated to that individual.

A possible approach towards avoiding identity disclosure is the one taken by k-anonymity, where each record in the published microdata set is made indistinguishable within a set of k records based on the quasi-identifiers. This way an intruder with access to an external non-anonymous data set that contains the quasi-identifiers in $T(A_1, \ldots, A_n)$ is unable to perform a re-identification of the records in the published data set. Given a specific individual in the external data set, the intruder can at most determine a set of k records in the published data set that must contain that individual. The original proposal to achieve k-anonymity [7] was based on generalization and suppression of the information contained in the quasi-identifier attributes. Another proposal is based on micro-aggregation ([8], [9]).

k-Anonymity does not in general protect against attribute disclosure. If all the individuals within a group of indistinguishable records have the same value for a confidential attribute, then an intruder will learn the value of that confidential attribute for those individuals, even without re-identification.

Further refinements on k-anonymity that try to address attribute disclosure have been proposed: l-diversity [10] requires the presence of l different values for the confidential attribute in every group of records sharing the same quasi-identifier values; t-closeness [11] requires the distribution of the confidential attribute in any group of records sharing the same quasi-identifier values to be close to its distribution in the overall data set.

To provide an accurate definition of k-anonymity, we first formalize and slightly generalize the definition of quasi-identifier. Usually a quasi-identifier is said to be a group of attributes that can be employed to unambiguously identify an individual. We note that any combination of attributes that may provide a level of re-identification beyond the admissible limits set by the data collector should also be treated as a quasi-identifier. For example, if we want to guarantee k-anonymity, any combination of externally available attributes that may be used to refer to a set containing less than k records must be considered as a quasi-identifier.

Definition 1 (Quasi-identifier)**.** A quasi-identifier QI of T is a subset of the set of attributes $\{A_1, \dots, A_n\}$ that satisfy the following two conditions: (i) the attributes in QI are available in an external, non-anonymous data set; (ii) the values of the attributes in QI may allow an intruder to determine the identity corresponding to a record in the published microdata set beyond an admissible level.

The goal of k-anonymity is to cloak the identity corresponding to a record by making each record indistinguishable within a set of k records. In other words, given a record in an external non-anonymous data set, an intruder must not be able to link it with certainty to a set of records in the published data set with cardinality less than k. This determines the criterion used by k-anonymity to define what is an admissible level of re-identification, and hence what is a k-anonymous quasi-identifier.

For k-anonymity to provide the desired level of protection, an intruder must not be able to link an external record in a non-anonymous data set to a group of less than k records in the published data set, *no matter the quasi-identifier used*. Note that, by adding more attributes to a quasi-identifier, we increase the level of certainty that the intruder may get. Therefore, to achieve protection against all possible quasi-identifiers $QI_1, \dots, QI_m$, it suffices to achieve k-anonymity for the quasi-identifier that results from the union $QI_1 \cup \dots \cup QI_m$.

Definition 2 (k-Anonymity [7])**.** A microdata set $T'(A_1, \dots, A_n)$ is said to satisfy k-anonymity if, for each record $t \in T'$, there are at least $k-1$ other records sharing the same values for all the quasi-identifier attributes.

In order to apply the previous definition of k-anonymity, the data collector needs to know which attributes are available externally in a non-anonymous data set. Since assuming such knowledge by the data collector is a strong assumption, we will consider only two scenarios:

1) Uninformed intruder scenario. The intruder does not know the value that any confidential attribute takes for any individual in the data set.
2) Informed intruder scenario. An informed intruder may know some of the confidential attributes for some of the individuals. This may happen, for example, if the intruder is acquainted with an individual that is included in the microdata set. If the confidential attributes known by the intruder were not deemed quasi-identifiers, the intruder might exploit his knowledge to obtain a more accurate re-identification.

There may be multiple informed intruders, each of them knowing a different subset of confidential attributes over a different subset of records. For the informed intruder scenario we assume that the number of intruders and confidential attributes is the same, and that each of the intruders knows the values of all confidential attributes for all individuals, except for one confidential attribute whose values are completely unknown to the intruder for all individuals. We also assume that the intruders do not collude, as in that case they would be able to learn all the confidential attributes for all individuals even without seeing the published microdata set. Our intruders are not the strongest possible ones: a stronger intruder would be one with total knowledge of all confidential attributes except one, *and partial knowledge* of the remaining confidential attribute (whose values would be known to the intruder for some individuals). However, we judge the proposed intruders to be reasonably strong.

III. Probabilistic k-anonymity

k-Anonymity guarantees that, for any combination of values of quasi-identifier attributes in the published microdata set $T'(A_1, \dots, A_n)$, there are at least k records sharing that combination of values. Therefore, given an individual in an external non-anonymous data set, the probability of performing the right linkage back to the corresponding record in the published microdata set, and thus the probability of learning its confidential attributes, is at most $1/k$. It is in this sense that probabilistic k-anonymity is defined.

A similar relaxation on the notion of k-anonymity was presented in [12], which partitioned the dataset and applied a permutation inside each of the partition components. We do the same in Section IV to achieve probabilistic k-anonymity. However, ours is a more general framework, not limited to permutations (even if permutations are convenient choice to simplify probability calculations). Moreover [12] was limited to a single confidential attribute, whereas we handle multiple confidential attributes that can also be quasi-identifiers.

Definition 3 (Probabilistic k-anonymity)**.** Let $T'(A_1, \dots, A_n)$ be a published data set generated from an original data set $T(A_1, \dots, A_n)$ using an anonymization mechanism M. The data set T' is said to satisfy probabilistic k-anonymity if, for any non-anonymous external data set E, the probability for an intruder I knowing T', M and E to correctly link any record $x \in E$ and its corresponding record (if any) in T' is at most $1/k$.

Note than any method used to achieve k-anonymity also leads to probabilistic k-anonymity. In this sense, it may be said that k-anonymity provides a stronger guarantee. However, from the point of view of the probability of re-identification, both provide the same level of protection.

The advantage of probabilistic k-anonymity in comparison to k-anonymity is that, by relaxing the requirements on the indistinguishability within groups of k records, the range of eligible methods to enforce it is wider, and therefore we may expect a reduction in the information loss.

As probabilistic k-anonymity is expressed in terms of probability of re-identification, it is natural to think of the released data set $T'(A_1, \ldots, A_n)$ as a perturbation of $T(A_1, \ldots, A_n)$. We use the notations in Figure 1. The records x_i in T have been split in two parts: the quasi-identifier attributes qi_i, and the confidential attributes c_i. The records in T' are obtained by applying a random perturbation to the corresponding record in T: $x'_i = X(x_i)$. This perturbation affects only the quasi-identifier attributes. For the sake of simplicity, we assume that the released records in T' correspond to the first $|T'|$ records in T. If $|T| = |T'|$, then all the records are released. The data set E links the quasi-identifiers qi_i to the identifier id_i. The functions Id and Rid assign a record in T' to the records in E, thus performing the re-identification of the records in T'. The function Rid is the re-identification function used by the intruder, while Id is assumed to be the correct re-identification function. If there is no record in T' corresponding to the identity (*i.e.* the identified record) $e_i \in E$, then Id returns the empty set.

The goal of probabilistic k-anonymity is to limit the probability of performing the right linkage to at most $1/k$. With the above notations this requirement can be stated as: for all $e_i \in E$ and for all $Rid()$

$$P(Rid(e_i) = Id(e_i)) \leq \frac{1}{k}$$

This formula catches the essence of the definition of probabilistic k-anonymity: the probability of performing the right re-identification is not greater than $1/k$. However, by having the intruder use any possible function $Rid()$ to perform the re-identification, the details on how a rational intruder will proceed are hidden. Given a record e_i, a rational intruder selects the record x_r in T' that has the greatest probability given the knowledge of T', E and M. The following examples will clarify how a rational intruder acts. All examples assume that E contains identities for all records in T, which is the best possible knowledge that an intruder can have.

TABLE I
DATA SETS IN EXAMPLE 4

T	T'	E
$x_1 = (qi_1, c_1)$ $x_2 = (qi_2, c_2)$	$x'_1 = (qi'_1, c_1)$	$e_1 = (qi^E_1, id_1)$ $e_2 = (qi^E_2, id_2)$

TABLE II
DATA SETS IN EXAMPLE 5

T	T'	E
$x_1 = (qi_1, c_1)$ $x_2 = (qi_2, c_2)$	$x'_1 = (qi'_1, c_1)$ $x'_2 = (qi'_2, c_2)$	$e_1 = (qi^E_1, id_1)$ $e_2 = (qi^E_2, id_2)$

Example 4. Let us assume that T contains two records, and that only the first one is included in the anonymized data set. This situation is shown in Table I. From the intruder's point of view, x'_1 corresponds to either the individual in e_1 or e_2. The best the intruder can do is to select the one that has the greatest probability given the knowledge of T', E, and the mechanism M used to generate T' from T.

The probability that x'_1 corresponds to e_i equals the probability of obtaining qi'_1 from qi^E_i, over the total probability of obtaining qi'_1 from any other record in E:

$$P(X'(qi^E_i) = qi'_1 | T', E, M)$$

$$= \frac{P(X'(qi^E_i) = qi'_1 | M)}{\sum_{(qi^E_j, id_j) \in E} P(X'(qi^E_j) = qi'_1 | M)}$$

The intruder selects e_1 as his guess if $P(X'(qi^E_1) = qi'_1 | T', E, M) \geq P(X'(qi^E_2) = qi'_1 | T', E, M)$, and e_2 otherwise.

In the previous example we have seen that, given a record in E, the linkage is performed to the record in T' that has greatest probability. If that probability is smaller than $1/k$, then the probability of performing the right linkage will also be smaller than $1/k$, as any other linkage will indeed result in a yet smaller probability. Therefore, to achieve probabilistic k-anonymity, we must have for all $qi^E \in E$ and all $qi' \in T'$

$$P(X'(qi^E) = qi' | T', E, M) \leq \frac{1}{k} \qquad (1)$$

Example 5. In this example the amount of information in T' has been increased, by adding the record x'_2. The new data sets are shown in Table II. As E is assumed to exactly contain the identities for the individuals in T, the intruder knows that if one identity in E corresponds to a specific record in T', the other identity in E must correspond to the other record in T'. This must be taken into account when computing the probabilities. For example, the probability $P(X'(qi^E_1) = qi'_1 | T', E, M)$ that qi^E_1 corresponds to qi'_1 equals $P(X'(qi^E_1) = qi'_1, X'(qi^E_2) = qi'_2 | T', E, M)$, which can be computed as

$$\frac{P(X'(qi^E_1) = qi'_1, X'(qi^E_2) = qi'_2 | M)}{\sum_{\{i,j\} = \{1,2\}} P(X'(qi^E_i) = qi'_1, X'(qi^E_j) = qi'_2 | M)}$$

The next example shows how the correct re-identification probability would be computed in the most general case.

Example 6. Assume data sets T, T' and E as in Table III. Contrary to Example 5, fixing a correspondence between a record in T' and a record in E does not completely fix the rest of the correspondences. We still have to consider all the possible combinations. The probability $P(X'(qi^E_1) = qi'_1 | T', E, M)$ that qi^E_1 corresponds to qi'_1 equals $\sum P(X'(qi^E_1) = qi'_1, X'(qi^E_{i_2}) =$

$$\begin{array}{ccccc} T(A_1,\dots,A_n) & \overset{X'}{\longrightarrow} & T'(A_1,\dots,A_n) & \overset{Id,\,Rid}{\longleftarrow} & E \\ \hline x_1=(qi_1,c_1) & & x'_1=(qi'_1,c_1) & & e_1=(qi_1,id_1) \\ \vdots & & \vdots & & \vdots \\ x_{|T|}=(qi_{|T|},c_{|T|}) & & x'_{|T'|}=(qi'_{|T'|},c_{|T'|}) & & e_{|E|}=(qi_{|E|},id_{|E|}) \end{array}$$

Fig. 1. Notations for probabilistic k-anonymity

TABLE III
DATA SETS IN EXAMPLE 6

T	T'	E
$x_1 = (qi_1, c_1)$ $\vdots$ $x_N = (qi_N, c_N)$	$x'_1 = (qi'_1, c_1)$ $\vdots$ $x'_M = (qi'_M, c_M)$	$e_1 = (qi^E_1, id_1)$ $\vdots$ $e_N = (qi^E_N, id_N)$

TABLE IV
DATA SETS IN THE UNINFORMED INTRUDER SCENARIO

T	T'	E
$x_1 = (qi_1, c_1)$ $\vdots$ $x_N = (qi_N, c_N)$	$x'_1 = (qi'_1, c_1)$ $\vdots$ $x'_N = (qi'_N, c_N)$	$e_1 = (qi^E_1, id_1)$ $\vdots$ $e_N = (qi^E_N, id_N)$

$qi'_{j_2}, \ldots, X'(qi^E_{i_M}) = qi'_{j_m}|T', E, M)$, where $1 < i_2 < \ldots < i_M \leq N$, and $\{j_2, \cdots, j_M\} = \{2, \cdots, M\}$. This probability can be computed as

$$\frac{\sum P(X'(qi^E_1) = qi'_1, X'(qi_{i_2}) = qi'_{j_2} \ldots X'(qi_{i_M}) = qi'_{j_M}|M)}{\sum P(X'(qi_{r_1}) = qi'_{s_1}, \ldots, X'(qi_{r_M}) = qi'_{s_M}|M)}$$

where $1 \leq r_2 < \ldots < r_m \leq N$, and $\{s_2, \cdots, s_M\} = \{2, \cdots, M\}$.

We have said that, to have probabilistic k-anonymity, Inequality (1) must hold. However, the previous examples show that the computation of the re-identification probability in Inequality (1) for an arbitrary mechanism M may be complex. In the following section, we propose to use data swapping as M, which has the advantage of making the computation of the re-identification probability very simple.

IV. PROBABILISTIC k-ANONYMITY VIA MICROAGGREGATION AND SWAPPING

The proposed method consists of two main steps: (i) partition the records in T into groups of size k and (ii) apply a permutation to the quasi-identifier attributes within each of the groups. This method can accommodate many variations, depending on how the partition step (i) is done.

Note that, as the same permutation is applied to all quasi-identifier attributes, the identity of the individual is not masked. However, the quasi-identifier attributes are dissociated from the confidential attributes, and therefore intruders can only guess the actual values corresponding to a confidential attribute with probability at most $1/k$. If leaking the mere presence of an individual in the data set is itself disclosive, then some of the quasi-identifier attributes must be considered confidential, which takes us to the informed intruder scenario.

We introduce first the method that offers protection against uninformed intruders. In other words, we assume that the attributes may be quasi-identifier attributes or confidential attributes, but not both. Later we extend the method to the scenario with informed intruders proposed in Section I.

A. Uninformed intruders

In presence of uninformed intruders there is a clear separation between quasi-identifier and confidential attributes. Assuming that all records in T are masked and included in T', we have the data sets in Table IV.

Selecting a random sample from T to create T' is a sensible approach, as it introduces uncertainty on whether an individual whose data was collected has been included in the published data set. However, by assuming that all the individuals in T have been included in T', we provide the intruder with the best information available. Therefore, if we achieve probabilistic k-anonymity in this scenario, then we will also achieve it in a scenario where a random sample from T is selected.

It is easy to see that the partition and swapping method described above satisfies probabilistic k-anonymity because

$$P(X'(qi^E_i) = qi|T', E, M) = \begin{cases} 1/k & \text{if } qi \in G(id(qi^E_i)) \\ 0 & \text{otherwise} \end{cases}$$

where $G(id(qi^E_i))$ is the group of records of T that contains the record corresponding to qi^E_i.

The key point in the method is the partition step. A first approach is to partition the data set T into random groups. This leads indeed not only to probabilistic k-anonymity, but to probabilistic $|T|$-anonymity, as the quasi-identifiers of a record can be swapped with the quasi-identifiers of any other record. Moreover, the risk of attribute disclosure is small. However, the impact on data quality can be substantial, because very different records may be swapped.

To achieve better data quality, the groups of records must be selected to be as homogeneous as possible, although this increases the risk of attribute disclosure. Our proposal is to generate the groups using a microaggregation algorithm ([8], [9]) over the quasi-identifier attributes. Microaggregation is a cardinality-constrained form of clustering in which the number of clusters (groups) is not fixed beforehand but the minimum cardinality of each group is required to be k. In the section devoted to informed intruders, there are some experimental results obtained by using the MDAV microaggregation algorithm ([9], [13]); MDAV attempts to maximize intra-group

TABLE V
QUASI-IDENTIFIERS AND CONFIDENTIAL ATTRIBUTE FOR EACH INFORMED INTRUDER

Intruder	Quasi-identifier attributes	Confidential attribute
I_1	$A_0, A_2, \ldots, A_n$	A_1
I_2	$A_0, A_1, A_3 \ldots, A_n$	A_2
$\vdots$	$\vdots$	$\vdots$
I_n	$A_0, A_1, \ldots, A_{n-1}$	A_n

homogeneity using the least squares criterion and it yields groups with size k, except perhaps one group which has size between k and $2k-1$.

Other options in the selection of the groups of records are possible. For example, a variant of MDAV, known as V-MDAV ([14], [15]), may be used that performs clustering in groups of variable size and that is known to reduce the information loss in clustered data sets. The μ-Approx microaggregation heuristic [16] offers also variable-sized groups and is proven to yield a clustering within a bound of the optimal clustering. Another possibility is to select the groups of records in such a way that the risk of attribute disclosure is reduced, by ensuring a certain diversity in the values of the confidential attributes within each group.

B. *MDAV microaggregation for informed intruders*

In the scenario for informed intruders presented in Section I we assumed the number of informed intruders to be the same as the number of confidential attributes in the data set. To be more specific, we consider the attributes: $A_0, A_1, \ldots, A_n$, with A_0 being a non-confidential quasi-identifier attribute, and $A_1, \ldots, A_n$ being confidential quasi-identifier attributes. Intruder I_i, for $i = 1$ to n, is assumed to know the values of all attributes except A_i.

To achieve the desired level of protection against all informed intruders, we apply the method presented for uninformed intruders once for each informed intruder, in order to dissociate the value of the confidential attribute unknown to this intruder from the rest of attributes. For each informed intruder, we use the quasi-identifiers and the confidential attribute shown in Table V.

One difficulty that we face with the previous approach is that dealing with informed intruders in sequence requires applying different permutations over different but overlapping sets of attributes of the original data set T (the quasi-identifiers for each informed intruder). To overcome this difficulty we take the reverse approach: instead of performing the permutation over the quasi-identifier attributes, we apply the reverse permutation to the single confidential attribute unknown to the current intruder. In this way, each permutation acts over a different attribute and there are no overlaps.

C. *Individual ranking microaggregation for informed intruders*

The above observation regarding the application of the inverse permutation on the single unknown confidential attribute leads to single-attribute microaggregation, also called individual ranking microaggregation. Instead of multivariate microaggregation of quasi-identifier attributes, we do individual ranking microaggregation on the unknown confidential attribute. By doing so, the data quality of the published data set is increased, as the confidential attributes are only swapped across records with similar values (see [17] on the low information loss caused by individual ranking microaggregation). It may be argued that there is an increase in the attribute disclosure risk; however, this increase can be mitigated by increasing k.

One extra benefit from this approach is that, since microaggregation is performed on a single attribute, there is no need to normalize attributes as required by multivariate microaggregation to avoid scale problems.

V. EXPERIMENTAL RESULTS

We have implemented the following three methods:

- *MDAV-ID*. MDAV microaggregation is run on the quasi-identifier attributes to partition the data set in groups of size k records. Within each group, quasi-identifiers are replaced by the group centroid in order to have identical quasi-identifiers for all records in the group. This is the procedure suggested in [9] and it achieves the standard notion of k-anonymity proposed in [7] in the sense that all quasi-identifiers within a group are made indistinguishable.
- *MDAV-SWAP*. This is the method described in Section IV-A for probabilistic k-anonymity: MDAV microaggregation on the quasi-identifier attributes plus swapping within groups.
- *IR-SWAP*. This is the method described in Section IV-B above for probabilistic k-anonymity: individual ranking microaggregation on each confidential attribute plus swapping within groups.

The above methods have been tested with the "Census" and "EIA" reference data sets proposed in the European project CASC [18].

A. *"Census" data set*

The "Census" data set contains 1080 records with 13 continuous attributes. Following the approach in [9] we consider the first 6 attributes in "Census" to be non-confidential quasi-identifiers, and the last 7 attributes to be confidential.

To assess the data quality, we evaluate the correlations from all attributes to the confidential attributes. As the proposed methods for probabilistic k-anonymity do not modify non-confidential attributes, correlations between the latter have the same value as in the original data set. Means and variances also remain unchanged for all attributes, because swapping does not change the values taken by each original attribute.

As an example, we computed the correlations for: i) the original data set (see Table VI); ii) the k-anonymous data set resulting from MDAV-ID with $k = 12$ (see Table VII); iii) the probabilistically k-anonymous data set resulting from MDAV-SWAP with $k = 12$ (see Table VIII); and the probabilistically k-anonymous data set resulting from IR-SWAP with $k = 12$

TABLE VI
CORRELATIONS TO THE CONFIDENTIAL ATTRIBUTES IN THE ORIGINAL "CENSUS" DATA SET

	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
A_1	.0038	-.027	-.024	.031	.032	.039	.036
A_2	.98	.14	.2	.73	.71	.72	.7
A_3	.44	-.12	-.058	.56	.55	.56	.55
A_4	.98	.2	.28	.73	.69	.71	.69
A_5	.78	.27	.27	.9	.85	.88	.86
A_6	.79	.13	.22	.59	.57	.57	.56
A_7	1	.17	.23	.72	.7	.71	.69
A_8		1	.45	-.17	-.19	-.17	-.17
A_9			1	.072	.061	.70	.075
A_{10}				1	.96	.98	.96
A_{11}					1	.91	.89
A_{12}						1	.97
A_{13}							1

TABLE VII
CORRELATIONS TO THE CONFIDENTIAL ATTRIBUTES IN THE DATA SET OBTAINED USING MDAV-ID WITH $k = 12$ ("CENSUS" DATA SET)

	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
A_1	-.0035	-.035	-.055	.034	.035	.042	.04
A_2	1	.18	.39	.8	.81	.8	.78
A_3	.79	-.17	.084	.89	.9	.89	.89
A_4	.99	.23	.45	.82	.8	.81	.8
A_5	.86	.18	.4	.94	.92	.94	.93
A_6	.95	.2	.43	.77	.76	.76	.75
A_7	1	.2	.41	.8	.8	.79	.78
A_8		1	.68	-.15	-.18	-.15	-.16
A_9			1	.18	.14	.17	.16
A_{10}				1	.98	1	.99
A_{11}					1	.97	.97
A_{12}						1	1
A_{13}							1

(see Table IX). The values in these tables must be taken with caution: they are results from a single execution of the algorithms, and may change in another execution. Despite these words of caution, we observe that MDAV-SWAP and IR-SWAP result in correlation values closer to the original data set than those obtained with MDAV-ID. The results of IR-SWAP are closest to the original correlations.

To obtain results with more statistical significance, we ran MDAV-ID, MDAV-SWAP and IR-SWAP 100 times. In Table X we report the mean and the standard deviation of the absolute value of the difference between the correlations to the confidential attributes in the anonymized data set and the original data set. The better the data quality of the anonymized data set, the closer the mean and standard deviation to zero. A value close to one for the mean means that most of the dependencies between attributes have been lost.

Table X confirms what had been observed from the previous tables based on a single run: MDAV-SWAP offers better quality than MDAV-ID, but IR-SWAP clearly offers the best quality among the three methods compared. For example, for the data set tried, similar data quality is obtained using MDAV-ID with $k = 11$, MDAV-SWAP with $k = 25$ and IR-SWAP with $k = 300$. Hence, probabilistic k-anonymity turns out to be much more information-preserving than k-anonymity.

TABLE VIII
CORRELATIONS TO THE CONFIDENTIAL ATTRIBUTES IN THE PROBABILISTICALLY k-ANONYMOUS DATA SET OBTAINED USING MDAV-SWAP WITH $k = 12$ ("CENSUS" DATA SET)

	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
A_1	-.0011	-.028	-.034	.032	.033	.036	.032
A_2	.81	.089	.17	.69	.67	.69	.67
A_3	.42	-.020	.091	.48	.47	.48	.43
A_4	.77	.093	.18	.68	.65	.68	.67
A_5	.72	.086	.16	.80	.76	.79	.77
A_6	.64	.086	.14	.54	.52	.54	.52
A_7	1	.12	.17	.69	.67	.66	.65
A_8		1	.19	-.013	-.022	-.042	-.011
A_9			1	.11	.10	.10	.13
A_{10}				1	.76	.81	.87
A_{11}					1	.72	.70
A_{12}						1	.77
A_{13}							1

TABLE IX
CORRELATIONS TO THE CONFIDENTIAL ATTRIBUTES IN THE PROBABILISTICALLY k-ANONYMOUS DATA SET OBTAINED USING IR-SWAP WITH $k = 12$ ("CENSUS" DATA SET)

	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}
A_1	.0041	-.017	-.018	.031	.038	.039	.038
A_2	.98	.13	.20	.73	.71	.72	.70
A_3	.44	-.12	-.041	.56	.55	.56	.55
A_4	.98	.19	.27	.73	.68	.71	.69
A_5	.78	.26	.26	.90	.85	.88	.86
A_6	.79	-.12	.21	.59	.57	.57	.56
A_7	1	.16	.23	.72	.69	.71	.69
A_8		1	.42	-.17	-.19	-.17	-.17
A_9			1	.077	.063	.075	.080
A_{10}				1	.95	.98	.96
A_{11}					1	.91	.89
A_{12}						1	.97
A_{13}							1

B. "EIA" data set

Due to space constraints, empirical results for the "EIA" data set are more succinctly presented. Table XI reports an evaluation for the "EIA" data set analogous to the one reported in Table X for the "Census" data set. Like before, we observe

TABLE X
MEAN AND STANDARD DEVIATION OF THE ABSOLUTE VALUE OF THE DIFFERENCE BETWEEN THE CORRELATIONS IN THE ORIGINAL AND THE ANONYMIZED DATA SETS ("CENSUS" DATA SET)

	MDAV-ID		MDAV-SWAP		IR-SWAP	
k	mean	st.dev.	mean	st.dev.	mean	st.dev.
5	.055	.064	.037	.045	.0021	.0041
7	.062	.071	.048	.056	.0022	.0039
9	.069	.078	.055	.064	.0028	.0049
11	.078	.085	.061	.070	.0038	.0068
25	.11	.11	.091	.093	.0061	.012
50	.14	.13	.13	.12	.010	.020
100	.17	.15	.19	.17	.020	.030
200	.29	.27	.31	.28	.044	.047
300	.38	.39	.37	.34	.087	.071

TABLE XI
MEAN AND STANDARD DEVIATION OF THE ABSOLUTE VALUE OF THE DIFFERENCE BETWEEN THE CORRELATIONS IN THE ORIGINAL AND THE ANONYMIZED DATA SETS ("EIA" DATA SET)

	MDAV-ID		MDAV-SWAP		IR-SWAP	
k	mean	st.dev.	mean	st.dev.	mean	st.dev.
5	.018	.017	.017	.035	.00064	.00075
7	.02	.017	.024	.05	.0012	.0018
9	.034	.031	.028	.053	.0015	.0018
11	.039	.036	.029	.052	.0019	.0023
25	.085	.078	.043	.081	.0063	.0072
50	.13	.12	.053	.089	.011	.011
100	.15	.14	.058	.092	.029	.037
200	.19	.18	.09	.11	.093	.074
300	.2	.18	.12	.13	.14	.091

that MDAV-SWAP performs better than MDAV-ID, but IR-SWAP is clearly the best of the three methods.

VI. CONCLUSIONS AND FUTURE RESEARCH

k-Anonymity is a broadly used privacy property that focuses on protecting against identity disclosure. In a k-anonymous data set, for each record there are at least $k-1$ other records sharing the same values for all the quasi-identifier attributes. Hence, enforcing k-anonymity implies variability loss and therefore quality loss. This is especially serious in a scenario with informed intruders, who know the values of some confidential attributes: the confidential attributes known by the informed intruder can be viewed as additional quasi-identifiers. The more quasi-identifier attributes, the more data quality loss is caused by k-anonymity.

To mitigate the above problem, we have introduced the notion of probabilistic k-anonymity. Like standard k-anonymity, probabilistic k-anonymity guarantees that the probability of correct re-identification is at most $^1/_k$, but without explicitly requiring that the quasi-identifier attributes take identical values within each group of k records. We have presented two computational methods to reach probabilistic k-anonymity, based on microaggregation and swapping. Experimental work shows that, for a fixed re-identification probability $^1/_k$, the new methods are much more quality-preserving than standard k-anonymity enforcement.

Future research will combine probabilistic k-anonymity with other properties like l-diversity or t-closeness in view of reducing the quality loss incurred to protect against attribute disclosure.

DISCLAIMER AND ACKNOWLEDGMENTS

The authors are with the UNESCO Chair in Data Privacy, but the views expressed in this article do not necessarily reflect the position of UNESCO nor commit that organization. This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS", TIN2011-27076-C03-01 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", by the Government of Catalonia under grant 2009 SGR 1135 and by the European Commission under FP7 project "DwB". The second author is partly supported as an ICREA Acadèmia Researcher by the Government of Catalonia.

REFERENCES

[1] J. Domingo-Ferrer. A survey of inference control methods for privacy-preserving data mining. In *Privacy-Preserving Data Mining*, volume 34 of *Advances in Database Systems*, pages 53–80. Springer, 2008.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4):14:1–14:53, June 2010.

[3] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. Microdata protection. In *Secure Data Management in Decentralized Systems*, volume 33 of *Advances in Information Security*, pages 291–321. Springer, 2007.

[4] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, VLDB '05, pages 901–909. VLDB Endowment, 2005.

[5] L. Sweeney. *Uniqueness of Simple Demographics in the U.S. Population*. LIDAP-WP4, Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh PA, 2000.

[6] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, and P.-P. DeWolf. *Handbook on Statistical Disclosure Control (version 1.2)*. ESSNET SDC Project, 2010. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf.

[7] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Tech. rep. SRI-CSL-98-04, SRI Computer Science Laboratory, Palo Alto, CA, 1998.

[8] J. Domingo-Ferrer and J. M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201, 2002.

[9] J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogenerous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 2005.

[10] A. Machanavajjhala, D. Kiefer, J. Gehrke, and M. Venkitasubramaniam. l-Diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

[11] N. Li and T. Li. t-Closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of IEEE 23rd Int'l Conf. on Data Engineering (ICDE'07)*, pages 106–115, 2007.

[12] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. Aggregate query answering on anonymized tables. In *Proceedings of IEEE Int'l Conf. on Data Engineering (ICDE 2007)*, pages 116–125, 2007.

[13] A. Hundepool, A. Van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. *u-ARGUS version 3.2 Software and User's Manual*. Statistics Netherlands, Voorburg NL, 2003.

[14] A. Solanas and A. Martinez-Balleste. V-mdav: a multivariate microaggregation with variable group size. In *Proceedings of COMPSTAT 2006*, pages 917–925. Physica-Verlag, September 2006.

[15] A. Solanas, Ú. Gonzalez-Nicolas, and A. Martínez-Ballesté. A variable-mdav-based partitioning strategy to continuous multivariate microaggregation with genetic algorithms. In *IEEE International Joint Conference on Neural Networks-IJCNN 2010*, pages 1–7, 2010.

[16] J. Domingo-Ferrer, F. Sebé, and A. Solanas. A polynomial-time approximation to optimal multivariate microaggregation. *Computers & Mathematics with Applications*, 55(4):714–732, 2008.

[17] J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz, editors, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 111–134. North-Holland, Amsterdam, 2001.

[18] R. Brand, J. Domingo-Ferrer, and J. M. Mateo-Sanz. *Reference data sets to test and compare SDC methods for protection of numerical microdata*. European FP5 Project IST-2000-25069 CASC, 2002. http://neon.vb.cbs.nl/casc/CASCtestsets.htm.