Data Mining and Knowledge Discovery, 11, 195–212, 2005

DOI: 10.1007/s10618-005-0007-5

Ordinal, Continuous and Heterogeneous *k*-Anonymity Through Microaggregation

JOSEP DOMINGO-FERRER josep.domingo@urv.net
Department of Computer Engineering and Maths,Rovira i Virgili University of Tarragona, Av. Països Catalans 26, E-43007 Tarragona, Catalonia

VICENÇ TORRA vtorra@iiia.csic.es
Institut d'Investigació en Intel·ligència Artificial-CSIC, Campus UAB, E-08193, Bellaterra, Catalonia

Editor: Geoff Webb

Received October 27, 2004; Accepted April 14, 2005

Published online: 8 September 2005

Abstract. *k*-Anonymity is a useful concept to solve the tension between data utility and respondent privacy in individual data (microdata) protection. However, the generalization and suppression approach proposed in the literature to achieve *k*-anonymity is not equally suited for all types of attributes: (i) generalization/suppression is one of the few possibilities for nominal categorical attributes; (ii) it is just one possibility for ordinal categorical attributes which does not always preserve ordinality; (iii) and it is completely unsuitable for continuous attributes, as it causes them to lose their numerical meaning. Since attributes leading to disclosure (and thus needing *k*-anonymization) may be nominal, ordinal and also continuous, it is important to devise *k*-anonymization procedures which preserve the semantics of each attribute type as much as possible. We propose in this paper to use categorical microaggregation as an alternative to generalization/suppression for nominal and ordinal *k*-anonymization; we also propose continuous microaggregation as *the* method for continuous *k*-anonymization.

Keywords: *k*-anonymity, microdata privacy, database security, microaggregation

1. Introduction

Whenever data from respondents are collected and then released for general or research use (e.g. in official statistics, in e-commerce or in e-health), there is a tension between privacy for respondents and data utility for users. This tension is the *raison d'être* of Statistical Disclosure Control (SDC, (Willenborg and DeWaal, 2001)).

The protection provided by SDC techniques normally entails some degree of data modification. The challenge for SDC is to modify data in such a way that both the risk of disclosing private respondent information and the information loss caused are acceptably low. SDC-protected data should stay useful for data mining purposes.

SDC can be applied to information in several formats: tabular data, dynamically queryable databases and microdata (individual respondent data). We will concentrate here on microdata protection and specifically on how to deal with the aforementioned tension between low disclosure risk and low information loss.

1.1. Contribution and plan of this paper

Section 2 recalls the basics of microdata protection. Section 3 discusses approaches to trading off information loss for disclosure risk and analyzes their strengths and limitations; in particular *k*-anonymity is identified as a clean approach to conciliating information loss and disclosure risk. In Section 4, a critique of the current generalization/suppression approach to *k*-anonymity is made. Section 5 shows how to use microaggregation to achieve *k*-anonymity for continuous, ordinal and nominal data. Section 6 presents empirical results. Section 7 contains a conclusion.

2. Fundamental concepts of microdata protection

A microdata set **V** can be viewed as a file with n records, where each record contains m attributes on an individual respondent. The attributes in an original unprotected dataset can be classified in four categories which are not necessarily disjoint:

- *Identifiers*. These are attributes that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc. Since our objective is to prevent confidential information from being linked to specific respondents, we will assume in what follows that, in a pre-processing step, identifiers in **V** have been removed/encrypted.
- *Quasi-identifiers*. Borrowing the definition from Dalenius (1986) and Samarati (2001), a quasi-identifier is a set of attributes in **V** that, in combination, can be linked with external information to re-identify (some of) the respondents to whom (some of) the records in **V** refer. Unlike identifiers, quasi-identifiers cannot be removed from **V**. The reason is that any attribute in **V** potentially belongs to a quasi-identifier (depending on the external data sources available to the user of **V**). Thus one would need to remove all attributes (!) to make sure that the dataset no longer contains quasi-identifiers.
- *Confidential outcome attributes*. These are attributes which contain sensitive information on the respondent. Examples are salary, religion, political affiliation, health condition, etc.
- *Non-confidential outcome attributes*. Those attributes which contain non-sensitive information on the respondent. Examples are town and country of residence, etc. Note that attributes of this kind cannot be neglected when protecting a data set, because they can be part of a quasi-identifier. For instance, if 'Job' and 'Town of residence' can be considered non-confidential outcome attributes, but their combination can be a quasi-identifier, because everyone knows who is the doctor in a small village.

The purpose of microdata SDC mentioned in the previous section can be stated more formally by saying that, given an original microdata set **V**, the goal is to release a protected microdata set $\mathbf{V}'$ in such a way that:

1. Disclosure risk (i.e. the risk that a user or an intruder can use $\mathbf{V}'$ to determine confidential attributes on a specific individual among those in **V**) is low.

2. User analyses (regressions, means, etc.) on $\mathbf{V}'$ and $\mathbf{V}$ yield the same or at least similar results. This is equivalent to requiring that information loss caused by SDC should be low, i.e. that the utility of the SDC-protected data should stay high.

Microdata protection methods can generate the protected microdata set $\mathbf{V}'$

- either by *masking original data*, i.e. generating a modified version $\mathbf{V}'$ of the original microdata set $\mathbf{V}$;
- or by *generating synthetic data* $\mathbf{V}'$ that preserve some statistical properties of the original data $\mathbf{V}$

Masking methods can in turn be divided in two categories depending on their effect on the original data (see Willenborg and DeWaal (2001) and Domingo-Ferrer and Torra (2001a) for more details on the methods mentioned below):

- *Perturbative*. The microdata set is distorted before publication. In this way, unique combinations of scores in the original dataset may disappear and new unique combinations may appear in the perturbed dataset; such confusion is beneficial for preserving statistical confidentiality. The perturbation method used should be such that statistics computed on the perturbed dataset do not differ significantly from the statistics that would be obtained on the original dataset. Microaggregation and additive noise are examples of perturbative methods.
- *Non-perturbative*. Non-perturbative methods do not distort data. Rather, they rely on the principles of generalization and suppression. For a categorical attribute V_i, generalization combines several categories to form new (less specific) categories; for a continuous attribute, generalization means replacing that attribute by another attribute which is a discretized version of the former. Suppression can be applied to the values of a few attributes in few records (local suppression) or can be applied to entire records. The latter is equivalent to obtaining the protected data set as a sample of the original data set (the sample formed by the non-suppressed records).

The alternative to masking methods is synthetic data generation, which seems to have the philosophical advantage of circumventing the re-identification problem: since published records are invented and do not derive from any original record, some authors claim that no individual having supplied original data can complain from having been re-identified. Other authors (e.g., Winkler (2004) and Reiter (2004) remark that, at a closer look, even synthetic data might contain some records allowing for re-identification of confidential information. In short, synthetic data overfitted to original data might lead to disclosure just as original data would.

So far in this section, we have classified microdata protection methods by their operating principle. If we consider the type of data on which they can be used, a different dichotomic classification applies:

- *Continuous*. An attribute is considered continuous if it is numerical and arithmetical operations can be performed on it. Examples are income and age. When designing

methods to protect continuous data, one has the advantage that arithmetical operations are possible, and the drawback that every combination of numerical values in the original dataset is likely to be unique, which leads to disclosure if no action is taken.

- *Categorical*. An attribute is considered categorical when it takes values over a finite set and standard arithmetical operations do not make sense. Thus, SDC techniques based on arithmetical manipulation cannot be used on categorical data. Two main types of categorical attributes can be distinguished:
  - *Ordinal*. An ordinal attribute takes values in an ordered range of categories. Thus, the $\leq$, max and min operators are meaningful and can be used by SDC techniques for ordinal data. The instruction level and the political preferences (left-right) are examples of ordinal attributes.
  - *Nominal*. A nominal attribute takes values in an unordered range of categories. The only possible operator is comparison for equality, which restricts the range of applicable SDC techniques. The eye color and the address of an individual are examples of nominal attributes.

Although most attributes in a quasi-identifier can be expected to be nominal or ordinal, continuous attributes can also be present. Indeed, sometimes numerical outcome attributes give enough clues for re-identification. Thus an intruder can use such continous attributes as (part of) a quasi-identifier. As an example, if respondents are companies and turnover is an outcome attribute, everyone in a certain industrial sector knows which is the company with largest turnover.

3. Approaches to trading off information loss and disclosure risk

There exist a plethora of methods to protect microdata in addition to the ones mentioned above (see Domingo-Ferrer and Torra (2005), Domingo-Ferrer and Torra (2001a) and Willenborg and DeWaal (2001). To complicate things further, most of such methods are parametric, so the user must go through two choices rather than one: a primary choice to select a method and a secondary choice to select parameters for the method to be used. To guide those choices, several approaches have been proposed which are summarized in this section.

3.1. Score construction

The mission of SDC to modify data in such a way that sufficient protection is provided at minimum information loss suggests that a good SDC method is one achieving a good tradeoff between disclosure risk and information loss.

Following this idea, (Domingo-Ferrer and Torra, 2001b) proposed a score for method performance rating based on the average of information loss and disclosure risk measures. For each method M and parameterization P, the following score is computed:

$$Score(\mathbf{V}, \mathbf{V}') = \frac{IL(\mathbf{V}, \mathbf{V}') + DR(\mathbf{V}, \mathbf{V}')}{2}$$

where *IL* is an information loss measure, *DR* is a disclosure risk measure and $\mathbf{V}'$ is the protected dataset obtained after applying a specific method with a specific parameterization to an original dataset $\mathbf{V}$.

Domingo-Ferrer and Torra (2001b) and Domingo-Ferrer et al. (2001) computed *IL* and *DR* using a weighted combination of a set of information loss and disclosure risk measures they defined. With the resulting score, a ranking of masking methods (and their parameterizations) was obtained. Yancey et al. (2002) followed the line of the above two papers and ranked a different set of methods using a slightly different score.

Using a score allows the selection of a masking method and its parameters to be regarded as an optimization problem. This idea was first used in Sebé et al. (2002). In that paper, a masking method was applied to the original data file and then a post-masking optimization procedure was applied to increase the score obtained.

On the negative side, no specific score weighting can do justice to all methods. Thus, when ranking methods, the values of all measures of information loss and disclosure risk should be supplied along with the overall score.

3.2. R-U maps

A tool which may be enlightening when trying to construct a score or, more generally, optimize the tradeoff between information loss and disclosure risk is a graphical representation of pairs of measures (disclosure risk, information loss) or their equivalents (disclosure risk, data utility). Such maps are called R-U confidentiality maps (Duncan et al., 2001a; Duncan et al., 2001b)). Here, *R* stands for disclosure risk and *U* for data utility. According to Duncan et al. (2001b), 'in its most basic form, an R-U confidentiality map is the set of paired values (*R, U*), of disclosure risk and data utility that correspond to various strategies for data release' (e.g., variations on a parameter). Such (*R,U*) pairs are typically plotted in a two-dimensional graph, so that the user can easily grasp the influence of a particular method and/or parameter choice.

3.3. k-Anonymity

A different approach to facing the conflict between information loss and disclosure risk is the following concept proposed by Samarati and Sweeney (1998), Samarati (2001) and Sweeney (2002a, 2002b).

Definition 3.1 (k-anonymity). A dataset is said to satisfy k-anonymity for $k > 1$ if, for each combination of values of quasi-identifiers (e.g. name, address, age, gender, etc.), at least k records exist in the dataset sharing that combination.

Note that, if a protected dataset $\mathbf{V}'$ satisfies k-anonymity, an intruder trying to link $\mathbf{V}'$ with an external non-anonymous data source will find at least k records in $\mathbf{V}'$ that match any value of the quasi-identifier the intruder uses for linkage. Thus re-identification, i.e. mapping a record in $\mathbf{V}'$ to a non-anonymous record in the external data source, is not

possible; the best the intruder can hope for is to map groups of k records in $\mathbf{V}'$ to each non-anonymous external record.

If, for a given k, k-anonymity is assumed to be enough protection, one can concentrate on minimizing information loss with the only constraint that k-anonymity should be satisfied. This is a clean way of solving the tension between data protection and data utility. In Samarati (2001) and Sweeney (2002b), the approach suggested to reach k-anonymity is to combine generalization and local suppression (described in Section 2 above); thus, with that approach, minimizing information loss usually translates to minimizing the number and/or the magnitude of suppressions and also minimizing the granularity loss caused by generalizations.

For the sake of concreteness, we will assume in what follows a single quasi-identifier, i.e. we will deal with a single intruder or intruder coalition able to link the protected dataset with external datasets through the quasi-identifier.

4. A critique of generalization/suppression for *k*-anonymity

Satisfying k-anonymity with minimal data modification using generalization (recoding) and local suppression has been shown to be NP-hard in Meyerson and Williams (2004) and Aggarwal et al. (2004). In fact, even how to optimally combine generalization and local suppression is an open issue. Unless carefully combined, those two non-perturbative methods may cause a substantial loss of data utility.

Furthermore, the use of generalization to ensure k-anonymity poses several practical problems. One of them is the computational cost of finding the optimal recoding. This is partly related to the exponential number of generalizations that can be defined for each attribute:

Lemma 4.1. *For an attribute with c categories, there are $2^c - c - 1$ possible generalizations.*

Proof: Generalization is replacing a subset of categories by a new general category. Thus the number of generalizations equals the number of subsets of categories containing more than one category. There are 2^c subsets of categories, of which c consist of a single category and one is the empty subset. Thus there are $2^c - c - 1$ subsets containing more than one category. □

Another problem is determining the subset of appropriate generalizations, i.e. which are the new categories and which is the subset of old categories that can be recoded into each of such new categories. Not all recodings are appropriate because the semantics of the categories and the intended data uses must be taken into account. For example, when generalizing ZIP codes, recoding 08201 and 08205 into 0820* makes sense as long as 0820* is meaningful as a location (e.g. corresponds to a city, a county or another geographical area). For the same reason, it is probably not meaningful to recode 08201 and 05201 into $0 * 201$ because the set of regions represented by $0 * 201$ might lack any geographical significance. The need for significance makes automatic generation of recodings a thorny issue.

Table 1. Records consisting of attributes V_1 and V_2 with ranges $D(V_1) = \{a, b, c, d, e\}$ and $D(V_2) = \{r, s, t, u, v\}$.

Record	V_1	V_2
r_1	a	r
r_2	b	r
r_3	c	r
r_4	e	r
r_5	e	s
r_6	e	t
r_7	e	v
r_8	d	v
r_9	c	v
r_{10}	a	v
r_{11}	a	u
r_{12}	a	t

Given a set of possible generalizations, the methods in the literature diverge on how the generalization is applied. This is, once a particular generalization rule $c_i \rightarrow C$ is considered, methods diverge on which records containing c_i are recoded. For example, μ-Argus (Hundepool et al., 2003) and (Domingo-Ferrer and Torra, 2001b) recode all occurrences of c_i (this is known as global recoding) while Sweeney (2002a) and Samarati (2001) only replace some of the occurrences (this is known as local recoding).

Neither global nor local recoding are free from disadvantages:

- Global recoding implies greater information loss because it may recode some records that do not need it. A related drawback is that the recoding that might be suitable for a set of records might be completely unsuitable for another set. Records in Table 1 illustrate this process. Note that, if we consider 3-anonymization of this table, the best recoding for records r_1, r_2 and r_3 turns out to be the rule $\{a, b, c\} \rightarrow ABC$. At the same time, the best recoding for records r_4, r_5 and r_6 is the rule $r, s, t \rightarrow RST$. Nevertheless, such recodings would force r_9 to be recoded as (ABC, v) and r_{12} to be recoded as (ABC, RST). It can be seen from the graphical representation of records given in Figure 1 that such a

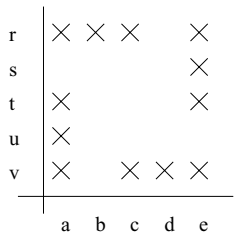

Figure 1 Graphical representation of records in Table 1.

transformation is very awkward when r_9 and r_{12} are not considered in isolation but with the rest of records.

- Local recoding is quite difficult to use in an automated way and makes data analysis more complex as both original values c_i and recoded values C appear in the protected file. Thus, local recoding requires the system to consider a larger number of generalizations and it is possible that several recodings for a category c_i exist in the file. Such is the case when categories c_i and c_j are recoded for some records into category C and at the same time categories c_i and c_k are recoded into C' for some other records in the same dataset. This complicates data analyses on the protected dataset.

In Samarati (2001) and Sweeney (2002a), the use of local suppression is suggested to avoid too much recoding. Local suppression has several drawbacks:

- As mentioned above, it is not known how to optimally combine generalization and local suppression.
- The use of suppression is not homogeneous in the literature:
  - While e.g. Sweeney (2002a) applies suppression at the tuple level (a tuple can be suppressed only in its entirety), others, such as Hundepool et al. (2003) and Aggarwal et al. (2004), suppress only some particular attributes for some particular records. In fact, Aggarwal et al. (2004) provide a polynomial 1.5-approximation to the optimal attribute suppression.
  - Suppression can consist of either blanking a value or replacing it with a locally neutral value (e.g. some kind of average).
- Whatever the type of local suppression used, it is very unclear how the user of protected data can analyze them without highly specific software (e.g. imputation software or software for dealing with censored data).

For ordinal attributes, using generalization and local suppression to achieve k-anonymity is far from perfect for the above reasons, but could still be considered. However, for continuous attributes in a quasi-identifier, generalization and local suppression are definitely unsuitable. Using such non-perturbative methods on a continuous attribute causes this attribute to become categorical and lose its numerical semantics. In other words, if any value in a continuous range [a, b] is replaced by a label $L_{[a,\ b]}$, then information on the position of original numerical values within [a, b] vanishes: for example, one cannot infer from $L_{[a,\ b]}$ whether the original numerical values were mostly in the lower half of [a, b] or in its upper half. It would be thoroughly unacceptable if, on the grounds of using k-anonymity, data protectors denied continuous attributes to data users. Therefore, a method to provide k-anonymity for continuous attributes is definitely needed.

5. Microaggregation for *k*-anonymity

For the reasons sketched in Section 4, it is very interesting to find alternatives to generalization/suppression for satisfying k-anonymity. Microaggregation stands out as a natural

approach to satisfy k-anonymity. Microaggregation is a family of perturbative SDC methods originally defined for continuous data (Defays and Nanopoulos, 1993; Domingo-Ferrer and Mateo-Sanz, 2002) and recently extended for categorical data (Torra, 2004). Whatever the data type, microaggregation can be operationally defined in terms of the following two steps:

Partition: The set of original records is partitioned into several clusters in such a way that records in the same cluster are *similar* to each other and so that the number of records in each cluster is at least k.

Aggregation: An aggregation operator (for example, the mean for continuous data or the median for categorical data) is computed for each cluster and is used to replace the original records. In other words, each record in a cluster is replaced by the cluster's prototype.

In the remainder of this paper, we will show how to use microaggregation for k-anonymity in order to circumvent most of the problems of generalization/suppression listed above:

- Microaggregation is a unified approach, unlike the dual method combining generalization and suppression;
- Even if optimal microaggregation is also *NP*-hard (Oganian and Domingo-Ferrer, 2001)—like generalization/suppression—, near-optimal heuristics exist –unlike for generalization/suppression—; one of those will be described below;
- Microaggregation does not complicate data analysis by adding new categories to the original scale —unlike global recoding—;
- Microaggregation does not result in suppressed data, which makes analysis of k-anonymized data easy with standard software;
- Microaggregation is perfectly suitable to protect continuous data without removing their numerical semantics.

5.1. Multivariate microaggregation

We give next an algorithm for the partition step in multivariate microaggregation called MDAV-generic. MDAV-generic is a generic variant of the algorithm of the MDAV (Maximum Distance to Average Vector) that we implemented in Hundepool et al. (2003) as an evolution of the multivariate fixed-size microaggregation described in Domingo-Ferrer and Mateo-Sanz (2002). The common and distinctive feature of this algorithm series is that single-axis projection of multivariate data is not required. The difference between MDAV-generic and its predecessors (Hundepool et al., 2003) and (Domingo-Ferrer and Mateo-Sanz, 2002) is that the former can work with any type of attribute, aggregation operator and distance (continuous, ordinal or nominal), whereas the precedessors were designed for continuous data only and used the arithmetical mean and the Euclidean distance.

Algorithm 5.1 *(MDAV-generic) (R: dataset, k: integer).*

1. While $|R| \geq 3k$ do

(a) Compute the average record $\tilde{x}$ of all records in R. The average record is computed attribute-wise.
(b) Consider the most distant record x_r to the average record $\tilde{x}$ using an appropriate distance.
(c) Find the most distant record x_s from the record x_r considered in the previous step.
(d) Form two clusters around x_r and x_s, respectively. One cluster contains x_r and the $k-1$ records closest to x_r. The other cluster contains x_s and the $k-1$ records closest to x_s.
(e) Take as a new dataset R the previous dataset R minus the clusters formed around x_r and x_s in the last instance of Step 1d.

end while

2. If there are between $3k-1$ and $2k$ records in R:

(a) compute the average record $\tilde{x}$ of the remaining records in R
(b) find the most distant record x_r from $\tilde{x}$
(c) form a cluster containing x_r and the $k-1$ records closest to x_r
(d) form another cluster containing the rest of records.

else (less than $2k$ records in R) form a new cluster with the remaining records.

In the description of MDAV-generic, the term 'record' stands for the projection of an actual record on the attributes of the quasi-identifier.

Implementation of MDAV-generic for a particular attribute type requires specifying how the average record is computed and what distance is used. This is detailed below for the three different attribute types described above: continuous, ordinal and nominal.

5.2. *Continuous attributes*

Following Domingo-Ferrer and Mateo-Sanz (2002) and Hundepool et al. (2003), the average operator used is the arithmetical mean and the distance used is the Euclidean one. Before applying MDAV-generic, attributes are standardized (by subtracting their mean and dividing by their standard deviation), so that they have equal weight when computing distances.

After application of MDAV-generic, attributes are destandardized and the original units are recovered. By construction, MDAV-generic exactly preserves the means of original attributes in the k-anonymized dataset. A simple rescaling transformation is applied to the k-anonymized dataset for exact preservation of the variances of the original attributes. This rescaling does not violate k-anonymity and for the j-th k-anonymized attribute is

$$\left(\frac{\left(x_{ij}-m_1^0(j)\right)\sqrt{\mu_2(j)}}{\sqrt{m_2(j)}}\right)+\mu_1^0(j)$$

where x_{ij} is the value of the k-anonymized j-th attribute for the i-th record, $(m_1^0(j), m_2(j))$ are the mean and the variance of the k-anonymized j-th attribute and $(\mu_1^0(j), \mu_2(j))$ are the mean and the variance of the original j-th attribute.

5.3. *Ordinal attributes*

A possible distance between two ordinal categories a and b of an attribute V_i, with $a \leq b$, is

$$d_{ORD}(a, b) = \frac{|\{i | a \leq i < b\}|}{|D(V_i)|} \tag{1}$$

that is, the number of categories separating a and b divided by the number of categories in the range of the attribute (the division is used to standardize the distance between 0 and 1).

The average operators we use for ordinal attributes are the median and the convex median.

Definition 5.2 (Median). Given an ordinal scale $C = \{c_1 < c_2 < \cdots < c_o\}$, the median of the set $S = \{a_1, a_2, \ldots, a_N\}$ (with $a_i \in C$) is the category that occupies the central position in S once S is ordered. In terms of frequencies, the median is a category such that its predecessors and successors in the ordered S have equal frequency.

For example, the median of $S = \{1, 2, 2, 5, 6\}$ is 2.

Definition 5.3 (Convex median). If the frequency function f on categories is transformed into a convex function f'

$$f'(c_i) = \min\left(\max_{c_j \leq c_i}(f(c_j)), \max_{c_j \geq c_i}(f(c_j))\right)$$

then the median over f' is called convex median.

Figure 2 illustrates the computation of frequencies for the median and the convex median for the set $S = \{1, 2, 2, 5, 6\}$ in the ordinal scale $C = \{0, 1, 2, 3, 4, 5, 6, 7\}$.

The advantage of the convex median over the median is that the former allows for compensation: aggregation using the median can only yield one of the aggregated categories (i.e. a category with nonzero frequency), whereas using the convex median does not present this limitation. Like the arithmetic mean, the convex median can yield a value which, while being different from all aggregated values, is more central to them. This can be readily seen if we take integer categories: the median of $\{1, 2, 7\}$ is 2, while the convex median is 4; clearly, 4 is more central to the set, because it is closer to the arithmetic mean 3.3.

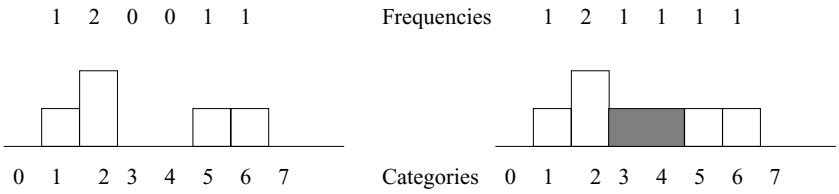

Figure 2. Frequencies for computing the median and the convex median of the set $S = \{1, 2, 2, 5, 6\}$ defined on the ordinal scale $C = \{0, 1, 2, 3, 4, 5, 6, 7\}$.

5.4. *Nominal attributes*

Distance for nominal attributes is defined using the equality predicate. Thus, the distance between two values of a nominal attribute is 0 if they are equal and 1 if they are not.

The plurality rule (or mode) is used as the average operator. This is, for a set $S = \{a_1, a_2, \ldots, a_N\}$, the most frequent value is selected as the average.

MDAV-generic for categorical data is similar to the algorithm proposed in (Torra, 2004), in that both use the median and the mode as average operators. What is different is the partition step, which does not require the number of clusters to be specified as an input parameter.

6. Empirical results

As mentioned in Section 3.3, using k-anonymity has the advantage that security against disclosure risk is no longer an empirical outcome: it becomes an input parameter. Indeed, once the data protector adopts a specific quasi-identifier and a parameter k suitable for her/his disclosure scenario, there is no need to worry about disclosure risk any more, so our empirical work can concentrate on assessing information loss. Even though MDAV-generic can at the same time work on continuous and categoric attributes, we report information loss results separately so that comparison with previous work is easier.

6.1. *Continuous attributes*

MDAV-generic was first tried on a continuous dataset extracted from the U. S. Current Population Survey (1995) using the Data Extraction System (DES) of the U. S. Census Bureau. This continuous dataset, described in more detail in Domingo-Ferrer and Torra (2001b), contains 1080 records described by 13 continuous attributes. We have computed k-anonymous versions of the dataset for $k = 3, \ldots 9$ and for quasi-identifiers consisting of: (i) only the first 6 attributes, and (ii) all 13 attributes. The longer quasi-identifier is a worst-case scenario, because one assumes that the intruder or the coalition of intruders can use all attributes for re-identification. Consistently with Section 5.2 above, the average operator used for k-anonymization was the arithmetic mean and the distance used was Euclidean.

In Table 2, the information loss measures described in Domingo-Ferrer and Torra (2001a) and Domingo-Ferrer et al. (2001) are given for the various combinations of quasi-identifier length (6, 13) and $k = 3, 6, 9, 12$. Information loss is only computed for attributes in the quasi-identifier, because these are the only attributes modified by k-anonymization. Information loss metrics are as follows:

1. IL_1 is the mean variation of individual attribute values in the original and k-anonymous data sets;
2. IL_2 is the mean variation of attribute means in both datasets;
3. IL_3 is the mean variation of attribute variances;
4. IL_4 is the mean variation of attribute covariances;

Table 2. Information loss measures (according to (Domingo-Ferrer and Torra, 2001a; Domingo-Ferrer et al., 2001) for continuous data depending on k and the quasi-identifierlength.

Quasi-identifier length	k	IL_1	IL_2	IL_3	IL_4	IL_5	IL
6	3	0.131	0	0	0.036	0.007	3.48
6	6	0.174	0	0	0.075	0.013	5.24
6	9	0.203	0	0	0.129	0.017	6.98
6	12	0.185	0	0	0.166	0.020	7.42
13	3	0.907	0	0	0.058	0.016	19.62
13	6	1.389	0	0	0.134	0.032	31.10
13	9	1.535	0	0	0.161	0.039	34.70
13	12	1.564	0	0	0.164	0.046	35.48

5. IL_5 is the mean variation of attribute Pearson's correlations;
6. IL is 100 times the average of IL_1, IL_2, IL_3, IL_4 and IL_5.

IL in Table 2 is comparable to IL reported in the empirical results in Domingo-Ferrer and Torra (2001b) and Domingo-Ferrer et al. (2001). The ten best masking methods identified in those papers result in IL between 13.37 and 25.81, which is much higher than IL reported in Table 2 for quasi-identifier length 6, and similar to the IL reported for quasi-identifier length 13. Of course, the ranking in Domingo-Ferrer and Torra, (2001b) and Domingo-Ferrer et al. (2001) is based not only on information loss, but also on disclosure risk; nonetheless, comparing that ranking with our results shows that information loss caused by microaggregation-based k-anonymity is quite moderate.

In Table 3, the new probabilistic information loss measures described in the companion paper (Mateo-Sanz et al., 2005) are given for the various combinations of quasi-identifier

Table 3. Information loss measures (according to (Mateo-Sanz et al., 2005)) for continuous data depending on k and the quasi-identifier length.

Quasi-identifier length	k	$PIL(Q)$	$PIL(m_1^0)$	$PIL(m_2)$	$PIL(m_{11})$	$PIL(r)$
6	3	0.142	0	0	0.161	0.255
6	6	0.187	0	0	0.270	0.371
6	9	0.194	0	0	0.332	0.430
6	12	0.198	0	0	0.373	0.471
13	3	0.498	0	0	0.363	0.534
13	6	0.536	0	0	0.548	0.674
13	9	0.577	0	0	0.598	0.705
13	12	0.618	0	0	0.628	0.731

length and k. These measures are bounded in the $[0, 1]$ interval and correspond to the previous ones as follows:

1. $PIL(Q)$ is average impact on quantiles from 5% to 95% in 5% increments over all attributes;
2. $PIL(m_1^0)$ is the average impact on means over all attributes;
3. $PIL(m_2)$ is the average impact on variances over all attributes;
4. $PIL(m_{11})$ is the average impact on covariances over all attribute pairs;
5. $PIL(r)$ is the average impact on Pearson's correlation coefficients over all attribute pairs.

From analysis of Tables 2 and 3, it can be seen that:

- By construction, MDAV-generic exactly preserves means and variances.
- The impact on the non-preserved statistics (individual values, quantiles, covariances and correlations) grows with the quasi-identifier length, as one would expect: the more intruder's knowledge, the more distortion is required to reach k-anonymity.
- For a fixed quasi-identifier length, the impact on the non-preserved statistics grows with k: the higher the anonymity, the more distortion is required.

6.2. *Categorical attributes*

MDAV-generic was also tried on a categorical dataset extracted from the U.S. Housing Survey (1993) using the Data Extraction System (DES) of the U.S. Census Bureau. Three attributes were ordinal and the remaining eight were nominal (this dataset is described in more detail in Torra (2004)). We have computed k-anonymous versions of the dataset for $k = 2, \ldots 9$ and for quasi-identifiers consisting of 3, 4, 8 and 11 attributes. For ordinal attributes, the median was used as average operator. For each combination of k and quasi-identifier length, we computed the information loss measures similar to those defined in Domingo-Ferrer et al. (2001) and Domingo-Ferrer and Torra (2001b) for categorical data. Information loss is only computed for the attributes in the quasi-identifier. These are:

Table 4. Information loss depending on k for a categorical quasi-identifier of length 8.

k	*Dist*	*CTBIL′*	*ACTBIL′*	*EBIL*
2	102.0	154.29	0.058	514.019
3	131.0	162.60	0.061	601.595
4	171.0	189.52	0.071	747.785
5	197.0	208.77	0.079	831.610
6	208.0	227.75	0.086	891.143
7	224.0	233.93	0.088	916.754
8	245.0	236.33	0.089	964.685
9	243.0	244.09	0.092	940.932

- *Dist*: Direct comparison of categorical original and protected values using a categorical distance;
- *CTBIL′*: Mean variation of frequencies in contingency tables for original and protected data (similar to *CTBIL* in Domingo-Ferrer and Torra (2001b) but dividing absolute differences among cells by the cell count in the contingency table of the original dataset);
- *ACTBIL′*: This is *CTBIL′* divided by the total number of cells in all considered tables;
- *EBIL*: Entropy-based information loss, measuring the reduction of uncertainty (e.g. information) in the protected categorical attributes with respect to the original attributes); this corresponds to *EBILMF* in Domingo-Ferrer and Torra (2001b).

Tables 4 and 5 correspond to two different quasi-identifiers: a quasi-identifier consisting of 8 attributes in the case of Table 4 and a quasi-identifier consisting of 4 attributes in the case of Table 5. Both quasi-identifiers contain a mixture of nominal and ordinal attributes, and the attributes in the shorter quasi-identifier are a subset of those in the longer one. As expected, the longer the quasi-identifier, the higher the information loss; also, for a fixed quasi-identifier length, the larger k, the higher the information loss.

Table 5. Information loss measures depending on k for a categorical quasi-identifier of length 4.

k	*Dist*	*CTBIL′*	*ACTBIL′*	*EBIL*
2	39.0	24.26	0.071	121.171
3	52.0	24.41	0.072	233.515
4	67.0	24.97	0.073	252.571
5	69.0	28.15	0.082	292.166
6	75.0	28.11	0.082	300.846
7	86.0	30.44	0.089	342.690
8	104.0	30.62	0.089	377.067
9	112.0	30.70	0.090	424.762

Table 6. Information loss measures for the experiments on categorical data depending on k for a quasi-identifier consisting of three ordinal attributes. Average operator used: median.

k	*Dist*	*CTBIL′*	*ACTBIL′*	*EBIL*
2	117.0	106.15	0.112	481.821
3	162.0	109.42	0.115	694.442
4	211.0	142.19	0.150	856.971
5	218.0	150.93	0.159	908.302
6	243.0	164.71	0.173	982.631
7	269.0	165.23	0.174	1037.38
8	283.0	171.68	0.181	1078.13
9	312.0	186.40	0.196	1160.37

Table 7. Information loss measures depending on k for a quasi-identifier consisting of three ordinal attributes. Average operator used: convex median.

k	*Dist*	$CTBIL'$	$ACTBIL'$	*EBIL*
2	111.0	100.66	0.106	461.089
3	151.0	108.95	0.115	639.920
4	198.0	141.87	0.149	828.326
5	229.0	145.75	0.153	938.592
6	240.0	157.82	0.166	975.469
7	269.0	188.18	0.198	1002.10
8	288.0	175.41	0.185	1067.81
9	296.0	179.31	0.189	1103.61

We then investigated whether replacing the median by the convex median as average operator leads to better results. Tables 6 and 7 compare both operators for a quasi-identifier consisting of the three ordinal attributes in the dataset. The results show that the convex median leads to better results in terms of the information loss measures selected than the median.

7. Conclusion

In statistical disclosure control of microdata, k-anonymity is an elegant way of dealing with the conflict between disclosure risk and data utility. In particular, it avoids the computational burden of disclosure risk assessment inherent to the classical approach to microdata SDC — first mask and then assess information loss and disclosure risk, usually via record linkage. If a protected dataset satisfies k-anonymity, disclosure is no longer possible. Thus the challenge is to satisfy k-anonymity with little information loss (damage to data utility) and high computational efficiency.

The algorithms to reach k-anonymity proposed in the literature are based on generalization and suppression. We have discussed the drawbacks of that approach, which include the difficulty of adequately combining generalization and suppression, the difficulty of processing partially suppressed data, the inability to deal with continuous microdata, etc. We have proposed microaggregation as a natural, unified and efficient way of reaching k-anonymity for any type of attributes (continuous and categorical).

Acknowledgments

Francesc Sebé's help in obtaining the results reported for continuous data is gratefully acknowledged. Comments by William Winkler were also particularly useful to improve this paper. This work was partly funded by the Spanish Ministry of Science and Technology and the European FEDER Fund under project TIC2001-0633-C03-01/03 "STREAMOBILE" and also by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-01/02 "PROPRIETAS".

References

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A. 2004. k-Anonymity: Algorithms and hardness. Technical report, Stanford University.

Dalenius, T. 1986. Finding a needle in a haystack - or identifying anonymous census records. Journal of Official Statistics, 2(3):329–336.

Defays, D. and Nanopoulos, P. 1993. Panels of enterprises and confidentiality: the small aggregates method. In Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys. Ottawa, Statistics Canada, pp.195–204.

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

Domingo-Ferrer, J., Mateo-Sanz, J.M., and Torra, V. 2001. Comparing sdc methods for microdata on the basis of information loss and disclosure risk. In Pre-proceedings of ETK-NTTS'2001 (vol. 2). Luxemburg. Eurostat, pp. 807–826.

Domingo-Ferrer, J. and Torra, V. 2001a. Disclosure protection methods and information loss for microdata. In P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, Amsterdam. North-Holland. http://vneumann.etse.urv.es/publications/bcpi pp. 91–110.

Domingo-Ferrer, J. and Torra, V. 2001b. A quantitative comparison of disclosure control methods for microdata. In P. Doyle, J.I. Lane, J.J.M. Theeuwes, and L. Zayatz (Eds.), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, Amsterdam. North-Holland. http://vneumann.etse.urv.es/publications/bcpi, pp. 111–134.

Domingo-Ferrer, J. and Torra, V. 2005. Privacy in statistical databases: Methods and performance metrics for microdata protection. manuscript.

Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., and Roehrig, S.F. 2001a. Disclosure limitation methods and information loss for tabular data. In P. Doyle, J.I. Lane, J.J. Theeuwes and L.V. Zayatz (Eds.), Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam. North-Holland, pp. 135–166.

Duncan, G.T., Keller-McNulty, S.A., and Stokes, S.L. 2001b. Disclosure risk vs. data utility: The r-u confidentiality map.

Hundepool, A., de Wetering, A.V., Ramaswamy, R., Franconi, L., Capobianchi, A., DeWolf, P.-P., Domingo-Ferrer, J., Torra, V., Brand, R., and Giessing, S. 2003. μ-ARGUS version 3.2 Software and User's Manual. Statistics Netherlands, Voorburg NL. http://neon.vb.cbs.nl/casc://neon.vb.cbs.nl/casc.

Mateo-Sanz, J.M., Domingo-Ferrer, J., and Sebé, F. 2005. Probabilistic information loss measures in confidentiality protection of continuous microdata. Data Mining and Knowledge Discovery, this issue.

Meyerson, A. and Williams, R. 2004. On the complexity of optimal k-Anonymity. In Proc. of the ACM Symposium on Principles of Database Systems-PODS'2004. Paris, France. ACM, pp. 223–228.

Oganian, A. and Domingo-Ferrer, J. 2001. On the complexity of optimal microaggregation for statistical disclosure control. Statistical Journal of the United Nations Economic Comission for Europe, 18(4):345–354.

Reiter, J.P. 2004. Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. Journal of the Royal Statistical Society, Series A, page forthcoming.

Samarati, P. 2001. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, 13(6):1010–1027.

Samarati, P. and Sweeney, L. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International.

Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J.M., and Torra, V. 2002. Post-masking optimization of the trade-off between information loss and disclosure risk in masked microdata sets. In J. Domingo-Ferrer (ed.), Inference Control in Statistical Databases, volume 2316 of LNCS, Berlin Heidelberg, Springer, pp. 163–171.

Sweeney, L. 2002a. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10(5):571–588.

Sweeney, L. 2002b. k-anonimity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, 10(5):557–570.

Torra, V. 2004. Microaggregation for categorical variables: A median based approach. In J. Domingo-Ferrer and V. Torra (Eds.), Privacy in Statistical Databases, volume 3050 of LNCS, Berlin Heidelberg. Springer, pp. 162–174.

Willenborg, L. and DeWaal, T. 2001. Elements of Statistical Disclosure Control. Springer-Verlag, New York.

Winkler, W. E. 2004. Re-identification methods for masked microdata. In J. Domingo-Ferrer and V. Torra (Eds.), Privacy in Statistical Databases, volume 3050 of LNCS, Berlin Heidelberg, Springer, pp. 216–230.

Yancey, W.E., Winkler, W.E., and Creecy, R.H. 2002. Disclosure risk assessment in perturbative microdata protection. In J. Domingo-Ferrer (Eds.), Inference Control in Statistical Databases, volume 2316 of LNCS, Berlin Heidelberg. Springer, pp. 135–152.