# Semantic Anonymisation of Categorical Datasets

**Sergio Martínez, Aida Valls and David Sánchez**

**Abstract** The exploitation of microdata compiled by statistical agencies is of great interest for the data mining community. However, such data often include sensitive information that can be directly or indirectly related to individuals. Hence, an appropriate anonymisation process is needed to minimise the risk of disclosing identities and/or confidential data. In the past, many anonymisation methods have been developed to deal with numerical data, but approaches tackling the anonymisation of non-numerical values (e.g. categorical, textual) are scarce and shallow. Since the utility of this kind of information is closely related to the preservation of its meaning, in this work, the notion of semantic similarity is used to enable a semantically coherent anonymisation. The knowledge modelled in ontologies is used as the basic pillar to propose semantic operators that enable an accurate management and transformation of categorical attributes. These operators are then used in three anonymisation mechanisms: Semantic Recoding, Semantic and Adaptive Microaggregation and Semantic Resampling. The three algorithms are compared in terms of semantic utility, privacy disclosure risk and runtime, with encouraging results.

## 1 Introduction

Inference control in statistical databases or Statistical Disclosure Control (SDC) aims to disseminate statistical data while preserving confidentiality. SDC is focused mainly on the preservation of privacy in structured databases [1–6]. Statistical Disclosure Control methods transform the original database into a new database by means on an anonymisation procedure, which takes into account that the protected data satisfies

S. Martínez (✉) · A. Valls · D. Sánchez
Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili,
Avda. Països Catalans, 26, 43007 Tarragona, Catalonia, Spain
e-mail: sergio.martinezl@urv.cat

A. Valls
e-mail: aida.valls@urv.cat

D. Sánchez
e-mail: david.sanchez@urv.cat

© Springer International Publishing Switzerland 2015

G. Navarro-Arribas and V. Torra (eds.), *Advanced Research in Data Privacy*,
Studies in Computational Intelligence 567, DOI 10.1007/978-3-319-09885-2_7

simultaneously utility and security conditions. The dataset will be useful if it is representative of the original dataset and it will be secure if it doesnot allow the re-identification of the original data.

In this chapter we focus only on the protection of *microdata* files, which consist on records detailing the information of an individual (e.g. person or company). The release of midrodata implies a risk of disclosure of the information of that individual. Given an original microdata set *D*with *n* records (corresponding *n* individuals) and *m* values in each record (corresponding to *m* attributes that are not identifiers), the goal is to release a protected microdata set $D^A$ (with also *n* records an *m* attributes) in such a way that:

1. Disclosure risk (i.e. the risk that a user or an intruder can use $D^A$ to determine confidential attributes on a specific individual among those in *D*) is low.
2. Data analysis (regressions, means, data mining, etc.) on $D^A$ and on *D* yield the same or at least similar results.

Most privacy-preserving mechanisms classify attributes in a dataset as: *Identifiers* (attributes that unambiguously identify the individual), *Quasi-identifiers* (attributes that may identify some of the respondents if they are combined with the other attributes available in external sources) and *Confidential*(attributes that contain sensitive information).

A well-known privacy model that relies on such classification to offer an a priori privacy guarantee is the *k*-anonymity [7]. *A dataset is said to satisfy k-anonymity for* $k > 1$ *if, for each combination of values of quasi-identifier attributes (e.g. name, address, age, gender, etc.), at least k records exist in the dataset sharing that combination.* Once a value for *k* is fixed (which should be a value that keeps the re-identification risk low enough), the goal of the masking method is to find an anonymisation that minimises the information loss (i.e. maximises the quality of the dataset $D^A$).

Anonymisation mechanisms enforcing *k*-anonymity convert an original set *D* in a publishable set $D^A$ through a masking process. These methods distort quasi-identifiers in a way that unique combinations of values in the original dataset disappear and data becomes more homogenous.

Notice, that the posterior use of the data plays an important role in the anonymisation process because the masked version should enable extracting the same conclusions from data analysis than the original one. This is especially important with data mining analysis, such as clustering, rule induction, profiling, or prediction, among others. In fact, privacy preserving data mining is a new research field that attempts to develop tools to study in an integrated way how to deal with privacy issues while performing data analysis [8]. The quality of the data can be measured as a function of the distribution of the values in the datasets. Even though data distribution is a dimension of data utility, we argue, as it has been stated by other authors [9], that retaining the semantics of the dataset with non-numerical data plays a more important role when one aims to extract conclusions by means of data analysis techniques.

Most masking method enforcing *k*-anonymity focus on the protection of numerical attributes. However, applying these methods to non-numerical attributes is not

straightforward because of the limitations on defining appropriate operators to manage categorical values. Although some recent masking methods are able to manage categorical data, they do not make an appropriate use of semantic techniques to preserve the *meaning* of the original dataset.

Methods available in the literature which consider the semantics of categorical attributes are mainly based on data *recoding*. To do so, they generalise input values by relying on tailor-made hierarchical structures: Value Generalization Hierarchies (VGHs). This approach presents three main drawbacks:

1. VGHs are manually constructed from each attribute in function of the input data (categorical values correspond to leaves). Human intervention is required and the VGH is only valid for a concrete dataset. This fact may not be assumable when dealing with large sets of categories.
2. VGHs produce ad-hoc and small hierarchies with a much reduced taxonomical detail offering a rough and biased knowledge model.
3. Generalisations based on VGHs usually produce a high information loss due to their coarse granularity. Moreover, the quality of the results heavily depends on the structure of VGH.

In this chapter we present three masking methods enforcing $k$-anonymity that are well suited for the anonymisation of categorical attributes from a semantic point of view. They are able to build an anonymised dataset that is semantically similar to the original one, thus preserving its analytical utility. To do so, we rely on the notion of *ontology-based semantic similarity*, which enables a semantically-coherent management of categorical data.

An ontology, in Information Science, can be defined as a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations. It is used to describe a domain of knowledge and to reason about the properties of the objects in that domain [10]. Ontologies are machine-interpretable so that it can be queried. In this sense, some standard languages have been designed to codify ontologies. They are usually declarative languages based on either first-order logic or on description logic [11]. Thanks to initiatives such as the Semantic Web, which brought the creation of thousands of domain ontologies [12], ontologies have been extensively exploited to compute semantic similarity between entities.

In the remainder of the chapter we describe three semantically grounded masking methods that exploit available ontologies to preserve the semantics of anonymised data. They are based on classic techniques found in Statistical Disclosure Control: Recoding, Microaggregation and Re-sampling. The three methods have been applied to enforce $k$-anonymity of Electronic Health Records. An empirical comparison of their performance is done in terms of semantic utility, privacy disclosure risk and runtime.

### 1.1 Problem Formalisation

A structured database consists of $n$ records corresponding to individuals, each one containing $m$ attribute values. In the simplest case, let us take anunivariate dataset with a single categorical attribute with $n$ records. On the contrary to continuous-scale numerical values, categorical attributes take values from a finite set of modalities (e.g., medical diagnoses); hence, they tend to repeat in the different records of the database. To explicitly consider repetitions, we represent the dataset in the form of $V = \{< v_1, \omega_1 >, \ldots, < v_p, \omega_p >\}$, where each $< v_i, \omega_i >$ tuple states the number $\omega_i$ of repetitions of each *distinct value* $v_i$ found in $V$. Note that, typically, $p$ (i.e., the number of distinct values) would be significantly lower than $n$ (i.e., the total amount of records).

*Example 1* Given the univariate dataset $V$ with $n = 12$ patient diagnosis: {*asbestosis*,*degenerative disorder*, *amyotrophia*, *myofibrosis*, *asbestosis*, *allergy*, *myofibrosis*, *allergy*, *squint*, *amyotrophia*, *degenerative disorder*, *allergy*}, we represent this dataset as a tuple with $p = 6$ elements:

$$\begin{aligned} V = \{ & < \textit{asbestosis}, \quad 2 >, < \textit{degenerative disorder}, \quad 2 >, \\ & < \textit{amyotrophia}, \quad 2 >, < \textit{myofibrosis}, \quad 2 >, < \textit{allergy}, \quad 3 >, \\ & < \textit{squint}, \quad 1 > \} \end{aligned}$$

This formalisation can be generalised for multivariate datasets with $m > 1$ attributes as follows. Let $MV = \{< \{v_{11}, \ldots, v_{1m}\}, \omega_1 >, \ldots, < \{v_{p1}, \ldots v_{pm}\}, \omega_p >\}$ be the representation of the dataset, where each tuple $\{v_{i1}, \ldots, v_{im}\}$ represents a distinct combination of $m$ attribute values, and $\omega_i$ states its number of occurrences (i.e., the frequency).

*Example 2* Given the multivariate dataset *MV* with two attributes describing conditions and treatments of a set of $n = 8$ patients: {{*lumbago*, *rehabilitation*}, {*colic*, *antibiotic*}, {*lumbago*, *rehabilitation*}, {*migraine*, *aspirin*}, {*lumbago*, *rehabilitation*}, {*lumbago*, *codeine*}, {*colic*, *hospitalisation*}, {*lumbago*, *codeine*}}, we represent this multivariate dataset as a tuple with $p = 5$ elements:

$$\begin{aligned} MV = \{ & < \{lumbago, rehabilitation\}, 3 >, < \{colic, antibiotic\}, 1 >, \\ & < \{migraine, aspirin\}, 1 >, < \{lumbago, codeine\}, 2 >, \\ & < \{colic, hospitalisation\}, 1 > \} \end{aligned}$$

Given that our goal to fulfil $k$-anonymity, if $\omega_i$ is equal or greater than a given value of $k$, the corresponding records in this tuple are already $k$-anonymous since they fulfil desired level of privacy. Hence, the goal of the masking process consists of generating a dataset where $\omega_i \geq k, \forall i$.

By managing the input dataset as formalised above we will be able to improve the computational cost of the algorithms, since they will be a function of $p$ instead of $n$. Moreover, given that categorical data is characterised by a limited and usually reduced set of modalities, it is expected that $p \ll n$.

## 2 Semantic Data Recoding

Data recoding, also known as *generalization* [13], is a making method that combines several categories in a new (less specific) one. For continuous attributes, global recoding means replacing an attribute by its discretized version, but the discretization leads very often to an unaffordable loss of information. For categorical attributes the methods rely on hierarchies of terms covering the categorical values observed in the sample, in order to replace a value by another more general one.

Previous works used ad-hoc small VGHs to find suitable generalisations. On the contrary, our ontology-based recoding takes advantage of the wide coverage of ontologies in different domains, such as WordNet or SNOMED CT. We present here a new recoding method capable of mapping categorical attribute values into ontological nodes that do not necessarily represent leaves of a hierarchy. As a result, semantically related concepts can be retrieved by going through the ontological hierarchies to which the value belongs. These ontological hierarchies are designed in a much more general and fine-grained fashion than VGHs and as a result of the consensus between domain knowledge experts rather than of the input data. Since we do not restrict the replacement to strict generalisations but to semantically similar entities, we enable a much wider and semantically-coherent set replacements. To ensure scalability with regards to the ontology size and input data, we bind the space of valid replacements to the set of value combinations that are present in the input dataset. When changing one value of a record for another, we can substitute an element by a taxonomical subsumer (this is the only case covered by the classic generalisation methods) but also with a hierarchical sibling (with the same taxonomical depth) or a specialisation (located at a lower level). In fact, in many situations a specialisation can be more similar than a subsumer,because highly specific concepts belonging to the lowest levels of a hierarchy have less differentiated meanings. As a result, the value change would result in a higher preservation of the semantic of data. This is an interesting characteristic and an improvement over the more restricted data transformations supported by VGH-based generalisation methods.

The proposed method is based on the substitution of all quasi-identifier values of each record with the values of another record [14]. To ensure the scalability of the method and guide the anonymisation process towards the minimisation of information loss, we have designed two heuristics ($H$) that ensure the selection, at each iteration, of the best set of records to transform:

($H_1$) From the input dataset, select the set of records $T$ with the lowest number of repetitions.

($H_2$) For each record $t \in T$ find the least distant record $v$ in the input dataset.

The goal of the first heuristic is to start the process with the records that fulfil $k$-anonymity the least,whereas, the aim of the second heuristic is minimise the information loss resulting from each substitution. To select such substitution, a semantic comparison operator has been defined. As a result of the replacement, quasi-identifier values for both records (the one to anonymise $t$ and the most semantically similar one $v$) will take the same values and become indistinguishable; therefore, the $k$-anonymity level for both records will increase.

The recoding algorithm presented in [14] proceeds as follows:

1. Select the tuple $t$ with the minimum number of repetitions.
2. As long as this number is lower than $k$, the dataset is not $k$-anonymous. To increase the number of repetitions, select the tuple $v$ that is the least distant to $t$ and that has the lowest amount of repetitions.
3. The original values of $t$ are replaced by the ones in $v$, increasing, in this manner, their anonymity level because the value in $v$ will have a higher number of repetitions.
4. Go to step 1.

This algorithm stops when in step 2 we have that the tuples with the minimum number of repetitions have a number of repetitions equal or higher than $k$. Consequently, with this iterative procedure that is applied to each non-anonymous record, the input dataset will fulfil $k$-anonymity.

In [15] the applicability and quality of the anonymised data has been demonstrated with several tests conducted with real data. Results indicate that, compared with a classical approach based on optimisation of the distribution of the data, ours retains the quality and utility of data better from a semantic point of view. This was studied by means of comparing the results of a clustering process. The partitions generated from the original dataset and the anonymised data are more similar with our semantic method than with classical approaches.

## *2.1 Semantic Similarity*

To compare the categorical values and to calculate the least distant tuples from a semantic perspective, the following similarity function is proposed [16]. This similarity function takes into account both the data distribution as well as the semantic similarity of the terms.

On the one hand, the semantic similarity of the values is estimated by mapping them into ontological concepts. Then, any of the edge-counting measures that can be found in the literature to evaluate the distance between concept pairs according to the length of the path connecting the two concepts in the ontology can be used [17]. The main advantage of the edge-counting measures is their simplicity and low computational cost. However, they require rich and consistent ontologies to work properly.

On the other hand, to consider also the distribution of data during the comparison of attribute values, their frequency of appearance is also taken into account. Since each distinct value $v_i$ appears $\omega_i$ times in the dataset, we propose to count the semantic distance between a given value $v_i$ and a base value $b$ as many times as indicated by its frequency of appearance $\omega_i$. Since this is equivalent to multiplying the semantic distance by the frequency, $\omega_i$ acts as a weighting factor. In that way, the accumulated distances resulting from grouping together a base value $b$ and all the records with the value $v_i$ can be minimised. Moreover, since all repetitions $\omega_i$ of $v_i$ are treated as a unit, sets of identical records will be grouped together, obtaining more cohesive groups.

Formally, the comparison operator used to group records with respect to a base value $b$ is defined as follows.

**Definition 1** The weighted semantic distance (*wsd*) between anunivariate reference value $b$ and a univariate set of records $< v_i, \omega_i >$ is defined as:

$$wsd\,(b, < v_i, \omega_i >) = \omega \cdot sd(b,\ v_i) \tag{1}$$

This measure can be generalised to multivariate data as follows:

**Definition 2** The distance between a multivariate reference value with $m$ attributes $\{b_1, \ldots, b_m\}$ and a multivariate set of records $< \{v_{i1}, \ldots, v_{im}\}, \omega_i >$ is defined as the *average* of the weighted semantic distances of the individual attribute values:

$$wsd\,(\{b_1, \ldots, b_m\}, < \{v_{i1}, \ldots v_{im}\}, \omega_i >) = \sum_{j=1}^{m} \frac{wsd\left(b_j, < v_{ij}, \omega_i >\right)}{m} \tag{2}$$

Note that this makes that the computational cost of our algorithms uniquely depend on the number of different tuples ($p$), unlike related works, which depend on the total size of the dataset ($n$) and on the depth and branching factor of the hierarchy (which represent an exponentially large generalisation space of substitutions to evaluate).

## 3 Semantic Microaggregation

Among the plethora of anonymisation methods, *microaggregation* stands as a natural approach to satisfy $k$-anonymity in statistical databases [18]. It builds clusters of at least $k$ original records according to a similarity function; then, each record of each cluster is replaced by the centroid of the cluster to which it belongs. As a result, each combination of values is repeated at least $k$ times and, hence, the masked dataset becomes $k$-anonymous. The goal of microaggregation is to find the partition that minimises the information loss. Because the search for the optimal partition when considering multivariate data is NP-hard [19], sub-optimal heuristic methods have been proposed. One the most popular ones is the MDAV (Maximum Distance

Average Vector) method [20], because it provides high quality aggregations without being constrained by some configuration parameters, as other methods do [21]. Our proposal focuses on adapting this method to categorical data.

The MDAV method is based on generating clusters of at least $k$ elements. First, the centroid of the dataset is calculated and the most distant object $r$ is selected by means of a distance measure appropriate for the type of data. Then, a cluster is constructed with the $k-1$ nearest objects to $r$. After that, the most distant record $s$ to $r$ is selected and a new cluster is constructed. The whole process is repeated until no objects remain ungrouped. As a result of the microaggregation process, all clusters have a fixed size of $k$, except the last cluster that may have a cardinality between $k$ and $2k-1$, because the initial number of records may not be divisible by $k$. Finally, all the elements in each cluster are replaced by the centroid of the cluster, becoming $k$-anonymous.

The MDAV method, however, presents some limitations that may hamper the utility of categorical data.

- The fact of relying on fixed-size clusters is a hard restriction that hampers the quality of the clusters in terms of cohesion. A low cohesion increases the information loss resulting from the replacement of the individual records by the cluster centroid [22]. The possibility of varying the size of the clusters ensuring a minimum cardinality of $k$ to fulfil $k$-anonymity would be preferable because it allows a better adaptation of the clusters to the data distribution. This is especially relevant for *categorical* data like job or city of living because, due to their discrete nature, modalities tend to repeat and, hence, it would be desirable to put as many repetitions as possible into the same cluster to maximise its cohesion.
- The results are influenced by the two operators needed during the microaggregation: the *distance measure*, used to compare records and centroids, and the *centroid construction,* needed to calculate the global centroid at each iteration and to select the representative record for each cluster. Since arithmetic functions cannot be applied to this kind of data, a straightforwardway to apply MDAV to categorical data consists on using Boolean equality/inequality operators [18, 23] or to use the common abstraction of a set of values in an ontology as the centroid [24].

In our proposal, some algorithmic and design modifications are made by considering the distributional characteristics of categorical attributes with the purpose of minimising the information loss. Two main aspects have been considered: (1) the interpretation of the semantics of non-numerical values during the whole microaggregation process, and (2) the consideration of the distribution categorical attributes to define adaptive-sized clusters, producing more cohesive results while fulfilling $k$-anonymity. The goal is to aggregate data into highly cohesive clusters, in order to minimise the information loss resulting from the masking process. The proposed modifications are the following [25]:

- Adaptive microaggregation: we propose a modification of the MDAV algorithm that, while ensuring the $k$-anonymity property, creates clusters of different sizes according to the data distribution. The algorithm will put all the records that have

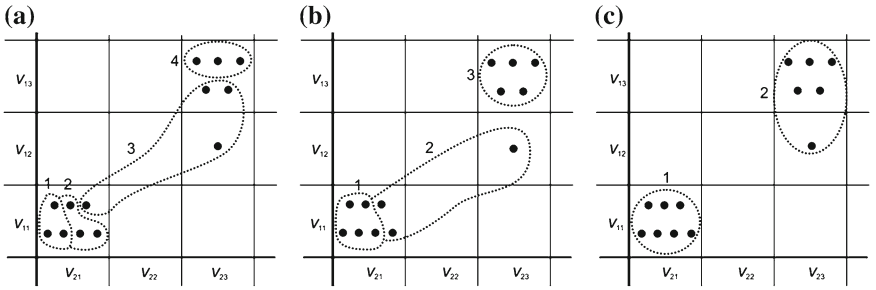



**Fig. 1** An example of microaggregation with $k = 3$. **a** Fixed-sized clustering. **b** Variable-sized clustering with a maximum size of $2k - 1$. **c** Adaptive clustering without maximum size restriction

the same values in the same cluster, while ensuring that the cluster has, at least, $k$ elements to fulfil the $k$-anonymity property. In this manner, the clustering construction process is guided by the data distribution, as illustrated in Fig. 1.

- Semantic distance: given that categorical data should be interpreted according to their underlying semantics, we propose to use the same weighted semantic similarity operator define above (Eq. 2) to guide the cluster construction process. It considers both the meaning of the value according to the background ontology and the distribution of those values.
- Semantic centroid: the construction of an appropriate and representative centroid is crucial to guide the microaggregation process and to minimise the information loss. We proposed in [25] a new procedure to calculate the centroid of multivariate categorical datasets by exploiting ontologies as well as considering the data distribution.

As a result of incorporating this adaptive behaviour during the clustering construction, the SA-MDAV (*Semantic Adaptive MDAV*) is presented [26]:

1. Compute the centroid *X*of all tuples in input data.
2. Select the most distant tuple $r$ to the centroid $X$ and the most distant tuple $s$ to $r$.
3. Form a cluster with the tuple $r$ and calculate the centroid $C$ of the cluster.
4. Add to this cluster the closest tuples to centroid $C$, recalculating the centroid whenever a tuple is added, until at the number of records is at least $k$.
5. Form a cluster with the tuple $s$ and do the same process as for $r$.
6. Remove the added tuples from input data and repeat the entire process until no tuples remaining in the dataset.
7. Replace each tuple by the centroid of the cluster that it belongs to.

We evaluate this method over two different datasets with categorical (details in [15, 16]). Results proved that SA-MDAV, even though being heuristic and subject to sub-optimal choices to preserve its scalability, improves related works by a considerable margin, both when considering the absolute information loss and also when evaluating the balance between information loss and disclosure risk.

### 3.1 The Centroid of Categorical Values

Centroid calculus for numerical data relies on standard averaging operators (e.g. arithmetic mean) [2]. However, the accurate centroid calculus for non-numerical data is challenging due to the lack of semantic aggregation operators and the necessity of considering a discrete set of centroid values. Related works propose methods to compute centroids for non-numerical data *either* relying on the distributional features of data, where the centroid is the *modal* value [23], *or* on background semantic, where the centroid is the term that generalises all aggregated values in a taxonomy [24]. Since only one dimension of data (distribution or semantics) is considered, both approaches result in suboptimal results [25].

In this section, we introduce a centroid calculation method for multivariate non-numerical data that considers, in an integrated manner, *both* semantics and distribution of data. To obtain accurate centroids, ontologies are exploited not only to semantically compare terms, as proposed in the previous section, but also to retrieve the centroid candidates.

The first issue concerns the search space of centroid candidates ($c$). Since $c$ must be necessarily a discrete value, some approaches (like the ones based on taking the modal value of a sample [23]) bound the set of possible candidates to those values appearing in the input dataset. Hence, the centroid accuracy would depend on the granularity and suitability of the input data. So, we extend the centroid search space to all terms of the taxonomy related to the input data. Centroid candidates will be all input terms together with their taxonomical ancestors. For semantic comparison, the weighted semantic distance defined in the previous section is used.

Formally, let us take $V = \{< v_1, \omega_1 >, \ldots, < v_p, \omega_p >\}$ as an input dataset with a single non-numerical attribute. The first step maps the terms $v_i$ of the set $V$ in a background ontology and extracts the minimum hierarchy $H$ that taxonomically models all $v_i$ values. All terms in $H$, which include both values in $V$ and their taxonomical ancestors, are considered as centroid candidates. The centroid will be the term $c$ in $H$ that minimises the weighted semantic distance (Eq. 2) to all $v_i$ in $V$. Note that, in this case, each centroid candidate $c$ acts as the base value in Eq. (2).

**Definition 3** The centroid of a set of non-numerical values $v_i$ in $V$ is defined as the term $c_j$ that minimises the weighted semantic distance *wsd* with respect to all the values $v_i$ in the space $V$.

$$centroid\,(V) = \{\arg\min_{\forall c_j \in H}(\sum_{i=1}^{p} wsd(c_j, < v_i, \omega_i >))\} \tag{3}$$

## 4 Semantic Resampling

The third masking method is based on dataresampling. Resampling was originally proposed for protecting tabular data, but later it has been used for microdata [27]. Although resampling has not gained as much research attention as other masking

algorithms like microaggregation [22], it has demonstrated to be fast, which is an interesting feature when dealing with large-scale datasets and to retain more utility than other methods for numerical data [28, 29].

Briefly, being $n$ the number of records in the dataset, the resampling method takes $t$ samples with replacement (i.e. values can be taken more than once). Each sample is sorted in increasing order. Then, the masked dataset is obtained by taking, as first value, the average of the first values of all samples, as second value, the average of the second values of all samples, and so on.

Comparative studies [28, 29] show that Heer's resampling achieves a high utility preservation with respect to other masking techniques, but with a slightly higher disclosure risk. This is related to the fact that, unlike other masking methods [22–24], the Heer's approach was designed without considering $k$-anonymity (formalized years later in [13]). Hence, resampled results cannot guarantee an a priori level of privacy.

A new version of resampling is presented here, named S$k$Resampling [30]. The new resampling method fulfils $k$-anonymity while it is also able to deal with non-numerical data from a semantic perspective. Two issues not considered in previous works.

The S$k$Resampling algorithm is focused on minimizing the information loss when masking categorical data while ensuring the fulfilment of $k$-anonymity. It is based on the Heer's resampling method [27] with the following modifications:

- *k-anonymous resampling*: the original sampling method has been modified so that masked records fulfil $k$-anonymity.
- *Semantic resampling of categorical data*: in order to semantically interpret non-numerical data during the resampling process, we have applied semantic operators, such as the weighted semantic similarity function and a sorting operator presented below.

Let $D$ be the input dataset, $k$ is the desired level of $k$-anonymity and $n$ is the number of records in $D$. The algorithm proceeds as follows:

1. Create $k$ set of $n/k$ records, by a sampling procedure without replacement (i.e. each record is taken only once). The aim it that $k$ records can be replaced by their average in a later stage and become $k$-anonymous.
2. Sort each of these samples with the same semantic criterion.
3. Create $P_i$ ordered sets with the records at the $i$th position of all sorted samples. The idea is that, by sorting the samples, similar records appear at similar positions of different samples.
4. The anonymised dataset is obtained by replacing all records of each $P_i$ by the centroid of $P_i$.

With this algorithm $P_i$ contains at least $k$ records that are all substituted by the same centroid, hence, we guarantee that the masked dataset is $k$-anonymous. Note that, in step 1, when taking $n/k$ records per sample, the remaining $n \bmod k$ records should be also treated. These records are added to the set with the closest centroid (in step 3), recalculating the centroid before executing step 4.

In comparison with other methods, resampling is faster since the sampling process is done randomly, which makes it especially suitable for very large datasets. In contrast, this randomness may negatively influence the information loss. The evaluation results sustain our hypothesis, since the method is able to minimise the information loss in comparison with non-semantic approaches [16].

### *4.1 Sorting Non-numerical Data*

In the resampling method, a sorting operator is required. In [16] we proposed a procedure to arrange the records with non-numerical data according to their semantic similarity.

Sorting categorical values is not straightforward since, in general, they are not ordinal (i.e. they do not take values in an *ordered* range). In this section, we detail a sorting algorithm for categorical data.

To sort a set of values, a reference point is needed so that values could be arranged according to their similarity/distance to that reference. Numerically, this is done according to the max/min value (i.e. the most extreme value) of the set. To define a sorting procedure for categorical data, we also rely on the notion of the most extreme value, which corresponds to the one that, globally, is the *most distant* to all other values (conceptually, this is the opposite of the centroid as computed in Sect. 3.1). Once this reference value is obtained, other values are sorted by iteratively picking those that are least distant to that extreme value.

To set the reference value/record as well as to compare it to other elements in the set, while considering both the semantic and distributional features of data, we rely on the weighted semantic distance and the centroid calculus procedures explained before.

## 5 A Comparative Study in the Medical Domain

This section compares the three methods detailed above when applied to the anonymisation of real medical data. In particular, we address the problem of protecting the privacy of Electronic Health Records of patients. This kind of dataset contains confidential information regarding clinical outcomes that should not be disclosed. At the same time, the analysis of the health care experience captured in clinical databases is very important because it may lead to improved continuity in patient assessment, improved treatment, avoidance of adverse drugs reactions, and in ensuring that people at risk receive appropriate support services [31, 32]. Thus, utility of anonymised data is critical.

Moreover, due to the importance of terminology in clinical assessment, the medical domain has been very prone to the development of large and detailed ontologies such SNOMED CT [33, 34], which is used in our tests to protect medical terms.

**Table 1** Example of clinical dataset provided by OSHPD

| ID | Age range | Patient ZIP code | Principal diagnosis cause of admission | Other condition that coexist at the time of admission |
|---|---|---|---|---|
| * | 50–54 | 916** | Abstinent alcoholic | Metabolic acidosis due to salicylate |
| * | 65–69 | 913** | Infected spinal fixation device | Uric acid renal calculus |
| * | 65–69 | 903** | Aneurysm of thoracic aorta | Cardiac oedema |
| * | >=85 | 902** | Fibroma of ovary | Chronic osteoarthritis |
| * | 30–34 | 917** | Appendicitis | Severely obese |

## *5.1 The Dataset*

The evaluation has been carried out over a database containing inpatient information provided by the California Office of Statewide Health Planning and Development (OSHPD) and collected from licensed hospitals in California. Specifically, we used the patient discharge dataset corresponding to the 4th quarter of 2009 for the hospital with the largest amount of records (i.e., Cedars Sinai Medical Center, Los Angeles County).

Prior to publication, the OSHPD has masked or removed some attributes that resulted in unique combinations of certain demographic variables (see an example on the first three columns of Table 1). For evaluation purposes, two categorical attributes were considered in our tests: *principal diagnosis* and *other conditions* of the patient at the time of the admission (i.e., $m = 2$), which are stored as ICD-9 codes in the original data file. After removing records with missing information, a total of 3,006 individual records is available for testing (i.e., $n = 3,006$). A total of 2,331 different combinations of values (i.e., $p = 2,331$ tuples) can be found, from which a significant amount (2,073) are unique.

## *5.2 Comparing Anonymisation Methods*

The three algorithms have been compared under the perspectives of information loss, disclosure risk and runtime. The information loss ($L$) measure estimates the utility of anonymised dataset by quantifying the *semantic information loss* caused by replacing original values by their masked versions. It is measured as the *Sum of Squared Errors* (SSE). To measure the information loss from a *semantic* perspective, we compare the categorical values using the semantic weighted distance defined in Eq. (2) on the basis of the knowledge available in the SNOMED CT ontology.

$$L = \sum_{i=1}^{n} \left( \frac{\sum_{j=1}^{m} sd(r_{ij}, r_{ij}^{A})}{m} \right)^{2} \tag{4}$$

where $n$ is the number of records in the dataset, each one composed by $m$ attributes, $r_{ij}$ is the original value of the $j$th attribute of the $i$th record and $r_{ij}^{A}$ denotes its masked version.

On the other hand, the disclosure risk is calculated by means as the percentage of correct linkages between the original and masked datasets. The Record Linkage (RL) is assessed also by the same semantic similarity measure and SNOMED CT as knowledge base. The specific record linkage method is detailed in [35].

To evaluate the balance score between the information loss and the disclosure risk of the different methods, we use the score function (Eq. 5). The lower the score is, the higher the quality of the method because both low information loss and low disclosure risk are achieved.

$$score = \alpha L + (1 - \alpha) RL \tag{5}$$

First, as shown in Fig. 2, we consider an equal balance between the data utility and the disclosure risk, that is, $\alpha = 0.5$. The conclusion is that the recoding method provides the best results for low $k$ levels ($k <= 5$) and the microaggregation method works better for high $k$ levels ($k > 5$). Thus, depending on the level of $k$-anonymity it would be convenient to choose a method or another. It is also relevant to note that the score is maintained almost constant as $k$-values grow, stating that the quality of the methods scales well as the privacy requirements increase. On the other hand, resampling provides the worst balance between information loss and disclosure risk.

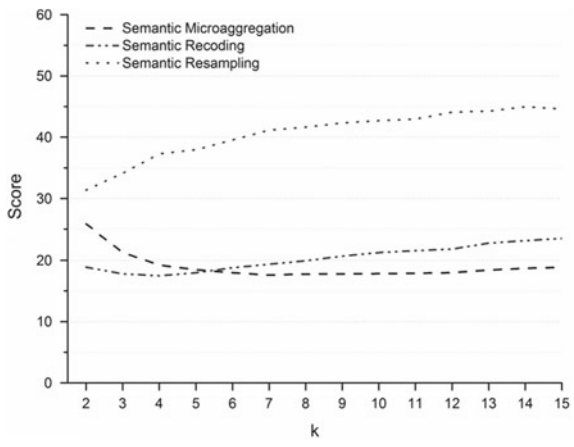



**Fig. 2** Score with an equilibrated balance ($\alpha = 0.5$) between information loss and disclosure risk

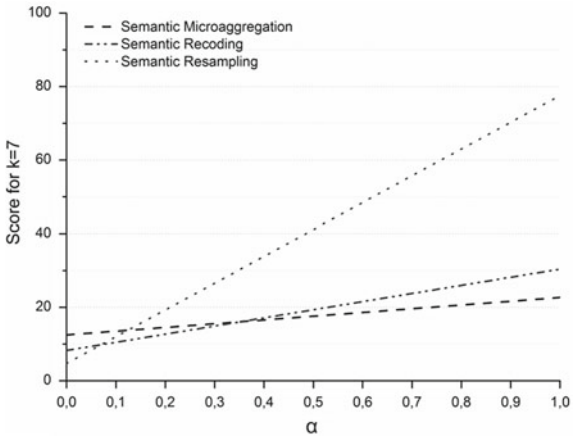



**Fig. 3** Score values when varying the trade-off between information loss and disclosure risk

We have also studied the behaviour of the overall score when varying the parameter $\alpha$ between 0 to 1. With $\alpha = 0$, the score is based solely on the disclosure risk measure, while with $\alpha = 1$, the score is based only on the information loss. In this analysis, an intermediate level of anonymity ($k = 7$) has been fixed.

As it can be seen in Fig. 3, the recoding method achieves the best results for low values of $\alpha$ whereas themicroaggregation method obtains the best results for high values of $\alpha$. This means that if we need to give more importance to information loss with an adequate level of disclosure risk, the microaggregation method would be more suitable. On the contrary, if we give more weight to disclosure risk with a moderate level of information loss, the recoding method would be preferable. The resampling method obtains the best result only when data utility is not taken into account.

Finally, the runtime of the anonymisation algorithms is a relevant feature to consider when resources are limited, such as in EHRs, which are likely to contain large amounts of data. Figure 4 shows the comparison for the three SDC methods executed on a 2.4 GHz Intel Core processor with 4 GB RAM. The fastest method is *resampling*, with an almost negligible runtime. *Microaggregation* is the slowest for values of $k$ below 4, whereas it surpasses *recoding* for higher $k$-values, with an almost inverse logarithmic shape.

Given the above results, the semantically-grounded *microaggregation* method seems the best approach to anonymise data when the meaning of original data should be preserved as much as possible with a moderate level of disclosure risk. Moreover, it is especially efficient for high $k$-anonymity values. *Recoding* would be only considered if very low $k$-anonymisation levels are required (being more computationally efficient than *microaggregation*) or when analyses to be performed over masked data will be focused solely on data distribution rather than on their semantics. Finally, only when input EHRs are so large to be non-computationally feasibly anonymised

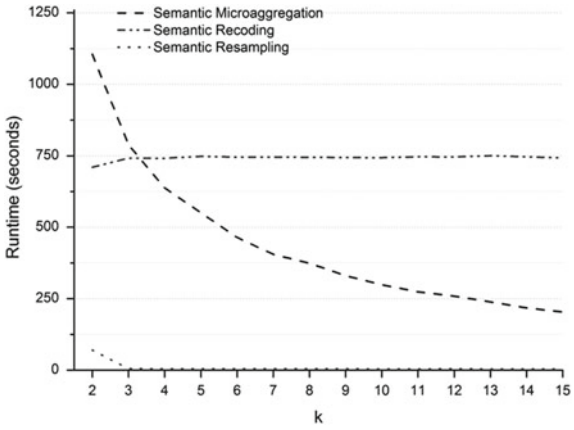



**Fig. 4** Runtime (in seconds) for different levels of *k*-anonymity

by means of *microaggregation* or *recoding* methods, the *resampling* method could be considered thanks to its high efficiency, even at the expenses of a higher information loss.

## 6 Conclusions and Future Work

In this chapter, we presented some methods for anonymisation for categorical attributes, showing that data semantics plays acrucial rolein order to retain the analytical utility. Given the importance of data semantics in anonymisation tasks, we have studied the definition of new operators needed in masking methods. These operators exploit the semantics provided by ontologies to enable a coherent comparison, aggregation and sorting of categorical data and without neglecting the distribution of the values.

The three masking methods apply different perturbation methods in order to achieve *k*-anonymity, namely by recording original values, by microaggregation of semantically similar records (that are substituted by a common centroid) or by sampling and sorting the original data into small sets (also replaced by their centroid).

The three methods have been evaluated with a real dataset from different perspectives, concluding that each method has its own advantages and limitations. We can also conclude that well-defined general purpose semantic structures, as ontologies, are a good source of information to interpret the semantics of terms and their use is crucial to retain the utility of data during the anonymisation process.

This work shows that it is possible to use ontologies with traditional masking methods if some operators are adapted. The three operators presented here appear in other anonymisation techniques available in the literature, which could be easily adapted to deal appropriately with categorical data.

As future work, it would be interesting to study how the different methods behave with other ontologies with different sizes and granularities. The possibility of combining several ontologies as background knowledge could be also considered. Regarding the privacy model, research on the application of the proposed semantic framework to other models and methods can be done. Related with the privacy model, recently, a more robust privacy model like *differential privacy* can be considered. *Differential privacy* [36] ensures that released data are insensitive to any individual's data. Hence, individual data remains uncertain for an attacker. Finally, we would also consider the adaptation of other numerical operators such as variance or co-variance for categorical values by means of ontologies.

**Acknowledgments** This work has been supported by the Spanish Ministry of Science and Innovation (through projects ICWT TIN2012-32757, ARES-CONSOLIDER INGENIO 2010 CSD2007-00004 and BallotNext IPT-2012-0603-430000) and by the Government of Catalonia under grants 2009 SGR 1135 and 2009 SGR-01523. Dr. Martínez was supported with research grants by the Universitat Rovira i Virgili and Ministerio de Educación y Ciencia (Spain).

## References

1. Willenborg, L., de Waal, T.: Elements of Statistical Diclosure Control. Lecture Notes in Statistics, vol. 155. p. 261. Springer, New York (261)
2. Domingo-Ferrer, J.: A Survey of inference control methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining, pp. 53–80. Springer, US (2008)
3. Jin, X., Zhang, N., Das, G.: ASAP: eliminating algorithm-based disclosure in privacy-preserving data publishing. Inf. Syst. **36**(5), 859–880 (2011)
4. Herranz, J., et al.: Classifying data from protected statistical datasets. Comput. Secur. **29**(8), 875–890 (2010)
5. Oliveira, S.R.M., Zaïane, O.R.: A privacy-preserving clustering approach toward secure and effective data analysis for business collaboration. Comput. Secur. **26**(1), 81–93 (2007)
6. Shin, H., Vaidya, J., Atluri, V.: Anonymization models for directional location based service environments. Comput. Secur. **29**(1), 59–73 (2010)
7. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertainty Fuzziness Knowl. Based Syst. **10**(5), 557–570 (2002)
8. Aggarwal, C.C., Yu, P.S.: Privacy-Preserving Data Mining: Models and Algorithms. Springer Publishing Company, Incorporated, Berlin (2008)
9. Torra, V.: Towards knowledge intensive data privacy. In: Proceedings of the 5th International Workshop on Data Privacy Management, and 3rd International Conference on Autonomous Spontaneous Security, Springer, Athens, Greece (2011)
10. Guarino, N.: Formal, ontology and information systems. In: 1st International Conference on Formal Ontology in Information Systems. IOS Press, Trento, Italy (1998)
11. Gomez-Perez, A., Fernandez-Lopez, M., Corcho, O.: Ontological Engineering, 2nd Printing. Springer, New York (2004)
12. Ding, L. et al.: Swoogle: a search and metadata engine for the semantic web. In: Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management. ACM, Washington, D.C., USA (2004)
13. Samarati, P., Sweeney, L.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory (1998)

14. Martínez, S., et al.: Privacy protection of textual attributes through a semantic-based masking method. Inf. Fusion **13**(4), 304–314 (2011)
15. Martínez, S.: Ontology based semantic anonimisation of microdata. Universitat Rovira i Virgili. PhD. Thesis (2013). http://www.tdx.cat/bitstream/handle/10803/108961/Tesi.pdf?sequence=1
16. Martínez, S., Sánchez, D., Valls, A.: A semantic framework to protect the privacy of electronic health records with non-numerical attributes. J. Biomed. Inform. **46**(2), 294–303 (2013)
17. Rada, R., et al.: Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. **19**(1), 17–30 (1989)
18. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Min. Knowl. Discov. **11**(2), 195–212 (2005)
19. Oganian, A., Domingo-Ferrer, J.: On the complexity of optimal microaggregation for statistical disclosure control. Stat. J. United Nations Econ. Comm. Eur. **18**(4), 345–353 (2001)
20. Hundepool, A. et al.: $\mu$-ARGUS version 3.2 software and user's manual. Statistics Netherlands, Voorburg NL (2003). http://neon.vb.cbs.nl/casc://neon.vb.cbs.nl/casc
21. Domingo-Ferrer, J., et al.: Efficient multivariate data-oriented microaggregation. VLDB J. **15**(4), 355–369 (2006)
22. Domingo-Ferrer, J., Mateo-Sanz, J.M.: Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans. Knowl. Data Eng. **14**(1), 189–201 (2002)
23. Torra, V.: Microaggregation for categorical variables: a median based approach. In: Domingo-Ferrer, J., Torra, V. (eds.) Privacy in Statistical Databases, pp. 518–518. Springer, Berlin (2004)
24. Abril, D., Navarro-Arribas, G., Torra, V: Towards semantic microaggregation of categorical data for confidential documents. In: Proceedings of the 7th International Conference on Modeling Decisions for Artificial Intelligence. Springer, Perpignan, France (2010)
25. Martínez, S., Valls, A., Sánchez, D.: Semantically-grounded construction of centroids for datasets with textual attributes. Knowl. Based Syst. **35**, 160–172 (2012)
26. Martínez, S., Sánchez, D., Valls, A.: Semantic adaptive microaggregation of categorical microdata. Comput. Secur. **31**(5), 653–672 (2012)
27. Heer, G.R.: A bootstrap procedure to preserve statistical confidentiality in contingency tables. In: International Seminar on Statistical Confidentiality. Eurostat, Luxembourg (1993)
28. Herranz, J., Nin, J., Torra, V.: Distributed privacy-preserving methods for statistical disclosure control data privacy management and autonomous spontaneous security. Int. Sci. **5939**, 33–47 (2010)
29. Karr, A.F., et al.: A framework for evaluating the utility of data altered to protect confidentiality. Am. Stat. **60**, 224–232 (2006)
30. Martínez, S., Sánchez, D., Valls, A.: Towards k-anonymous non-numerical data via semantic resampling. In: Greco, S. et al. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems, Catania, Italy (2012)
31. Elliot, M., Purdam, K., Smith, D.: Statistical disclosure control architectures for patient records in biomedical information systems. J. Biomed. Inform. **41**(1), 58–64 (2008)
32. Malin, B., Sweeney, L.: How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J. Biomed. Inform. **37**(3), 179–192 (2004)
33. Spackman, K.A., Campbell, K.E., Cote, R.A.: SNOMED RT: a reference terminology for health care. In: Proceedings of AMIA Annual Fall Symposium, pp. 640–644 (1997)
34. Nelson, S.J., Johnston, D., Humphreys, B.L.: Relationships in medical subject headings. In: Relationships in the Organization of Knowledge, pp. 171–184. K.A. Publishers, New York (2001)
35. Martínez, S., Sánchez, D., Valls, A.: Evaluation of the disclosure risk of masking methods dealing with textual attributes. Int. J. Innovative Comput. Inf. Control **8**(7(A)), 4869–4882 (2012)
36. Dwork, C.: Differential privacy. In: ICALP, Springer (2006)